Gene 511 (2012) 195–201

Contents lists available at SciVerse ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

First draft genome sequence of the Japanese eel, *Anguilla japonica*

Christiaan V. Henkel [a,1], Ron P. Dirks [a,*,1], Daniëlle L. de Wijze [b], Yuki Minegishi [b], Jun Aoyama [c], Hans J. Jansen [a], Ben Turner [d], Bjarne Knudsen [d], Martin Bundgaard [d], Kenneth Lyneborg Hvam [d], Marten Boetzer [e], Walter Pirovano [e], Finn-Arne Weltzien [f], Sylvie Dufour [g], Katsumi Tsukamoto [c], Herman P. Spaink [b], Guido E.E.J.M. van den Thillart [b]

[a] *ZF-screens B.V., Niels Bohrweg 11, 2333 CA Leiden, The Netherlands*
[b] *Institute of Biology, Leiden University, P.O. Box 9505, 2300 RA Leiden, The Netherlands*
[c] *Ocean Research Institute, University of Tokyo, 5-1-5, Kashiwanoha, Kashiwa, Chiba, 277–8564, Tokyo, Japan*
[d] *CLC bio, Finlandsgade 10–12, Katrinebjerg, DK-8200 Aarhus N, Denmark*
[e] *BaseClear B.V., Einsteinweg 5, 2333 CC Leiden, The Netherlands*
[f] *Norwegian School of Veterinary Science, P.O. Box 8146, 0033 Oslo, Norway*
[g] *UMR BOREA, CNRS 7208, Museum National d'Histoire Naturelle, 7, rue Cuvier, 75231 Paris Cedex 05, France*

ARTICLE INFO

Article history:
Accepted 11 September 2012
Available online 29 September 2012

Keywords:
Eel
Anguilla
Assembly pipeline
Illumina sequencing
De novo genome
Hox genes

ABSTRACT

The Japanese eel is a much appreciated research object and very important for Asian aquaculture; however, its genomic resources are still limited. We have used a streamlined bioinformatics pipeline for the *de novo* assembly of the genome sequence of the Japanese eel from raw Illumina sequence reads. The total assembled genome has a size of 1.15 Gbp, which is divided over 323,776 scaffolds with an N50 of 52,849 bp, a minimum scaffold size of 200 bp and a maximum scaffold size of 1.14 Mbp. Direct comparison of a representative set of scaffolds revealed that all the Hox genes and their intergenic distances are almost perfectly conserved between the European and the Japanese eel. The first draft genome sequence of an organism strongly catalyzes research progress in multiple fields. Therefore, the Japanese eel genome sequence will provide a rich resource of data for all scientists working on this important fish species.

1. Introduction

The first draft genome sequence of an organism has a major impact on research approaches in multiple fields, including molecular biology, physiology, ecology and evolution. Publication of the draft genomes of teleosts such as the Japanese pufferfish *Takifugu rubripes* (Aparicio et al., 2002), the green spotted pufferfish *Tetraodon nigroviridis* (Jaillon et al., 2004) and the medaka *Oryzias latipes* (Kasahara et al., 2007) has opened the window to the investigation of the entire genetic architecture of their biology. Combinations of complete genome information and other genome-wide scanning techniques, such as amplified fragment length polymorphisms (AFLP), quantitative traits loci (QTL) mapping and microarray or mRNA-Seq mediated transcriptome mapping, enable detailed investigation of the genetic mechanisms of specific phenomena like growth, osmoregulation, immune responses, stress tolerance and local adaptations. Those tools thus enable to "genomically" address central biological questions such as how fishes adapt to various water environments – including pollution – and how their remarkable morphological and ecological diversity has been controlled and evolved.

The freshwater eel, genus *Anguilla*, is a catadromous species, which migrates between growth habitats in freshwater and estuaries and spawning areas in the tropical and subtropical seas (Tesch, 2003). Previous studies have observed their migratory characteristics and revealed that anguillid species show various kinds and degrees of intra- and interspecific variation, such as migration scale, larval growth, spawning season and morphological characteristics (Aoyama, 2009); however, compared to e.g. salmon, little is known about the genetic backgrounds of eels (reviewed by Minegishi et al., 2012). Recently the draft genome sequence of the European eel has been determined (Henkel et al., 2012), and some mRNA-Seq studies of this species have been reported (Coppe et al., 2010) and are in progress (our unpublished data). In order to facilitate genome-wide studies of eels, comparative

Abbreviations: AFLP, amplified fragment length polymorphisms; MP, mate-pair; PE, paired-end; QTL, quantitative traits loci.

