Genetics and Molecular Research

De novo assembly and characterization of farmed blue fox (*Alopex lagopus*) global transcriptome using Illumina paired-end sequencing

P.C. Guo[1]*, S.Q. Yan[1]*, S. Si[1], C.Y. Bai[1], Y. Zhao[1], Y. Zhang[1], J.Y. Yao[2] and Y.M. Li[1]

[1]College of Animal Science, Jilin University, Changchun, China
[2]College of Animal Science and Technology, Jilin Agricultural University, Changchun, China

*These authors contributed equally to this study.
Corresponding author: Y.M. Li
E-mail: li_ym@jlu.edu.cn

Genet. Mol. Res. 15 (1): gmr.15017603
Received September 9, 2015
Accepted November 27, 2015
Published March 24, 2016
DOI http://dx.doi.org/10.4238/gmr.15017603

ABSTRACT. The blue fox (*Alopex lagopus*), a coat-color variant of the Arctic fox, is a domesticated fur-bearing mammal. In the present study, transcriptome data generated from a pool of nine different tissues were obtained with Illumina HiSeq2500 paired-end sequencing technology. After filtering from raw reads, 32,358,290 clean reads were assembled into 161,269 transcripts and 97,252 unigenes by the Trinity fragment assembly software. Of the assembled unigenes, 37,967 were annotated in the National Center for Biotechnology Information (NCBI) Non-Redundant (NR) protein database and 26,264 in the Swiss-Prot database. Among the annotated unigenes, 24,839 and 24,267 were assigned using the Gene Ontology (GO) and euKaryotic Orthologous Groups (KOG) databases, respectively. Altogether, 17,057 unigenes were mapped onto 227 pathways using the Kyoto Encyclopedia of Genes and Genomes database. In addition, 6394 simple sequence repeats were identified by examining

12,965 unigenes (>1 kb), which could contribute to the development of molecular markers. This study generated transcriptome data for the blue fox that will promote further progress in expression profiling studies, and provide a good annotation basis for genomic studies.

Key words: Blue fox; Transcriptome; Assembly

INTRODUCTION

The Arctic fox (*Alopex lagopus*) lives under some of the most frigid conditions on the planet. It builds up its fat reserves in the autumn, which provide insulation during the winter and a source of energy when food is scarce. The blue fox, a color variant of the Arctic fox, has a uniform dark blue coat in the summer. In winter, its coat is lighter than in the summer, and is far more colorful than that of the Arctic fox (Våge et al., 2005). The exact mechanisms involved in the functional regulation of pigment and fat metabolism associated with the seasonal variation are not well understood. Polymorphic simple sequence repeats (SSRs) are widely used as molecular markers for genetic diversity research (Zheng et al., 2012; Abebe et al., 2015), linkage mapping (Sargan et al., 2007; Jiang et al., 2013), and breeding studies (Bjørnstad and Røed, 2001; Baumung et al., 2006; Nakamura et al., 2006). Few SSRs are currently available for the blue fox, and traditional technologies are limited in identifying SSRs. Next-generation sequencing is a rapid and cost-effective technology for SSR development (Gan et al., 2015; Králová-Hromadová et al., 2015).

The transcriptome is the complete set of transcripts in a cell or tissue at a specific developmental stage or physiological condition. For species with no reference genome, transcriptome sequencing is a quick method of obtaining cDNA fragments that are assembled *de novo* into transcripts. Based on a unigene library constructed with transcript sequences, biological analyses, including structural gene annotation, gene expression analysis, and gene function annotation, can be conducted, providing data for molecular studies of non-model organisms.

Currently, relatively few expressed sequence tag (EST) or genomic sequences are available for the blue fox. In the present study, a transcriptome pool of the blue fox was sequenced using Illumina paired-end sequencing in order to generate a large amount of EST and EST-SSR data, which will facilitate expression profiling and provide a good annotation basis for genomic studies.

