The big-headed turtle, *Platysternon megacephalum*, as the sole member of the monotypic family Platysternidae, has a number of distinct characteristics including an extra-large head, long tail, flat carapace, and a preference for low water temperature environments. We performed whole genome sequencing, assembly, and gene annotation of an adult male big-headed turtle based on the Illumina HiSeq X genomic sequencing platform. We generated ~497.1 Gb of raw sequencing data (× 208.9 depth) and produced a draft genome with a total length of 2.32 Gb and contig and scaffold N50 sizes of 41.8 kb and 7.22 Mb, respectively. We also identified 924 Mb (39.84%) of repetitive sequences, 25,995 protein-coding genes, and 19,177 non-coding RNAs. We generated the first *de novo* genome of the big-headed turtle; these data will be essential to the further understanding and exploration of the genomic innovations and molecular mechanisms contributing to its unique morphology and physiological features.

Background & Summary

The big-headed turtle, *Platysternon megacephalum* (NCBI Taxonomy ID: 55544), as the sole member of the family Platysternidae, is native to Southeast Asian countries, including China, Cambodia, Laos, Myanmar, Thailand and Vietnam. The species inhabits flowing cool rocky mountain streams with water temperatures usually between 12–28 °C. It is an aquatic predator with strong climbing abilities, preying on lizard, frogs, fish, shrimps, crabs, snails, earthworms, insects, and even small birds and mammals, along with consuming some fruit and plant matter. The big-headed turtle has some defining features, such as an extra-large head, long tail, and flat carapace. Its head width is approximately equal to half of its carapace width and therefore cannot be retracted into the shell. This species has the longest tail relative to body size of any turtle, with tail length being more than half of the carapace length. In addition, it has a very flat carapace and an eagle-like hooked upper jaw (Fig. 1). The species prefers its body temperature to be around 25.3 °C. While big-headed turtles are unique in many morphological and physiological characteristics compared to other turtle species, they are not unique in the anthropogenic threats they face.

The Asian turtle crisis has resulted in population declines in many species, including the big-headed turtle. Over harvesting for the pet trade, traditional medicine, and food represent the main drivers of the population decline and have resulted in an over 89% reduction in population density in South China. The species was listed in the CITES (Convention on International Trade in Endangered Species of Wild Fauna and Flora) Appendix I in 2013 and as an endangered species in the IUCN (International Union for Conservation of Nature) Red List of Threatened Species in 2001. Conservation and restoration measures are needed to protect this endangered species, and a genomic resource will be an essential tool for conservation efforts. However, there is currently little knowledge about the big-headed turtle's genetic information, with previous research mainly focusing on mitochondrial DNA and the development of microsatellite markers.

In this study, we report the first sequencing, assembly, and annotation of the big-headed turtle genome. The final draft genome assembly was approximately 2.32 Gb with a contig N50 of 41.8 kb and scaffold N50 of 7.22 Mb. A total of 25,995 protein-coding genes and 19,177 non-coding RNAs (409 rRNAs, 2,089 tRNAs, 16,050 miRNAs, and 629 snRNAs) were predicted from the genome assembly. The genomic resource of the big-headed turtle will be a key tool in the study of conservation genetics for this species.
Methods

Sample collection and sequencing. The genomic DNA of the big-headed turtle was extracted from the leg muscles of a single adult male obtained from Heyuan, Guangdong Province, China. The sampled turtle was one of a number that the law enforcement agencies confiscated from the black market and then transferred to scientific research institutions for study and captive breeding. Our institute has government permission to use confiscated big-headed turtles for scientific research (e.g. conservation genetics, ecology). The sampled turtle was euthanized and the animal collection and utility protocols were reviewed and approved by the Animal Ethics Committee at the Guangdong Institute of Applied Biological Resources (No: GIABR20170103). Ten paired-end libraries including four short-insert libraries (250 bp × 2, and 500 bp × 2) and six long-insert libraries (2 kb × 2, 5 kb × 2, and 10 kb × 2) were constructed and sequenced on an Illumina HiSeq X platform according to the manufacturer’s instructions (Illumina, San Diego, California, USA). The sequenced read length was 150 bp for each library. A total of 497.1 Gb (208.9×) of raw sequences were eventually obtained (Table 1). Prior to assembly, quality control was performed for raw reads using a SOAPfilter to filter out the adaptor sequences, the reads containing more than 10% unidentified nucleotides, and low-quality reads containing more than 50% bases with Illumina phred quality score ≤8. We obtained approximately 469.2 Gb of clean reads for further assembly.

