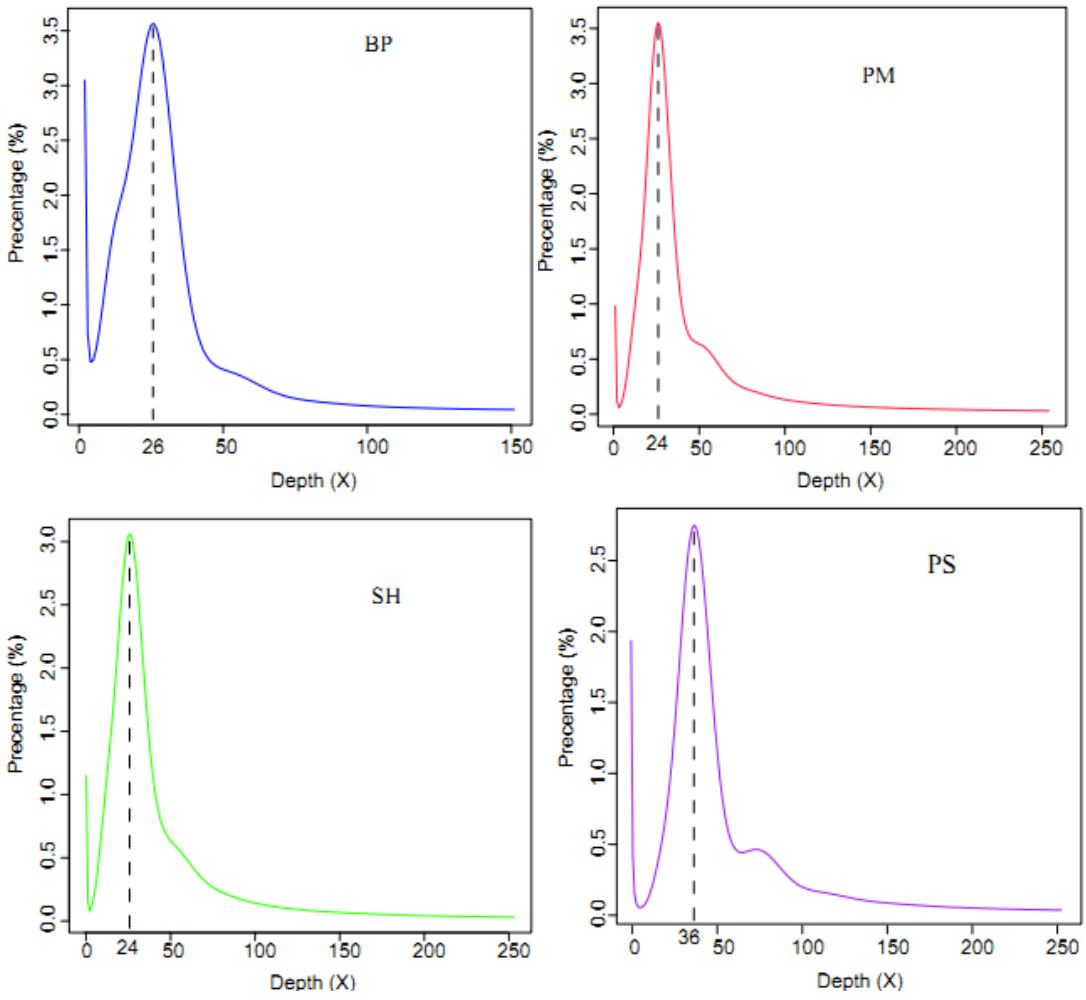
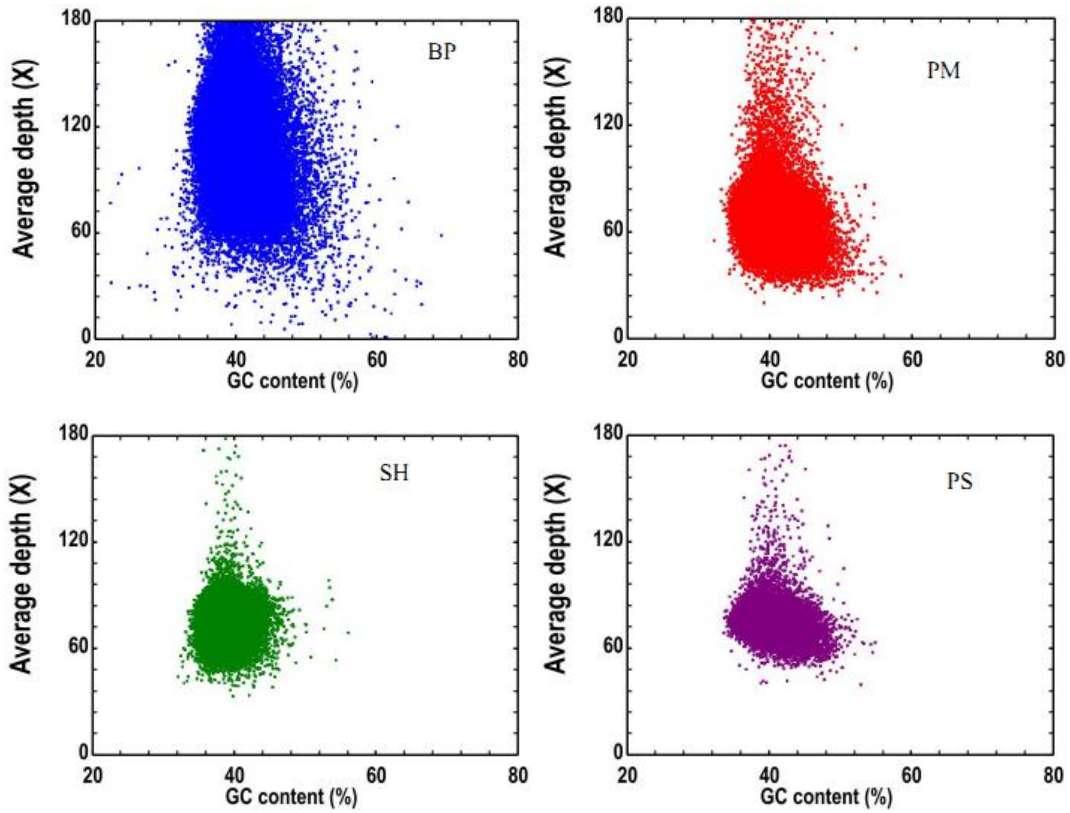
Supplementary Information

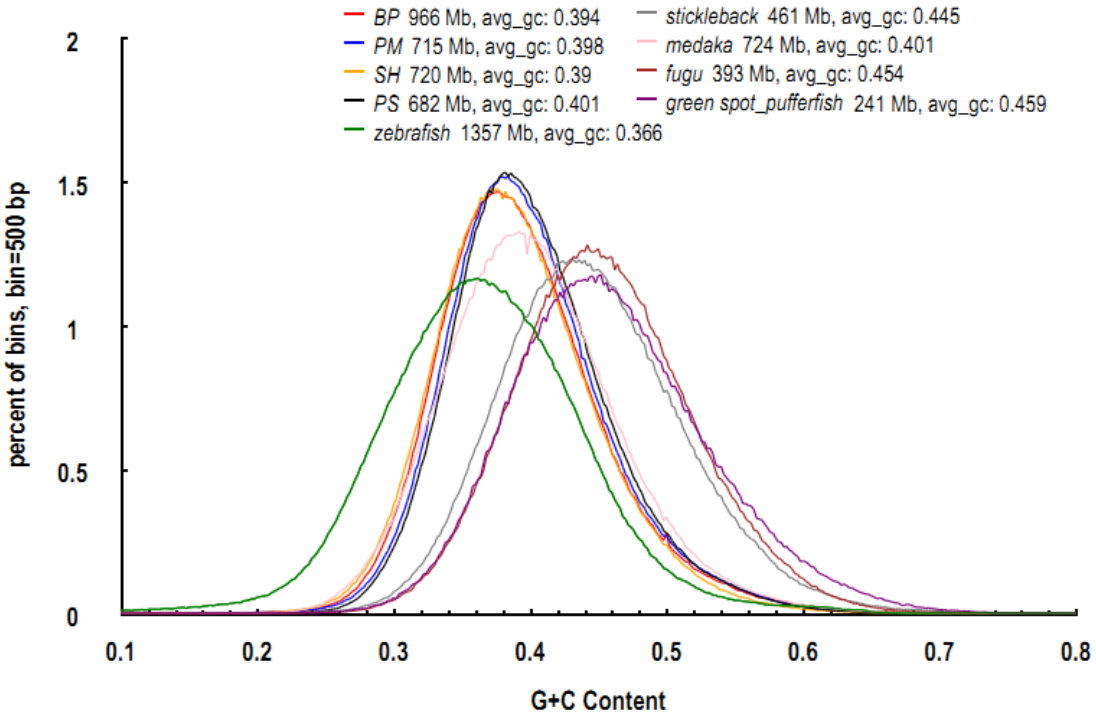
Supplementary Figures



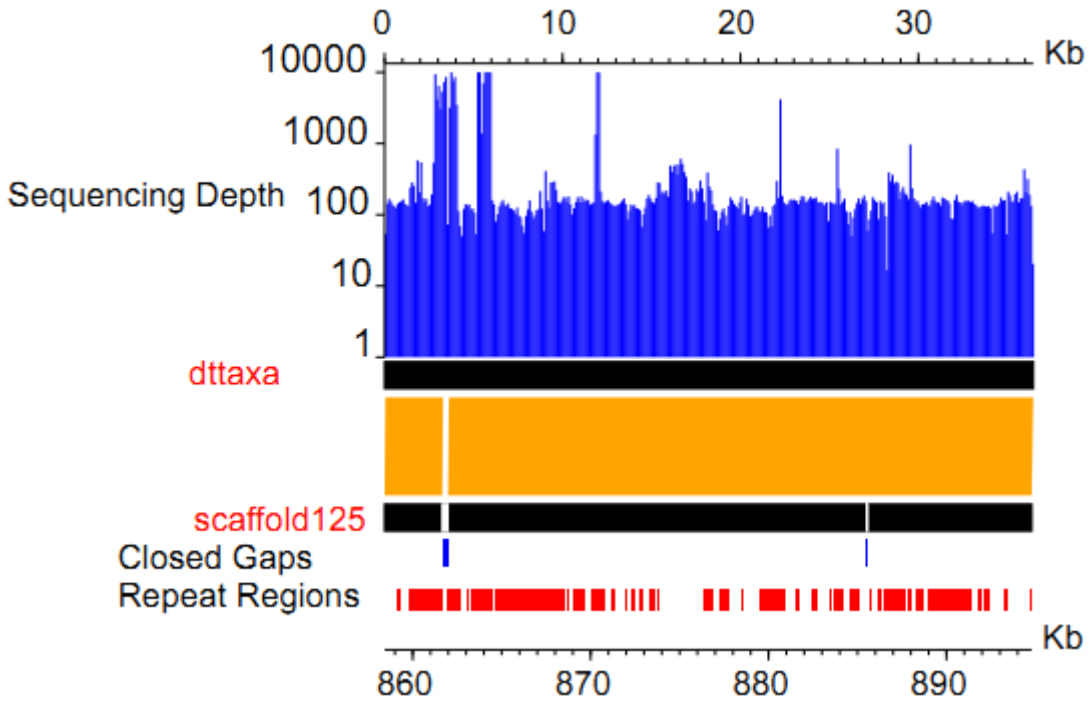
Supplementary Figure 1. **17-kmer analysis for estimation of genome size.** The dashed lines represent the coordinates of K_depth. We calculated the genome sizes of four sequenced mudskippers (BP 0.983 Gb, PM 0.780 Gb, SH 0.806 Gb and PS 0.739 Gb) on basis of the filtered reads from short-insert libraries.