MATERIAL AND METHODS

Total RNA was isolated using an RNAprep Pure Tissue Kit (Tiangen, Beijing, China) according to the manufacturer instructions. The RNA sample was quantified with an Agilent 2100 Bioanalyzer RNA Nanochip Kit, and RNA quality was assessed on 1.0% agarose gels. An equal amount of RNA from each of nine tissues (heart, liver, spleen, lung, kidney, skin, subcutaneous fat, muscle, and brain) was pooled, and the mixed RNA was subjected to Solexa sequencing by Biomarker Technologies Co. Ltd. (Beijing, China). Poly-(A) mRNA was isolated from total RNA with poly-(T) oligo-attached magnetic beads, and fragmented into short sequences using divalent cations at high temperature. The shortened RNA fragments were transcribed to first-strand cDNA using a random hexamer primer, and second-strand cDNA was subsequently synthesized using DNA Polymerase I and RNase H. In order to select cDNA fragments that were 150-200 bp in length, the fragments were purified with the AMPure XP system (Beckman Coulter, Danvers, MA, USA). The polymerase chain reaction (PCR) products were purified and library quality was assessed

using the Agilent Bioanalyzer 2100 and a Real-Time PCR System (Illumina, San Diego, CA, USA). A cDNA library was sequenced on a flow cell using an Illumina HiSeq2500 sequencing platform, and paired-end reads were generated. Data analysis and base calling were performed using the Illumina instrument software.

Clean reads were filtered from raw reads by the removal of adaptors, ambiguous reads, and low-quality reads. Subsequently, high-quality clean reads were assembled into unigenes by the Trinity software (Grabherr et al., 2011). The assembled unigenes were analyzed by searching the Non-Redundant (NR), Swiss-Prot, Gene Ontology (GO), euKaryotic Orthologous Groups (KOG), and Kyoto Encyclopedia of Genes and Genomes (KEGG) databases using the BLAST algorithm, with an E-value cutoff of 1.0E-5, and the transcripts were functionally annotated as the retrieved protein or nucleic acid with the highest sequence similarity. Amino acid sequences translated from unigenes were aligned with HMMER (http://hmmer.janelia.org/) to the Pfam database (E-value ≤ 1.0E-10) for function prediction (Chukkapalli et al., 2004; Finn et al., 2014).

The online tool MIcroSAtellite (MISA) (http://pgrc.ipk-gatersleben.de/misa/) was used to identify SSRs in a submitted sequence. The minimum number of repeats used for selecting the SSRs was ten for mononucleotide repeats, six for dinucleotide repeats, and five for tri-, tetra-, penta-, and hexanucleotide repeats. The number of unigenes containing SSRs, SSR motifs, and repeats was recorded.

RESULTS AND DISCUSSION

By high-quality deep sequencing and data filtering, 32,358,290 clean reads (8.15 Gb) with 85.1% Q30 bases and 54.14% GC content were obtained for further analysis. Using the Trinity *de novo* assembly program, all of the clean reads were assembled into 4,558,249 contigs with a mean length of 55.3 bp and an N50 length of 48 bp (Table 1). The majority of contigs were in the range 200-300 bp, which accounted for 98.91% of the total. A total of 161,269 transcripts with an N50 length of 1286 bp and a mean length of 749.74 bp were then connected through the overlap between the contigs (Table 1). Subsequently, 97,252 unigenes with an N50 length of 870 bp and a mean length of 555.95 bp were obtained (Table 1). The frequency distributions of contig, transcript, and unigene lengths and ratios are presented in **Figure S1**.

Table 1. Frequency distributions of *de novo* transcriptome assembly.

Length range	Contig	Transcript	Unigene
200-300	4,508,730 (98.91%)	59,962 (37.18%)	49,674 (51.08%)
300-500	21,374 (0.47%)	33,768 (20.94%)	20,690 (21.27%)
500-1000	13,402 (0.29%)	31,836 (19.74%)	13,923 (14.32%)
1000-2000	8360 (0.18%)	23,257 (14.42%)	8685 (8.93%)
2000+	6383 (0.14%)	12,446 (7.72%)	4280 (4.40%)
Total number	4558,249	161,269	97,252
Total length	252,081,687	120,909,061	54,066,996
N50 length	48	1286	870
Mean length	55.30	749.74	555.95

A total of 39,002 unigenes were successfully annotated into known genes, 10,598 of which were over 1000 bp long and 27,043 over 300 bp long (Table 2). Most (37,967 to 39,002) exhibited their highest homology with sequences in the NR database; of these, a large proportion of sequences (47.63%) matched those in *Canis lupus familiaris*, followed by *Mustela putorius* (3.50%), *Leptonychotes weddellii* (3.25%), and *Sus scrofa* (3.02%); the remainder (16,174 unigenes) exhibited less than 2.86% similarity to other species (Figure 1). In addition, 26,264, 24,839, 17,057, 24,267, and 19,448 unigenes were successfully matched in the Swiss-Prot, GO, KEGG, KOG, and Pfam databases, respectively. The comprehensive functional annotation is presented in Table 2. While most unigenes (58,250) were not annotated in the database, and longer unigenes were more likely to have homologs in the database, 82.4% of the unigenes over 1000 bp had homologs, and only 24.3% of unigenes shorter than 300 bp had homologs. Following the functional annotation of GO and KOG terms and KEGG pathways, some categories were used to discover important candidate genes for further investigation.

