Song et al. GigaScience (2016) 5:5
DOI 10.1186/s13742-016-0112-y

GigaScience

DATA NOTE Open Access

Draft genome of the Chinese mitten crab, *Eriocheir sinensis*

Linsheng Song[1,2†], Chao Bian[3†], Yongju Luo[4†], Lingling Wang[5†], Xinxin You[3], Jia Li[3], Ying Qiu[3], Xingyu Ma[6], Zhifei Zhu[6], Liang Ma[7], Zhaogen Wang[7], Ying Lei[7], Jun Qiang[1], Hongxia Li[1], Juhua Yu[1], Alex Wong[8], Junmin Xu[3,6*], Qiong Shi[3,6*] and Pao Xu[1*]

Abstract

Background: The Chinese mitten crab, *Eriocheir sinensis*, is one of the most studied and economically important crustaceans in China. Its transition from a swimming to a crawling method of movement during early development, anadromous migration during growth, and catadromous migration during breeding have been attractive features for research. However, knowledge of the underlying molecular mechanisms that regulate these processes is still very limited.

Findings: A total of 258.8 gigabases (Gb) of raw reads from whole-genome sequencing of the crab were generated by the Illumina HiSeq2000 platform. The final genome assembly (1.12 Gb), about 67.5 % of the estimated genome size (1.66 Gb), is composed of 17,553 scaffolds (>2 kb) with an N50 of 224 kb. We identified 14,436 genes using AUGUSTUS, of which 7,549 were shown to have significant supporting evidence using the GLEAN pipeline. This gene number is much greater than that of the horseshoe crab, and the annotation completeness, as evaluated by CEGMA, reached 66.9 %.

Conclusions: We report the first genome sequencing, assembly, and annotation of the Chinese mitten crab. The assembled draft genome will provide a valuable resource for the study of essential developmental processes and genetic determination of important traits of the Chinese mitten crab, and also for investigating crustacean evolution.

Keywords: Crab genome, Genomics, Assembly, Annotation

Data description

Genomic DNA was extracted from muscle tissue of a single female crab (*Eriocheir sinensis*; NCBI Taxonomy ID: 95602) after 3 generations of inbreeding that was obtained from a local farm in Panjin, Liaoning Province, China. We used the whole-genome shotgun sequencing strategy and constructed the subsequent short-insert libraries (170, 250, 500 and 800 bp) and long-insert libraries (2, 5, and 10 kb) using the standard protocol provided by Illumina (San Diego, USA). Paired-end sequencing was performed by the Illumina HiSeq 2000 system. In total, we generated 258.8 Gb of raw reads from all constructed libraries.

We extracted clean reads of the short-insert libraries (500 or 800 bp) to estimate the crab genome size by k-mer frequency distribution analysis [1]. A k-mer is related to an artificial sequence division of K nucleotides iteratively from sequencing reads. We defined the k-mer length as 17 bp; thus, a L bp-long clean sequence would include (L-17 + 1) k-mers. The frequency of each k-mer can be calculated from the genome sequence reads. Typically, k-mer frequencies were plotted against the sequence depth gradient following a Poisson distribution in any given dataset. The genome size (G), can be deduced from the formula:

$$G = N \times (L\text{-}17 + 1)/K\_depth$$

where N is the total number of reads, and K_depth indicates the frequency that occurrs more often than other frequencies. In our calculations, N was 789,326,187

* Correspondence: xujunmin@genomics.cn; shiqiong@genomics.cn; xup@ffrc.cn
†Equal contributors
[3]Shenzhen Key Lab of Marine Genomics, Guangdong Provincial Key Lab of Molecular Breeding in Marine Economic Animals, BGI, Shenzhen 518083, China
[1]Key Laboratory of Freshwater Fisheries and Germplasm Resources Utilization, Ministry of Agriculture, Freshwater Fisheries Research Center, Chinese Academy of Fishery Sciences, Wuxi 214081, China
Full list of author information is available at the end of the article