Assembly and evaluation. A total of 170 Gb high-quality reads from the short-insert reads (350 bp) were used to estimate the genomic information of the big-headed turtle, and 17-mer frequency information was generated based on the K-mer analysis as implemented in the software GEC12. The heterozygous ratio was also evaluated based on the frequency of the heterozygous k-mers and homozygous k-mers using GCE software12,13. According to 17-mer analysis the estimated genome size of big-headed turtle was 2,383.87 Mb (~2.38 Gb), and the estimated heterozygous and repeat sequencing ratios were calculated to be 0.33 and 53%, respectively (Fig. 2 and Table 2).

De novo assembly was performed from the generated clean reads using SOAPdenovo14, a de Bruijn graph algorithm-based de novo genome assembler. We assembled the big-headed turtle genome in three steps: contig construction, scaffolding, and gap filling. First, three K (45, 59, 71) values were used to assemble the genome, according to the N50 length of contig and the BUSCO assessments of three genomes. The genome of 59-mer was chosen as the final genome for subsequent analysis. The clean reads of short-insert libraries (250 bp and 500 bp) were used to construct the contigs with 59-mer. Then reads of long-insert libraries (2 kb, 5 kb and 10 kb) were implemented to link the contig sequences into scaffold sequences. To further improve assembly quality, GAPcloser14 and SSPACE15 were applied to reduce gap regions and raise scaffold length using a genome sketch assembled by SOAPdenovo. This last step improved the contig N50 and N90 sizes to 41,757 and 5,528 bp, and the scaffold N50 and N90 sizes to 7,221,511 and 257,323 bp, respectively, with the fragments being longer than 100 bp (Table 3). The final assembly of the big-headed turtle genome had a total length of 2.32 Gb, which was similar to the three previously published turtle genomes: Chrysemys picta bellii16, Chelonia mydas and Pelodiscus sinensis17.

Table 1. Statistics of big-headed turtle genome sequencing data. *Sequence coverage was calculated based on the genome size of 2.38 Gb according to k-mer analysis.

| Insert size | Libraries | Read length (bp) | Raw data | Clean data |
|------------|-----------|----------------|----------|------------|
|            |           |                | Total data (Gb) | Sequence coverage (×)* |
|            |           |                | Total data (Gb) | Sequence coverage (×)* |
| 250bp      | 2         | 150            | 120.9     | 50.8       | 112.6     | 47.3     |
| 500bp      | 2         | 150            | 103.6     | 43.5       | 96.5      | 40.5     |
| 2Kbp       | 2         | 150            | 98.9      | 41.6       | 92.2      | 38.7     |
| 5Kbp       | 2         | 150            | 78.6      | 33.0       | 75.5      | 31.7     |
| 10Kbp      | 2         | 150            | 95.1      | 40.0       | 92.4      | 38.8     |
| Total      | 10        |                | 497.1     | 208.9      | 469.2     | 197.1    |

Fig. 1 A representative big-headed turtle, *Platysternon megacephalum* in China.
The key assembly statistics of the big-headed turtle genome are comparable to or better than those of previously published turtle genomes (Table 4).

**Annotation**

**Repeat annotation.** There are two major types of repetitive sequences: tandem repeats and interspersed repeats. For the repeat annotation of the big-headed turtle genome, both homology-based predictions and de novo methods were performed. In the homolog-based methods, interspersed repeats were identified using RepeatMasker and RepeatProteinMask to search against the published RepBase sequences. In the de novo method, RepeatMasker and RepeatModeler were used to detect interspersed repeats in the genome. Tandem Repeats Finder (TRF) was subsequently used to search for tandem repeats. Overall, the results identified a total of 924 Mb of non-redundant repetitive sequences in the big-headed turtle genome, which account for 39.84% of the whole genome (Table 5). The most predominant elements were long interspersed nuclear elements (LINEs), which accounted for 27.47% (637 Mb) of the genome (Table 6).

