SCIENTIFIC REPORTS

OPEN

Received: 19 February 2016
Accepted: 14 June 2016
Published: 29 June 2016

Transcriptome analysis revealed chimeric RNAs, single nucleotide polymorphisms and allele-specific expression in porcine prenatal skeletal muscle

Yalan Yang[1,2,*], Zhonglin Tang[1,2,*], Xinhao Fan[1], Kui Xu[1], Yulian Mu[1], Rong Zhou[1] & Kui Li[1,2]

Prenatal skeletal muscle development genetically determines postnatal muscle characteristics such as growth and meat quality in pigs. However, the molecular mechanisms underlying prenatal skeletal muscle development remain unclear. Here, we performed the first genome-wide analysis of chimeric RNAs, single nuclear polymorphisms (SNPs) and allele-specific expression (ASE) in prenatal skeletal muscle in pigs. We identified 14,810 protein coding genes and 163 high-confidence chimeric RNAs expressed in prenatal skeletal muscle. More than 94.5% of the chimeric RNAs obeyed the canonical GT/AG splice rule and were trans-splicing events. Ten and two RNAs were aligned to human and mouse chimeric transcripts, respectively. We detected 106,457 high-quality SNPs (6,955 novel), which were mostly (89.09%) located within QTLs for production traits. The high proportion of non-exonic SNPs revealed the incomplete annotation status of the current swine reference genome. ASE analysis revealed that 11,300 heterozygous SNPs showed allelic imbalance, whereas 131 ASE variants were located in the chimeric RNAs. Moreover, 4 ASE variants were associated with various economically relevant traits of pigs. Taken together, our data provide a source for studies of chimeric RNAs and biomarkers for pig breeding, while illuminating the complex transcriptional events underlying prenatal skeletal muscle development in mammals.

Muscle fibers are the basic structural and functional units of skeletal muscle[1]. The number of muscle fibers determines the capacity for postnatal muscle fiber growth[2,3]. Porcine skeletal muscle development is a complex biological process, especially during prenatal developmental stages. All muscle fibers are formed during the prenatal stage, whereas postnatal skeletal muscle development is mainly associated with increased muscle fiber size[4]. In pigs, prenatal myogenesis exhibits two major waves of fiber generation: primary fiber formation at 35–60 days post coitus (dpc) and secondary myogenesis at 54–90 dpc[5]. The majority of muscle fibers are formed during secondary myogenesis using the primary fibers as templates[6]. Previous studies showed that the critical time point for the formation of secondary myogenesis fibers was at approximately 63 dpc[7], whereas the stages ranging from 49 to 77 dpc were pivotal for formation of various muscle phenotypes[8]. However, the molecular mechanisms underyling myofiber formation in mammals such as pigs remain unclear. Transcriptome profiling of prenatal skeletal muscle is an effective strategy for understanding the molecular events mediating myogenesis in pigs.

Gene expression profiles during tissue and organ development are complex. Multiple transcript types, including long non-coding RNA, chimeric RNA, and circular RNA, as well as transcriptional events, including alternative splicing and allele-specific expression (ASE), contribute to the complexity of the transcriptome and provide significant obstacles to the achievement of a comprehensive understanding of the genetic basis of skeletal muscle development[9,10]. Transcriptomic research on porcine skeletal muscle has mainly focused on mRNA[7,8,11],

[1]The State Key Laboratory for Animal Nutrition, Institute of Animal Science, Chinese Academy of Agricultural Sciences, Beijing 100193, P.R.China. [2]Agricultural Genome Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, 518124, P.R.China. *These authors contributed equally to this work. Correspondence and requests for materials should be addressed to R.Z. (email: zhourong03@caas.cn) or K.L. (email: likui@caas.cn)

miRNA[12–16], and lncRNA[17]. No report exists regarding chimeric RNA, single nucleotide polymorphisms (SNPs), and allele-specific expression analysis in pig skeletal muscle.

Chimeric RNA molecules, also known as fusion transcripts, are composed of exons from two genes located at different genomic loci[18,19]. In the human genome, at least 4–5% of tandem genes are occasionally transcribed into chimeric proteins, suggesting that chimeric RNAs production is a common event with the potential to generate hundreds of additional proteins[20]. The presence of chimeric RNAs augments the number of transcriptional events and complexity of a given genome. Chimeric RNAs are suspected to function in cancer cells[21,22], as well as in normal cells and tissues[18,23,24]. In a recent study, we identified a set of chimeric RNAs in pigs[19]. To our knowledge, our report was the first study on chimeric RNAs in mammalian skeletal muscle.

Biomarkers and information regarding allele-specific expression (ASE) associated with muscle growth are important in animal breeding. SNPs are the most abundant type of DNA sequence polymorphism and serve as powerful genetic markers in pig breeding[25–28]. A well-known example of a porcine SNP is the nonconservative R200Q substitution mutation in the protein kinase, AMP-activated, gamma 3 non-catalytic subunit (*PRKAG3*) gene, which is associated with high glycogen content in pig skeletal muscle[29]. ASE analysis is used to detect allelic imbalance in transcription and assess *cis*-regulatory variation[30,31]. At least 30% of genes are influenced by to ASE, which has a considerable impact on gene expression[32]. The RNA-seq approach provides an effective method for comprehensively identifying SNPs and ASE variants in transcribed regions of the genome.

In this study, we used high-throughput transcriptome sequencing to systematically explore transcriptional events associated with prenatal skeletal muscle development in pigs. We first carried out systematic identification and characterization of protein coding genes and chimeric RNAs. Subsequently, we analyzed SNPs and ASE in prenatal skeletal muscle of Tongcheng pigs. This study provides a resource of chimeric RNAs, SNPs, and ASE that illuminates the molecular events underlying prenatal porcine skeletal muscle development and allows the development of molecular markers for pig breeding.

Results and Discussion

Global expression analysis of protein coding genes in prenatal skeletal muscle. Samples of prenatal skeletal muscle from Tongcheng pigs were analyzed using RNA-seq with a paired-end sequencing strategy on an Illumina HiSeq 2000 instrument. A total of 55.02 million 90-bp pair-end high-quality reads were obtained, of which 83.9% were mapped to *Sus scrofa* genome assembly 10.2. RPKM values were calculated to allow measurement of the expression levels of protein coding genes. Using RPKM >0.1 as a threshold, we detected 14,810 protein coding genes (PCGs) (Table S1), accounting for 68.54% of the PCGs included in the Ensembl release 78 mart database, indicating that most known PCGs were expressed in prenatal porcine skeletal muscle, while confirming that RNA-seq was an effective method for identifying PCGs with low expression levels. The read coverage of the RNA-seq data and the expression levels of the PCGs in the *Sus scrofa* reference genome are shown in Fig. 1. The PCG expression distribution is shown in Fig. 2A. In prenatal skeletal muscle, 69.9% of PCGs (10,347/14,810) were weakly expressed with RPKM <5, while only 1.4% (211/14,810) of PGCs were abundantly expressed with RPKM ≥ 100. Additionally, 25 highly expressed PCGs with RPKM values greater than 1000 were detected (Table 1). Gene ontology (GO) analysis of the 200 PCGs with the greatest transcript abundance revealed that genes associated with muscle development and contraction, such as *ACTC1*, *TNNC2*, *ACTA1*, *TNNC1*, *MYL3*, *ACTA2, MYH3*, and *MYL1*, were significantly enriched as expected; this phenomenon could be explained by the formation of the majority of muscle fibers during secondary myogenesis[6]. Genes involved in translational elongation (*EEF1G*, *EEF1B2*, *EEF2, EIF4G2*), ribosome biogenesis (*RPS* and *RPL* family genes), and regulation of ATPase activity (*NDUFA4*, *ND4L*, *NDUFB10*, *COX3*, *ND5*, *ND2*, *ND3*, *CYTB*, *ATP6*), which play essential roles in protein synthesis and fulfilling the energy requirements of prenatal skeletal muscle development, were significantly enriched (Fig. 2B). Two widely used housekeeping genes, β-actin (*ACTB*) and glyceraldehyde-3-phosphate dehydrogenase (*GAPDH*)[33], were also highly expressed in prenatal skeletal muscle.

