RESEARCH ARTICLE

Characterizing the transcriptome and molecular markers information for roach, *Rutilus rutilus*

WEI CHI[1,2], XUFA MA[1,2], JIANGONG NIU[1,3] and MING ZOU[1,2]*

[1]*College of Fisheries, Huazhong Agricultural University, Wuhan 430070, People's Republic of China*
[2]*Key Laboratory of Freshwater Animal Breeding, Ministry of Agriculture, Beijing 430070, People's Republic of China*
[3]*Fisheries Research Institute of Xinjiang Uygur Autonomous Region, Urumqi 830000, People's Republic of China*

Abstract

Rutilus rutilus (roach) is native to most of Europe and western Asia, and the Irtysh River basin in Sinkiang, northwest China is the marginal area of their natural distribution. The wide distribution and unique characteristic of this species makes it an ideal model for analysing ecological and comparative genomics. However, the limited genome sequences available for this species have hindered these investigations. Transcriptomes from the brains and livers of five individuals collected from the Irtysh River basin were sequenced using Illumina paired-end sequencing technology. A collection of 132,289 unigenes for this species were obtained using a *de novo* assembly method based on nearly 120 million clean reads encompassing more than 14 Gb data. Approximately 37.5% (49,656), 27.1% (35,867) and 21.2% (27,987) of the transcriptome had homologues deposited in Nt, Nr and Swiss-Prot, respectively; 12.3% (16,328) were assigned to eukaryotic orthologous groups of proteins classifications, and 21.5% (28,429) harboured Interpro domains. On the basis of the assembled transcriptome, we detected 177,493 single-nucleotide variation resident in 39.3% (52,029) of the sequences and 20.8% (27,497) of the sequences harbouring 36,639 simple sequence repeats. The identified molecular markers are a basis for further ecological analysis, and the transcriptome reported here allows for more extensive evolutionary analyses of the Cyprinidae, the most species-rich family of freshwater fishes.

[Chi W., Ma X., Niu J. and Zou M. 2016 Characterizing the transcriptome and molecular markers information for roach, *Rutilus rutilus*. *J. Genet*. **95**, 45–51]

Introduction

The common roach, *Rutilus rutilus*, is a species of small-scaled, ray-finned fish in the subfamily Leuciscinae and the family Cyprinidae. It is a fresh-water and brackish-water fish native to most of the Europe and western Asia. The Irtysh River basin, located in northwest China, is the marginal area of this species' natural distribution.

The widespread distribution and environmental heterogeneity of this species can produce many local populations, making it an ideal model for population genetics and phylogeographic analyses, and many such studies have been reported (Bouvet *et al.* 1991; Hanfling *et al.* 2004; Keyvanshokooh *et al.* 2007; Keyvanshokooh and Kalbassi 2006). However, no such analyses are reported for *R. rutilus* distributed in the Irtysh River basin. In fact, there are many issues concerning the genetic diversity of central and marginal populations. Some researchers have suggested that peripheral populations exhibit low genetic diversity, while others have opposed this view (Eckert *et al.* 2008). Analyses of the genetic structure of *R. rutilus* distributed in its marginal area, the Irtysh River basin in northwest China, may provide additional insights into this issue. However, the limited number of molecular markers has hindered such investigations. Further, molecular events underlying different local populations (e.g., divergences between populations adapted to freshwater and brackish water) should be partially revealed by comparing their expression profiles. *R. rutilus* has recently become a model fish for ecotoxicology, and many studies have focussed on it (Jobling *et al.* 1998, 2002; Lange *et al.* 2008, 2011). This trend may be due to a high incidence of the simultaneous presence of both male and female gonadal features in the same individual, which seems to have resulted from exposure to ambient levels of chemicals. We hypothesized that digital expression profiling or microarrays should be good alternative approaches to resolve this problem.

A transcriptome consists of all the transcripts expressed in one cell or a population of cells at a certain time. Obtaining part of the genome sequences and developing molecular markers for *R. rutilus*, RNA-Seq using Illumina's next-generation

*For correspondence. E-mail: zoumingr@163.com.

Keywords. transcriptome; Cyprinidae; single-nucleotide variation; *Rutilus rutilus*.

sequencing technology may prove to be efficacious. The technology using the sequencing-by-synthesis method can generate millions of short reads that can be used in many analyses such as the identification of differentially expressed genes and the development of molecular markers. A number of studies have shown that it is rapid and cost-effective method for the development of single-nucleotide polymorphism (SNP)/single-nucleotide variation (SNV) and SSRs markers and is especially suitable for nonmodel species including plants (Blanca *et al.* 2011; Zhang *et al.* 2013) and animals (Jung *et al.* 2011; Qu *et al.* 2012; Zheng *et al.* 2014). A number of reports have suggested that the assembled transcriptome could also be used for analyses of phylogenomics (Zou *et al.* 2012) and comparative genomics (Irie and Kuratani 2011; Yang *et al.* 2012).