Table 1 Summary of genome annotations

		Number	Average transcript length (bp)	Average coding sequence length (bp)	Average exons per gene	Average exon length (bp)	Average intron length (bp)
De novo	AUGUSTUS	14,436	10,104	1,195	4.97	240	2,245
	Genescan	29,097	13,045	1,022	5.01	203	2,995
Homolog	*H. sapiens*	5,646	4,752	922	3.74	246	1,398
	C. gigas	9,470	3,067	641	2.69	238	1,432
	C. elegans	3,142	3,913	819	3.27	250	1,361
	D.melanogaster	4,369	6,178	981	4.31	227	1,571
D. pulex		14,183	2,887	628	2.48	252	1,521
Transcriptome		14,123	11,161	2,223	6.83	325	1,532
GLEAN		7,549	12,742	1,470	6.36	230	2,101

and K_depth was 40; therefore, the crab genome size was estimated to be 1.66 Gb.

For whole-genome assembly, we employed Platanus [2] with optimized parameters (–k 27, –m 200) to construct contigs and original scaffolds. All reads were mapped onto contigs for scaffold building by utilizing the paired-end information. This paired-end information was subsequently applied to link contigs into scaffolds using a stepwise approach. Some intra-scaffold gaps were filled by local software using read-pairs in which one end uniquely mapped to a contig and the other end was located within a gap. Final genome assembly of the Chinese mitten crab is 1.12 Gb in total length, which is about 67.5 % of the estimated genome size. The contig N50 size (i.e., 50 % of the genome is in fragments of this length or longer) is 6.02 kb, and the scaffold (>2 kb) N50 is 224 kb.

We constructed a *de novo* repeat library using RepeatModeller (Version 1.04, default parameter) and LTR_FINDER [3]. To identify known and *de novo* transposable elements (TEs), we employed RepeatMasker (Version 3.2.9) [4] against the Repbase TE library [5] (Version 14.04) and the *de novo* repeat library. In addition, we used RepeatProteinMask (Version 3.2.2) implemented in RepeatMasker to detect the TE-relevant proteins. We also predicted tandem repeats utilizing Tandem Repeat Finder [6, 7] (Version 4.04) with parameters set as "Match = 2, Mismatch = 7, Delta = 7, PM = 80, PI = 10, Minscore = 50, and MaxPerid = 2000". Finally, we confirmed that the repeat sequences occupy approximately 50.4 % of the crab genome. Among them, the long interspersed elements, occupying 19.0 % of the crab genome, are the most predominant type of repeat sequences.

Subsequently, we performed annotation analysis containing four major steps. (1) The homology-based gene prediction: We aligned *Homo sapiens*, *Crassostrea gigas*, *Caenorhabditis elegans*, *Drosophila melanogaster* and *Daphnia pulex* proteins (Ensembl release 75) to the crab genome using TblastN with an E-value ≤ 1E-5, and then made use of GeneWise2.2.0 [7] for precise spliced alignment and predicting gene structures. Short genes (<150 bp) and premature or frame-shifted genes were removed. (2) The *ab initio* prediction: Genome sequences of the crab were repeat-masked, and 1500 full-length, randomly selected genes from their homology gene sets were used to train the model parameters for AUGUSTUS2.5 [8]. We then utilized AUGUSTUS2.5 and GENSCAN1.0 [9] for *de novo* prediction on repeat-masked genome sequences. Short genes were discarded using the same filter threshold that was used for homology prediction. (3) Gene structure identification using transcriptome reads: We mapped the mixed RNA reads (from hepatopancreas tissue taken from four molting stages) reported in Huang's study [10] on the crab genome using TopHat1.2 [11]. Subsequently, we sorted and merged the TopHat mapping results and then applied Cufflink [12] software to identify gene structures to assist gene annotation. (4) Gene set integration: All of the above gene sets were merged to form a comprehensive and non-redundant gene set using GLEAN [13]. We obtained a final gene set containing 7,549 genes (Table 1), which is more than the gene number (5,775) identified for horseshoe crab [14]. Meanwhile, the CEGMA [15] evaluation demonstrated the annotation completeness to be 66.9 % (166 of 248 core eukaryote genes were aligned).