**Structural annotation of genes.** Three methods (homology-based, ab initio and transcriptome-based predictions) were used to predict gene structure in the big-headed turtle genome. In the homology-based method, the protein repertoires of *Alligator sinensis*, *Chelonia mydas*, *Chrysemys picta bellii*, *Deinagkistrodon acutus*,...
**Gallus gallus, Gekko japonicus, Nanorana parkeri, Parus major, Pelodiscus sinensis, Philomachus pugnax and Xenopus laevis** were downloaded from the NCBI database and mapped onto the big-headed turtle genome using TBLASTn with an E-value cutoff of $1 \times 10^{-5}$. Then, homologous genome sequences were aligned against the matching proteins using GeneWise to define gene models. In the ab initio prediction, Augustus, GlimmerHMM and SNAP were used to predict the coding regions of genes. To optimize the genome annotation, seven RNA tissue libraries (liver, skin, lung, intestine, heart, muscle and stomach) were constructed according to the manufacturer's instructions (Illumina, San Diego, California, USA) and a total of 61.27 Gb of sequence data was generated. RNA-seq reads were used in de novo assembly with Trinity. The unique transcriptional sequences were employed to predict gene models using PASA. In total, 25,995 non-redundant protein-coding genes were annotated in the big-headed turtle genome (Table 7).

### Functional annotation of genes.

The functional annotation of protein-coding genes of the big-headed turtle genome were predicted by aligning protein sequences against public databases including SwissProt, KEGG and TrEMBL by searching against the Rfam database. We identified a total of 409 rRNA, 2,089 tRNA, 16,050 miRNA, and 629 snRNA genes in the big-headed turtle genome (Table 9).

### Non-coding RNA annotation.

The non-coding RNAs were predicted in the big-head genome based on four categories: ribosomal RNA (rRNA), transfer RNA (tRNA), microRNAs (miRNA) and small nuclear RNA (snRNA). rRNA was applied to identify tRNA with eukaryotic parameters according to the characteristics of tRNA. Because of the highly conserved characteristics of rRNA, BLASTN was used to predict rRNA sequences by aligning with a human template with an E-value of $1 \times 10^{-5}$. The miRNA and snRNA sequences were identified using INFERNAL by searching against the Rfam database. We identified a total of 409 rRNA, 2,089 tRNA, 16,050 miRNA, and 629 snRNA genes in the big-headed turtle genome (Table 9).
Data Records
This Whole Genome Shotgun project including the assembled genome sequence and the structural and functional annotation of genes has been deposited at DDBJ/ENA/GenBank under the accession QXTE00000000.33. The version described in this paper is version QXTE01000000. Repeat annotation and Non-coding RNA annotation have been deposited at Figshare34. Raw read files have been deposited at NCBI Sequence Read Archive under the accession SRP15641935.
To assess the quality of the genome assembly, we performed three independent evaluations as described below. First, the base content was counted with scaffolds longer than 100 bp and the results showed that the GC content for the big-headed turtle was 44.63% (Table 10), which was comparable to those of *Chrysemys picta bellii*, *Chelonia mydas* and *Pelodiscus sinensis* (Table 4). Second, the short-insert paired-end reads (250 bp and 500 bp) were mapped to the genome with BWA. The mapping rate was 98.84% and the genome coverage was 99.57% (Table 11), indicating high reliability of genome assembly. Third, the SNPs (Single Nucleotide Polymorphisms) were counted to validate the uniformity of the genome using SAMtools37, and we found the ratio of homozygous SNPs was only 0.014% (Table 12), indicating that the assembly had a high base accuracy. Finally, CEGMA (http://korflab.ucdavis.edu/dataset/cegma/) and BUSCO (http://busco.ezlab.org/) were used to evaluate the completeness of the assembly. CEGMA assessment showed that our assembly captured 226 (91.13%) of the 248 ultra-conserved core eukaryotic genes, of which 202 (81.45%) were complete (Table 13). BUSCO analysis showed that 95.2% and 2.6% of the expected vertebrate genes were identified as complete and fragmented, respectively, while 2.2% were considered missing in the assembly (Table 14).