In this study, we have sequenced the transcriptomes of brains and livers of five *R. rutilus* individuals collected from the Irtysh River basin based on the Illumina Hiseq 2000 platform. A large number of SNVs and SSRs identified from the assembled transcriptome could be used for subsequent genetic marker development. Moreover, the reported transcriptome also provides an invaluable genomic resource for future research in other fields such as ecotoxicology and comparative genomics.

Materials and methods

Sample collections

Live *R. rutilus* were collected using gill nets downstream from the Irtysh River (near Kazakhstan) located in Sinkiang, northwest China. Total RNAs from the brains and livers of five individuals were extracted using TRIzol reagent (Transgene Company, Illkirch Graffenstaden Cedex, Kensington, France) according to the manufacturer's protocol. All specimens were euthanized with 300 mg/L tricaine methanesulphonate (MS 222) before tissue collection. A degradation and contamination was preliminary monitored on 1% agarose gels. A NanoPhotometer spectrophotometer (IMPLEN, Westlake Village, USA) was then used to check the purity of the RNA and a Qubit RNA Assay Kit in Qubit 2.0 Flurometer (Life Technologies, Carlsbad, USA) was employed to measure the concentration of RNA. The RNA integrity number (RIN) of each sample was assessed using the RNA Nano 6000 Assay Kit of the Agilent Bioanalyzer 2100 system (Agilent Technologies, Santa Clara, USA) and were all larger than 8.0. The total RNAs from the brains and livers of the five individuals were then mixed and used for subsequent library construction and sequencing.

Library construction and sequencing

First, mRNA was purified from the total RNA using poly-T oligo-attached magnetic beads. Then, fragmentation of mRNA was carried out using divalent cations under elevated temperature in NEBNext first strand synthesis reaction buffer (5×). Random hexamer primer was then used to reverse transcribe the mRNA fragments and cDNA fragments of preferentially 150–200 bp in length were selected. Following PCR amplification and purification, the quality of the cDNA library was assessed on an Agilent Bioanalyzer 2100 system. Sequencing was performed on an Illumina Hiseq 2000 platform, and millions of 125 bp paired-end reads were generated. The following analyses were all based on the generated short reads, and a flowchart summarizes the main steps as shown in figure 1 in electronic supplementary material at http://www.ias.ac.in/jgenet/.

Quality control and de novo assembly

Raw reads in FASTQ format were first processed through a series of in-house perl scripts to remove reads containing adapter sequences, or containing more than 10% of undefined bases (N), or containing more than half of low quality bases (Q score <5). The *de novo* assembly of the transcriptome was accomplished using Trinity (Grabherr *et al.* 2011) with default settings, except that min_kmer_cov was set to 2. The longest isoform of each gene (subcomponent) was selected as a unigene to form a nonredundant transcriptome.

Annotations and classifications

The annotations and classifications for the transcriptome sequences were based on sequence similarities deposited in the following five databases: Nr (NCBI nonredundant protein sequences), Nt (NCBI nonredundant nucleotide sequences), Swiss-Prot (a manually annotated and reviewed protein sequence database), KOG (euKaryotic Orthologous Groups of proteins), and Pfam (protein family database). NCBI blast 2.2.28+ (Altschul *et al.* 1990) was used to compare with sequences in Nr, Nt, Swiss-Prot and KOG with the e-value cutoff set to 1e-5 for the first three databases and 1e-3 for the last one. Based on the Pfam classification, the prediction of the protein domains were performed using *hmmscan* included in the HMMER 3.0 (Eddy 2009) software package. According to the priorities from Nr to Swiss-Prot and homologues obtained from these two databases, we predicted open reading frames (ORFs) for each unigene. For sequences found to have no homologues, ESTScan 3.0.3 (Iseli *et al.* 1999) was used to predict their ORFs.

Identification of SNVs, SSRs and other repetitive elements

All sequences of the nonredundant transcriptome were screened as potential SNVs, SSRs and other repetitive elements. All cleaned reads were mapped back to the transcriptome using Bowtie (Langmead *et al.* 2009) with default settings. Then, samtools v0.1.18 (Li *et al.* 2009) and Picardtools v1.41 (http://broadinstittute.github.io/picard) were used to sort and remove duplicated reads. The SNV calling (including indels) was implemented using GATK2 software (McKenna *et al.* 2010), the generated raw vcf files were filtered according to the GATK standard filter method, and only SNVs with a Q score >30 and distance >5 were retained.