Chimeric RNAs expressed in prenatal skeletal muscle. Based on our transcriptome sequencing data, we identified chimeric RNAs associated with prenatal skeletal muscle development using the ChimeraScan[34] and FusionMap[35] programs. We detected 535 and 351 potential chimeric RNAs (including 163 RNAs detected by both programs) using ChimeraScan (Table S2) and FusionMap (Table S3), respectively (Fig. 3A). Of the 163 chimeric RNAs detected by both programs, 36.8% ($n=60$) were intrachromosomal fusions, 62.0% ($n=101$) were adjacent fusions, and only 1.2% ($n=2$) were interchromosomal fusions (Fig. 3B). According to a previous study[36], we classified the 101 adjacent fusions into four categories: 10 read-through transcripts, 45 convergent transcripts, 36 divergent transcripts, and 10 overlapping transcripts (Fig. 3B). The distribution of chimeric RNAs in the *Sus scrofa* genome is shown in Fig. 1. We found that 94.5% (154/163) of chimeric RNAs had canonical splice sites and obeyed the GT/AG rule, implying that chimeric RNAs were mainly formed by trans-splicing and had properties similar to those of protein coding genes to some extent. Indeed, previous studies have demonstrated that chimeric RNAs have the potential to be translated into functional proteins[18,37]. GO analysis showed that the parental genes of the chimeric RNAs were mainly involved in regulation of cellular process, system development, positive regulation of biological process, cell differentiation, and regulation of cell proliferation (Fig. 3C). These findings suggest that the identified chimeric RNAs likely play important roles in prenatal porcine skeletal muscle development.

To determine whether homologues of the chimeric RNAs identified in the current study exist in other species, we aligned them to chimeric transcripts from the human, mouse, and fruit fly genomes in the ChiTaRS2.1 database[38]. The alignment sequences were retained only when at least 20 nt of either side of the fusion junction could be mapped. Unfortunately, we found that only 10 and 2 of the chimeric RNAs identified in pigs had homologues in the human and mouse transcriptomes, respectively (Table S4), while no chimeric RNA homologues were identified in the fruit fly transcriptome. These findings suggest that chimeric RNAs in pigs exhibit high species specificity.

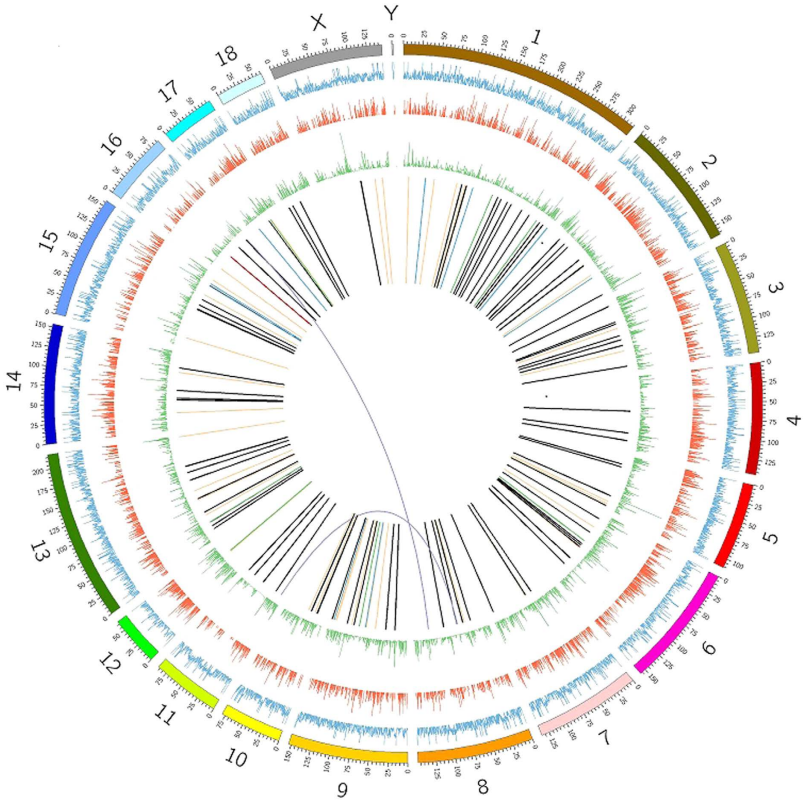

Figure 1. Transcriptome sequencing in prenatal porcine skeletal muscle. Chromosome ideograms are shown in the outer layer. The transcriptome sequencing coverage is shown in the first middle layer. Expression levels of genes are shown in the second middle layer. The SNP distribution is shown in the third middle layer. Chimeric RNAs are shown in the central layer. The chimeric RNA ssc-chimeric-113 is shown in red.

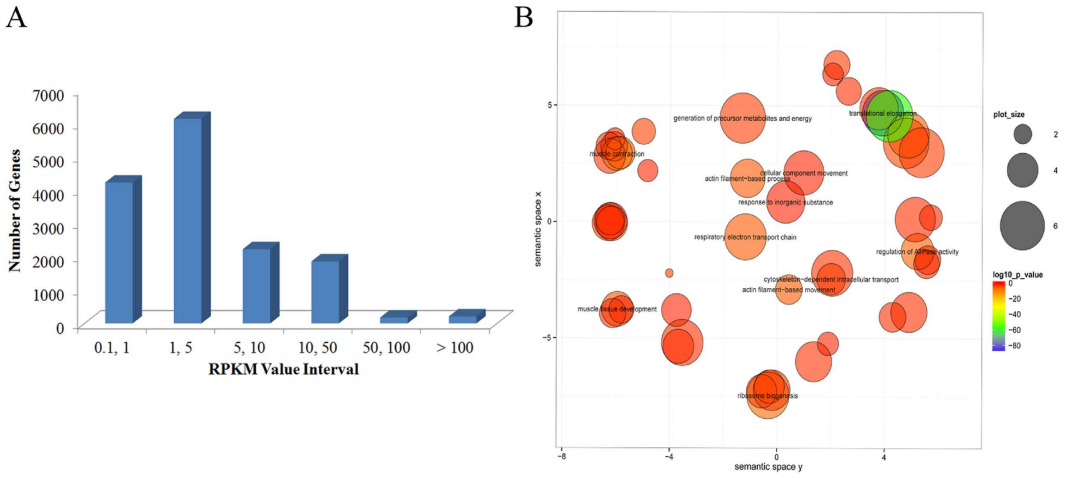

Figure 2. Analysis of protein coding genes in prenatal porcine skeletal muscle. (**A**) Distribution of detected protein coding genes with different expression levels. (**B**) GO biological process categories of the 200 most highly expressed genes.

