DNA Research **18**, 343–351, (2011)
Advance Access Publication on 5 July 2011

doi:10.1093/dnares/dsr021

Analysis of Muscle and Ovary Transcriptome of *Sus scrofa*: Assembly, Annotation and Marker Discovery

Qinghua Nie[1,2], Meixia Fang[3], Xinzheng Jia[1,2], Wei Zhang[1,2], Xiaoning Zhou[1,2], Xiaomei He[1,2], and Xiquan Zhang[1,2,*]

Department of Animal Genetics, Breeding and Reproduction, College of Animal Science, South China Agricultural University, Guangzhou, Guangdong 510642, People's Republic of China[1]; Guangdong Provincial Key Lab of Agro-Animal Genomics and Molecular Breeding, Guangzhou, Guangdong 510642, People's Republic of China[2] and Department of Laboratory Animal Science, Medical College of Jinan University, Guangzhou, Guangdong 510632, People's Republic of China[3]

*To whom correspondence should be addressed. Tel. +86 20-85285703. Fax. +86 20-85280740. Email: xqzhang@scau.edu.cn

Edited by Minoru Ko
(Received 13 April 2011; accepted 8 June 2011)

Abstract

Pig (*Sus scrofa*) is an important organism for both agricultural and medical purpose. This study aims to investigate the *S. scrofa* transcriptome by the use of Roche 454 pyrosequencing. We obtained a total of 558 743 and 528 260 reads for the back-leg muscle and ovary tissue each. The overall 1 087 003 reads give rise to 421 767 341 bp total residues averaging 388 bp per read. The *de novo* assemblies yielded 11 057 contigs and 60 270 singletons for the back-leg muscle, 12 204 contigs and 70 192 singletons for the ovary and 18 938 contigs and 102 361 singletons for combined tissues. The overall GC content of *S. scrofa* transcriptome is 42.3% for assembled contigs. Alternative splicing was found within 4394 contigs, giving rise to 1267 isogroups or genes. A total of 56 589 transcripts are involved in molecular function (40 916), biological process (38 563), cellular component (35 787) by further gene ontology analyses. Comparison analyses showed that 336 and 553 genes had significant higher expression in the back-leg muscle and ovary each. In addition, we obtained a total of 24 214 single-nucleotide polymorphisms and 11 928 simple sequence repeats. These results contribute to the understanding of the genetic makeup of *S. scrofa* transcriptome and provide useful information for functional genomic research in future.

Key words: transcriptome; *Sus scrofa*; 454 sequencing; SNPs; SSRs

1. Introduction

As a predominant domestic animal, pig (*Sus scrofa*) not only provides us plenty meat, but also is an important model organism for medical research. Understanding the genetic principle of growth and reproduction traits is helpful for *S. scrofa* production and also has scientific significance for basic biology and human medicine.[1] Notable progress has been achieved to identify causative genes or single-nucleotide polymorphisms (SNPs) underlie complex traits in the past decades. A recessive missense mutation of *S. scrofa* ryanodine receptor (*RYR1*) gene was proved to induce malignant hyperthermia and halothane sensitivity in *S. scrofa*, which was effectively applied in selection program by many breeding farms.[2] In 2003, an SNP in intron 3 of insulin-like growth factor 2 was identified as a causative quantitative trait nucleotide for porcine muscle growth.[3]

With the accomplishment of *S. scrofa* genome sequencing, it is realistic and efficient to identify genes or genetic markers underlying complex traits at whole-genome or transcriptome level. In 2005, a 0.66× coverage *S. scrofa* genome including 3.84 million shotgun sequences was generated from Hampshire, Yorkshire, Landrace, Duroc and ErHuaLian pigs, giving rise to 2.08 billion nucleotide bases, indicating that the *S. scrofa* genome is much closer to human than mouse does.[4] Later, the complete *S. scrofa* genome was available online in 2009 (November 2009, SGSC Sscrofa9.2/susScr2; http://genome.ucsc.edu/cgi-bin/hgBlat). From 4.8 million whole-genome shotgun sequences, 98 151 SNPs were predicted with one sequence representing the polymorphism, and most SNPs were confirmed by testing in three purebred boar lines and wild boar.[5] Recently, a high-density SNP chip (Illumina Porcine 60K + SNP iSelect Beadchip) was designed and supplied for commercial use, which included 64 232 SNPs after reliable validation.[6] The developed SNP chip was subsequently used in whole-genome association analysis to identify genes for body composition and structural soundness traits in *S. scrofa*.[7]

Based on genomic background, investigation of *S. scrofa* transcriptome is realistic and extremely useful for identification of candidate genes account for quantitative traits at the global level. The cDNA microarray or gene chip is a common tool for transcriptomic analysis. Confirmed by real-time RT–PCR and association analyses, some candidate genes and SNP markers for *S. scrofa* reproduction traits have been identified with microarray profiling by Affymetrix Porcine Genechip^{TM}.[8] With the use of *S. scrofa* whole-genome 70-mer oligonucleotide microarray, 62 expression quantitative traits loci (eQTLs) were successfully identified from loin muscle tissue through global genome-wide linkage analysis.[9] Nevertheless, cDNA microarray has some limitations as it fails to recognize new genes (transcripts) and sequence variations.