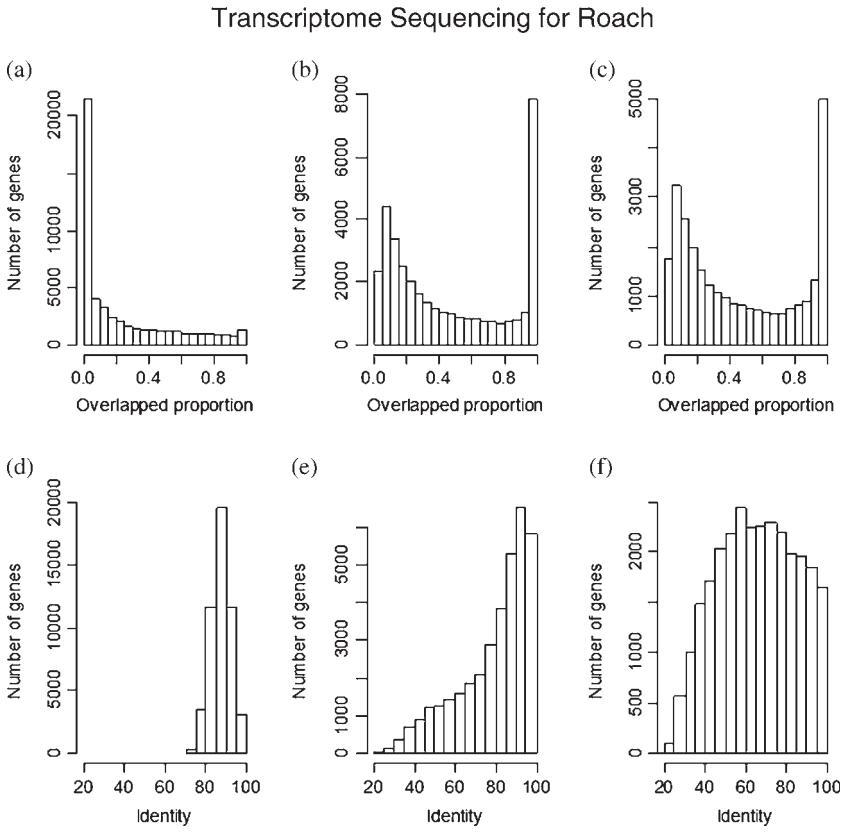

Figure 1. Distribution of overlapped proportion (overlapped region accounting for the total length of its best hit homologues deposited in the database) and identity (the proportion of identical characters accounting for the overlapped region) for the *R. rutilus* transcriptome. a, b, and c are overlapped proportions for Nt, Nr and Swiss-Prot, respectively. d, e, and f are identity proportions for Nt, Nr and Swiss-Prot, respectively.

SSRs in the transcriptome were identified using MISA (http://pgrc.ipk-gatersleben.de/misa/misa.html) with default settings and primers for each SSR were designed using Primer3 (primer3.sourceforge.net/releases.php). Other repetitive elements such as DNA transposons, retroelements and others residing in the transcriptome were identified using RepeatMasker (www.repeatmasker.org). All experimental procedures were approved by the Animal Care and Use Committee of Huazhong Agricultural University. Because *R. rutilus* is not endangered or protected in this area, specific permission was not required for the field study.

Results and discussion

Illumina paired-end sequencing and de novo assembly

In this study, a total of more than 2 × 60 million paired-end raw reads were generated from a 150–200 bp insert library. After stringent quality assessment and data filtering, more than 2 × 59 million clean reads encompassing more than 14 Gb of data were selected as high-quality reads. High-quality bases (Q score >20) accounting for the total bases in the retained reads were 96.0% and 93.8% for the forward and reverse reads, respectively. All subsequent analyses were based on these filtered clean reads. The short reads have been deposited in NCBI Sequence Read Archive database (SRA) with accession number SRR1776878. All short sequences were assembled into 241,409 transcripts using a *de novo* assembly method, and 132,289 unigenes were retained to form a nonredundant transcriptome. The size distribution of the transcriptome is shown in figure 2 supplementary material. It shows that the length of most unigenes are shorter than 500 bp, while a number of sequences are longer than 2000 bp. The average length, median length and N50 sequence length of the transcriptome were 670, 327 and 1215 bp, respectively. To validate the assembly of the transcriptome, 10 unigenes were randomly selected and eight of them were confirmed by using RT-PCR and following Sanger sequencing (table 1 in electronic supplementary material, GenBank accession numbers KR935251–KR935258).