Validation of chimeric RNAs. To validate the reliability of the group of identified chimeric RNAs, we selected 29 chimeric RNAs for RT-PCR verification in the same prenatal porcine skeletal muscle used for RNA sequencing analysis. The primers were designed to span the fusion junction of the chimeric RNAs. The vast majority of selected chimeric RNAs (20/29) were amplified by RT-PCR and confirmed by direct sequencing (Figure S1, Table S5). The consistency of the RT-PCR and prediction results suggests that the group of identified chimeric RNAs is sufficiently reliable for further research.

Ensembl Gene ID	Gene symbol	Description	RPKM
ENSSSCG00000018082	MT-CO3	Cytochrome c oxidase subunit 3	7,133.76
ENSSSCG00000018075	MT-CO1	Cytochrome c oxidase subunit 1	6,221.94
ENSSSCG00000018081	ATP6	ATP synthase subunit a	5,381.04
ENSSSCG00000018078	MT-COII	Cytochrome c oxidase subunit 2	4,344.78
ENSSSCG00000021943			3,431.54
ENSSSCG00000004803	ACTC1	Actin, alpha, cardiac muscle 1	2,598.12
ENSSSCG00000004489	EEF1A	Eukaryotic translation elongation factor 1 alpha 1	2,469.99
ENSSSCG00000007799	HUMMLC2B	Myosin light chain, phosphorylatable, fast skeletal muscle	2,409.61
ENSSSCG00000017581	COL1A1	Collagen type I alpha 1	2,348.98
ENSSSCG00000016157	MYL1	Myosin light chain 1/3, skeletal muscle isoform	2,247.39
ENSSSCG00000018087	MT-ND4	NADH-ubiquinone oxidoreductase chain 4	2,207.78
ENSSSCG00000018094	CYTB	Cytochrome b	2,166.85
ENSSSCG00000007424	TNNC2	Cytochrome c oxidase subunit 1	1,824.34
ENSSSCG00000010190	ACTA1	Actin, alpha 1, skeletal muscle	1,821.53
ENSSSCG00000018084	ND3	NADH-ubiquinone oxidoreductase chain 3	1,686.45
ENSSSCG00000000694	GAPDH	Glyceraldehyde-3-phosphate dehydrogenase	1,423.94
ENSSSCG00000018065	MT-ND1	NADH-ubiquinone oxidoreductase chain 1	1,403.79
ENSSSCG00000014540	FTH1	Ferritin, heavy polypeptide 1	1,303.34
ENSSSCG00000018092	NADH6	NADH-ubiquinone oxidoreductase chain 6	1,189.82
ENSSSCG00000018069	MT-ND2	NADH-ubiquinone oxidoreductase chain 2	1,166.88
ENSSSCG00000006558	RPS27	Ribosomal protein S27	1,138.51
ENSSSCG00000025883	RPL37	Ribosomal protein L37	1,065.38
ENSSSCG00000005316	TPM2	Tropomyosin 2 (beta)	1,030.23
ENSSSCG00000005003	LOC100736624	40S ribosomal protein S29	1,026.09
ENSSSCG00000015103	RPS25	Ribosomal protein S25	1,022.35

Table 1. The most highly expressed protein coding genes (RPKM >1000) in prenatal skeletal muscle.

Subsequently, we focused on ssc-chimeric-113, a chimeric product generated from ENSSSCG00000024947 and *NDUFS4*. ssc-chimeric-113 was highly expressed in the results from the FusionMap (ranking 6th with 151 seed counts) and ChimeraScan (ranking 8th with a score of 139) analyses. The *NDUFS4* gene (NADH dehydrogenase (ubiquinone) Fe-S protein 4, 18kDa (NADH-coenzyme Q reductase)) is highly expressed in skeletal muscle and potentially related to intramuscular fat deposition in pigs[39]. A genome-wide association study (GWAS) showed that a single nucleotide polymorphism site in *NDUFS4* was significantly associated with loin muscle area[40], implying that *NDUFS4* might play an important role in skeletal muscle development. The ENSSSCG00000024947 and *NDUFS4* genes are both located on chromosome 16, but on different strands. Our transcriptome sequencing data confirmed that ssc-chimeric-113 was abundantly expressed, as evidenced by 37 spanning reads across the fusion junction (Fig. 4A). This fusion junction was also confirmed using a dataset containing the transcriptome sequences of 9 different tissues in Guizhou pigs (data not shown). To verify the bioinformatics results, we performed PCR amplification of the prenatal skeletal muscle RNA used in transcriptome analysis, yielding a fragment 363 bp in length (Fig. 4B). Sanger sequencing showed that this PCR product was a fragment of ssc-chimeric-113 cDNA (Fig. 4C). BLAT of this sequence to *S. scrofa* genome assembly 10.2 showed that nucleotides 1–192 mapped onto the plus-strand of chromosome 16 at positions 34854150–34854341 in exon 2 of ENSSSCG00000024947, whereas nucleotides 191–363 mapped onto the minus-strand of chromosome 16 at positions 34963257–34963429 in exon 2 of *NDUFS4* (Fig. 4D). These results verified the existence of ssc-chimeric-113.

Identifying SNPs in prenatal skeletal muscle. Whole-transcriptome RNA sequencing is an effective strategy for identifying polymorphisms in the genome, especially in transcribed regions. This approach has been used to identify candidate SNPs in exonic regions associated with traits of interest, including growth and meat quality[41,42]. To our knowledge, no such studies have been performed in porcine skeletal muscle at any developmental stage.

We identified 106,457 high quality SNPs in transcripts expressed in prenatal skeletal muscle (Table S6). The number of SNPs within each chromosome was directly proportional to chromosome length and gene number. Chromosome 1 contained the most SNPs, whereas chromosome 16 contained the fewest SNPs (Fig. 5A). The proportion of substitution transitions (A/G and C/T, 73.91%) was much higher than the proportion of transversions (A/C, A/T, G/C, and G/T; 26.09%). The frequency of A/G transition (37.2%) was similar to that of C/T transition (36.6%). Among transversions, the frequency of each type was approximately 7%, with the exception of A/G transition, for which the frequency was 4.6%. The transition:transversion ratio was 2.83:1 (Fig. 5B), which was similar to values reported in other species[41,43]. We found that 12,643 annotated genes contained one or more SNPs. The average number of SNPs per gene was 10.2, while 71.0% of genes had fewer than 10 SNPs. Interestingly,