**Technical Validation**

**Assessing the completeness of the genome assembly.** To assess the quality of the genome assembly, we performed three independent evaluations as described below. First, the base content was counted with scaffolds longer than 100 bp and the results showed that the GC content for the big-headed turtle was 44.63% (Table 10), which was comparable to those of *Chrysemys picta bellii*, *Chelonia mydas* and *Pelodiscus sinensis* (Table 4). Second, the short-insert paired-end reads (250 bp and 500 bp) were mapped to the genome with BWA. The mapping rate was 98.84% and the genome coverage was 99.57% (Table 11), indicating high reliability of genome assembly. Third, the SNPs (Single Nucleotide Polymorphisms) were counted to validate the uniformity of the genome using SAMtools, and we found the ratio of homozygous SNPs was only 0.014% (Table 12), indicating that the assembly had a high base accuracy. Finally, CEGMA (http://korflab.ucdavis.edu/dataset/cegma/) and BUSCO (http://busco.ezlab.org/) were used to evaluate the completeness of the assembly. CEGMA assessment showed that our assembly captured 226 (91.13%) of the 248 ultra-conserved core eukaryotic genes, of which 202 (81.45%) were complete (Table 13). BUSCO analysis showed that 95.2% and 2.6% of the expected vertebrate genes were identified as complete and fragmented, respectively, while 2.2% were considered missing in the assembly (Table 14).

**Annotation filtering and validation.** The EVM software was used to merge the above results of gene annotation and 39,212 genes were obtained from the merged set. To further revise the genome annotation, we removed the following types of genes: (1) genes with overlap regions of TE ≥ 20%; (2) premature termination genes; (3) genes with only de novo predictive support. After filtering, 25,995 genes were retained. In addition, total RNA-seq reads of 7 tissues were mapped onto the big-headed turtle genome to further identify exon regions and splice positions using Tophat and Cufflinks, and 20,028 (77.05%) genes had evidence supports of RNA data (RPKM value > 1) cross above 7 tissues.

### Table 10. Base content statistics of the genome. *GC content of the genome without N.*

| Type | Number (bp) | % of genome |
|------|-------------|-------------|
| A    | 632,158,059 | 27.25       |
| T    | 631,959,166 | 27.25       |
| C    | 509,623,076 | 21.97       |
| G    | 509,248,147 | 21.97       |
| N    | 36,532,422  | 1.57        |
| Total| 2,319,520,870| —           |
| GC*  | 1,018,871,223| 44.63       |

### Table 11. Statistics of mapping ratio in genome.

| Type    | Value                                      |
|---------|--------------------------------------------|
| Reads   | Mapping rate (%)                          |
| Genome  | Average sequencing depth (x)              |
|         | Coverage (%)                              |
|         | Coverage at least 4x (%)                  |
|         | Coverage at least 10x (%)                 |
|         | Coverage at least 20x (%)                 |
|         |                                            |
|         | 98.84                                      |
|         | 72.05                                     |
|         | 99.57                                     |
|         | 98.73                                     |
|         | 97.25                                     |
|         | 95.38                                     |

### Table 12. Number and density of SNPs in big-headed turtle genome.

| Type               | Number | Proportion (%) |
|--------------------|--------|----------------|
| All SNPs           | 5,319,363| 0.233%         |
| Heterozygous SNPs  | 4,999,745| 0.219%         |
| Homozygous SNPs    | 319,618 | 0.014%         |

### Table 13. Assessment of CEGMA.

| Species            | Complete Proteins | Completeness (%) | Complete + Partial Proteins | Completeness (%) |
|--------------------|-------------------|------------------|----------------------------|------------------|
| Big-headed turtle  | 202               | 81.45            | 226                        | 91.13            |
| Species                  | Size       | BUSCO notation assessment results* |
|-------------------------|------------|-----------------------------------|
| Big-headed turtle       | 2320 Mb    | C: 95.2% [S: 94.2%, D: 1.0%], F: 2.6%, M: 2.2%, n: 2586 |

Table 14. Assessment of BUSCO. *C: Complete BUSCOs; S: Complete and single-copy BUSCOs; D: Complete and duplicated BUSCOs; F: Fragmented BUSCOs; M: Missing BUSCOs; n: Total BUSCO groups searched.

Code Availability

The execution of this work involved using many software tools, whose settings and parameters are described below.