A total of 14,942 unigenes had coordinates and represented 14,816 gene loci in the *D. rerio* genome according to the reciprocal best blast hit analyses. Given the similar genome size (roughly one-third of the human genome, www.genomesize.com) and ploidy between *R. rutilus* and *D. rerio* (Klose *et al.* 1969), we hypothesized that a similar number of genes reside in the two cyprinid genomes. The total number of genes with intact ORFs in *D. rerio* was 26,459 (www.

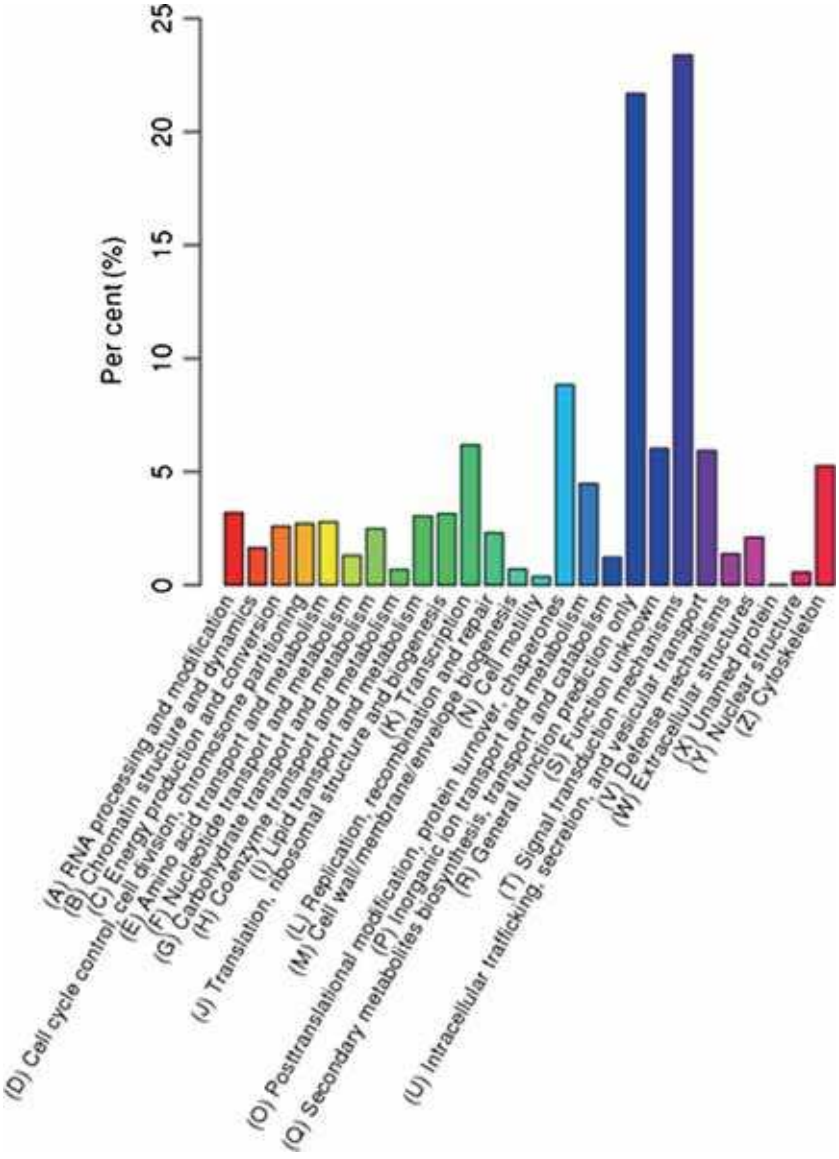

Figure 2. Summary of unigenes classified according to each KOG category.

ensembl.org), thus ∼56% of the gene loci in *R. rutilus* were represented in the newly generated transcriptome.

Approximately 81% (more than 95 million) of the clean reads could map back to the *de novo* assembled transcriptome, with more than one million reads mapped back to each of the top two sequences. According to the annotations obtained from homologues deposited in Nr, the two sequences with the most reads mapped back were 'NADH dehydrogenase subunit 5' and 'apolipoprotein A-I-1'. However, many factors, such as the length of the sequences, depth of sequencing and expression level, could influence the number of reads mapped back to unigenes. To avoid contamination by these factors, we calculated the expected number of fragments per kilobase of transcript sequence per millions of base pairs sequenced (FPKM values) to assess the expression pattern of these *de novo* assembled sequences. Figure 3 in electronic supplementary material shows the distribution of FPKM values calculated for the *R. rutilus* transcriptome. It reveals that most sequences were poorly expressed, with only a small number of highly expressed genes according to our high-throughput sequencing data. In particular, three members of the apolipoprotein multigene family, apolipoprotein C-I, apolipoprotein A-I-1 and 14-kDa apolipoprotein, reported to be mainly expressed in the liver (Delcuve *et al.* 1992; Lauer *et al.* 1988; Zhou *et al.* 2005), are among the top three highly-expressed genes according to the FPKM values.