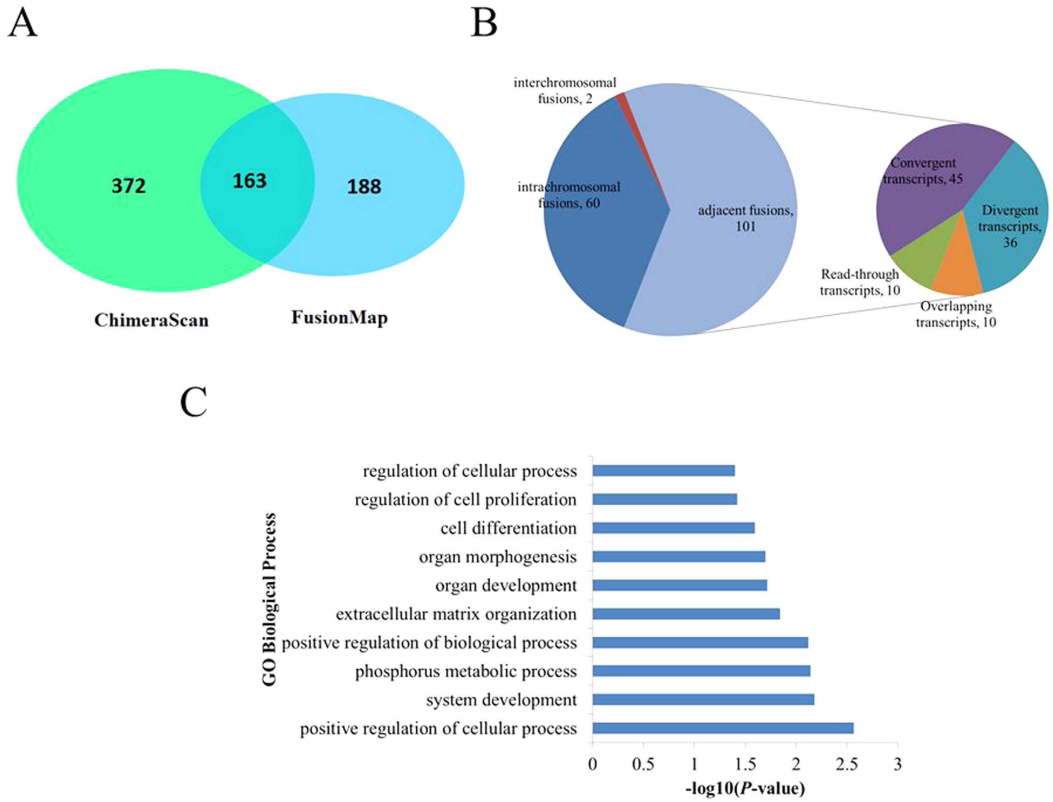

Figure 3. Identification of chimeric RNAs in prenatal porcine skeletal muscle. (**A**) Numbers of chimeric RNAs identified by ChimeraScan and FusionMap. (**B**) Classification of chimeric RNAs. (**C**) GO biological process analysis of the parental genes of chimeric RNAs.

we found that 808 genes harbored more than 25 SNPs, implying that these genes exhibited high diversity. These results suggest that these 808 genes might be particularly susceptible to artificial selection and were helpful for understanding population diversity (Fig. 5C). We also compared the identified SNPs with the *S. scrofa* dbSNP database (Build 140); 93.6% of the variants (99,602 SNPs) were deposited in the dbSNP database, indicating the high quality and reliability of our SNP analysis. At the same time, we detected 6,955 novel SNPs. Our results have increased the number of known SNPs in *S. scrofa*.

SNP annotation and function analysis. The distribution of the discovered SNPs within various genomic features was analyzed using Ensembl's Variant Effect Predictor[44]. Of the SNPs present in coding regions, 6,095 were nonsynonymous, whereas 16,237 were synonymous. The ratio of nonsynonymous to synonymous SNPs was approximately 0.37 (6,095/16,237). We also identified 23,047 SNPs located at 5′- or 3′-UTR regions and 29,147 SNPs in intronic regions (Fig. 5D). In addition, we detected 26 SNPs in termination codons and 222 SNPs in splice sites, which may affect transcript splicing and thus potentially affect protein products and their functions (Table 2, Table S6). A large proportion of SNPs identified fell into the intronic and intergenic regions, providing evidence for the incomplete annotation status of the current swine reference genome and suggesting that comprehensive exploration of the transcriptome profiles of pigs is merited.

Non-synonymous coding SNPs were further analyzed because they might result in amino acid substitution and thus affect protein activity. We carried out GO and KEGG enrichment analysis to investigate the putative functions of 1804 genes containing nonsynonymous SNPs. The results of these analyses revealed that 132 GO biological process terms were significantly enriched in the set of 1804 genes containing nonsynonymous SNPs ($p < 0.05$). These genes containing nonsynonymous SNPs were mainly involved in the response to DNA damage stimulus, DNA repair, the cellular response to stress, DNA metabolic processes, and the cell cycle (Fig. 6A). Interestingly, muscle development-related GO terms, including muscle cell development, skeletal muscle organ development, skeletal muscle tissue development, and muscle fiber development, were also significantly enriched in the set of genes containing nonsynonymous SNPs. This finding might be explained by the high expression levels of muscle development-related genes in prenatal skeletal muscle. These results demonstrate that our strategy is a powerful method of identifying SNP biomarkers associated with growth and meat quality traits. We found that a set of 641 genes containing nonsynonymous SNPs was significantly enriched in 14 KEGG pathways, including ECM-receptor interaction, focal adhesion, butanoate metabolism, and fatty acid metabolism ($p < 0.01$) (Fig. 6B). Moreover, we identified 1,046 SNPs in the set of chimeric RNAs, of which 988 SNPs (94.5%), including 95 nonsynonymous SNPs and 295 synonymous SNPs, were annotated in the dbSNP database and thus might be considered as candidate markers for studying the functions of chimeric RNAs in pigs.