Table 2. Functional annotation of the unigenes.

Annotated databases	Unigene	≥300 nt	≥1000 nt
COG	8659	7025	3789
GO	24,839	18,796	8542
KEGG	17,057	12,479	5486
KOG	24,267	17,904	7827
Pfam	19,448	15,917	8608
Swiss-Prot	26,264	20,047	9178
NR	37,967	26,743	10,559
All	39,002	27,043	10,598

NR = Non-Redundant; GO = Gene Ontology; KOG = euKaryotic Orthologous Groups; KEGG = Kyoto Encyclopedia of Genes and Genomes; COG = Clusters of Orthologous Groups.

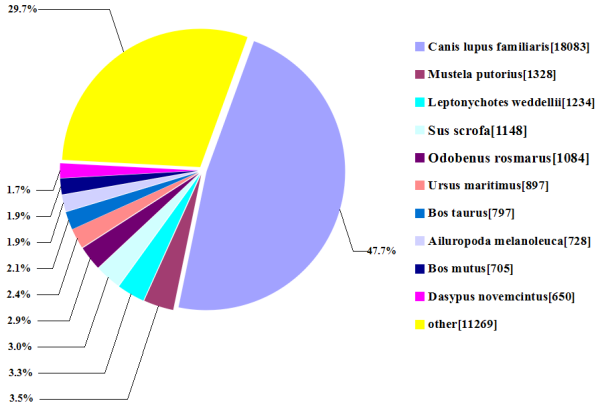

Figure 1. Species frequency distribution of unigene homology against the Non-Redundant (NR) database on the best BLAST hits (E-value ≤ 1.0 E-5). Different colors represent different species.

Based on the NR annotation, the unigenes were searched against the GO terms that provided a dynamic, controlled vocabulary and hierarchical relationships for the representation of information on cellular components, molecular functions, and biological processes (Figure 2); 24,839 unigenes were assigned to these three GO terms, which were subdivided into 19, 20, and 22 subcategories, respectively. Because some unigenes were assigned to one or more GO term, 86,005, 76,487, and 32,678 unigenes were assigned to biological processes, cellular components, and molecular functions, respectively. For the cellular components, cell part, cell, and organelle were three major subcategories, while virion part, virion, and nucleoid were minor subcategories. For the molecular functions, binding, catalytic, and transporter activity were the main subcategories; morphogen activity, protein tag, and chemorepellent activity were the smallest subcategories. For the biological processes, genes involved in cellular processes, single-organism processes, and metabolic processes were highly represented, but cell aggregation, cell killing, and biological phase were poorly represented.