RNA sequencing (RNA-seq) is a new but efficient technology for the thorough investigation on transcriptome. With the rapid development of second-generation sequencing, RNA-seq becomes more efficient and less costive by some latterly developed platforms, i.e. Roche 454, Illuminate Solexa GA IIx, Life Technology SOLID, Helicos Biosciences tSMS and others.[10] The Roche 454 can generate long reads and is generally used in transcriptome analysis in human,[11,12] mammals,[13] insects,[14] fish,[15] plants[16,17] and microorganisms.[18]

Until now, reports on *S. scrofa* transcriptome by RNA-seq technology are very limited. In this study, we performed 454 pyrosequencing of muscle and ovary tissues to characterize *S. scrofa* transcriptome and to identify potential markers for growth and reproduction traits.

2. Materials and methods

2.1. *Animal and RNA preparation*

One landrace female at 6-month age was subject to transcriptomic analysis of *S. scrofa*. Landrace is a typical commercial *S. scrofa* strain and widely used in domestic livestock production. The animal was slaughtered quickly to collect two tissues of the ovary and back-leg muscle. The fresh tissues were steeped in liquid nitrogen immediately after collection, and then kept at −80°C refrigerator (Thermo Forma, USA) for preservation before use. Trizol (Invitrogen, CA, USA) was used to isolate total RNA following the manufacturer's protocols.

2.2. *Construction of cDNA library and 454 sequencing*

Approximately 10-μg total RNA was delivered to Shanghai Majorbio Bio-pharm Biotechnology Co., Ltd. (Shanghai, China) for the construction of a cDNA library. RNA quality was assessed by 260/280 and 260/230 ratios with the Agilent 2100 Bioanalyzer. The SMART cDNA library construction kit (Clontech, Mountain View, CA, USA) was used to construct the cDNA library of the muscle and ovary tissues following the manufacturer's protocol step by step. cDNA was sheared by nebulization and DNA bands (500–800 bp) were extracted from gel after agarose gel electrophoresis. The obtained DNA was purified, blunt ended, ligated to adapters and finally small fragments were removed. The quality control of a double DNA library was performed using High Sensitivity Chip (Agilent Technologies). The concentration was examined by TBS 380 Fluorometer. One-plate whole run sequencing was performed on the GS FLX Titanium chemistry (Roche Diagnostics, Indianapolis, IN, USA) by Shanghai Majorbio Bio-pharm Biotechnology Co., Ltd. following the manufacturer's protocol.

2.3. *Bioinformatic analysis*

2.3.1. Reads trimming and assemble For each of the sequencing reads, low-quality bases and the sequencing adapter were trimmed using LUCY and SeqClean. The remained 454 reads of the ovary and back-leg muscle were first assembled using the Newbler software using default parameters. The combined reads of the ovary and back-leg muscle were also assembled using the Newbler software.

2.3.2. Assemble of EST To improve the assembler quality, we collected the ESTs of *S. scrofa* from PigEST database (http://pigest.ku.dk/download/index.html), which included 398 837 Pub EST sequences and 823 871 Sino-Danish *S. scrofa* EST sequences. All the ESTs from Roche 454 and PigEST databases were used to run the final assemble of *S. scrofa*. We chosen the trans-Abyss software for the final assemble, which assembled the ESTs through combining the results of different Kmer parameters. The unigenes with >100 bp in length were used for the subsequent analysis. Moreover, all ESTs were mapped to the *S. scrofa* draft genome version 9 (download from

Ensembl ftp server http://asia.ensembl.org/info/data/ftp/index.html). ESTs were considered to be mapped successfully if it had over 95% identities to the corresponding genome sequences.

2.3.3. Transcriptome annotation The unigenes were compared with the protein non-redundant database using BlastX[19] with E-value $< 1.0 \times 10^{-5}$ (E-values $<1.0 \times 10^{-5}$ were considered as a significant level). Gene ontology (GO) terms[20] were extracted from the best hits obtained from the BlastX against the nr database (E-value $\leq 1.0 \times 10^{-6}$) using blast2go and then were sorted for the GO categories using in-house perl scripts. The metabolic pathway was performed using Kyoto Encyclopaedia of genes and genomes.

2.3.4. Expression analysis EST reads from the ovary and back-leg muscle were mapped to a unigene sequence using the SSAHA software, respectively. The expression of each unigene was calculated using the numbers of reads with a specific match. Genes with different expression were identified using R package DGEseq.[21]

2.3.5. Bioinformatic mining of microsatellites and SNP makers The unigene sequences were screened for microsatellites using software MISA (MicroSAtellite, http://pgrc.ipk-gatersleben.de/misa/). All the 454 reads were mapped to the unigenes using SSAHA. The SNPs were extracted using VarScan with the default parameter only when both alleles were detected from 454 reads. The released *S. scrofa* genome (download from Ensembl ftp server http://asia.ensembl.org/info/data/ftp/index.html) was used to confirm and locate the SNPs. Only those SNPs that both specifically match a certain genomic region and have a minor allelic frequency no less than 20% are included in analysis.