* Corresponding author at: ZF-screens B.V., Niels Bohrweg 11, 2333 CA Leiden, The Netherlands. Tel.: +31 71 3322270; fax: +31 84 8319969.
E-mail addresses: henkel@zfscreens.com (C.V. Henkel), dirks@zfscreens.com (R.P. Dirks), d.l.de.wijze@biology.leidenuniv.nl (D.L. de Wijze), y.minegishi@biology.leidenuniv.nl (Y. Minegishi), jaoyama@aori.u-tokyo.ac.jp (J. Aoyama), jansen@zfscreens.com (H.J. Jansen), bturner@clcbio.com (B. Turner), bknudsen@clcbio.com (B. Knudsen), mbundgaard@clcbio.com (M. Bundgaard), khvam@clcbio.com (K.L. Hvam), Marten.Boetzer@baseclear.com (M. Boetzer), Walter.Pirovano@baseclear.nl (W. Pirovano), Finn-Arne.Weltzien@nvh.no (F.-A. Weltzien), dufour@mnhn.fr (S. Dufour), ktpc@aori.u-tokyo.ac.jp (K. Tsukamoto), h.p.spaink@biology.leidenuniv.nl (H.P. Spaink), g.van.den.thillart@biology.leidenuniv.nl (G.E.E.J.M. van den Thillart).
[1] Contributed equally.

0378-1119
http://dx.doi.org/10.1016/j.gene.2012.09.064

genomics will be necessary. The Japanese eel *Anguilla japonica* is widely distributed from the northern Philippines to North Japan (Ege, 1939). They spawn near Western Mariana Ridge, and eggs and larvae born there are transported to growth habitats in eastern Asian countries by the North Equatorial Current and then the Kuroshio Current (Shinoda et al., 2011; Tsukamoto, 2009). After metamorphosis to glass eels, they migrate upstream to enter their growth phase for several years, and migrate back to the spawning area located at 2000–3000 km away (Aoyama, 2009). In addition to migratory ecology, much attention has been paid to the physiology of the Japanese eel. Endocrinological changes and hormone dynamics of, especially reproduction and osmoregulation, have been investigated (e.g. reviewed by Kazeto et al., 2011; Miura and Miura, 2011) Although research on this species has been intensively conducted and a genetic linkage map has been recently developed (Nomura et al., 2011), it is still a long way to understand the genomic mechanisms of their biology.

Fresh water eels, especially Japanese, European and American eels (respectively *Anguilla japonica*, *A. anguilla* and *A. rostrata*), are also a very nutritious and popular food source and their high economic value has caused overfishing. Together with other potential causes such as habitat destruction, pollution and diseases, this has resulted in a more than 90% decline in global stocks over the past decades (Dekker et al., 2003). A considerable percentage of consumption eels are grown to a marketable size in specific eel farms; however, eel farming is still fully dependent on wild caught juvenile eels (glass eels). Attempts toward artificial reproduction of fresh water eels already started in the 1930s (Fontaine, 1936). Although artificial maturation has now been shown for multiple eel species (Lokman and Young, 2000; Ohta et al., 1997; Oliveira and Hable, 2010; Pedersen, 2003), complete artificial reproduction has so far only been achieved for the Japanese eel (Tanaka et al., 2003) and has recently resulted in the first artificial F2 generation of eels (Nomura et al., 2011). For the first time, selective breeding of fresh water eels is now within reach. In the near future, this will allow the selection of eel brood stocks based on multiple parameters, including growth rate, speed of maturation, resistance against diseases, taste and fat content of the final product, etcetera. Traditional selective breeding has much to gain from molecular biology and genetics. Molecular techniques will contribute to more precise breeding programs and reduced inbreeding. Therefore, proper genetic resources for fresh water eels will be indispensible for selective breeding. Vice versa, artificial breeding of the Japanese eel will allow the generation of improved genetic linkage maps, which will again result in improved quality of these genetic resources. Here we used a versatile *de novo* assembly protocol to generate the first draft genome of the Japanese eel, *Anguilla japonica*. The Japanese eel genome will be an invaluable source of information, both for future applications in aquaculture and for scientific purposes.