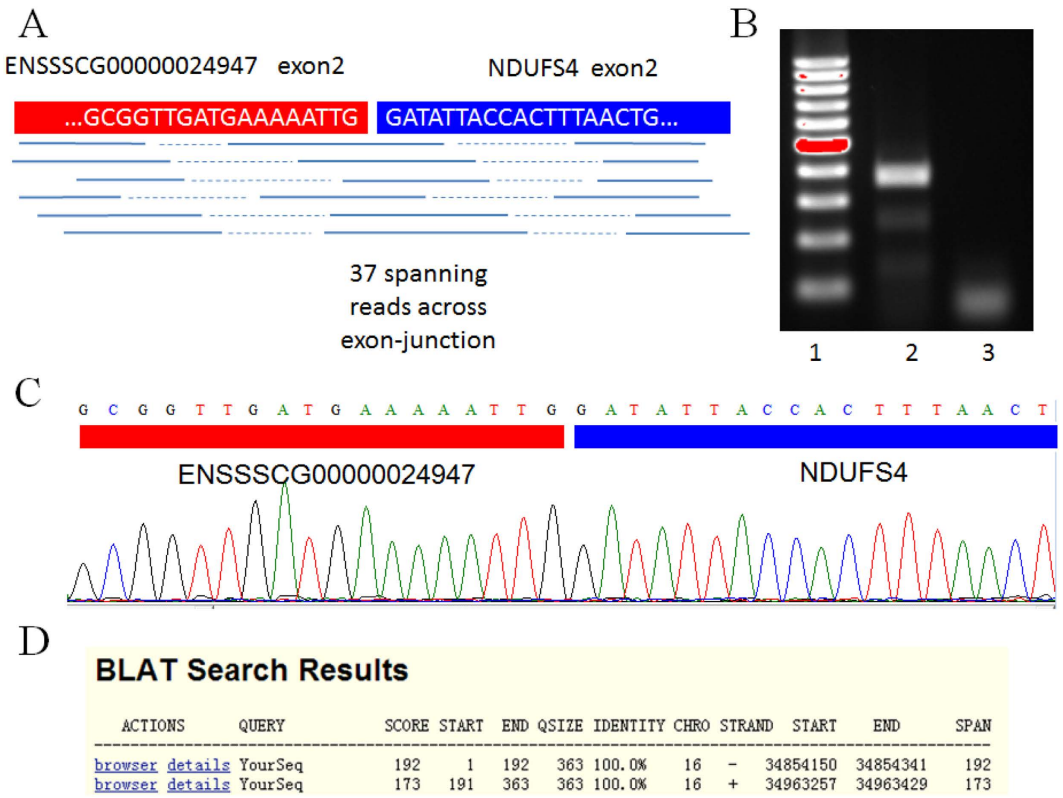

Figure 4. Validation of ssc-chimeric-113. (**A**) Detection of ssc-chimeric-113 via transcriptome sequencing. The "grep" command was used to identify 37 reads spanning the exon-junction. (**B**) Validation of ssc-chimeric-113 by PCR amplification and electrophoresis. Lane 1: marker. Lane 2: electrophoresis result. Lane 3: no template control. (**C**) Validation of the ssc-chimeric-113 breakpoint using Sanger sequencing. (**D**) BLAT of the Sanger sequencing result on *Sus scrofa* genome assembly 10.2 (http://genome.ucsc.edu/cgi-bin/hgBlat).

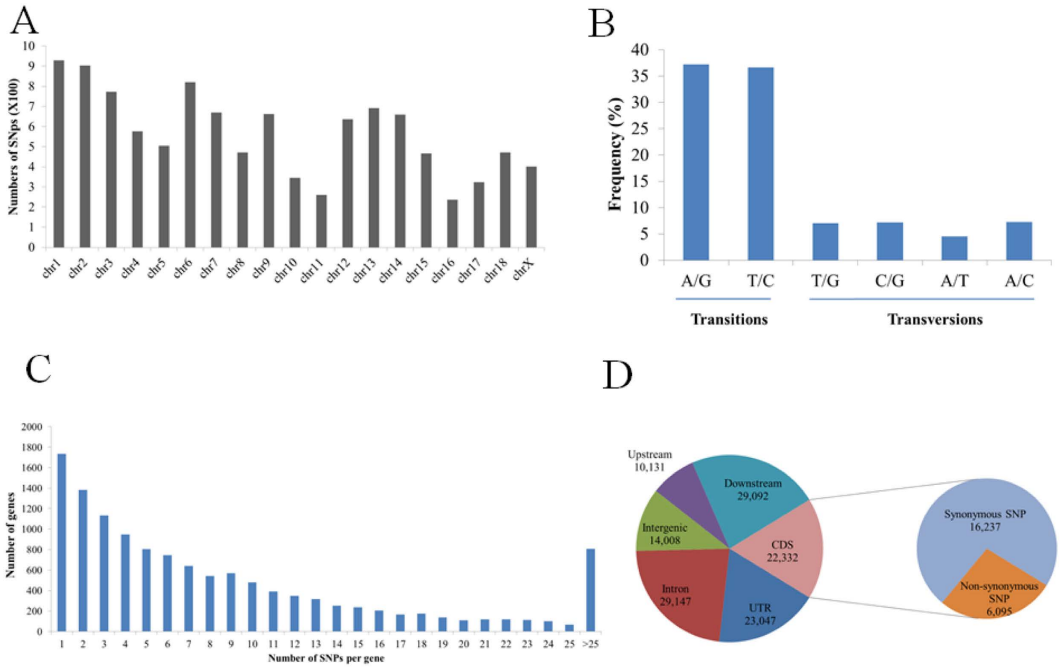

Figure 5. SNP identification in porcine prenatal skeletal muscle. (**A**) SNP distribution in porcine chromosomes. (**B**) Frequency of different substitution types in the identified SNPs. (**C**) Distribution of the number of SNPs per gene. (**D**) Distribution of SNPs in different genomic regions.

SNP annotation	Number of SNPs
Total SNPs	106,457
SNPs in annotated genes	12,643
UTR	23,047
Intron	29,147
Intergenic	14,008
Non-synonymous SNPs	6,095
Synonymous SNP	16,237
CDS	22,332
Splice region	222
Termination codons	26
Upstream	10,131
Downstream	29,092

Table 2. Annotation and classification of putative SNPs.

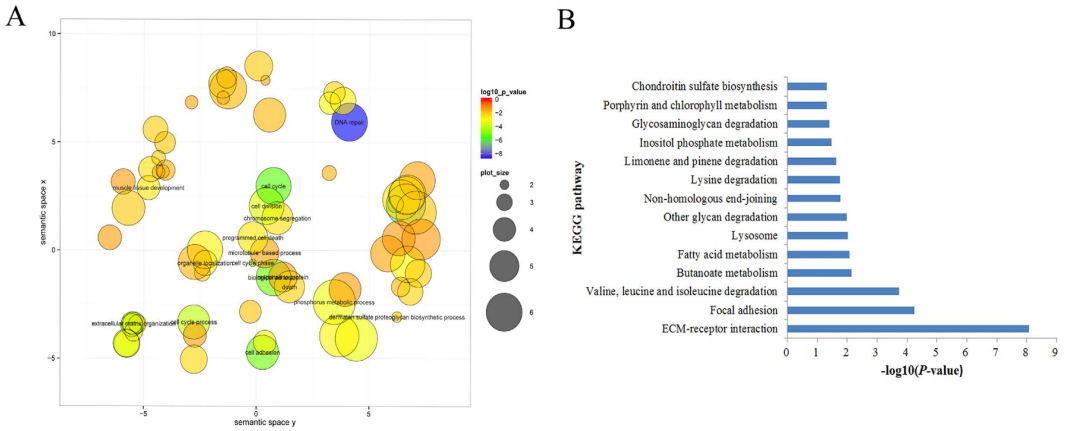

Figure 6. Functional annotation of genes containing nonsynonymous SNPs. (**A**) GO biological process analysis results. (**B**) KEGG pathway analysis.

Next, we queried the set of 106,457 high-quality SNPs to determine their presence in *S. scrofa* quantitative trait loci (QTLs) deposited in the AnimalQTLdb[45]. We counted the numbers of SNPs located in QTLs associated with production traits. There were 94,839 SNPs (89.09%) located within 685 QTL regions related to 90 production-related traits (Table S7). For example, 46,301 SNPs were located in 78 QTL regions for body weight at birth, whereas 45,809 SNPs were located in 181 QTL regions for average daily gain. The high proportion of SNPs located within QTLs for production-related traits indicates that our analysis is an effective strategy for detecting candidate quantitative trait nucleotides responsible for genetic variability influencing production traits.

ASE analysis in prenatal skeletal muscle. Gene expression is influenced by *cis*- and *trans*-regulatory genetic variation. Genome-wide ASE analysis is an effective method for inferring the existence of *cis*-regulatory variants[30,46]. In this study, the ASEReadCounter tool was used to retrieve allele counts. Subsequently, a binomial test and Benjamini-Hochberg false discovery rate (FDR) correction were performed to identify ASE variants. The allelic distribution ratios, defined as the ratio of the abundance of the non-reference allele to the sum of the abundance of the non-reference allele and that of the reference allele, are shown in Fig. 7A. The analysis revealed that 11,300 heterozygous SNPs exhibited allelic imbalance (allelic ratios >0.65 or <0.35 and FDR <0.05) (Fig. 7B, Table S8), of which 845 SNPs were heterozygous-derived nonsynonymous variants, including 138 SNPs classified by Sift[47] as "deleterious". We then tested whether sites exhibiting ASE were more likely to be nonsynonymous SNPs, revealing a significant difference in the proportion of nonsynonymous SNPs with significant ASE and that of the entire set of analyzed SNPs (Fisher's exact test, $p < 0.001$), which suggested an enrichment of nonsynonymous variants in ASE. In addition, we detected 131 ASE SNPs located in the chimeric RNAs.