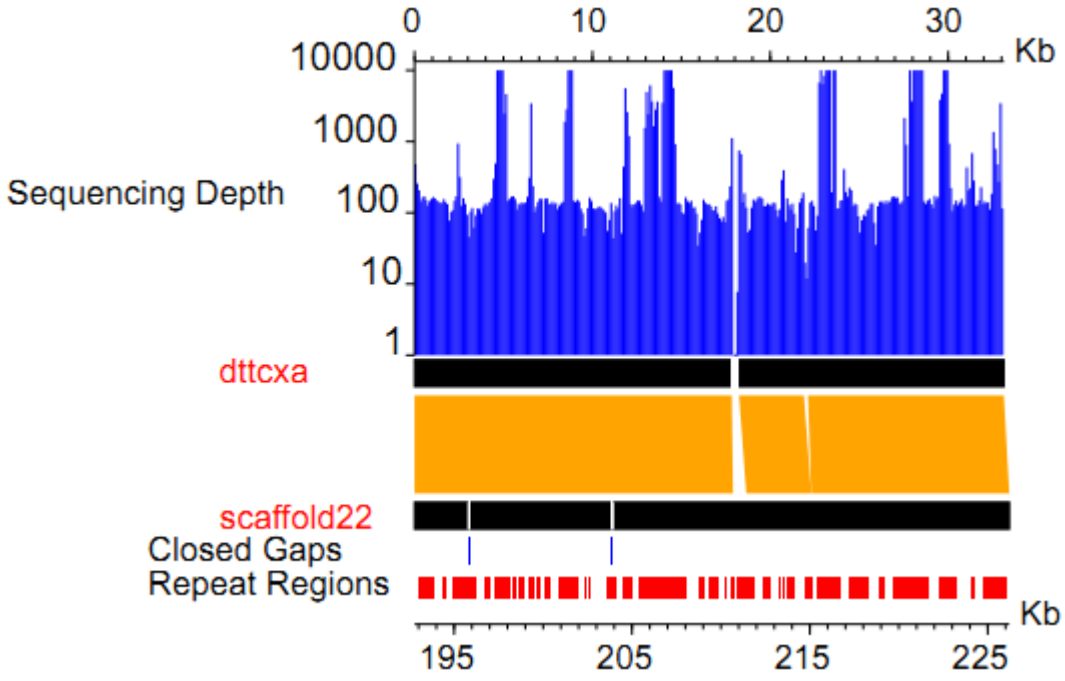


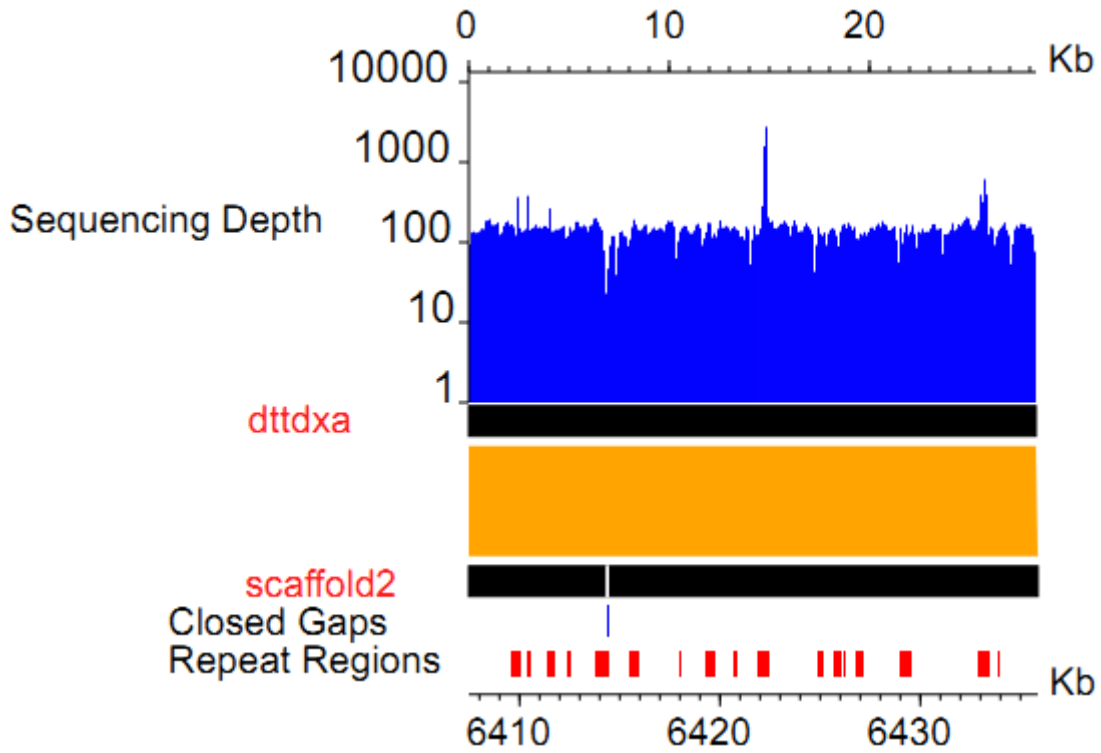
Supplementary Figure 2. **GC ratio and corresponding sequencing depth of four mudskippers.** The GC contents and average depths were calculated within the Non-overlapping 10-kb sliding window. Only one main GC cluster was identified for each species, implying high quality and clean reads were obtained from each library. The GC contents for the four examined mudskippers are around 40%.

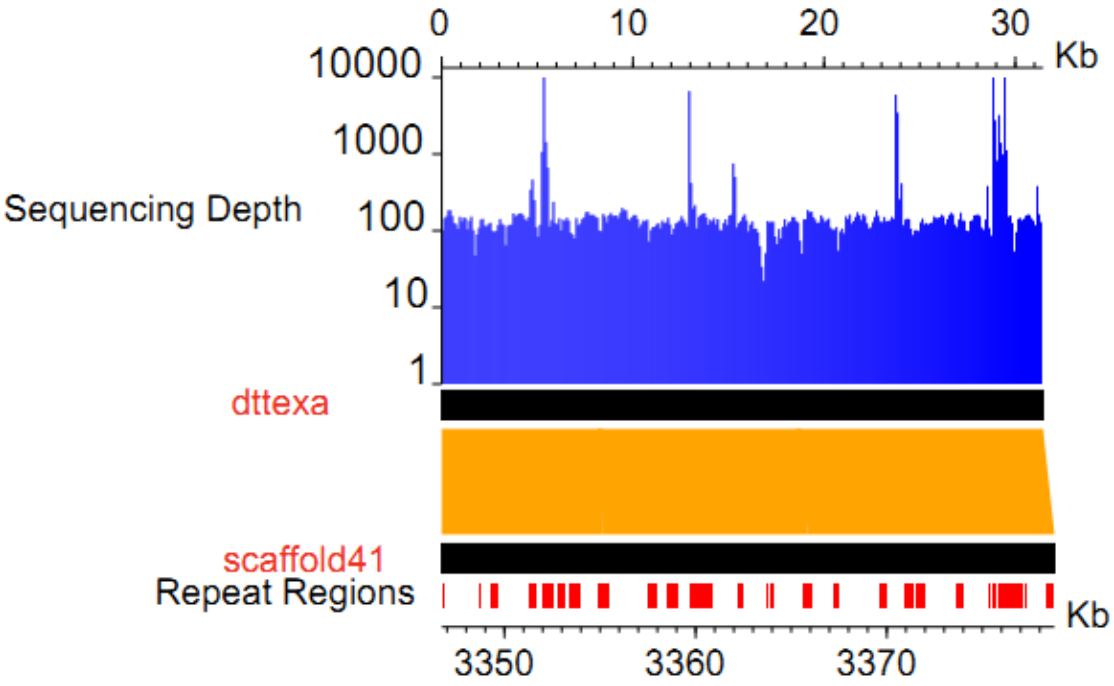


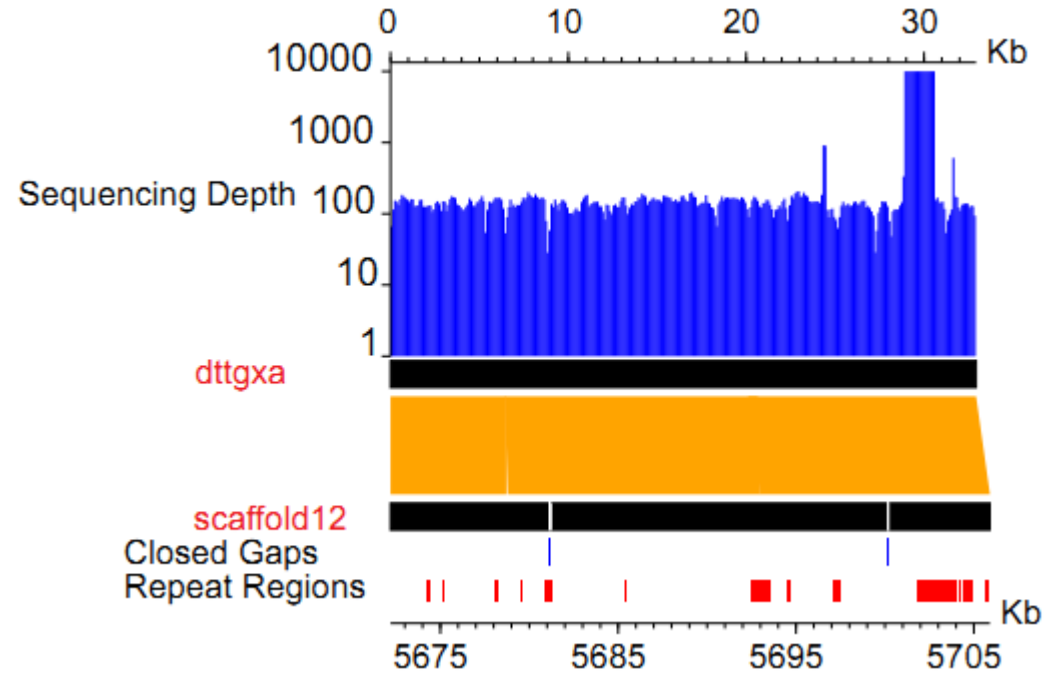
Supplementary Figure 3. GC content distributions in the nine sequenced teleost genomes. The X-axis represents GC content and the Y-axis is the proportion of sliding windows for a given GC content. Intriguingly, we sorted the GC ratio of each teleost from small to large and discovered that the order is the same as their evolutionary placement: (zebrafish, (((SH, BP), (PM, PS)), (medaka, (stickleback, (fugu, greenpuffer))))). The 9 sequenced fish include zebrafish, mudskipper, medaka, stickleback, and two pufferfishes (Ensembl release 64).

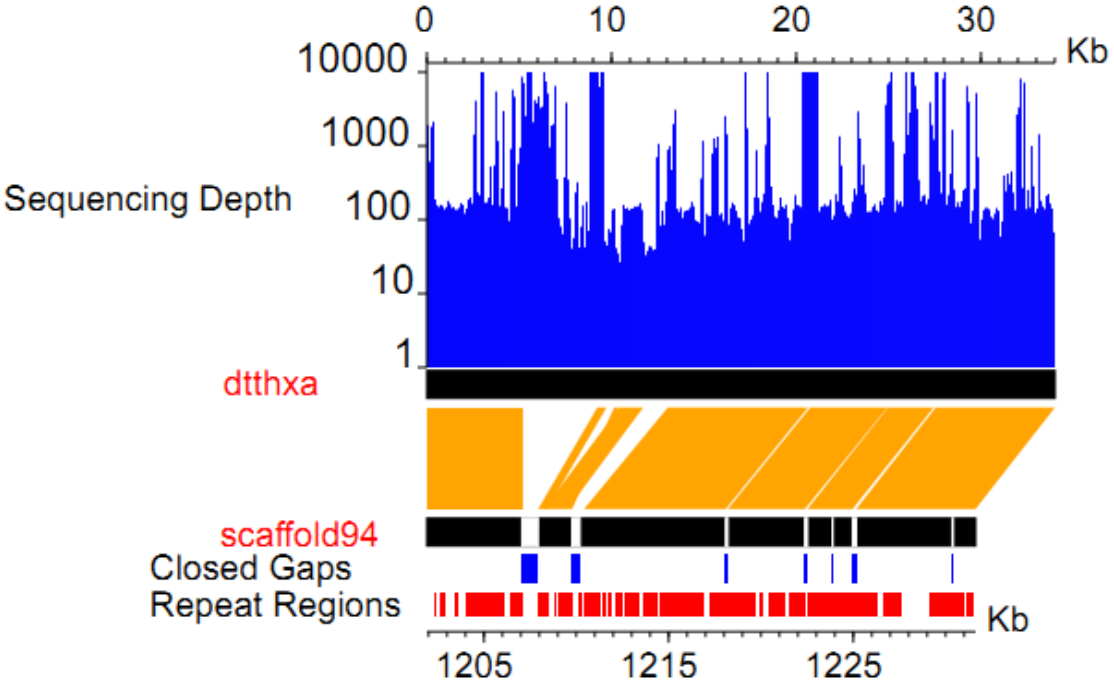

0
10
20
30
Kb
10000
1000
100
10
1
Sequencing Depth
dttaxa
scaffold125
Closed Gaps
Repeat Regions
Kb
860
870
880
890