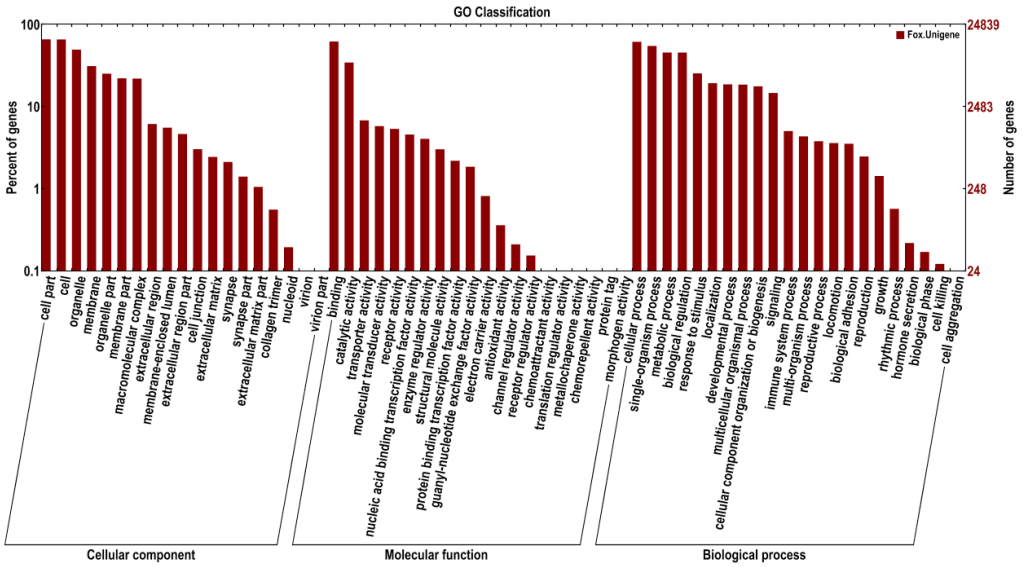

Figure 2. Functional annotation of assembled unigenes based on gene ontology (GO) categorization.

In addition, all of the assembled unigenes were searched against the KOG database for functional prediction and classification (Figure 3). A total of 24,267 unigenes were assigned to 25 different clusters. Because some unigenes were assigned to one or more KOG clusters, 4818 unigenes were assigned to the functional cluster of signal transduction mechanisms, which was marginally higher than the number of unigenes assigned to only the general function prediction, both of which were significantly greater than the number of unigenes assigned to other clusters. A few unigenes were assigned to cell wall/membrane/envelope biogenesis, nuclear structure, structure, cell motility, coenzyme transport, and metabolism, all of which contained fewer than 100 unigenes.

All of the assembled unigenes were searched against the KEGG database, which is an alternative approach to categorize gene functions that is focused on biochemical pathways. A total of 17,057 unigenes were assigned to 227 different pathways. Because some unigenes were

assigned to one or more pathways, 19,609 unigenes were assigned to metabolism, genetic information processing, environmental information processing, cellular processes, and organismal systems.

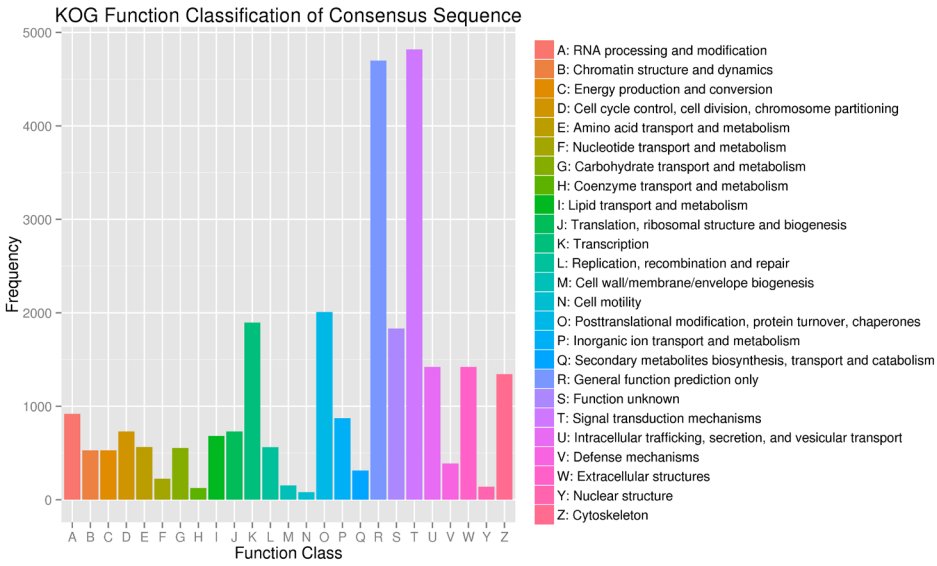

Figure 3. KOG (euKaryotic Orthologous Groups) classification.