GWASs have reported a large number of SNPs associated with phenotypes of various economic traits in pigs. To illuminate the functional impacts of SNPs, we examined whether SNPs reported by previous GWASs exhibited ASE in our study. Surprisingly, we identified 4 ASE variants reported by previous GWASs. Of these, SNP rs335265740 in the 3′-UTR of nuclear receptor subfamily 3, group C, member 1 (glucocorticoid receptor) (*NR3C1*) was associated with relative flare fat[48]. SNPs rs45433464 (located in stearoyl-CoA desaturase (delta-9-desaturase) (*SCD*)) and rs81215882 (located in phosphoglucomutase 1 (*PGM1*)) were significantly associated with average daily weight gain[49,50]. SNP rs80863153 in the 5′-UTR of aldehyde dehydrogenase 18 family, member A1 (*ALDH18A1*) was associated with hematological traits[51]. The analysis of allelic imbalance suggested that *cis*-regulatory variations might be associated with phenotypic divergence in pigs. Additionally, the

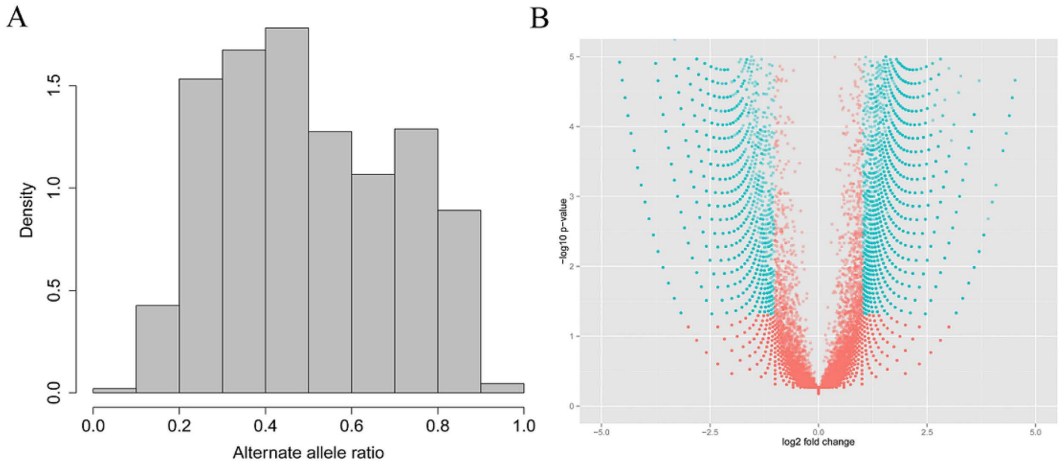

Figure 7. Allele-specific expression analysis of heterozygous SNPs in porcine prenatal skeletal muscle. (**A**) Distribution of the alternate allele ratio for all heterozygous sites expressed in prenatal porcine skeletal muscle. (**B**) Volcano plot analysis of heterozygous SNPs with allelic imbalance. The blue points are SNPs showing significant ASE, whereas red points are SNPs with no significant ASE.

rs340729607 (T/A) variant introduced a premature stop codon in exon 7 of mitochondrial ribosomal protein L1 (*MRPL1*), indicating nonsense-mediated decay. Unfortunately, we did not detect ASE variants in the chimeric RNAs generated from genes reported to influence economically important traits in the GWASs.

Conclusion

In this study, we first performed a comprehensive analysis of chimeric RNAs, SNPs, and ASE variants in prenatal skeletal muscle using RNA-seq. We identified 163 high-confidence chimeric RNAs potentially associated with porcine prenatal skeletal muscle development. The existence of chimeric RNAs in pigs broadened our knowledge of the complexity of mammalian transcriptomes and illuminated the gene interaction network that functions during skeletal muscle development. The newly discovered SNPs and ASE variants expand the catalog of genetic variants in pigs and will facilitate molecular marker-assisted selection in pig breeding and relevant GWASs. This study provides a foundation for studies aimed at revealing the complex transcriptional mechanisms underlying prenatal skeletal muscle development in mammals, as well as a molecular marker resource that can be utilized in pig breeding. However, further studies are needed to decipher the biological functions of the chimeric RNAs, SNPs, and ASE variants identified in this study.

Materials and Methods

Animals and sample collection. All animal experiments were performed according to the procedures defined by national and local animal welfare bodies and were approved by the Institutional Animal Care and Use Committee at the Institute of Animal Science, Chinese Academy of Agricultural Sciences. The *longissimus dorsi* muscle samples were isolated from Tongcheng two pig fetuses (one male and one female) at 5 time points (gestational days 50, 55, 60, 65, and 75). All samples were maintained in liquid nitrogen until use.

RNA extraction and high-throughput paired-end RNA-sequencing. Total RNA was extracted using Trizol (Invitrogen, Carlsbad, CA, USA) following the manufacturer's protocol. RNA integrity was measured using an Agilent 2100 Bioanalyzer. Only samples with RNA Integrity Number (RIN) values greater than eight were used for sequencing. Library construction and Solexa sequencing were performed using methods described previously[17] according to the manufacturer's instructions (Illumina, USA). Briefly, total RNA from samples collected at five time points were pooled into a single sample in equal proportions. PolyA + RNA was purified from total RNA using magnetic oligo(dT) and fragmented. First-strand cDNA was generated using Random Primer p(dN)6 and Superscript III (Invitrogen, Carlsbad, CA, USA), after which second-strand cDNA synthesis and adaptor ligation were performed. cDNA fragments of 240–310 bp were isolated. The library was sequenced on the Illumina HiSeq 2000 platform to generate 90-bp paired-end reads.

Transcriptome mapping and expression quantification. After filtering low quality reads, clean reads were mapped against the *S. scrofa* reference genome (assembly 10.2)[52] using Tophat version 2.1.0[53] with default options. Assignment of reads to genes was performed using htseq-count[54]. The expression levels of protein coding genes were measured as numbers of reads per kilobase of exon per gene per million mapped reads (RPKM)[55].

Identification of *Sus scrofa* chimeric RNAs. ChimeraScan (version 0.4.3)[34] and FusionMap (version 2015-03-31)[35] software was used to identify chimeric RNAs with the Ensembl release 78 reference genome (*S. scrofa* assembly 10.2)[52] using default parameters. Classification of adjacent chimeric RNAs was performed as described in a previous study[36]: (1) read-through genes, adjacent genes in the same orientation; (2) diverging genes, adjacent genes in opposite orientations whose 5′ ends are in close proximity; (3) convergent genes, adjacent

genes in opposite orientations whose 3′ ends are in close proximity; (4) overlapping genes, adjacent genes who share common exons. For conservation analysis of the pig identified chimeric RNAs, we downloaded sets of human, mouse, and fruit fly chimeric transcripts from the ChiTaRS 2.1 database[38] and aligned the pig chimeric RNAs to those from other species using the BLAST program (Basic Local Alignment Search Tool, version 2.2.2 6+) with default parameters[56] (at least 20 nt of the sequence on either side of the fusion junction must have been mapped). The "grep" command was used to search the reads spanning the fusion junction sequences of the ENSSSCG00000024947-NDUFS4 chimeric RNA from the fastq files of the transcriptome data as described previously[57].