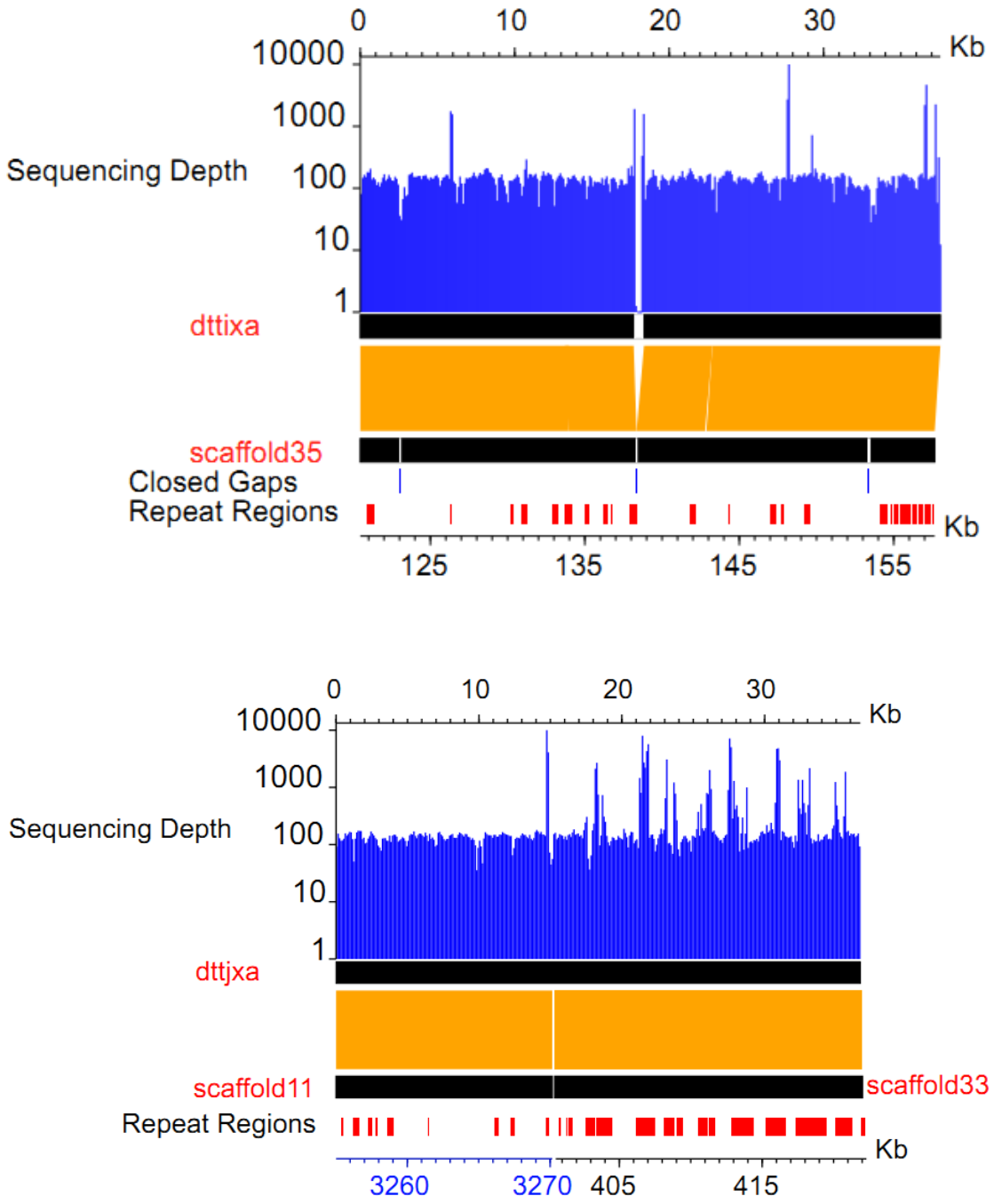

0
10
20
30
Kb
10000
1000
100
10
1
Sequencing Depth
dttcxa
scaffold22
Closed Gaps
Repeat Regions
Kb
195
205
215
225


0
10
20
Kb
10000
1000
100
10
1
Sequencing Depth
dttdxa
scaffold2
Closed Gaps
Repeat Regions
Kb
6410
6420
6430

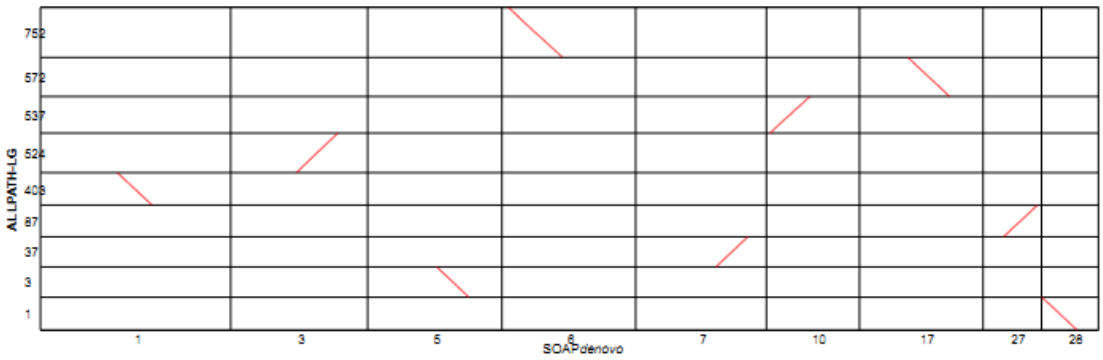

0
10
20
30
Kb
10000
1000
100
10
1
Sequencing Depth
dttexa
scaffold41
Repeat Regions
Kb
3350
3360
3370


0
10
20
30
Kb
10000
1000
100
10
1
Sequencing Depth
dttgxa
scaffold12
Closed Gaps
Repeat Regions
Kb
5675
5685
5695
5705

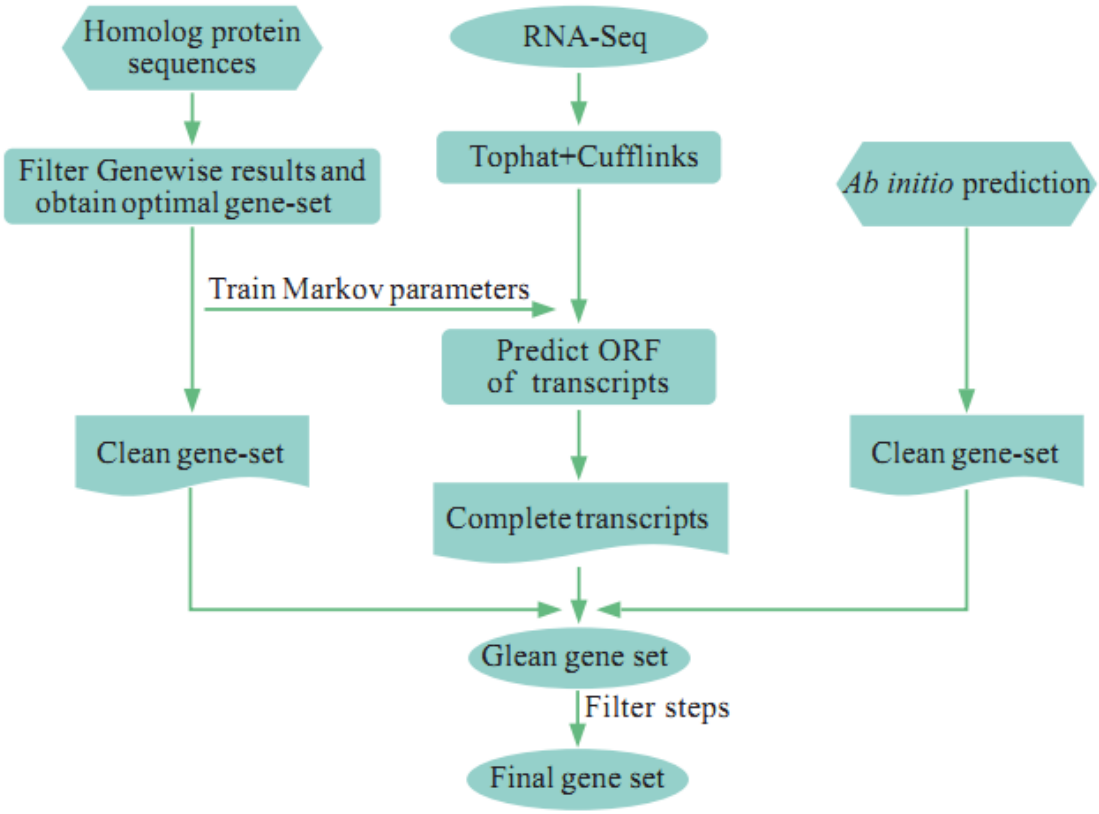

0
10
20
30
Kb
10000
1000
100
10
1
Sequencing Depth
dtthxa
scaffold94
Closed Gaps
Repeat Regions
Kb
1205
1215
1225



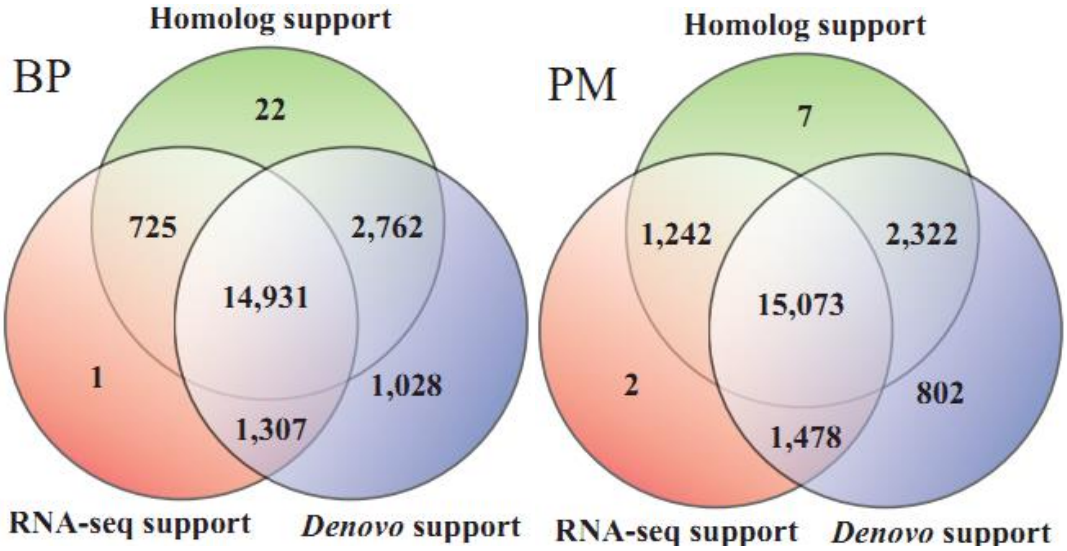
Supplementary Figure 4. Assembly assessment by Sanger-sequenced 40-kb fosmid clones. Eight sequences (dttaxa, dttcxa, dttdxa, dttexa, dttgxa, dtthxa, dttixa, dttjxa) were aligned to the whole genome of BP. The fosmid sequencing depth and annotated repeat regions are demonstrated in blue and red, respectively. Only dtthxa alignment had large gaps due to huge repeats, which may disturb the algorithm of SOAP*denovo* assembly in these regions.



Supplementary Figure 5. Comparison of the nine longest scaffolds derived from the ALLPATH-LG and SOAPdenovo assemblies. The X-axis stands for the scaffold ID of SOAP*denovo* and the Y-axis represents ALLPATH-LG scaffolds ID. Red lines indicate the alignments between both assemblies.

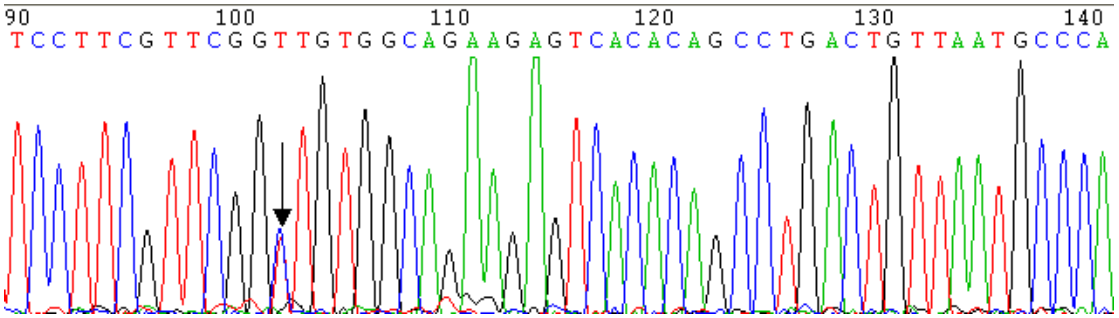


Supplementary Figure 6. Obtaining reference gene sets from this comprehensive annotation pipeline.