A summary of the unigene annotation information is presented in **Table S1**. Of the 17,057 unigenes assigned to pathways, 29.61% were classified in organismal systems, with most of them involved in the immune, digestive, and endocrine systems. The metabolism pathway accounted for 20.71% of the total, with most of the unigenes in this pathway involved in carbohydrate metabolism, amino acid metabolism, lipid metabolism, and glycan biosynthesis and metabolism. Environmental information processing accounted for 19.59% of the total, and was subdivided into membrane transport, signal transduction, and signaling molecules and interactions. Signal transduction was subdivided into 11 signaling pathways with 2351 unigenes, and the mitogen-activated protein kinase (MAPK) signaling pathway contained the greatest number of unigenes. Cellular processes accounted for 18.64% of the total and included transport and catabolism, cell motility, cell growth and death, and cellular community. The focal adhesion pathway, which belongs to the cellular community, was assigned the highest number of unigenes (598). Genetic information processing, including transcription, translation, folding, sorting, degradation, replication, and repair accounted for 14.28% of the total (Figure 4). A total of 190 unigenes were assigned to the melanogenesis pathway, which may be worthy of further investigation to search for candidate genes associated with coat color in the blue fox (**Table S2**). In addition, 9, 16, and 65 unigenes were assigned to the fatty acid biosynthesis, fatty acid elongation in mitochondria, and fatty acid metabolism pathways, respectively (**Table S3**), suggesting that further research should be conducted on fatty acid metabolism in the blue fox. The MAPK signaling pathway had the most unigenes assigned to it, suggesting that many biological processes are regulated by the MAPK signaling pathway, a possibility that requires further investigation.

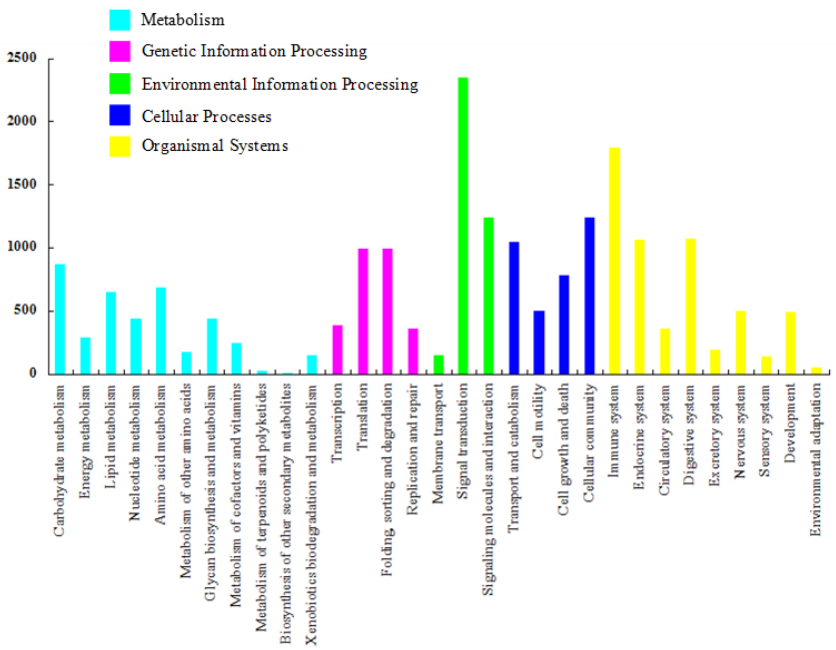

Figure 4. Kyoto Encyclopedia of Genes and Genomes (KEGG) classification.

For exploring SSR markers in the blue fox transcriptome, 12,965 unigenes (>1 kb) were filtered from the total. Subsequently, the unigenes were analyzed with MISA. Six types of SSR were found, and the percentages of mononucleotides, dinucleotides, trinucleotides, tetranucleotides, pentanucleotides, and hexanucleotides were 67.94, 13.83, 18.81, 1.39, 0.22, and 0.17%, respectively (**Table S4**). Statistical analysis revealed that 5907 unigenes contained SSRs, 412 of which contained more than one identical type of SSR, and 14 unigenes were present in a compound formation with several different types of SSR. Regarding mononucleotides, A/T was the most common repeat motif (55.97%), and was significantly more frequent than C/G (9.97%). The most abundant repeat motif in dinucleotides was AC/GT/CA/TG, while the rarest was CG/GC. Of the17 categories of trinucleotides, the CCG/CGG/GCC/GGC repeat motif was the most common. Thirteen categories of pentanucleotides were only owned by one unigene, as were nine categories of hexanucleotides, except for two unigenes that contained the GGGTCC/CCTGGG repeat motif. A total of 5907 unigenes contained EST-SSRs, which will provide a good resource for mining and use in molecular marker-assistant breeding.