Reverse transcription polymerase chain reaction. To validate the identified chimeric RNAs, total RNA from prenatal porcine skeletal muscle was reverse-transcribed into cDNA using the RevertAid First Strand cDNA Synthesis Kit (MBI Fermentas, Vilnius, Lithuania) according to the manufacturer's protocols. The chimeric cDNA containing the fusion junction was amplified by PCR as follows: an initial denaturation at 94 °C for 3 min, followed by 34 cycles of denaturation at 95 °C for 15 s, annealing at 60 °C for 30 s, and elongation at 72 °C for 20 s, and a final extension for 5 min at 68 °C. The PCR products were confirmed by direct sequencing.

SNP identification and annotation. The Genome Analysis Toolkit (GATK version 3.3) package[58] was used for SNP discovery according to the best practice recommendations regarding the RNA-seq variant analysis workflow of the Broad Institute (https://www.broadinstitute.org/gatk/guide/best-practices?bpm=RNAseq). Stringent parameters were used to minimize detection of false-positive SNPs. Clusters of at least 3 SNPs within a window of 35 bases were filtered out. Hard filtering values, including Fisher strand values (FS > 30.0), qual by depth values (QD < 2.0), and read depth value (DP < 5), were selected. SNPs located on unplaced scaffolds and mitochondria were not included in this study. SNP annotation was performed using in-house Perl scripts and Ensembl's Variant Effect Predictor[44].

ASE analysis. The ASEReadCounter tool[59] in the GATK package was used to retrieve the allele counts of heterozygous SNP sites. Heterozygous sites with individual allele read depth less than 3 and total (both alleles) read depth less than 10 were filtered out. A binomial test and Benjamini-Hochberg FDR correction were performed. Cut-off criteria of allele ratio >0.65 or <0.35 and FDR <0.05 were used to identify significant allelic imbalances.

Gene ontology and KEGG pathway analysis. Gene ontology (GO) and KEGG pathway enrichment analyses were performed with the Database for Annotation, Visualization, and Integrated Discovery (DAVID) website (http://david.abcc.ncifcrf.gov/)[60]. Because of the poor pig Ensembl annotations in the DAVID database, we converted the pig Ensembl gene IDs into human gene symbol IDs with Biomart (http://www.biomart.org/) before performing the GO and KEGG pathway analyses. We set the EASE value to 0.05 for the enrichment analysis. Significantly enriched GO biological process terms were summarized and visualized using REVIGO (http://revigo.irb.hr/)[61].

References

1. Huard, J., Li, Y. & Fu, F. H. Muscle injuries and repair: current trends in research. *The Journal of Bone & Joint Surgery* **84,** 822–832 (2002).
2. Dwyer, C., Fletcher, J. & Stickland, N. Muscle cellularity and postnatal growth in the pig. *Journal of Animal Science* **71,** 3339–3343 (1993).
3. Picard, B., Lefaucheur, L., Berri, C. & Duclos, M. J. Muscle fibre ontogenesis in farm animal species. *Reproduction Nutrition Development* **42,** 415–431 (2002).
4. Rehfeldt, C., Fiedler, I., Dietl, G. & Ender, K. Myogenesis and postnatal skeletal muscle cell growth as influenced by selection. *Livestock Production Science* **66,** 177–188 (2000).
5. Wigmore, P. & Stickland, N. Muscle development in large and small pig fetuses. *Journal of Anatomy* **137,** 235 (1983).
6. Du, M. *et al.* Fetal programming of skeletal muscle development in ruminant animals. *Journal of Animal Science* **88,** E51–E60 (2010).
7. Zhao, Y. *et al.* Dynamic transcriptome profiles of skeletal muscle tissue across 11 developmental stages for both Tongcheng and Yorkshire pigs. *BMC genomics* **16,** 377 (2015).
8. Zhao, X. *et al.* Comparative analyses by sequencing of transcriptomes during skeletal muscle development between pig breeds differing in muscle growth rate and fatness. *PLoS One* **6,** e19774 (2011).
9. Lindberg, J. & Lundeberg, J. The plasticity of the mammalian transcriptome. *Genomics* **95,** 1–6 (2010).
10. Gustincich, S. *et al.* The complexity of the mammalian transcriptome. *The Journal of physiology* **575,** 321–332 (2006).
11. Tang, Z. *et al.* LongSAGE analysis of skeletal muscle at three prenatal stages in Tongcheng and Landrace pigs. *Genome Biol* **8,** R115 (2007).
12. Qin, L. *et al.* Integrative analysis of porcine microRNAome during skeletal muscle development. *PLoS One* **8,** e72418 (2013).
13. Hou, X. *et al.* Discovery of MicroRNAs associated with myogenesis by deep sequencing of serial developmental skeletal muscles in pigs. *PLoS One* **7,** e52123 (2012).
14. Zhou, B., Liu, H., Shi, F. & Wang, J. MicroRNA expression profiles of porcine skeletal muscle. *Animal genetics* **41,** 499–508 (2010).
15. Hou, X. *et al.* Comparison of skeletal muscle miRNA and mRNA profiles among three pig breeds. *Molecular Genetics and Genomics* 1–15 (2015).
16. Tang, Z. *et al.* Integrated analysis of miRNA and mRNA paired expression profiling of prenatal skeletal muscle development in three genotype pigs. *Scientific Reports* **5,** 15544 (2015).
17. Zhao, W. *et al.* Systematic identification and characterization of long intergenic non-coding RNAs in fetal porcine skeletal muscle development. *Sci Rep* **5,** 8957 (2015).
18. Frenkel-Morgenstern, M. *et al.* Chimeras taking shape: potential functions of proteins encoded by chimeric RNA transcripts. *Genome Res* **22,** 1231–1242 (2012).
19. Ma, L. *et al.* Identification and analysis of pig chimeric mRNAs using RNA sequencing data. *BMC genomics* **13,** 429 (2012).
20. Parra, G. *et al.* Tandem chimerism as a means to increase protein complexity in the human genome. *Genome Res* **16,** 37–44 (2006).
21. Gingeras, T. R. Implications of chimaeric non-co-linear transcripts. *Nature* **461,** 206–211 (2009).
22. Pane, F. *et al.* BCR/ABL genes and leukemic phenotype: from molecular mechanisms to clinical correlations. *Oncogene* **21,** 8652–8667 (2002).
23. Akiva, P. *et al.* Transcription-mediated gene fusion in the human genome. *Genome research* **16,** 30–36 (2006).