In summary, we report the first genome sequencing, assembly, and annotation of the Chinese mitten crab. The draft genome will provide a valuable resource for studying essential developmental processes in the Chinese mitten crab, investigating crustacean evolution, and improving the molecular breeding of this economically important species.

Availability of supporting data

Supporting data are available in the GigaDB database [16], and the raw data were deposited in the PRJNA305216.

Abbreviations

Gb: Gigabase; TE: Transposable element.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

LS, QS and PX conceived the project. LW, YL, XM, LM, ZW, YL, JQ, HL, JQ, JY and ZZ collected the samples and extracted the genomic DNA. CB led the genome analysis, conducted the genome assembling, and predicted gene structure and repeat sequences. CB, QS, XY, JL, YQ, JX and AW wrote the article. All authors participated in discussion of the project and data. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by the China 863 Project (No. 2014AA093501), the Special Project on the Integration of Industry, Education and Research of Guangdong Province (No. 2013B090800017), and the Shenzhen Scientific R & D Grant (No. CXB201108250095A).

Author details

[1]Key Laboratory of Freshwater Fisheries and Germplasm Resources Utilization, Ministry of Agriculture, Freshwater Fisheries Research Center, Chinese Academy of Fishery Sciences, Wuxi 214081, China. [2]College of Fisheries and Life Science, Dalian Ocean University, Dalian 116023, China. [3]Shenzhen Key Lab of Marine Genomics, Guangdong Provincial Key Lab of Molecular Breeding in Marine Economic Animals, BGI, Shenzhen 518083, China. [4]Guangxi Academy of Fisher Sciences, Nanning 530021, China. [5]Key Laboratory of Experimental Marine Biology, Institute of Oceanology, Chinese Academy of Sciences, Qingdao 266071, China. [6]BGI Zhenjiang Institute of Hydrobiology, Zhenjiang 212000, China. [7]Zhenjiang Agriculture Committee, Zhenjiang 212000, China. [8]BGI-Hong Kong, Hong Kong 999077, China.

Received: 10 December 2015 Accepted: 12 January 2016
Published online: 28 January 2016

References

1. Li R, Fan W, Tian G, Zhu H, He L, Cai J, et al. The sequence and de novo assembly of the giant panda genome. Nature. 2010;463:311–7.
2. Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, et al. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Res. 2014;24:1384–95.
3. Xu Z, Wang H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. 2007;35:W265–8.
4. Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. Current protocols in bioinformatics/editoral board, Andreas D Baxevanis [et al.] 2004; Chapter 4:Unit 4 10.
5. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 2005;110:462–7.
6. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27:573–80.
7. Birney E, Clamp M, Durbin R. GeneWise and Genomewise. Genome Res. 2004;14:988–95.
8. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. 2006;34:W435–9.
9. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. Journal Mol Bio. 1997;268:78–94.
10. Huang S, Wang J, Yue W, Chen J, Gaughan S, Lu W. Transcriptomic variation of hepatopancreas reveals the energy metabolism and biological processes associated with molting in Chinese mitten crab, Eriocheir sinensis.
11. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25:1105–11.
12. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotech. 2010;28:511–5.
13. Elsik CG, Mackey AJ, Reese JT, Milshina NV, Roos DS, Weinstock GM. Creating a honey bee consensus gene set. Genome Biol 2007;8:R13.
14. Nossa CW, Havlak P, Yue JX, Lv J, Vincent KY, Brockmann HJ, et. al. Joint assembly and genetic mapping of the Atlantic horseshoe crab genome reveals ancient whole genome duplication. GigaScience 2014;3:9.
15. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 2007;23:1061-7.
16. Song L, Bian C, Luo Y, Wang L, You X, Li J, Qiu Y, Ma X, Zhu Z, Ma L, Wang Z, Lei Y, Qiang J, Li H, Yu J, Wong A, Xu J, Shi Q, Xu P. Supporting data for the "Draft genome of the Chinese mitten crab, Eriocheir sinensis". GigaScience Database. 2016. http://dx.doi.org/10.5524/100186.