- **(1) GCE:** version1.0.0, parameters: -H 1; **(2) SOAPdenovo:** version2, k-mer size of 59; **(3) GAPcloser:** version1.12, parameters: -l 150 -p 31; **(4) SSPACE:** version3.0, default parameters; **(5) RepeatMasker:** Repeat Masker-open-4-0-6, parameters: -a -nolow -no_is -norna -parallel 1; **(6) RepeatModeler:** RepeatModeler-open-1.0.11, parameters: -database genome -engine ncbi -pa 15; **(7) Tandem Repeats Finder:** TRF-407b, parameters: 2 7 7 80 10 50 2000 -d -h; **(8) TBLASTn:** blast-2.2.26, parameters: -p tblastn -e 1e-05 -F T -m 8 -d; **(9) GeneWise:** version2.4.1, parameters: -tfor -genef -gff; **(10) Augustus:** version3.2.3, parameters: -uniqueGeneId = true -noInFrameStop = true -gff3 = on -genemodel = complete -strand = both; **(11) GlimmerHMM:** version3.0.1, parameters: -g -f; **(12) SNAP:** snap-2013-11-29, default parameters; **(13) Trinity:** trinityrnaseq-2.1.1, parameters: -seqType fq-CPU 20-max_memory 200G-normalize_reads-full_cleanup-min_glue 2-min_kmer_cov 2-KMER_SIZE 25; **(14) PASA:** PASA_r20140417, default parameters; **(15) InterPro:** version29.0, perl-based version4.8, default parameters; **(16) rRNAscan-SE:** rRNAscan-SE-1.3.1, default parameters; **(17) INFERNAL:** version1.1c4 (June 2013); **(18) BLASTp:** blast-2.2.26, parameters: -p blastn -e 1e-10 -v 10000 b 10000; **(19) BWA:** bwa-0.7.10, parameters: -a mem -k 31 -w 10 -b 3 -O 1 -I 4 -t 10; **(20) SAMTools:** samtools-0.1.19, parameters: mpileup mpileup -m 2 -u; **(21) EVM:** VidenceModeler-1.1.1, parameters: -segmentSize 20000–overlapSize 20000; **(22) TopHat:** tophat-2.0.13, parameters: -p 6–max-intron-length 500000 -m 2-library-type fr-unstranded; **(23) Cufflinks:** cufflinks-2.1.1, parameters: -l 50000 -p 1--library-type fr-unstranded -I 1 -CUFF; **(24) BUSCO:** version3.0.2, OrthoDBv9_vertebrata.