2.4. Data deposition

The Roche 454 reads of *S. scrofa* by this study are now available from Animal Genomics databases by National Animal Genome Research Program (NAGRP, USA) with URL of http://www.animalgenome.org/repository/pub/SCAU2011.0502/.

3. Results

3.1. 454 sequencing and assembly

In this study, we constructed two cDNA libraries and subsequently obtained two sets of transcriptomic reads for the back-leg muscle and ovary each. The schematic of 454 EST analyses is showed by Fig. 1. For the back-leg muscle, Roche 454 sequencing yielded a total of 558 743 reads with a total nucleotide size of 219 021 745 bp, giving rise to average 392 bp per read (Table 1; Supplementary Fig. S1). For the ovary, meanwhile, we obtained 528 260 reads for 202 745 596 bp in total, giving rise to average 384 bp per read (Table 1; Supplementary Fig. S1).

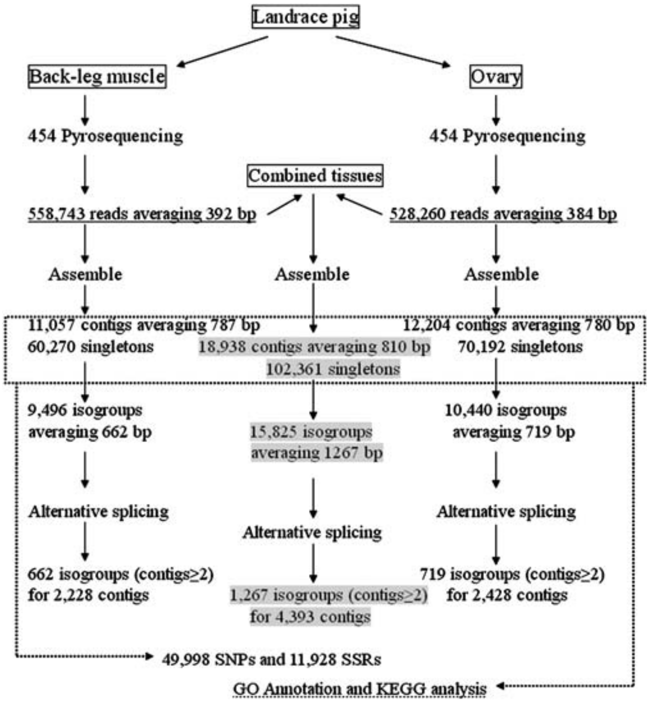

Figure 1. Schematic of 454 EST analyses. The steps include 454 sequencing, assembly of reads into contigs and isogroups, GO annotation, KEGG analysis and discovery of SNPs and SSRs.

Table 1. Draft sequence data by 454 sequencing

Types	Muscle (RL4)	Ovary (RL10)	Combined tissues
Number of reads	558 743	528 260	1 087 003
Total residues (bp)	219 021 745	202 745 596	421 767 341
Smallest (bp)	27	23	23
Largest (bp)	772	813	813
Average length (bp)	392	384	388

By assemble analysis, we obtained 71 327 ESTs (11 057 contigs averaging 787 bp and 60 270 singletons) for the back-leg muscle and 82 396 ESTs (12 204 contigs averaging 780 bp and 70 192 singletons) for the ovary, as well as 121 299 ESTs (18 938 contigs averaging 810 bp and 102 361 singletons) for combined tissues (Table 2; Supplementary Fig. S2). Most of these contigs distributed in the 401–1500-bp region, and about half of them distributed in 401–700 bp for each of the back-leg muscle (51.35%), ovary (51.79%) and combined tissue (49.62%) (Table 3; Supplementary Fig. S2).

Contigs were further assembled to 9496 isogroups averaging 662 bp for the back-leg muscle, 10 440 isogroups averaging 719 bp for the ovary and 15 825 isogroups averaging 1267 bp for combined tissues (Table 2). In average, each isogroup has 1.16, 1.17 and 1.20 contigs for the back-leg muscle (11 057 contigs for 9496 isogroups), ovary (12 204 contigs for 10 440 isogroups) and combined tissues (18 938 contigs for 15 825 isogroups), respectively. As far as all contigs were concerned, the concentration of A, T, C and G was 29.0, 28.7, 21.1 and 21.2%, respectively, giving rise to the overall GC content of 42.3% for *S. scrofa* transcriptome.