In summary, we generated 97,252 unigenes from a pool of nine tissues of a farmed blue fox using the Illumina platform. This is the first comprehensive transcriptome for the blue fox, which was aligned with databases to annotate functional information; 39,002 unigenes were annotated to putative functions, and six types of SSR were found in 5907 unigenes. All of these data provide a valuable resource for further research on this species.

Conflicts of interest

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

Research supported by a project from the National Natural Science Foundation of China (#31401979).

REFERENCES

Abebe AS, Mikko S and Johansson AM (2015). Genetic diversity of five local Swedish chicken breeds detected by microsatellite markers. *PLoS One* 10: e0120580.http://dx.doi.org/10.1371/journal.pone.0120580

Baumung R, Cubric-Curik V, Schwend K, Achmann R, et al. (2006). Genetic characterisation and breed assignment in Austrian sheep breeds using microsatellite marker information. *J. Anim. Breed. Genet.* 123: 265-271.http://dx.doi.org/10.1111/j.1439-0388.2006.00583.x

Bjørnstad G and Røed KH (2001). Breed demarcation and potential for breed allocation of horses assessed by microsatellite markers. *Anim. Genet.* 32: 59-65.http://dx.doi.org/10.1046/j.1365-2052.2001.00705.x

Chukkapalli G, Guda C and Subramaniam S (2004). SledgeHMMER: a web server for batch searching the Pfam database. *Nucleic Acids Res.* 32: W542-544.http://dx.doi.org/10.1093/nar/gkh395

Finn RD, Bateman A, Clements J, Coggill P, et al. (2014). Pfam: the protein families database. *Nucleic Acids Res.* 42: D222-D230.http://dx.doi.org/10.1093/nar/gkt1223

Gan C, Love C, Beshay V, Macrae F, et al. (2015). Applicability of next generation sequencing technology in microsatellite instability testing. *Genes (Basel)* 6: 46-59.http://dx.doi.org/10.3390/genes6010046

Grabherr MG, Haas BJ, Yassour M, Levin JZ, et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29: 644-652.http://dx.doi.org/10.1038/nbt.1883

Jiang L, Chu G, Zhang Q, Wang Z, et al. (2013). A microsatellite genetic linkage map of half smooth tongue sole (*Cynoglossus semilaevis*). *Mar. Genomics* 9: 17-23.http://dx.doi.org/10.1016/j.margen.2012.07.002

Králová-Hromadová I, Minárik G, Bazsalovicsová E, Mikulíček P, et al. (2015). Development of microsatellite markers in *Caryophyllaeus laticeps* (Cestoda: Caryophyllidea), monozoic fish tapeworm, using next-generation sequencing approach. *Parasitol. Res.* 114: 721-726.http://dx.doi.org/10.1007/s00436-014-4239-4

Nakamura A, Kino K, Minezawa M, Noda K, et al. (2006). A method for discriminating a Japanese chicken, the Nagoya breed, using microsatellite markers. *Poult. Sci.* 85: 2124-2129.http://dx.doi.org/10.1093/ps/85.12.2124

Sargan DR, Aguirre-Hernandez J, Galibert F and Ostrander EA (2007). An extended microsatellite set for linkage mapping in the domestic dog. *J. Hered.* 98: 221-231.http://dx.doi.org/10.1093/jhered/esm006

Våge DI, Fuglei E, Snipstad K, Beheim J, et al. (2005). Two cysteine substitutions in the MC1R generate the blue variant of the Arctic fox (*Alopex lagopus*) and prevent expression of the white winter coat. *Peptides* 26: 1814-1817.http://dx.doi.org/10.1016/j.peptides.2004.11.040

Zheng JY, Wang H, Chen XX, Wang P, et al. (2012). Microsatellite markers for assessing genetic diversity of the medicinal plant *Paris polyphylla* var. *chinensis* (Trilliaceae). *Genet. Mol. Res.* 11: 1975-1980.http://dx.doi.org/10.4238/2012.August.6.1

Supplementary material

Figure S1. Length distributions of the assembled contigs, transcripts, and unigenes.

Table S1. Kyoto Encyclopedia of Genes and Genomes (KEGG) biochemical mapping for the blue fox (*Alopex lagopus*).

Table S2. Summary of unigenes assigned to the melanogenesis pathway.

Table S3. Summary of unigenes assigned to fatty acid pathways.

Table S4. Summary of simple sequence repeat (SSR) types in the blue fox (*Alopex lagopus*) transcriptome.