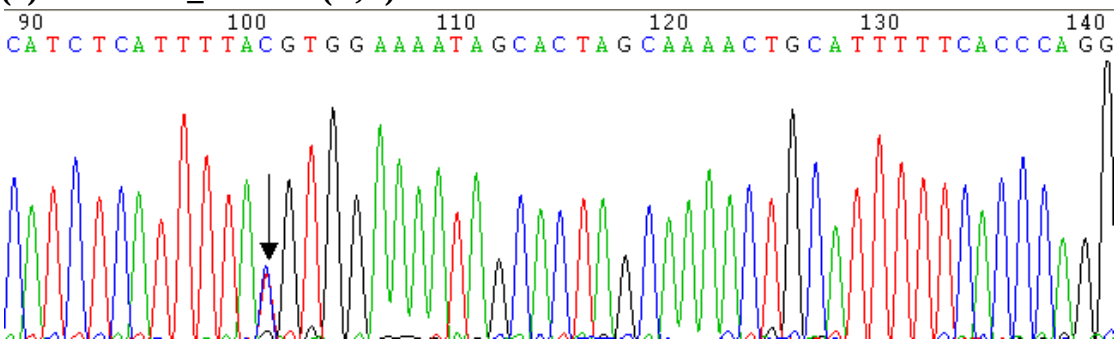


Supplementary Figure 7. Gene models of BP and PM are supported by three types of evidence. Over 70% of the genes are supported by all the three types of evidence (homolog-based methods, RNA-seq and *Denovo* prediction), confirming that the generated gene sets are highly reliable and elaborate.

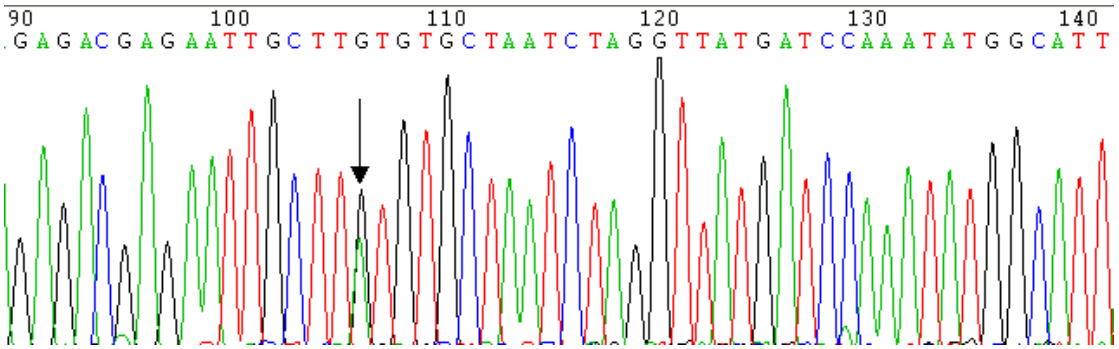
(1) scaffold31_2866146 (C,T)



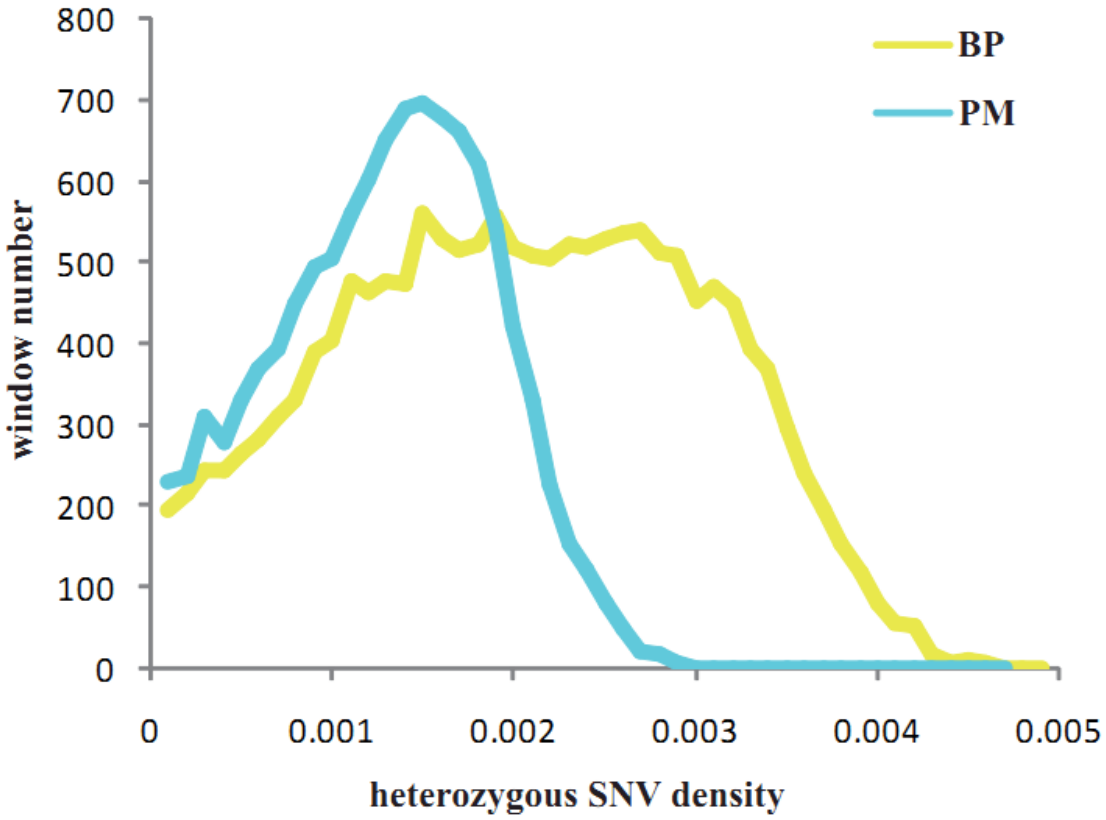
(2) scaffold32_1102764 (C,T)



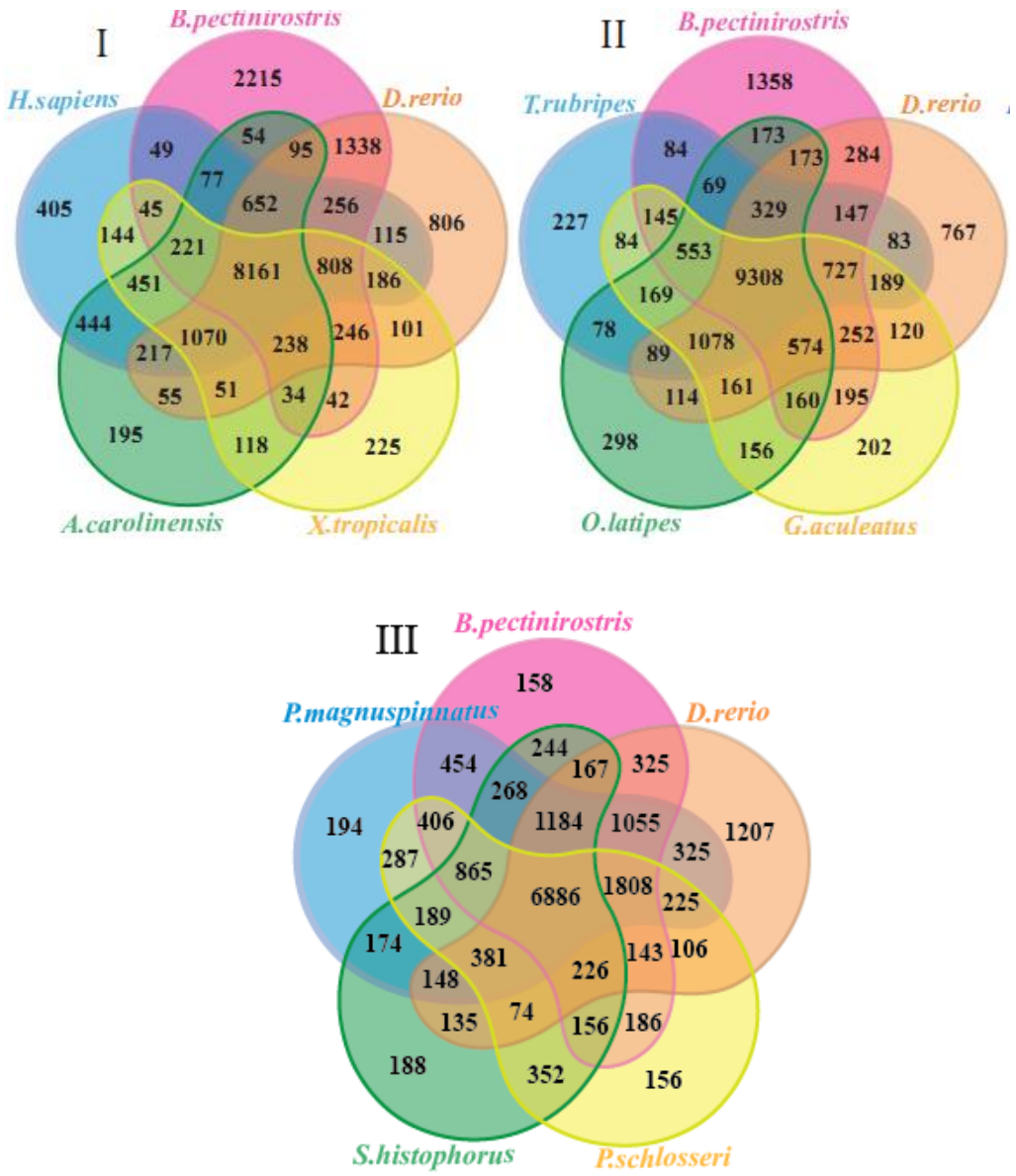
(3) scaffold34_3464355 (A,G)



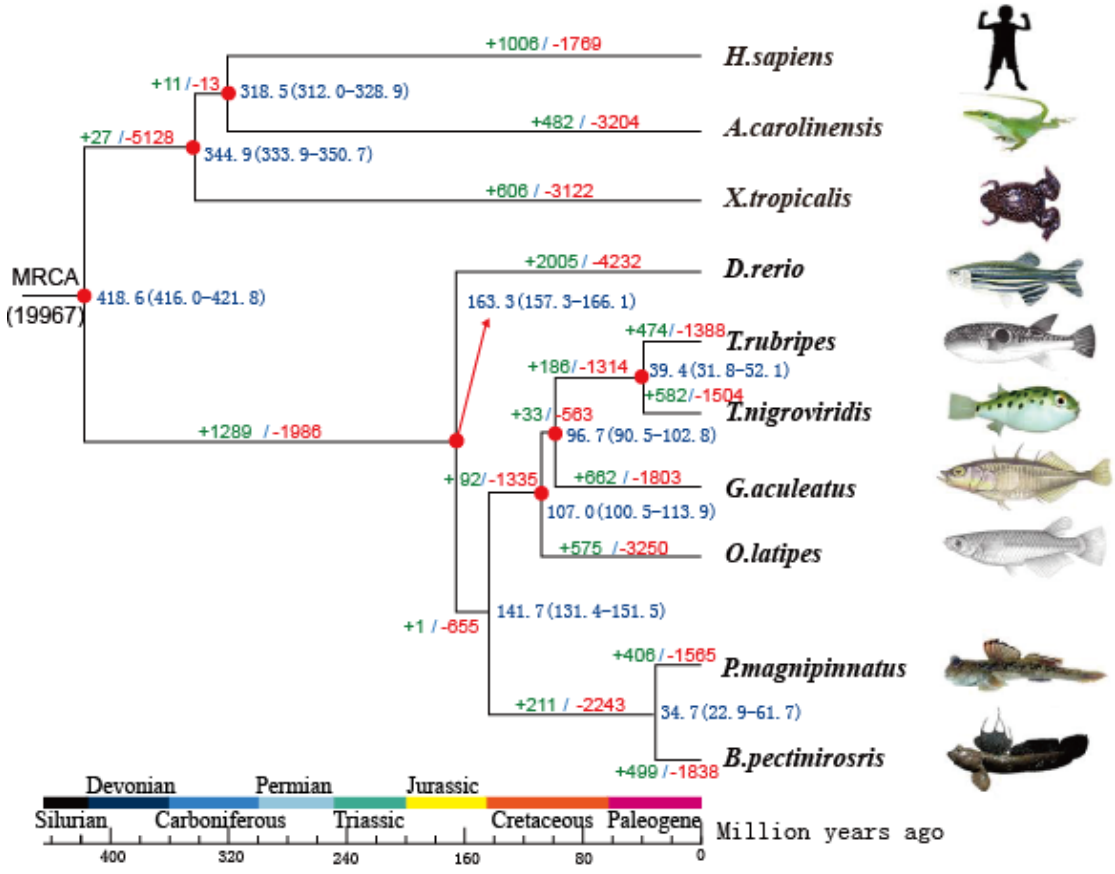
Supplementary Figure 8. Validation of heterozygous SNVs by Sanger sequencing. The black arrows indicate the real heterozygous SNV sites that contain double half-depth sequence peaks.