By comparison, our original 454 sequence with the *S. scrofa* genome, 246 628 ESTs could match the genome, and among them, 216 739 specifically locate on one region and 29 889 locate on two or more regions. As far as different chromosomes are compared, chromosome 1 (23 237 or 10.72%), 4 (17 087 or 7.88%), 2 (16 957 or 7.82%), 14 (16 724 or 7.72%) and 13 (15 738 or 7.26%) contained more transcripts than others, whereas X (4495 or 2.07%) and mitochondrial DNA (mtDNA; 1665 or 0.77%) had very fewer EST (Supplementary Fig. S3).

Table 2. Summary on assemble analysis

Types	Muscle (RL4)	Ovary (RL10)	Combined tissues
Num of contigs	11 057	12 204	18 938
Smallest (bp)	42	44	42
Largest (bp)	3540	3462	4218
Total length (bp)	8 703 645	9 517 255	15 332 944
Average length (bp)	787	780	810
Num of isogroups	9496	10 440	15 825
Num of isogroups (contigs $\geq$ 2)	662	719	1267
Singleton	60 270	70 192	102 361
Total	71 327	82 396	121 299

3.2. *Alternative splicing*

More contigs than isogroups are found because some contigs or called isocontigs are attributed to the same isogroups due to alternative splicing (Tables 2 and 4). There are 7.0% (662 of 9496), 6.9% (719 of 10 440) and 8.0% (1267 of 15 825) isogroups have no less than two contigs in the back-leg muscle, ovary and combined tissues, respectively. The alternative-spliced isogroups in the back-leg muscle, ovary and combined tissues averagely have 3.4, 3.5, and 3.5 isocontigs, respectively (Table 4).

3.3. *GO assignments*

A total of 56 589 transcripts of *S. scrofa* were assigned for GO analysis based on matches with sequences whose functions were known previously. Among these transcripts, 45 846 transcripts were successfully annotated with confident matches. As many as 38 563 transcripts are involved in biological process, including cellular process (34 217 transcripts with percentages of 16.98%), metabolic process (29 471; 14.62%), biological regulation (16 944; 8.41%), regulation of biological process (16 002; 7.94%), multicellular organismal process (10 148; 5.04%), response to stimulus (10 023; 4.97%), localization (9680; 4.80%), cellular component organization or biogenesis (9276; 4.60%), developmental process (8610; 4.27%), establishment of localization (8393; 4.16%), signalling (7083; 3.51%), positive regulation of biological process (6576; 3.26%), negative regulation of biological process (6487; 3.22%) and signalling process (5517; 2.74%), as well as other activities (23 100; 11.46%) (Fig. 2A).

Moreover, 35 787 transcripts are subject to a cellular component and could be divided into cell (33 486; 24.06%), cell part (33 483; 24.06%), organelle (25 767; 18.52%), organelle part (16 894; 12.14%), macromolecular complex (13 151; 9.45%), membrane-enclosed lumen (6469; 4.65%), extracellular region (5378; 3.86%), extracellular region part (3457; 2.48%) and others (1082; 0.78%)

Table 3. Statistics of contigs by 454 sequencing

Length	Muscle		Ovary		Combined tissues	
	Numbers	Per cent (%)	Numbers	Per cent (%)	Numbers	Per cent (%)
1–100	12	0.11	11	0.09	12	0.06
101–400	225	2.03	309	2.53	411	2.17
401–700	5678	51.35	6320	51.79	9397	49.62
701–1000	2813	25.44	3050	24.99	4760	25.13
1001–1500	1756	15.88	1928	15.80	3187	16.83
1501–2000	449	4.06	474	3.88	864	4.56
>2000	124	1.12	112	0.92	307	1.62
Total	11 057	100	12 204	100	18 938	100

(Fig. 2B). GO analysis also showed that 40 916 transcripts had potential molecular function, such as binding (36 550; 48.15%), catalytic activity (21 865; 28.8%), structural molecule activity (4385; 5.78%), transporter activity (3188; 4.2%), molecular transducer activity (2379; 3.13%), enzyme regulator activity (2337; 3.08%), transcription regulator activity (2224; 2.93%), nucleic acid binding transcription factor activity (1317; 1.74%), electron carrier activity (948; 1.25%) and others (714; 0.94%) (Fig. 2C).

Table 4. Variant transcripts by assemble analysis

No.[a]	Numbers of isogroups (contigs)		
	Muscle	Ovary	Combined tissues
2	452 (904)	498 (996)	857 (1714)
3	71 (142)	76 (152)	143 (286)
4	64 (128)	58 (116)	115 (230)
5	20 (40)	12 (24)	29 (58)
6	18 (36)	20 (40)	29 (58)
7	6 (12)	8 (16)	17 (34)
8	5 (10)	10 (20)	15 (30)
9	4 (8)	8 (72)	14 (28)
≥10	22 (529)	29 (644)	48 (1106)
In total (≥2)	662 (2228)	719 (2488)	1267 (4393)

[a]Numbers of contigs per isogroup.