Functional annotation and classification

Using a BLAST algorithm, homologues deposited in the five databases obtained for *R. rutilus* transcriptome are shown in table 1.The proportions of annotated sequences in the

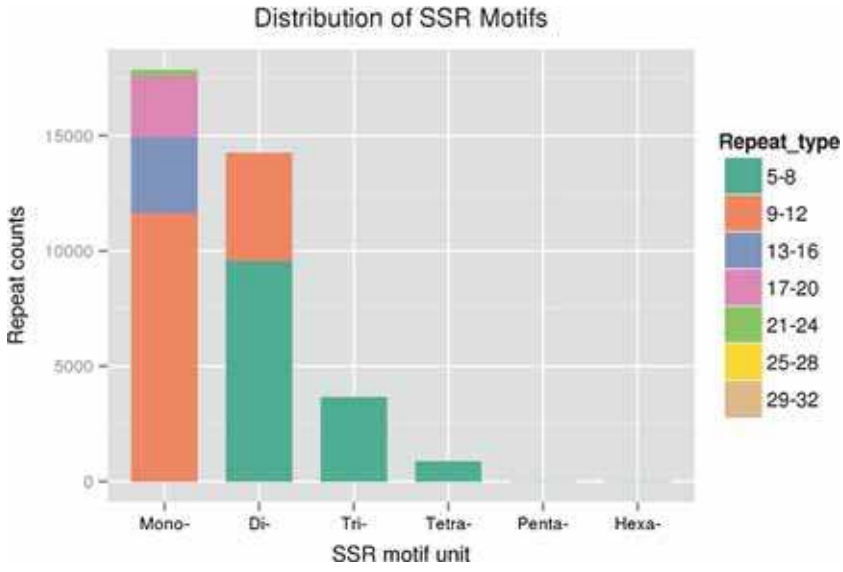

Figure 3. Summary of all the SSRs identified from the *R. rutilus* transcriptome.

R. rutilus transcriptome are largely different among the five databases, with genes annotated in KOG being the smallest (12.3%) and genes annotated in Nt the largest (37.5%). Only 30% of the unigenes were annotated when sequence lengths ranged from 200 to 300 bp; however, when their lengths ranged from 900 to 1000 bp, the percentages of annotated genes reached 76%. Generally, the length of a gene is directly correlated to its likelihood of being annotated. We selected unigenes with lengths ranging from 200 to 1000 bp, which represented a majority of the transcriptome, to scrutinize the correlation between the proportions of annotated genes and their lengths, and we found that they were significantly positively correlated (r=0.99). We have proposed that the short genes may represent chimeras resulting from assembly errors and fragmented transcripts, as well as noncoding RNAs that were less likely to be annotated.

In general, 37.5% (49,656), 27.1% (35,867) and 21.2% (27,987) of the transcriptome found homologues deposited in Nt, Nr and Swiss-Prot, respectively. The number of sequences overlapped more than 50% of the length of their best hit homologues deposited in Nt, Nr and Swiss-Prot were 9903, 15,094 and 12,086, of which 2032, 8903 and 6330 overlapped more than 90%, individually. The distribution of the overlapped region accounting for the total length of its best hit homologues deposited in Nt, Nr and Swiss-Prot for each sequence of the *R. rutilus* transcriptome (henceforth termed 'overlapped proportion' for convenience) and the proportion of identical characters accounting for the overlapped region (henceforth termed 'identity') are shown in figure 1. These results reveal that we retrieved the largest number of homologues with highly overlapped proportions from Nr. The number of retrieved homologues from Swiss-Prot is also large, but the number is very low for Nt. In contrast, the largest number of genes with high similarity was retrieved from Nt. The number of homologues retrieved from Nr is also very large, but the number retrieved from Swiss-Prot is not large, and their similarities are evenly distributed. This outcome may be due to that those homologues retrieved from Nt were based on nucleotide similarity, while the homologues retrieved from Nr and Swiss-Prot were based on the similarity of the amino acids. Moreover, proteins deposited in Swiss-Prot were calibrated and may have some limitations when used for annotation, especially for nonmodel species.

Table 1. Summary of homologues deposited in the five databases retrieved for *R. rutilus* transcriptome.

Database	Number of homologue sequences found in the database	Percentage accounting for total genes (%)
Nr	35,867	27.1
Nt	49,656	37.5
Swiss-Prot	27,987	21.2
KOG	16,328	12.3
Pfam	28,429	21.5
Total unigenes	132,289	100

Nr, NCBI nonredundant protein sequences; Nt, NCBI nonredundant nucleotide sequences; Swiss-Prot, a manually annotated and reviewed protein sequence database; KOG, euKaryotic Orthologous Groups of proteins; Pfam, protein family database.