Supplementary Figure 9. Distribution of heterozygous SNV density across the BP and PM genomes. Much wider distribution of SNV density was identified in BP than in PM, which showed that the mutations in the BP genome were more evenly distributed. We selected Non-overlapping 50-kb windows and calculated heterozygosity density in every window. Yellow and blue lines represent the information in BP and PM, respectively.

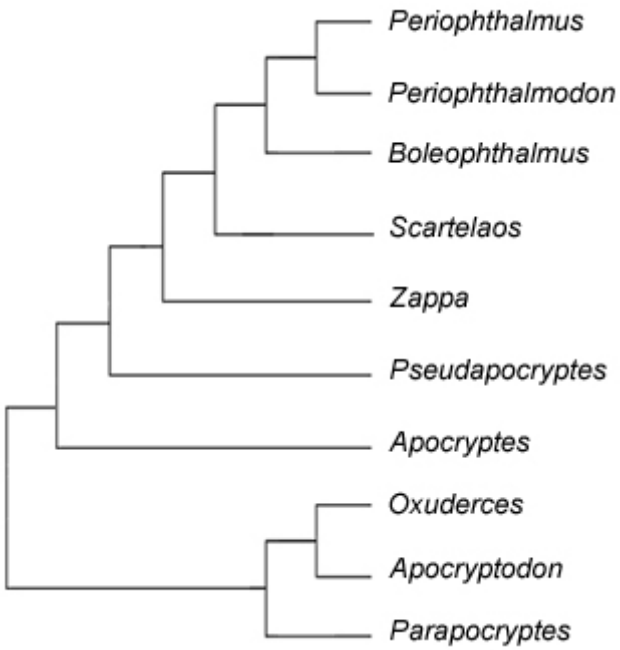


Supplementary Figure 10. Shared gene families of mudskippers with other vertebrates. I. BP and four representative vertebrates (*H. sapiens*, *D. rerio*, *X. tropicalis* and *A. carolinensis*), II. BP and four representative teleosts (*D. rerio*, *T. rubripes*, *G. aculeatus* and *O. latipes*), III. four mudskippers and *D. rerio*.

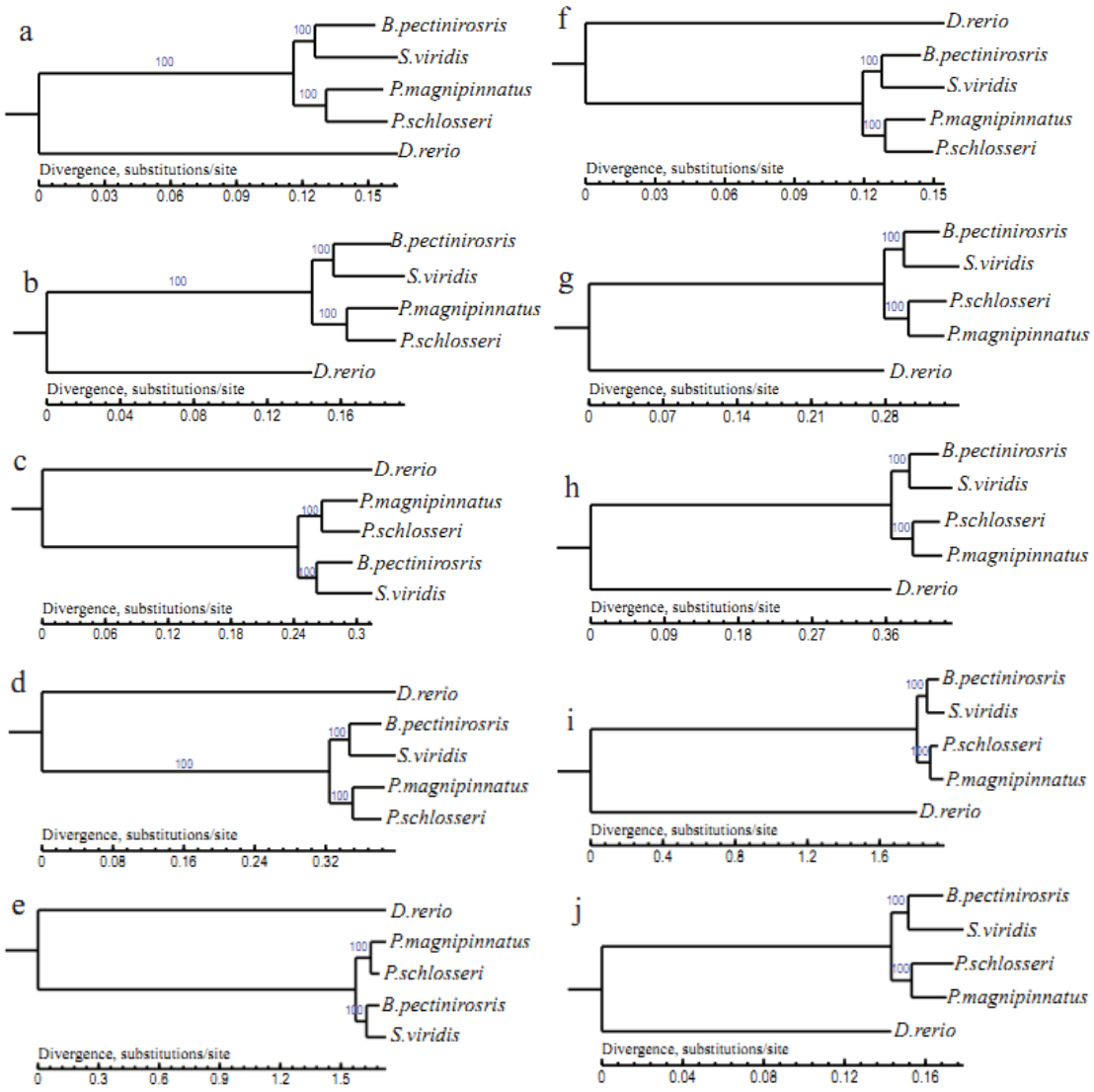


Supplementary Figure 11. Phylogenomic and gene family change analyses.

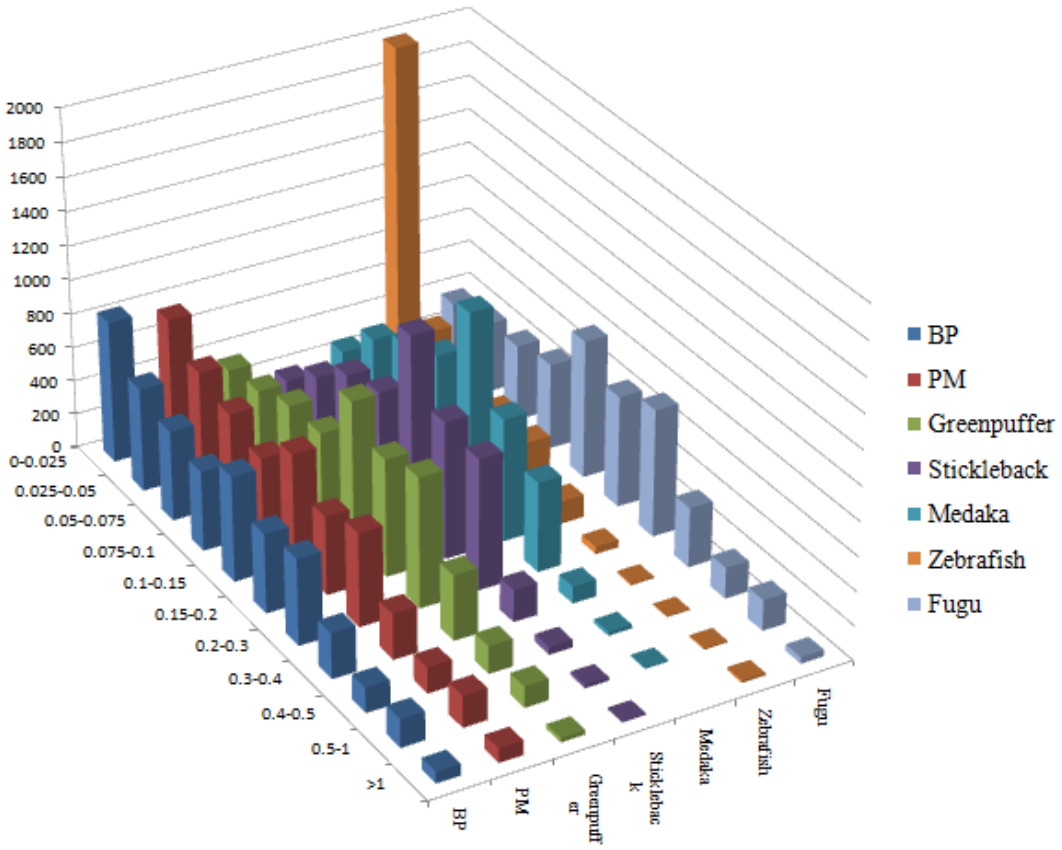
Mrbayes phylogenetictree was constructed by 3,445 single-copy genes from BP, PM, five teleosts (zebrafish, medaka, stickleback, fugu and greenpuffer) and three tetrapods (human, frog and lizard). The blue numbers are the estimated divergence time, the green numbers represent expanded gene families, and the red numbers stand for contracted gene families. MRCA represents most recent common ancestor.



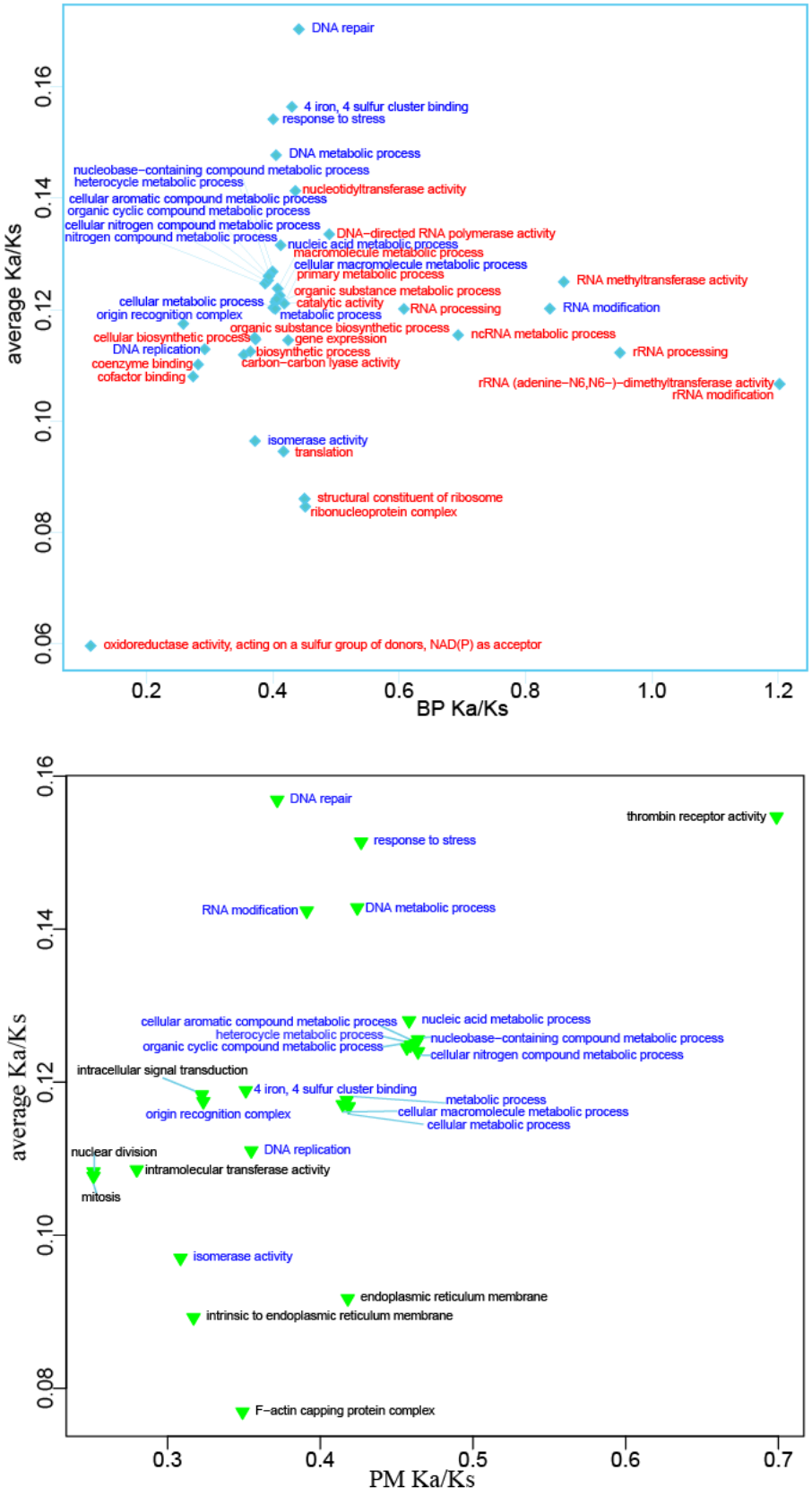
Supplementary Figure 12. Murdy's morphology-based cladistic revision of mudskippers. The image is adopted from Murdy (1989)[1].