3.4. *Metabolic pathways by KEGG analysis*

A total of 4268 transcripts are involved in 132 predicted KEGG metabolic pathways, and the numbers of transcripts in different pathways ranged from 1 to 1352. The top 20 pathways with EST numbers are shown in Table 5, and the highest of the number of transcripts is involved in the biosynthesis of secondary metabolites. Ten biosynthesis pathways included biosynthesis of alkaloids derived from histidine and purine (417), biosynthesis of alkaloids derived from ornithine, lysine and nicotinic acid (368), biosynthesis of alkaloids derived from the shikimate pathway (380), biosynthesis of alkaloids derived from terpenoid and polyketide (399), biosynthesis of secondary metabolites (1352), biosynthesis of plant hormones (625), biosynthesis of phenylpropanoids (466), biosynthesis of terpenoids and steroids (401) and biosynthesis of unsaturated fatty acids (104), as well as biosynthesis of ansamycins (7).

3.5. *Tissue-specific analysis for differentially expressed genes*

Comparison of gene expression by DEGseq showed that a total of 336 genes expressed in the back-leg muscle with a significantly higher level than that of the ovary. These genes are involved in biological process (160 genes), cellular components (96) and molecular function (80) (Supplementary Fig. S4). In addition, another 553 genes significantly expressed

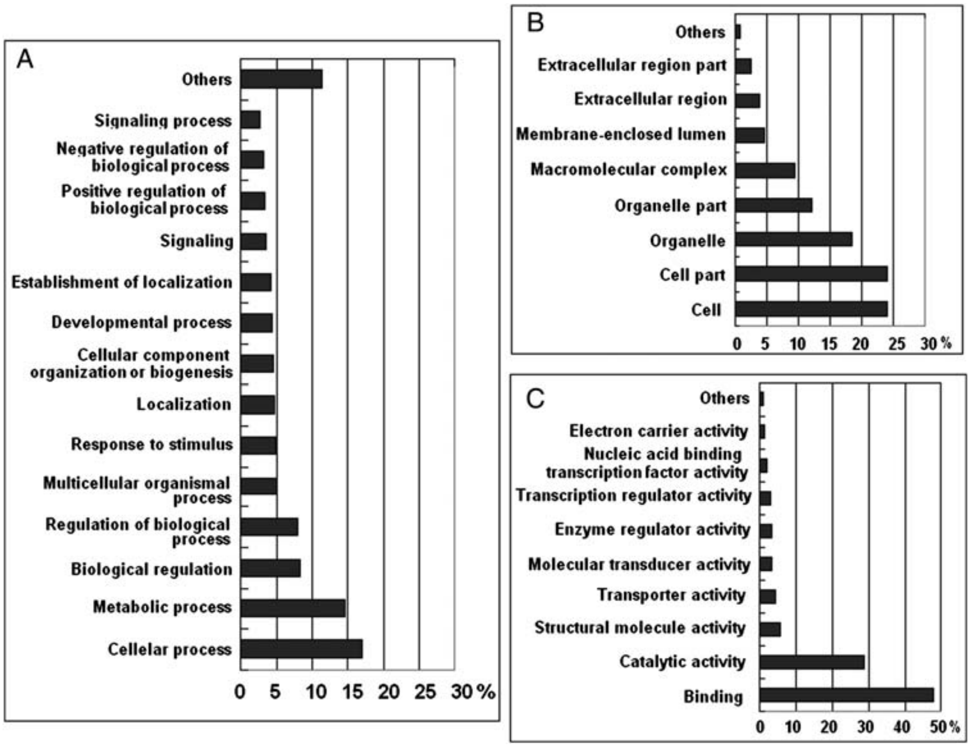

Figure 2. Functional classification of *S. scrofa* transcriptome. (A) GO: Biological process. (B) Cellular component. (C) GO: Molecular function. In some cases, one transcript or gene has multiple functions.

Table 5. The top 20 pathways with the highest EST numbers

No.	Pathways	Number of ESTs
1	Biosynthesis of secondary metabolites	1352
2	Oxidative phosphorylation	1111
3	Microbial metabolism in diverse environments	964
4	Purine metabolism	690
5	Biosynthesis of plant hormones	625
6	Biosynthesis of phenylpropanoids	466
7	Biosynthesis of alkaloids derived from histidine and purine	417
8	Pyrimidine metabolism	408
9	Biosynthesis of terpenoids and steroids	401
10	Biosynthesis of alkaloids derived from terpenoid and polyketide	399
11	Methane metabolism	399
12	Biosynthesis of alkaloids derived from shikimate pathway	380
13	Glycolysis/gluconeogenesis	379
14	Biosynthesis of alkaloids derived from ornithine, lysine and nicotinic acid	368
15	Glutathione metabolism	318
16	Pyruvate metabolism	268
17	Arginine and proline metabolism	261
18	Glycerophospholipid metabolism	222
19	Fatty acid metabolism	221
20	Valine, leucine and isoleucine degradation	218

in the ovary rather than the back-leg muscle. These genes are involved in biological process (295 genes), cellular components (130) and molecular function (128) (Supplementary Fig. S5).