24. Li, H., Wang, J., Mor, G. & Sklar, J. A neoplastic gene fusion mimics trans-splicing of RNAs in normal human cells. *Science* **321,** 1357–1361 (2008).
25. Schroyen, M. *et al.* The MC4R c.893G >A mutation: A marker for growth and leanness associated with boar taint odour in Belgian pig breeds. M*eat Sci* **101C,** 1–4 (2014).
26. Short, T. H. *et al.* Effect of the estrogen receptor locus on reproduction and production traits in four commercial pig lines. *J Anim Sci* **75,** 3138–3142 (1997).
27. Duijvesteijn, N. *et al.* A genome-wide association study on androstenone levels in pigs reveals a cluster of candidate genes on chromosome 6. *BMC Genet* **11,** 42 (2010).
28. Ren, J. *et al.* A 6-bp deletion in the TYRP1 gene causes the brown colouration phenotype in Chinese indigenous pigs. *Heredity (Edinb)* **106,** 862–868 (2011).
29. Milan, D. *et al.* A mutation in PRKAG3 associated with excess glycogen content in pig skeletal muscle. *Science* **288,** 1248–1251 (2000).
30. Pastinen, T. Genome-wide allele-specific analysis: insights into regulatory variation. *Nature Reviews Genetics* **11,** 533–538 (2010).
31. Crowley, J. J. *et al.* Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance. *Nature Genetics* **47,** 353–U102 (2015).
32. Ge, B. *et al.* Global patterns of cis variation in human cells revealed by high-density allelic expression analysis. *Nature genetics* **41,** 1216–1222 (2009).
33. Nygard, A.-B., Jørgensen, C. B., Cirera, S. & Fredholm, M. Selection of reference genes for gene expression studies in pig tissues using SYBR green qPCR. *BMC molecular biology* **8,** 67 (2007).
34. Iyer, M. K., Chinnaiyan, A. M. & Maher, C. A. ChimeraScan: a tool for identifying chimeric transcription in sequencing data. *Bioinformatics* **27,** 2903–2904 (2011).
35. Ge, H. *et al.* FusionMap: detecting fusion genes from next-generation sequencing data at base-pair resolution. *Bioinformatics* **27,** 1922–1928 (2011).
36. Maher, C. A. *et al.* Chimeric transcript discovery by paired-end transcriptome sequencing. *Proceedings Of the National Academy Of Sciences Of the United States Of America* **106,** 12353–12358 (2009).
37. Kannan, K. *et al.* Recurrent chimeric RNAs enriched in human prostate cancer identified by deep sequencing. *Proceedings of the National Academy of Sciences* **108,** 9172–9177 (2011).
38. Frenkel-Morgenstern, M. *et al.* ChiTaRS: a database of human, mouse and fruit fly chimeric transcripts and RNA-sequencing data. *Nucleic acids research* **41,** D142–D151 (2013).
39. Chen, Q. *et al.* Molecular characterization and expression analysis of NDUFS4 gene in m. longissimus dorsi of Laiwu pig (Sus scrofa). *Mol Biol Rep* **40,** 1599–1608 (2013).
40. Fan, B. *et al.* Genome-Wide Association Study Identifies Loci for Body Composition and Structural Soundness Traits in Pigs. *Plos One* **6,** e14726 (2011).
41. Djari, A. *et al.* Gene-based single nucleotide polymorphism discovery in bovine muscle using next-generation transcriptomic sequencing. *BMC Genomics* **14,** 307 (2013).
42. Fowler, K. E. *et al.* Genome wide analysis reveals single nucleotide polymorphisms associated with fatness and putative novel copy number variants in three pig breeds. *BMC Genomics* **14,** 784 (2013).
43. Xu, X. L. *et al.* Identification of somatic mutations in human prostate cancer by RNA-Seq. *Gene* **519,** 343–347 (2013).
44. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26,** 2069–2070 (2010).
45. Hu, Z. L., Fritz, E. R. & Reecy, J. M. AnimalQTLdb: a livestock QTL database tool set for positional QTL information mining and beyond. *Nucleic Acids Research* **35,** D604–D609 (2007).
46. Ramayo-Caldas, Y. *et al.* Liver transcriptome profile in pigs with extreme phenotypes of intramuscular fatty acid composition. *BMC genomics* **13,** 1 (2012).
47. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature protocols* **4,** 1073–1081 (2009).
48. Terenina, E. *et al.* Association study of molecular polymorphisms in candidate genes related to stress responses with production and meat quality traits in pigs. *Domest Anim Endocrinol* **44,** 81–97 (2013).
49. Li, X. *et al.* Analyses of porcine public SNPs in coding-gene regions by re-sequencing and phenotypic association studies. *Mol Biol Rep* **38,** 3805–3820 (2011).
50. Onteru, S. K. *et al.* Whole Genome Association Studies of Residual Feed Intake and Related Traits in the Pig. *PLoS One* **8,** e61756 (2013).
51. Zhang, F. *et al.* Genome-wide association studies for hematological traits in Chinese Sutai pigs. *BMC Genet* **15,** 41 (2014).
52. Groenen, M. A. *et al.* Analyses of pig genomes provide insight into porcine demography and evolution. *Nature* **491,** 393–398 (2012).
53. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14,** R36 (2013).
54. Anders, S., Pyl, P. T. & Huber, W. HTSeq–A Python framework to work with high-throughput sequencing data. *Bioinformatics*, btu638 (2014).
55. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods* **5,** 621–628 (2008).
56. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of molecular biology* **215,** 403–410 (1990).
57. Panagopoulos, I., Gorunova, L., Bjerkehagen, B. & Heim, S. The "Grep" Command But Not FusionMap, FusionFinder or ChimeraScan Captures the CIC-DUX4 Fusion Gene from Whole Transcriptome Sequencing Data on a Small Round Cell Tumor with t (4; 19)(q35; q13). *Plos One* **9,** e99439 (2014).
58. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20,** 1297–1303 (2010).
59. Castel, S. E., Levy-Moonshine, A., Mohammadi, P., Banks, E. & Lappalainen, T. Tools and best practices for allelic expression analysis. *bioRxiv* **016097,** (2015).
60. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols* **4,** 44–57 (2008).
61. Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PloS one* **6,** e21800 (2011).

Acknowledgements

We are grateful to Sanping Xu from the Tongcheng Animal Husbandry Bureau in Hubei Province for the animal preparation and sample collection. This work was supported by the National Natural Science Foundation of China (31172189 and 31501931), the Key Projects in the National Science & Technology Pillar Program during the Twelfth Five-year Plan Period (2015BAD03B02-2), the National Key Basic Research Program of China (2015CB943101 and 2014CB138504), the National Key Project (2016ZX08009003-006-003) and the Agricultural Science and Technology Innovation Program (ASTIP-IAS05, ASTIP-IAS16).

Author Contributions

R.Z. and K.L. conceived and designed the study project, Y.Y. and Z.T. analyzed the data, X.F. and K.X. performed the experiments, Y.M. contributed materials. Y.Y., R.Z. and Z.T. wrote the paper. K.L. contributed to result discussion and data interpretation. All authors read and approved the final manuscript.

Additional Information

Accession code: The RNA-seq raw data from this study have been deposited in the NCBI Sequence Read Archive with accession number SRP066035. (http://www.ncbi.nlm.nih.gov/Traces/sra/).

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Yang, Y. *et al.* Transcriptome analysis revealed chimeric RNAs, single nucleotide polymorphisms and allele-specific expression in porcine prenatal skeletal muscle. *Sci. Rep.* **6**, 29039; doi: 10.1038/srep29039 (2016).