Supplementary Figure 13. Phylogenetic trees generated using different methods support our new cladistic revision of the four mudskippers. A total of 4,306 single copy genes were used for these trees. Blue numbers at the nodes represent bootstrap values based on 1,000 replicates. a and b, NJ method; c-f, MrBayes, g-i, PhyML. a, e and d: protein sequences; b, d and h: coding sequences; e and i: 4D sites; f and i: first base of codons. Models: a, dm; b, mm; c, Jones+gamma; d-f, GTR+gamma; g, WAG+gamma; h-i, HKY85+gamma.

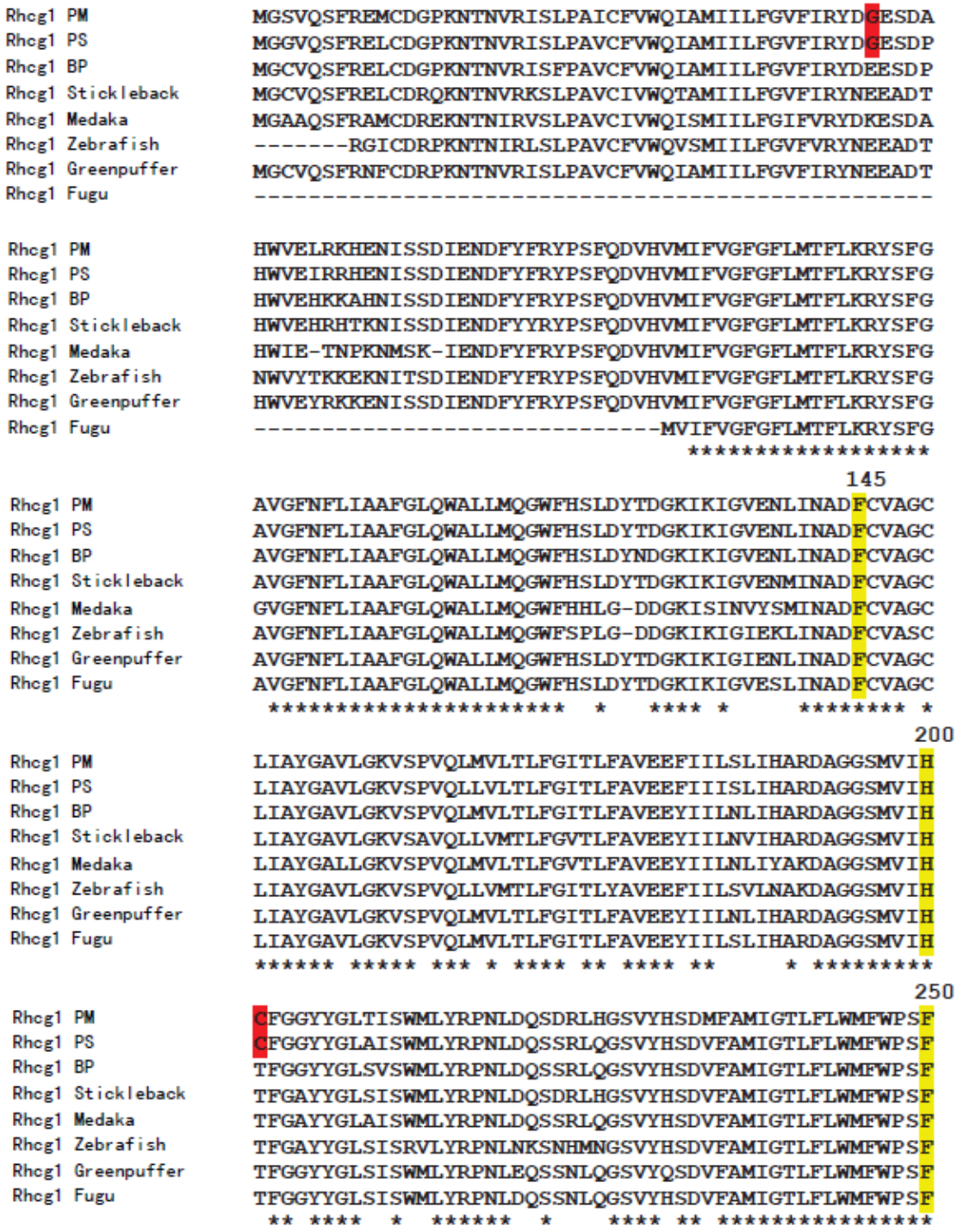


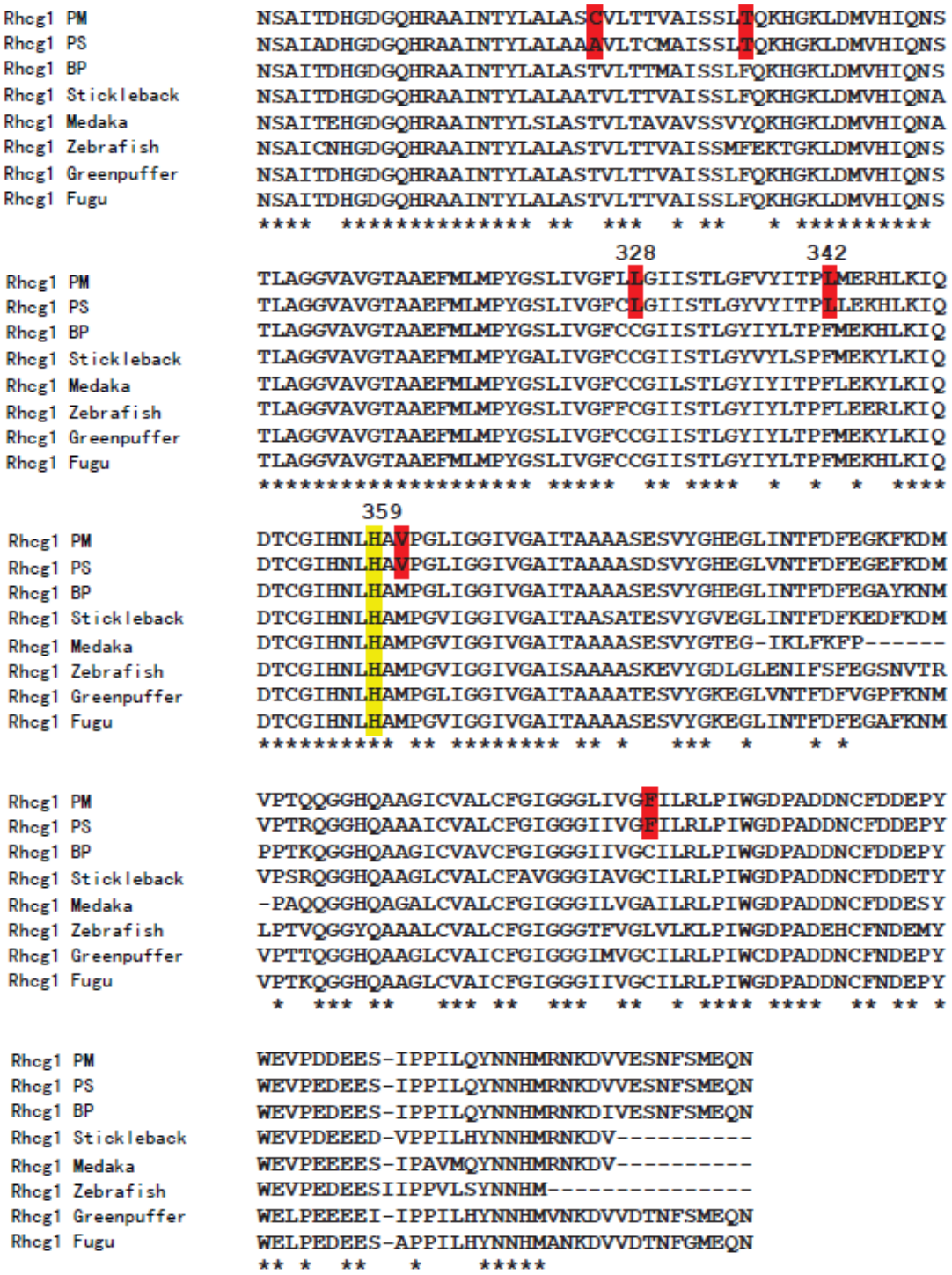
Supplementary Figure 14. Distribution of genes in the different dN/dS bins for all seven teleosts used in the dN/dS analysis. The average dN/dS values of BP, PM, zebrafish, medaka, stickleback, fugu and greenpuffer are 0.166, 0.181, 0.05, 0.12, 0.143, 0.184 and 0.176, respectively.



Supplementary Figure 15. Distribution of average of Ka/Ks ratios for enriched GO terms in BP and PM.

Common enriched GO terms of BP and PM were shown in blue.