2. Materials and methods

2.1. Genomic DNA libraries

Genomic DNA was isolated from blood of an adult female Japanese eel using the Blood and Tissue DNeasy kit (Qiagen GmbH, Hilden) according to the manufacturer's description. Paired-end libraries were prepared from 5 μg of isolated gDNA using the Paired-End Sequencing Sample Prep kit (Illumina Inc., San Diego) according to the manufacturer's description. Either a 280 bp band or a 600 bp band was cut from the gel (libraries PE280 and PE600). After amplification for 10 cycles the resulting libraries were analyzed with a Bioanalyzer 2100 DNA 1000 series II chip (Agilent, Santa Clara). Mate-pair libraries were prepared from 10 μg of isolated gDNA using the Mate Pair Library Prep Kit v2 (Illumina Inc., San Diego) according to the manufacturer's description. Agarose slices containing ~2-kb, ~4-kb, ~6-kb, ~8-kb or ~10-kb DNA fragments were cut from gel (libraries MP2K, MP4K, MP6K, MP8K, and MP10K). After the first gel purification the fragment length was analyzed using a Bioanalyzer 2100 DNA 12000 chip (Agilent). After circularization, shearing, isolation of biotinylated fragments, and amplification, the 400–600 bp fraction of the resulting fragments was isolated from gel. Finally, the libraries were examined with a Bioanalyzer 2100 DNA 1000 series II chip (Agilent).

2.2. Illumina sequencing

All libraries were sequenced using an Illumina GAIIx or HiSeq 2000 instrument according to the manufacturer's description. Genomic paired-end libraries were sequenced with a read length of 2×76 nucleotides or 2×151 nucleotides, and genomic mate-pair libraries with a read length of 2×76 nucleotides. Image analysis and base calling were done by the standard Illumina pipeline.

2.3. Genome assembly

Sequencing reads from both paired-end libraries were used in building the initial contigs. Both sets were preprocessed to eliminate low quality reads and nucleotides, as well as adapter contamination (mainly caused by insert sizes smaller than the read length). Because of the small insert size of the PE280 library, many read pairs from this library overlap at their 3' ends. When possible, these pairs were merged into longer single reads. This preassembly procedure has the dual advantage of producing long reads (which improve the quality and efficiency of the subsequent assembly) and providing confirmation for the identity of the 3' ends of the reads (which are generally determined with lesser confidence). We merged read pairs that exhibited at least eighteen nucleotides of unambiguous sequence overlap. Using this criterion, 29.6% of pairs could be merged, resulting in single reads with a mean length of 242 nt.

For initial contig assembly, we employed the CLC Assembly Cell de novo assembler (versions 3.2 and 4.0beta, CLC bio, Aarhus, Denmark). This is an efficient implementation of a De Bruijn graph-based assembler, which enables the assembly of the Japanese eel genome on a dual quad-core Xeon workstation with 48 GB of RAM installed in approximately eight hours. In the present study we included an extensive beta test in which we compared the performance of version 3.2 and the latest version 4.0 of the Assembly Cell. A total of 25 complete *de novo* assembly rounds were performed with k-mer values ranging from 21 to 31 nucleotides, and using bubble sizes of 50, 100, 250 or 400 bp. The k-mer parameter sets the length in nucleotides of nodes in the assembly graph. All individual sequencing reads are split into overlapping k-mers and connections between these. Smaller k-mers are less likely to be unique in the genome than high ones, resulting in a more tangled graph and therefore shorter contig sequences; however, using smaller k-mers, a larger number of connections between them can be extracted from sequencing reads. Therefore, at low coverage or using short reads, a lower k-mer setting may still yield larger contigs than a high one. For example, a single 76 nt read can yield a graph with fifty-six 21-mers and 55 connections, or one with forty-six 31-mers and 45 connections. It is difficult to predict which settings will yield the globally least tangled graph. The bubble parameter, new in CLC Assembly Cell 4.0, sets the maximum length of ambiguities in the k-mer graph (e.g. small repeats) for which the algorithm will attempt to resolve a path using long reads or pairs. Since our data include sequence information over distances far exceeding the k-mer lengths (i.e. merged reads up to 284 nt, pair inserts up to 750 nt), this addition is expected to increase the contiguity of the output.

Initial contigs were oriented in larger supercontigs (scaffolds) using SSPACE (Boetzer et al., 2011). Briefly, SSPACE aligns paired reads to the contigs (using Bowtie), and combines contigs if they are connected by at least a specified number of pairs within the limits set for the insert size of the pair library. The insert size is then used to estimate the size of the gap between the contigs. In addition, the algorithm can be forced