The KOG classifications were based on the similarity of conserved domains between *R. rutilus* and other eukaryotes with high quality whole genome sequences available. Based on the analysis, a total of 16,328 sequences were assigned to at least one KOG category (figure 2). Among the 26 KOG categories, the 'signal transduction mechanisms' was the largest group (3817, 23.4%), followed by 'general function prediction only' (3540, 21.7%), 'posttranslational modification, protein turnover, chaperones' (1444, 8.8%), 'transcription'

(1010, 6.2%), 'function unknown' (984, 6.0%), 'intracellular trafficking, secretion and vesicular transport' (968, 5.9%) and 'cytoskeleton' (859, 5.3%).

In conclusion, statistics of our functional annotations and classifications for the newly generated *R. rutilus* transcriptome are similar to other fishes (Fu and He 2012; Gao *et al*. 2012) and other eukaryotes including plants (Torre *et al*. 2014; Garcia-Seco *et al*. 2015). This may be because they are derived from a common ancestor. On the other hand, this may suggest that quality of the first reported *R. rutilus* transcriptome is receivable and could be used for subsequent analyses.

EST-SNV discovery

We identified 177,493 high-quality SNVs, of which 101,176 (57.0%) were transitions and 76,317 (43.0%) were transversions. These putative SNVs reside in 52,029 unigenes, most of them (38.8%) harbouring only one SNV, with the maximum number of SNVs residing in one gene being 62. Further analyses showed that 36,481 (20.6%) SNVs located in the possible coding regions, and a number of 170 changed the encoded amino acids, which may alter the activity of the protein and ultimately influence the fitness of the allele's owner. Using the mixture of the same amount of DNA from 10 individuals as template, one fragment of the possible orthologue of an intronless zebrafish gene harbouring five SNVs was PCR amplified (the forward primer was TAACGGGCTCCTTATGTCC, and the reverse was AATGAGGATGAAGGCAACA), and the sequencing chromatogram confirmed one of these SNVs. Moreover, we found 34,374 indels residing in 17,404 unigenes, and they may also influence the fitness of the owner, especially those located in coding regions.

Discovery of EST-SSR and other repetitive elements

A total of 36,639 SSRs were identified in 27,497 unigenes, of which 20,841 unigenes harboured only one SSR. A total of 2110 SSRs residing in 1791 unigenes were observed in compound formations (SSRs composed of two or more motifs separated by $<$100 bp). Summaries of all the SSRs identified are depicted in figure 3. The mononucleotide repeat motifs were the most abundant type, followed by dinucleotide, trinucleotide, tetranucleotide, pentanucleotide and hexanucleotide repeat motifs. Higher numbers of tandem repeats (9–12 times or more) are frequent in shorter SSR motifs (mononucleotide and dinucleotide repeats), while 5–8 fold tandem repeats are frequent in longer SSR motifs (dinucleotide repeats or longer). A/T repeats were the most frequent motifs (17,404) in mononucleotide repeats. GT/AC was the most common type of dinucleotide (4733), followed by TG/CA (3923). Among the trinucleotide repeats, the AAT/ATT (495) was the most frequent motifs, followed by TAT/ATA (353) and TTA/TAA (350). Moreover, primers were designed for 15,070 SSRs residing in 12,583 unigenes. SSRs have been widely used as molecular markers in many fields such as population genetics (Chapuis *et al.* 2014) and linkage analysis (Shirasawa *et al.* 2011), and SNVs should be more widely used in the near future since their genotyping can be automated more easily than other markers (Romay *et al.* 2013; Milano *et al.* 2014). In the present study, we provide information for these two types of molecular markers which should be important for future analyses.

Repeat masking revealed that ∼4.71% of the *R. rutilus* transcriptome consists of repetitive elements, and the profiles are shown in table 2 in electronic supplementary material. According to the number of repetitive elements, low complexity was the most represented, followed by DNA transposons, simple repeats and retroelements. Within DNA transposons, hobo-Activator, followed by En-Spm, Tourist/Harbinger, and Tc1-IS630-Pogo were the most represented, while, within retroelements, the order is LTR elements, LINEs and SINEs. Interestingly, a number of repetitive elements were found to be located in the predicted coding regions, the most represented being simple_repeat, low_complexity and LTR/Gypsy.