3.6. *Single-nucleotide polymorphisms*

By excluding those that either could not specifically match the *S. scrofa* genome or had minor allele frequencies lower than 20%, we totally obtained 24 214 SNPs which comprised various substitutions of A-G (8484), C-T (8697), A-C (1635), A-T (1991), C-G (1325) and G-T (2082). The ratio of transitions (17 181) to transversions (7033) is ~2.44. Except for 125 SNPs in mtDNA, all others (24 089) are in nucleic DNA including 1–18 autosomes (23 481) and X chromosome (608). The distribution of SNPs in each chromosome is described in Table 6.

Table 6. Distribution of SNPs in the *S. scrofa* genome

Chromosomes[a]	Counts
1	2358
2	1901
3	1467
4	1693
5	1133
6	1492
7	1922
8	1260
9	1415
10	968
11	533
12	1073
13	1418
14	2008
15	1086
16	580
17	680
18	494
X	608
MT	125
Total	24 214

[a]Chromosomes 1–18 and X indicate 18 autosomes and sex chromosome each, whereas MT indicates mtDNA.

3.7. *Simple sequence repeats or microsatellites*

We obtained 11 928 simple sequence repeats (SSRs), of which 45.72% were di-nucleotide repeats (5453), followed by 36.06% tri-nucleotide repeats (4301) and 14.85% tetra-nucleotide repeats (1771), as well as 3.38% penta-nucleotide repeats (403) (Table 7). There are six types of di-nucleotide repeats, and among them $(GT)_n$, $(AC)_n$ and $(AT)_n$ are three predominant types with frequencies of 28.5, 27.6 and 23.12%, respectively (Supplementary Table S1). Among tri-nucleotide repeats, the frequencies of 20 SSR types seem to vary moderately from 0.19 to 17.11%, and the most common repeats are $(GTT)_n$ (17.11%) and $(ACC)_n$ (10.14%) (Supplementary Table S2). As many as 45 SSRs present for tetra-nucleotide repeats, and five of them [15.70% for $(ATTT)_n$, 14.85% for $(AAAT)_n$, 14.00% for $(GTTT)_n$, 12.70% for $(CTTT)_n$ and 12.37% for $(AAAC)_n$] are major ones with frequencies over 10% (Supplementary Table S3). Among 21 penta-nucleotide repeats types, $(GTTTT)_n$ (34.99%) and $(AAAAC)_n$ (20.60%) are two predominant types, followed by 10.42% for $(ATTTT)_n$, and the rest less than 10% (Supplementary Table S4).

4. Discussion

The obtained transcriptomic sequences by this study are useful for us to understand the genetic makeup of *S. scrofa* whole transcriptome, and to our knowledge, it is very limited until now. Even though the draft *S. scrofa* genome was released on

Table 7. Summary on microsatellite loci in *S. scrofa* transcriptome

Number of repeats	Di-nucleotide repeats	Tri-nucleotide repeats	Tetra-nucleotide repeats	Penta-nucleotide repeats
4	—	2511	1134	303
5	—	1004	386	74
6	1858	434	118	18
7	986	189	31	6
8	592	80	20	—
9	384	23	26	1
10	257	22	17	—
11	202	16	9	1
12	187	11	6	—
13	184	6	13	—
14	131	2	11	—
15	115	1	—	—
16	100	—	—	—
17	93	—	—	—
18	49	—	—	—
19	63	2	—	—
20	54	—	—	—
21	48	—	—	—
22	41	—	—	—
23	29	—	—	—
24	18	—	—	—
25	18	—	—	—
26	9	—	—	—
27	10	—	—	—
28	14	—	—	—
29	11	—	—	—
Total	5453	4301	1771	403

November 2009, very fewer articles could be found regarding the *S. scrofa* genome or trancriptome based on large-scale sequence data. The 454 pyrosequencing yield 1 087 003 reads in total, including 558 743 for the back-leg muscle and 528 260 for the ovary and 421 Mb nucleotide residues for *S. scrofa* transcriptome by this study. A recent research obtained 1 253 361 454 sequences for the skeletal muscle (701 695) and heart (551 666) and showed the reproducibility within 454 sequencing and cDNA microarray; however, further analyses of *S. scrofa* transcriptome with 454 sequences were not reported thereby.[13] Most transcripts (78.7%) identified by this study could match the released *S. scrofa* genome, and they scattered on chromosome 1–18, X and mtDNA. The bigger chromosomes seem to contain more transcripts compared with others. The obtained overall GC content of *S. scrofa* transcriptome was 42.3%, which was lower than the reported GC content in 5′UTR (59.2%), coding (49.6%), but a little bit higher than 3′UTR (41.8%) in the *S. scrofa* genome.[4] It seemed that more 3′UTR than other regions (coding and 5′UTR) were included in analyses, as far as only assembled contigs rather than singletons were concerned. The genome-wide average GC content of the human genome is 41%, varying from different chromosomes and regions.[22] It is known that the mouse has a slightly higher overall GC content (42%), but the distribution is tighter.[23] Over 1 million 454 reads and 121 299 ESTs by this study are useful resource for further research on *S. scrofa* functional genomics.