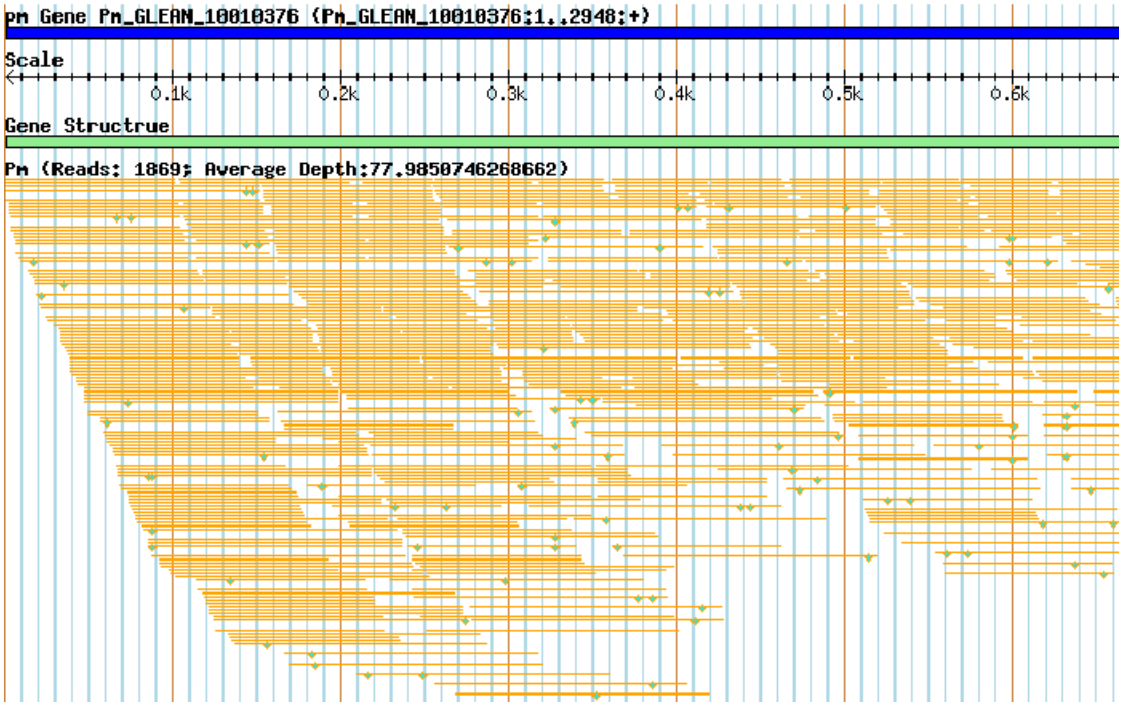

Rhcg1 PM MGSVQSFREMCDGPKNTNVRISLPAICFVWQIAMIILFGVFIRYDGESDA
Rhcg1 PS MGGVQSFRELCDGPKNTNVRISLPAVCFVWQIAMIILFGVFIRYDGESDP
Rhcg1 BP MGCVQSFRELCDGPKNTNVRISFPAVCFVWQIAMIILFGVFIRYDEESDP
Rhcg1 Stickleback MGCVQSFRELCDRQKNTNVRKSLPAVCIVWQTAMIILFGVFIRYNEEADT
Rhcg1 Medaka MGAAQSFRAMCDREKNTNIRVSLPAVCIVWQISMIILFGIFVRYDKESDA
Rhcg1 Zebrafish -------RGICDRPKNTNIRLSLPAVCFVWQVSMIILFGVFVRYNEEADT
Rhcg1 Greenpuffer MGCVQSFRNFCDRPKNTNVRISLPAVCFVWQIAMIILFGVFIRYNEEADT
Rhcg1 Fugu ----------
Rhcg1 PM HWVELRKHENISSDIENDFYFRYPSFQDVHVMIFVGFGFLMTFLKRYSFG
Rhcg1 PS HWVEIRRHENISSDIENDFYFRYPSFQDVHVMIFVGFGFLMTFLKRYSFG
Rhcg1 BP HWVEHKKAHNISSDIENDFYFRYPSFQDVHVMIFVGFGFLMTFLKRYSFG
Rhcg1 Stickleback HWVEHRHTKNISSDIENDFYYRYPSFQDVHVMIFVGFGFLMTFLKRYSFG
Rhcg1 Medaka HWIE-TNPKNMSK-IENDFYFRYPSFQDVHVMIFVGFGFLMTFLKRYSFG
Rhcg1 Zebrafish NWVYTKKEKNITSDIENDFYFRYPSFQDVHVMIFVGFGFLMTFLKRYSFG
Rhcg1 Greenpuffer HWVEYRKKENISSDIENDFYFRYPSFQDVHVMIFVGFGFLMTFLKRYSFG
Rhcg1 Fugu ----------MVIFVGFGFLMTFLKRYSFG
145
Rhcg1 PM AVGFNFLIAAFGLQWALLMQGWFHSLDYTDGKIKIGVENLINADFCVAGC
Rhcg1 PS AVGFNFLIAAFGLQWALLMQGWFHSLDYTDGKIKIGVENLINADFCVAGC
Rhcg1 BP AVGFNFLIAAFGLQWALLMQGWFHSLDYNDGKIKIGVENLINADFCVAGC
Rhcg1 Stickleback AVGFNFLIAAFGLQWALLMQGWFHSLDYTDGKIKIGVENMINADFCVAGC
Rhcg1 Medaka GVGFNFLIAAFGLQWALLMQGWFHHLG-DDGKISINVYSMINADFCVAGC
Rhcg1 Zebrafish AVGFNFLIAAFGLQWALLMQGWFSPLG-DDGKIKIGIEKLINADFCVASC
Rhcg1 Greenpuffer AVGFNFLIAAFGLQWALLMQGWFHSLDYTDGKIKIGIENLINADFCVAGC
Rhcg1 Fugu AVGFNFLIAAFGLQWALLMQGWFHSLDYTDGKIKIGVESLINADFCVAGC
200
Rhcg1 PM LIAYGAVLGKVSPVQLMVLTLFGITLFAVEEFIILSLIHARDAGGSMVIH
Rhcg1 PS LIAYGAVLGKVSPVQLLVLTLFGITLFAVEEFIIISLIHARDAGGSMVIH
Rhcg1 BP LIAYGAVLGKVSPVQLMVLTLFGITLFAVEEYIILNLIHARDAGGSMVIH
Rhcg1 Stickleback LIAYGAVLGKVSAVQLLVMTLFGVTLFAVEEYIILNVIHARDAGGSMVIH
Rhcg1 Medaka LIAYGALLGKVSPVQLMVLTLFGVTLFAVEEYIILNLIYAKDAGGSMVIH
Rhcg1 Zebrafish LIAYGAVLGKVSPVQLLVMTLFGITLYAVEEFIILSVLNAKDAGGSMVIH
Rhcg1 Greenpuffer LIAYGAVLGKVSPVQLMVLTLFGITLFAVEEYIILNLIHARDAGGSMVIH
Rhcg1 Fugu LIAYGAVLGKVSPVQLMVLTLFGITLFAVEEYIILSLIHARDAGGSMVIH
250
Rhcg1 PM CFGGYYGLTISWMLYRPNLDQSDRLHGSVYHSDMFAMIGTLFLWMFWPSF
Rhcg1 PS CFGGYYGLAISWMLYRPNLDQSSRLQGSVYHSDVFAMIGTLFLWMFWPSF
Rhcg1 BP TFGGYYGLSVSWMLYRPNLDQSSRLQGSVYHSDVFAMIGTLFLWMFWPSF
Rhcg1 Stickleback TFGAYYGLSISWMLYRPNLDQSDRLHGSVYHSDVFAMIGTLFLWMFWPSF
Rhcg1 Medaka TFGAYYGLAISWMLYRPNLDQSSRLQGSVYHSDVFAMIGTLFLWMFWPSF
Rhcg1 Zebrafish TFGAYYGLSISRVLYRPNLNKSNHMNGSVYHSDVFAMIGTLFLWMFWPSF
Rhcg1 Greenpuffer TFGGYYGLSISWMLYRPNLEQSSNLQGSVYQSDVFAMIGTLFLWMFWPSF
Rhcg1 Fugu TFGGYYGLSISWMLYRPNLDQSSNLQGSVYHSDVFAMIGTLFLWMFWPSF



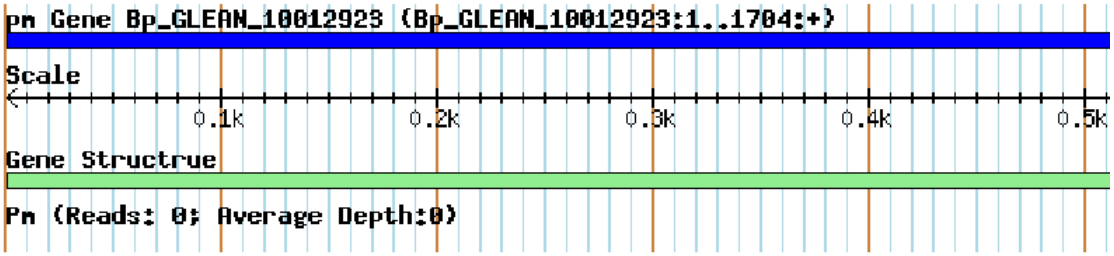
Supplementary Figure 16. Alignment of the protein sequences of Rhcg1 from PM, PS, BP, Stickleback, Medaka, Zebrafish, Greenpuffer, and Fugu.

Red rectangles indicate PM and PS-specific amino acid changes. Yellow rectangles indicate the conserved amino acids of Phe-Gate (F145, F250) and Twin-His (H200, H359). Asterisks represent conserved residues in all species.

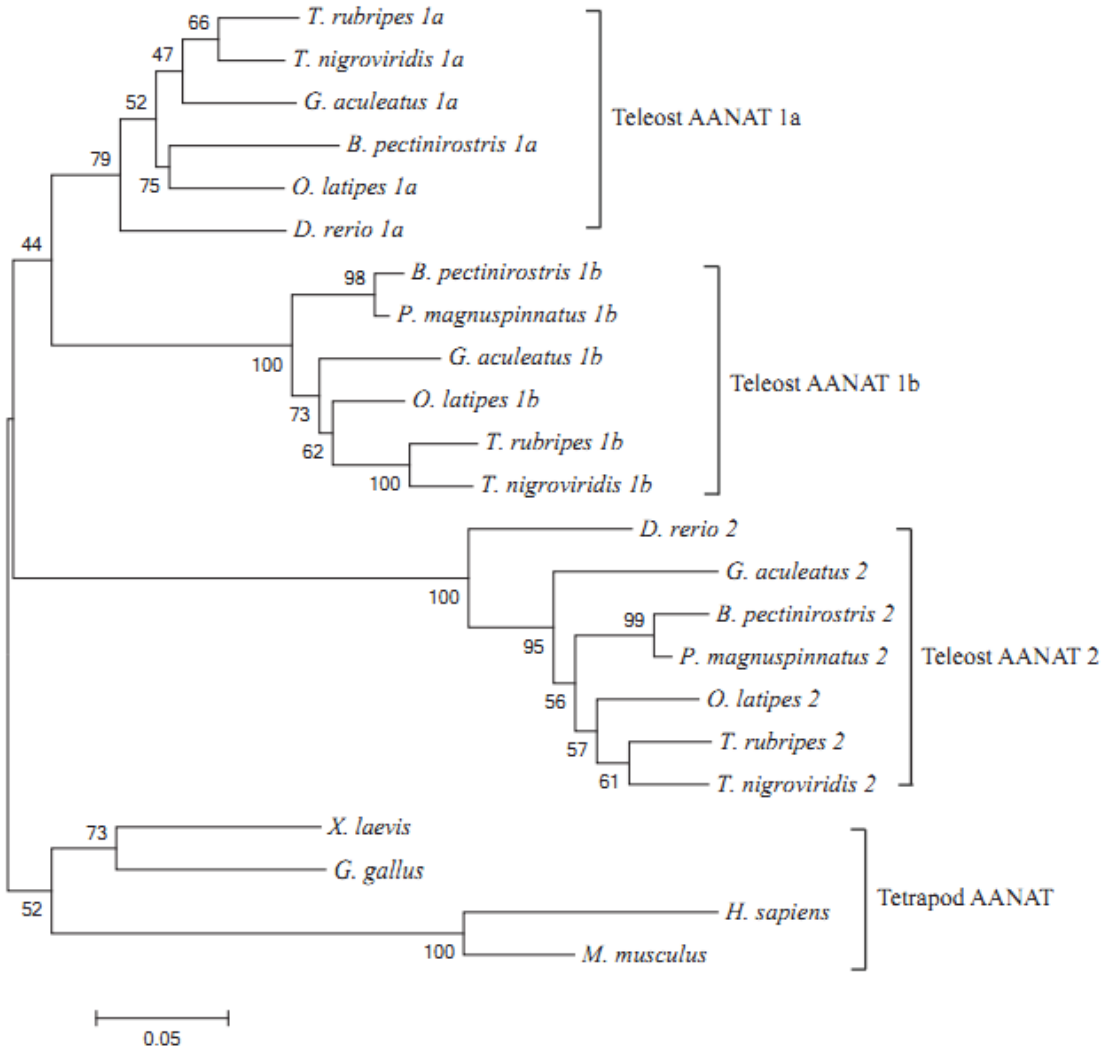
a.



b.