Conclusion

Using Illumina sequencing technology, we performed transcriptome sequencing for *R. rutilus*, a wide-spread fish native to Europe and western Asia. The *de novo* assembled transcripts were annotated and classified according to a series of public databases, and a large number of molecular markers were identified. This is the first report on the transcriptome of *R. rutilus*, and the markers we identified will enable genetics and molecular ecological studies on the cold-adapted species, and, more extensively, will facilitate evolutionary analyses such as phylogenomics and comparative genomics of the Cyprinidae, the most species-rich family of freshwater fishes.

Acknowledgements

We thank the two anonymous reviewers for their useful and detailed comments that greatly improved our report. This work was funded by a National Science and Technology support programme (no. 2012BAD25B06).

References

Altschul S. F., Gish W., Miller W., Myers E. W. and Lipman D. J. 1990 Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.

Blanca J., Canizares J., Roig C., Ziarsolo P., Nuez F. and Pico B. 2011 Transcriptome characterization and high throughput SSRs and SNPs discovery in *Cucurbita pepo* (Cucurbitaceae). *BMC Genomics* **12**, 104.

Bouvet Y., Soewardi K. and Pattee E. 1991 The discrimination of roach *Rutilus rutilus* (Linnaeus 1758) populations in different parts of a river system. An investigation using biochemical markers. *Hydrobiologia* **209**, 161–167.

Chapuis M. P., Plantamp C., Blondin L., Pages C., Vassal J. M. and Lecoq M. 2014 Demographic processes shaping genetic variation

of the solitarious phase of the desert locust. *Mol. Ecol.* **23**, 1749–1763.

Delcuve G. P., Sun J. M. and Davie J. R. 1992 Expression of rainbow trout apolipoprotein A-I genes in liver and hepatocellular carcinoma. *J. Lipid. Res.* **33**, 251–262.

Eckert C. G., Samis K. E. and Lougheed S. C. 2008 Genetic variation across species' geographical ranges: the central–marginal hypothesis and beyond. *Mol. Ecol.* **17**, 1170–1188.

Eddy S. R. 2009 A new generation of homology search tools based on probabilistic inference. *Genome Inform.* **23**, 205–211.

Fu B. and He S. 2012 Transcriptome analysis of silver carp (*Hypophthalmichthys molitrix*) by paired-end RNA sequencing. *DNA Res.* **19**, 131–142.

Gao Z. X., Luo W., Liu H., Zeng C., Liu X. L., Yi S. J. *et al.* 2012 Transcriptome analysis and SSR/SNP markers information of the blunt snout bream (*Megalobrama amblycephala*). *PLoS One* **7**.

Garcia-Seco D., Zhang Y., Gutierrez-Manero F. J., Martin C. and Ramos-Solano B. 2015 RNA-Seq analysis and transcriptome assembly for blackberry (Rubus sp. var. Lochness) fruit. *BMC Genomics* **16**, 5.

Grabherr M. G., Haas B. J., Yassour M., Levin J. Z., Thompson D. A., Amit I. *et al.* 2011 Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652.

Hanfling B., Durka W. and Brandl R. 2004 Impact of habitat fragmentation on genetic population structure of roach, *Rutilus rutilus*, in a riparian ecosystem. *Conserv. Genet.* **5**, 247–257.

Irie N. and Kuratani S. 2011 Comparative transcriptome analysis reveals vertebrate phylotypic period during organogenesis. *Nat. Commun.* **2**, 248.

Iseli C., Jongeneel C. V. and Bucher P. 1999 ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 138–148.

Jobling S., Nolan M., Tyler C. R., Brighty G. and Sumpter J. P. 1998 Widespread sexual disruption in wild fish. *Environ. Sci. Technol.* **32**, 2498–2506.

Jobling S., Beresford N., Nolan M., Rodgers-Gray T., Brighty G.C., Sumpter J. P. *et al.* 2002 Altered sexual maturation and gamete production in wild roach (*Rutilus rutilus*) living in rivers that receive treated sewage effluents. *Biol. Reprod.* **66**, 272–281.

Jung H., Lyons R. E., Dinh H., Hurwood D. A., McWilliam S. and Mather P. B. 2011 Transcriptomics of a giant freshwater prawn (*Macrobrachium rosenbergii*): de novo assembly, annotation and marker discovery. *PLoS One* **6**, e27938.

Keyvanshokooh S. and Kalbassi M. R. 2006 Genetic variation of *Rutilus rutilus caspicus* (Jakowlew 1870) populations in Iran based on random amplified polymorphic DNA markers: a preliminary study. *Aquac. Res.* **37**, 1437–1440.

Keyvanshokooh S., Ghasemi A., Shahriari-Moghadam M., Nazari R M. and Rahimpour M. 2007 Genetic analysis of *Rutilus rutilus caspicus* (Jakowlew 1870) populations in Iran by microsatellite markers. *Aquac. Res.* **38**, 953–956.