References

1. Rhodin, A. G. J. et al. Turtles of The World: Annotated Checklist And Atlas Of Taxonomy, Synonymy, Distribution, And Conservation Status 8th edn (Chelonian Research Foundation and Turtle Conservancy, 2017).
2. Shen, J. W., Pike, D. A. & Du, W. G. Movements and microhabitat use of translocated big-headed turtles (Platysternon megacephalum) in southern China. Chelonian Conserv. Bi. 9, 135–161 (2010).
3. Sun, Y. H., Hau, B. C. & Karraker, N. E. Diet of the endangered big-headed turtle Platysternon megacephalum. PeerJ 4, e2784 (2016).
4. Zhang, M. W., Zong Y. & Ma J. F. Fauna Sinica (Reptilia 1): General Accounts Of Reptilia, Testudines, And Crocodyliformes (Science Press, 1998).
5. Shen, J. W., Meng, F. W., Zhang, Y. P. & Du, W. G. Field body temperature and thermal preference of the big-headed turtle Platysternon megacephalum. Curr. Zool. 59, 626–632 (2013).
6. Gong, S. P. et al. Disappearance of endangered turtles within China’s nature reserves. Curr. Biol. 27, R170–R171 (2017).
7. Zheng, C. et al. Recombination and evolution of duplicate microsatellite loci in the mitochondrial genome of the Asian big-headed turtle Platysternon megacephalum. Plas One 8, e28854 (2013).
8. Peng, Q. L., Nie, L. W. & Pu, Y. G. Complete mitochondrial genome of Chinese big-headed turtle, Platysternon megacephalum, with a novel gene organization in vertebrate mtDNA. Gene 380, 14–20 (2006).
9. Parhame, J. F., Feldman, C. R. & Boore, J. L. The complete mitochondrial genome of the enigmatic bigheaded turtle (Platysternon): description of unusual genomic features and the reconciliation of phylogenetic hypotheses based on mitochondrial and nuclear DNA. BMC Evol. Biol. 6, 11 (2006).
10. Hua, L. et al. Isolation and characterization of 14 polymorphic microsatellite loci in the big-headed turtle (Platysternon megacephalum). Biochem. Genet. 52, 203–208 (2014).
11. Peng, J. et al. New microsatellite resources in Chinese big-headed turtle (Platysternon megacephalum). Conserv. Genet. Resour. 2, 55–57 (2010).
12. Liu, B. et al. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. Quant. Biol. 35, 62–67 (2013).
13. Li, X. & Waterman, M. S. Estimating the repeat structure and length of DNA sequences using L-tuples. Genome Res. 13, 1916–1922 (2003).
14. Li, R. et al. de novo assembly of human genomes with massively parallel short read sequencing. Genome Res. 20, 265–272 (2010).
15. Bozert, M. & Provanov, W. SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. BMC Bioinformatics 15, 211 (2014).
16. Shaffer, H. B. et al. The western painted turtle genome, a model for the evolution of extreme physiological adaptations in a slowly evolving lineage. Genome Biol. 14, R28 (2013).
17. Wang, Z. et al. The draft genomes of soft-shell turtle and green sea turtle yield insights into the evolution and development of the turtle-specific body plan. Nat. Genet. 45, 701–706 (2013).
18. Bhagavan, N. V. & Ha, C. E. Essentials of Medical Biochemistry. (Academic Press, 2011).
19. Maja, T.-G. & Nansheng, C. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr. Protoc. Bioinformatics 25, 4.10.11–4.10.14 (2009).
20. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27, 573–580 (1999).
21. Kent, W. J. BLAT—the BLAST-like alignment tool. Genome Res. 12, 656–664 (2002).
22. Birney, E., Clamp, M. & Durbin, R. Genome wide and Genomewise. Genome Res. 14, 988–995 (2004).
23. Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics 19(Suppl 2), ii215–225 (2003).
24. Majoros, W. H., Perteza, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics 20, 2878–2879 (2004).
25. Korf, I. Gene finding in novel genomes. BMC Bioinformatics 5, 59 (2004).
26. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29, 644–652 (2011).
27. Campbell, M. A., Haas, B. J., Hamilton, J. P., Mount, S. M. & Buell, C. R. Comprehensive analysis of alternative splicing in rice and comparative analyses with Arabidopsis. BMC Genomics 7, 327 (2006).
28. Mulder, N. & Apweiler, R. Inter Pro and Inter Pro Scan: tools for protein sequence classification and comparison. Methods Mol. Biol. 396, 59–70 (2007).
29. Ashburner, M. et al. Gene ontology: tool for the unification of biology. Nat. Genet. 25, 25–29 (2000).
30. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25, 955–964 (1997).
31. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. J. Mol. Biol. 215, 403–410 (1990).
32. Nawrocki, E. P., Kolbe, D. L. & Eddy, S. R. Infernal 1.0: inference of RNA alignments. Bioinformatics 25, 1335–1337 (2009).
33. Cao, D. N., Wang, M., Ge, Y. & Gong, S. P. Platysternon megacephalum isolate DO16091913, whole genome shotgun sequencing project. GenBank, http://identifiers.org/ncbi/insdc:QXTE01000000 (2018).
34. Cao, D. N., Wang, M., Ge, Y. & Gong, S. P. Repeat annotation and Non-coding RNA annotation of Platysternon megacephalum. figshare, https://doi.org/10.6084/m9.figshare.7586408.v1 (2019).
35. NCBI Sequence Read Archive. http://identifiers.org/ncbi/insdc.sra:SRP156419 (2018).
36. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760 (2009).
37. Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2078 (2009).
38. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. 9, R7 (2008).
39. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25, 1105–1111 (2009).
40. Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. 28, 511–515 (2010).

Acknowledgements
The authors would like to thank Dr. Daniel Gaillard for checking the manuscript and Mr. Yufeng Wei for his help in sample collection. This work was supported by the National Natural Science Foundation of China (31471966), Guangdong Natural Science Foundation (2015A030313903), and the GDAS Special Project of Science and Technology Development (2017GDASCX-0107; 2018GDASCX-0107).

Author Contributions
S.P.G. and D.N.C. designed and supervised the project. M.W. performed bioinformatics analyses. Y.G. collected the samples. D.N.C., S.P.G. and M.W. were involved in the data analyses and wrote the manuscript. All authors have read and approved the final manuscript.

Additional Information

Competing Interests: The authors declare no competing interests.

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

The Creative Commons Public Domain Dedication waiver http://creativecommons.org/publicdomain/zero/1.0/ applies to the metadata files associated with this article.

© The Author(s) 2019