Both gene annotation and pathway analyses are helpful for us to predict potential genes and their functions at a whole-transcriptome level. In *S. scrofa* transcriptome, as discovered by this study, the predominant gene clusters are involved in the cellular process and metabolic process of biological process, the binding and catalytic activity of molecular function, as well as the cell, cell part and organelle of a cellular component. Similar results are found in European eel,[24] rainbow trout[15] and Red bugs.[14] Whereas, in Chickpea transcriptome, genes are predominantly involved in the protein metabolism of biological process and the transferase activity of molecular function, as well as the chloroplast of cellular component, which indicated notable differences between animals and plants.[25] In addition, we also predicted overall 4268 ESTs (or transcripts) that are involved in 132 predicted KEGG metabolic pathways, and two major pathways (biosynthesis of secondary metabolites and oxidative phosphorylation) comprised over 1000 ESTs. The predicted pathways altogether with gene annotation are useful for further investigation on gene function in future.

The differentially expressed genes in the back-leg muscle and ovary tissues are probably related to their metabolism and functions. In this study, we found 336 and 553 genes that were significantly expressed in the back-leg muscle and ovary, and they were involved in biological process, cellular components and molecular function by further gene annotation. A recent study identified a total of 306 differently expressed genes between muscle and heart tissue based on 1 253 361 Roche 454 reads and confirmed most genes by a microarray approach.[13] In this study, more differently expressed genes were found between back-leg muscle and ovary based on 1 087 003 reads. As far as different tissues are used by both studies, these differently expressed genes still require for further confirmation. As the muscle and ovary are crucial tissues for growth and reproduction, the identified genes with different expression are the potential candidate for growth and reproduction traits of *S. scrofa*.

Plenty expression SNPs (eSNPs) as identified by this study are valuable molecular markers for further research on *S. scrofa*. A total of 24 214 SNPs (minor allele frequencies ≥ 0.2) were found to specifically match the *S. scrofa* genome, in which 608 and 125 are in X chromosome and mtDNA, and the rest (23 481) in 18 autosomes. In general, these SNPs should be reliable eSNPs and act as candidate markers for identification of eQTL. In the genome level, a total of 98 151 SNPs were predicted based on 4.8 million whole-genome shotgun sequences.[5] Another study, moreover, discovered as many as 372 886 SNPs by sequencing with Illumina's Genome Analyzer (GA), and 62 621 loci of them were used to design the Illumina Porcine 60K + SNP iSelect Beadchip.[6] This SNP chip is useful for identification of candidate genes or QTLs underling quantitative traits such as body composition.[7]

SSRs, or microsatellite, are neutral molecular markers that wildly distribute in a genome. It was formerly proved that SSRs comprise 3% of the human genome, with the greatest contribution from di-nucleotide repeats (0.5%).[22] In this study, 45.72% of 11 928 SSRs are di-nucleotide repeats, follows by tri-nucleotide repeats (36.06%) and tetra-nucleotide repeats (14.85%), as well as penta-nucleotide repeats (3.38%). In addition to $(AC)_n$ and $(AT)_n$ of di-nucleotide repeats, and $(AAT)_n$ and $(AAC)_n$ of tri-nucleotide, $(GT)_n$ and $(GTT)_n$ also have high frequencies as indicated by this study, which is different from that of the human genome.[22] It is probably because $(GT)_n$ and $(GTT)_n$ are equal to $(AC)_n$ and $(ACC)_n$ each, since the reverse strands for some ESTs are used in *de novo* analysis. In fact, slight differences are found for SSRs among human, mouse and dog, as well as *S. scrofa*.[23]

In conclusion, we have demonstrated the muscle and ovary transcriptome of *S. scrofa* by the use of high-throughout 454 pyrosequencing. Our study obtained a set of 121 299 transcripts or ESTs and demonstrated some important features of *S. scrofa* transcriptome, such as GC content, gene annotation and pathways across whole transcriptome. In addition, we identified reliable markers of 24 214 SNPs and 11 928 SSRs. This study is helpful for understanding the genetic architecture of *S. scrofa* transcriptome and provides useful resource and markers for functional genomic research in future.

Supplementary data: Supplementary data are available at www.dnaresearch.oxfordjournals.org.

Funding

This research was supported by the National Major Special Projects on New Varieties Cultivation for Transgenic Organisms (2008ZX08006-005 and 2009ZX08009-145B) and the Important Projects in Key Fields in Guangdong and Hongkong, 2008 (2008A02).