Klose J., Wolf U., Hitzeroth H., Ritter H. and Ohno S. 1969 Polyploidization in the fish family Cyprinidae, order Cypriniformes. II. Duplication of the gene loci coding for lactate dehydrogenase (E.C.: 1.1.1.27) and 6-phosphogluconate dehydrogenase (E.C.: 1.1.1.44) in various species of Cyprinidae. *Humangenetik* **7**, 245–250.

Lange A., Katsu Y., Ichikawa R., Paull G. C., Chidgey L. L., Coe T. S. *et al.* 2008 Altered sexual development in roach (*Rutilus rutilus*) exposed to environmental concentrations of the pharmaceutical 17α-ethinylestradiol and associated expression dynamics of aromatases and estrogen receptors. *Toxicol. Sci.* **106**, 113–123.

Lange A., Paull G. C., Hamilton P. B., Iguchi T. and Tyler C. R. 2011 Implications of persistent exposure to treated wastewater effluent for breeding in wild roach (*Rutilus rutilus*) populations. *Environ. Sci. Technol.* **45**, 1673–1679.

Langmead B., Trapnell C., Pop M. and Salzberg S. L. 2009 Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25.

Lauer S. J., Walker D., Elshourbagy N. A., Reardon C. A., Levy-Wilson B. and Taylor J. M. 1988 Two copies of the human apolipoprotein C-I gene are linked closely to the apolipoprotein E gene. *J. Biol. Chem.* **263**, 7277–7286.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N. *et al.* 2009 The sequence alignment/map format and SAM tools. *Bioinformatics* **25**, 2078–2079.

McKenna A., Hanna M., Banks E., Sivachenko A., Cibulskis K., Kernytsky A. *et al.* 2010 The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303.

Milano I., Babbucci M., Cariani A., Atanassova M., Bekkevold D., Carvalho G. R. *et al.* 2014 Outlier SNP markers reveal fine-scale genetic structuring across European hake populations (*Merluccius merluccius*). *Mol. Ecol.* **23**, 118–135.

Qu C., Liang X., Huang W. and Cao L. 2012 Isolation and characterization of 46 novel polymorphic EST-simple sequence repeats (SSR) markers in two sinipercine fishes (Siniperca) and cross-species amplification. *Int. J. Mol. Sci.* **13**, 9534–9544.

Romay M. C., Millard M. J., Glaubitz J. C., Peiffer J. A., Swarts K. L., Casstevens T. M. *et al.* 2013 Comprehensive genotyping of the USA national maize inbred seed bank. *Genome Biol.* **14**, R55.

Shirasawa K., Oyama M., Hirakawa H., Sato S., Tabata S., Fujioka T. *et al.* 2011 An EST-SSR linkage map of *Raphanus sativus* and comparative genomics of the Brassicaceae. *DNA Res.* **18**, 221–232.

Torre S., Tattini M., Brunetti C., Fineschi S., Fini A., Ferrini F. *et al.* 2014 RNA-seq analysis of *Quercus pubescens* leaves: *de novo* transcriptome assembly, annotation and functional markers development. *PLoS One* **9**, e112487.

Yang W., Qi Y., Bi K. and Fu J. 2012 Toward understanding the genetic basis of adaptation to high-elevation life in poikilothermic species: a comparative transcriptomic analysis of two ranid frogs, *Rana chensinensis* and *R. kukunoris*. *BMC Genomics* **13**, 588.

Zhang L., Yan H. F., Wu W., Yu H. and Ge X. J. 2013 Comparative transcriptome analysis and marker development of two closely related Primrose species (*Primula poissonii* and *P. wilsonii*). *BMC Genomics* **14**, 329.

Zheng X. H., Kuang Y. Y., Lu W. H., Cao D. C. and Sun X. W. 2014 Transcriptome-derived EST-SSR markers and their correlations with growth traits in crucian carp *Carassius auratus*. *Fish. Sci.* **80**, 977–984.

Zhou L., Wang Y., Yao B., Li C. J., Ji G. D. and Gui J. F. 2005 Molecular cloning and expression pattern of 14 kDa apolipoprotein in orange-spotted grouper, *Epinephelus coioides*. *Comp. Biochem. Physiol. B: Biochem. Mol. Biol.* **142**, 432–437.

Zou M., Guo B., Tao W., Arratia G. and He S. 2012 Integrating multi-origin expression data improves the resolution of deep phylogeny of ray-finned fish (Actinopterygii). *Sci. Rep.* **2**, 665.

Received 12 March 2015, in final revised form 3 June 2015; accepted 2 July 2015
Unedited version published online: 7 July 2015
Final version published online: 26 January 2016