References

1. Andersson, L. and Georges, M. 2004, Domestic-animal genomics: deciphering the genetics of complex traits, *Nat. Rev. Genet.*, **5**, 202–12.
2. Fujii, J., Otsu, K., Zorzato, F., et al. 1991, Identification of a mutation in porcine ryanodine receptor associated with malignant hyperthermia, *Science*, **253**, 448–51.
3. Van Laere, A.S., Nguyen, M., Braunschweig, M., et al. 2003, A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig, *Nature*, **425**, 832–6.
4. Wernersson, R., Schierup, M.H., Jørgensen, F.G., et al. 2005, Pigs in sequence space: a 0.66X coverage pig genome survey based on shotgun sequencing, *BMC Genomics*, **6**, 70.
5. Kerstens, H.H., Kollers, S., Kommadath, A., et al. 2009, Mining for single nucleotide polymorphisms in pig genome sequence data, *BMC Genomics*, **10**, 4.
6. Ramos, A.M., Crooijmans, R.P., Affara, N.A., et al. 2009, Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology, *PLoS One*, **4**, e6524.
7. Fan, B., Onteru, S.K., Du, Z.Q., Garrick, D.J., Stalder, K.J. and Rothschild, M.F. 2011, Genome-wide association study identifies loci for body composition and structural soundness traits in pigs, *PLoS One*, **6**, e14726.
8. Sun, X., Mei, S., Tao, H., et al. 2011, Microarray profiling for differential gene expression in PMSG-hCG stimulated preovulatory ovarian follicles of Chinese Taihu and Large White sows, *BMC Genomics*, **12**, 111.
9. Steibel, J.P., Bates, R.O., Rosa, G.J., et al. 2011, Genome-wide linkage analysis of global gene expression in loin muscle tissue identifies candidate genes in pigs, *PLoS One*, **6**, e16766.
10. Ozsolak, F. and Milos, P.M. 2011, RNA sequencing: advances, challenges and opportunities, *Nat. Rev. Genet.*, **12**, 87–98.
11. Sugarbaker, D.J., Richards, W.G., Gordon, G.J., et al. 2008, Transcriptome sequencing of malignant pleural mesothelioma tumors, *Proc. Natl Acad. Sci. USA*, **105**, 3521–6.
12. Wu, J.Q., Habegger, L., Noisa, P., et al. 2010, Dynamic transcriptomes during neural differentiation of human embryonic stem cells revealed by short, long, and paired-end sequencing, *Proc. Natl Acad. Sci. USA*, **107**, 5254–9.
13. Hornshøj, H., Bendixen, E., Conley, L.N., et al. 2009, Transcriptomic and proteomic profiling of two porcine tissues using high-throughput technologies, *BMC Genomics*, **10**, 30.
14. Bai, X., Mamidala, P., Rajarapu, S.P., Jones, S.C. and Mittapalli, O. 2011, Transcriptomics of the bed bug (*Cimex lectularius*), *PLoS One*, **6**, e16336.
15. Salem, M., Rexroad, C.E. 3rd., Wang, J., Thorgaard, G.H. and Yao, J. 2010, Characterization of the rainbow

trout transcriptome using Sanger and 454-pyrosequencing approaches, *BMC Genomics*, **11**, 564.

16. Bai, X., Rivera-Vega, L., Mamidala, P., et al. 2011, Transcriptomic signatures of ash (*Fraxinus* spp.) phloem, *PLoS One*, **6**, e16368.
17. Parchman, T.L., Geist, K.S., Grahnen, J.A., Benkman, C.W. and Buerkle, C.A. 2010, Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery, *BMC Genomics*, **11**, 180.
18. Poroyko, V., White, J.R., Wang, M., et al. 2010, Gut microbial gene expression in mother-fed and formula-fed piglets, *PLoS One*, **5**, e12459.
19. Altschul, S.F., Madden, T.L., Schaffer, A.A., et al. 1997, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, **25**, 3389–402.
20. Ashburner, M., Ball, C.A., Blake, J.A., et al. 2000, Gene ontology: tool for the unification of biology, *Nat. Genet.*, **25**, 25–9.
21. Wang, L., Feng, Z., Wang, X., Wang, X. and Zhang, X. 2010, DEGseq: an R package for identifying differentially expressed genes from RNA-seq data, *Bioinformatics*, **26**, 136–8.
22. International Human Genome Sequencing Consortium. 2001, Initial sequencing and analysis of the human genome, *Nature*, **409**, 860–921.
23. Mouse Genome Sequencing Consortium. 2002, Initial sequencing and comparative analysis of the mouse genome, *Nature*, **420**, 520–62.
24. Coppe, A., Pujola, J.M., Mase, G.E., et al. 2010, Sequencing, *de novo* annotation and analysis of the first *Anguilla anguilla* transcriptome: EeelBase opens new perspectives for the study of the critically endangered European eel, *BMC Genomics*, **11**, 635.
25. Garg, R., Patel, R.K., Tyagi, A.K. and Jain, M. 2011, *De novo* assembly of chickpea transcriptome using short reads for gene discovery and marker identification, *DNA Res.*, **18**, 53–63.