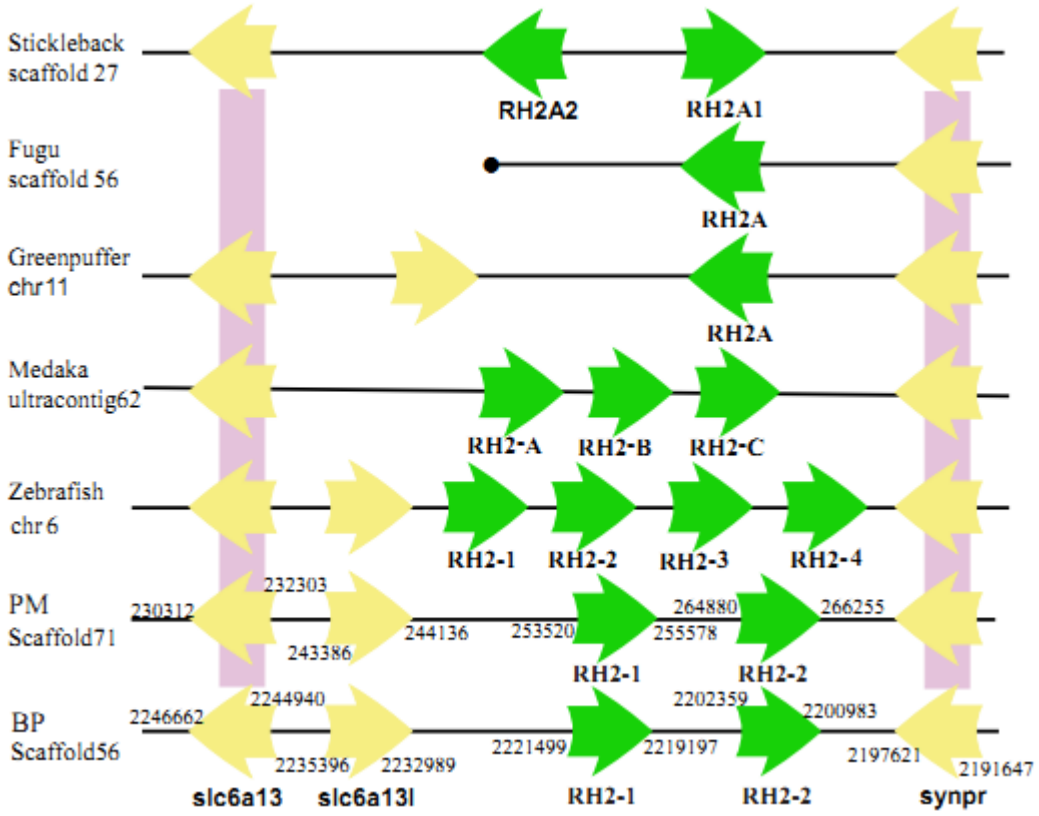


Supplementary Figure 17. An overview of PM reads mapped onto AANAT1a and 1b. a.AANAT1b (Pm_GLEAN_10010376) of PM showed massive reads aligned on it. **b.** No reads of PM was mapped on AANAT1a (Bp_GLEAN_10012923) of BP, confirming PM really lost its AANAT1a gene. Blue rectangles represented whole gene sequence and green rectangles represented exons of genes.

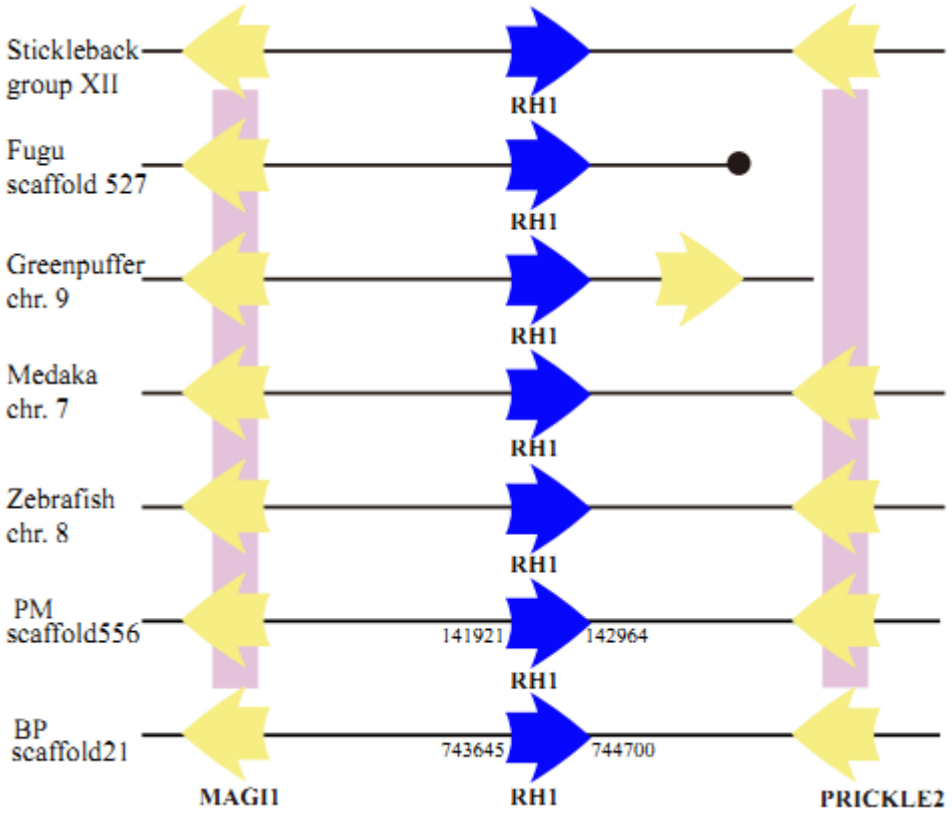


Supplementary Figure 18. Evolutionary relationships of vertebrate AANATs. Phylogenetic analyses were performed with NJ method in MEGA5[1]. The used sequences include *T. rubripes* AANAT2 (ENSTRUP00000029446), *T. nigroviridis* AANAT2 (ENSTNIP00000021755), *G. aculeatus* AANAT2 (ENSGACP00000018428), *B. pectinirostris* AANAT2 (Bp_GLEAN_10018012), *P. magnuspinnatus* AANAT2 (Pm_GLEAN_10016636), *O. latipes* AANAT2 (ENSORLP00000006487), *D. rerio* AANAT2 (ENSDARP00000002650), *B. pectinirostris* AANAT1b (Bp_GLEAN_10001218), *P. magnuspinnatus* AANAT1b (Pm_GLEAN_10010376), *O. latipes* AANAT1b (NP_001098330), *G. aculeatus* AANAT1b (ENSGACP00000009361), *T. rubripes* AANAT1b (ENSTRUP00000024942), *T.nigroviridis* AANAT1b (ENSTNIP00000006942), *X. laevis* AANAT (ENSXETP00000042273), *G. gallus* AANAT (AAB40942.1), *H. sapiens* AANAT (NP_001079.1), *M. musculus* AANAT (NP_033721.1), O. *latipes* AANAT1a (ENSORLP00000000892), *B. pectinirostris* AANAT1a (Bp_GLEAN_10012923), *G. aculeatus* AANAT1a (ENSGACP00000025480), *D. rerio* AANAT1a (ENSDARP00000053125), *T. rubripes* AANAT1a (ENSTRUP00000044871) and *T. nigroviridis* AANAT1a (ENSTNIP00000020660).

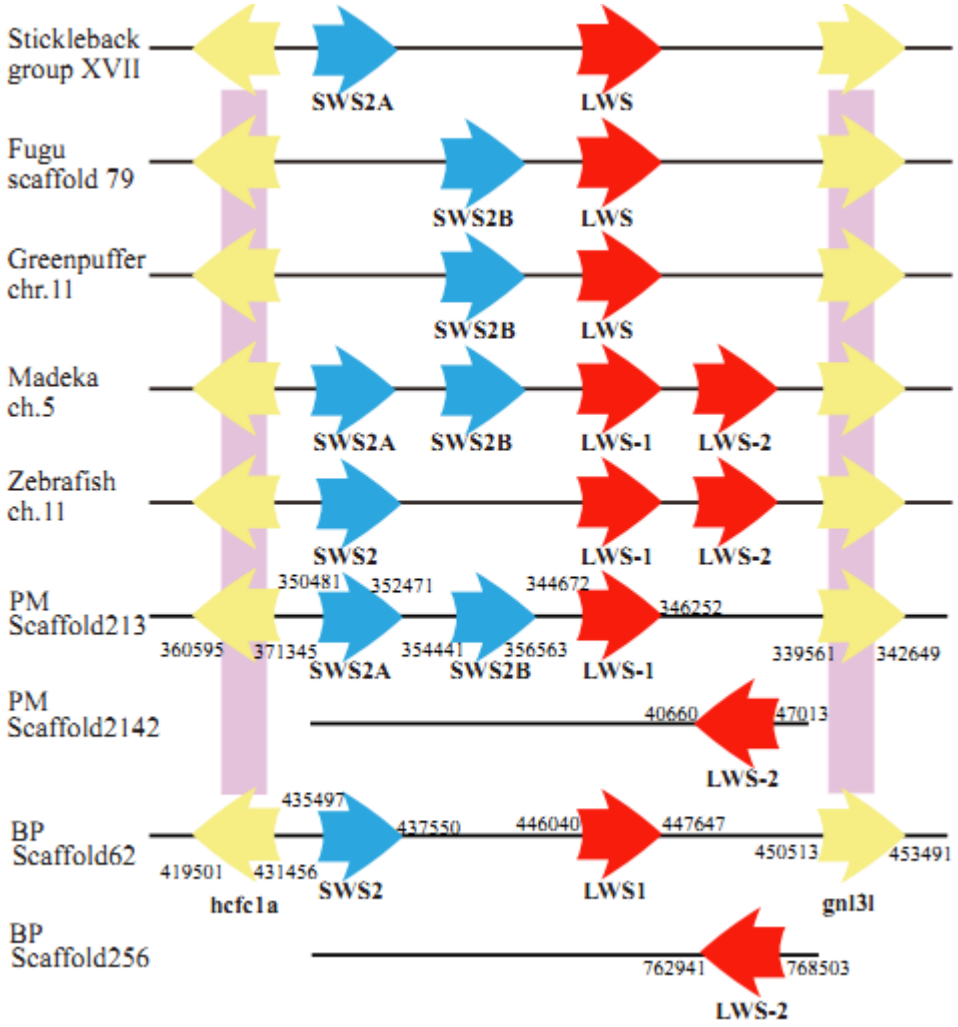
a.


Stickleback scaffold 27
RH2A2
RH2A1
Fugu scaffold 56
RH2A
Greenpuffer chr11
RH2A
Medaka ultracontig62
RH2-A
RH2-B
RH2-C
Zebrafish chr 6
RH2-1
RH2-2
RH2-3
RH2-4
PM Scaffold71
230312
232303
243386
244136
253520
264880
255578
266255
RH2-1
RH2-2
BP Scaffold56
2246662
2244940
2235396
2232989
2221499
2202359
2219197
2200983
2197621
2191647
slc6a13
slc6a13l
RH2-1
RH2-2
synpr

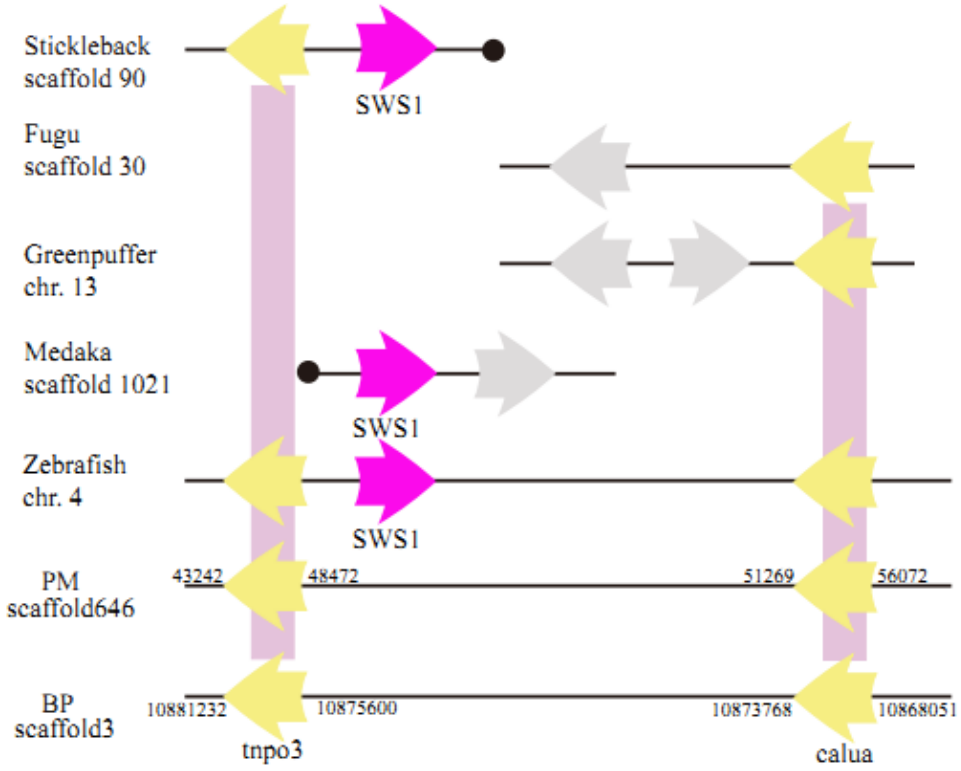
b.


Stickleback group XII
RH1
Fugu scaffold 527
RH1
Greenpuffer chr. 9
RH1
Medaka chr. 7
RH1
Zebrafish chr. 8
RH1
PM scaffold556
141921
142964
RH1
BP scaffold21
743645
744700
MAGI1
RH1
PRICKLE2

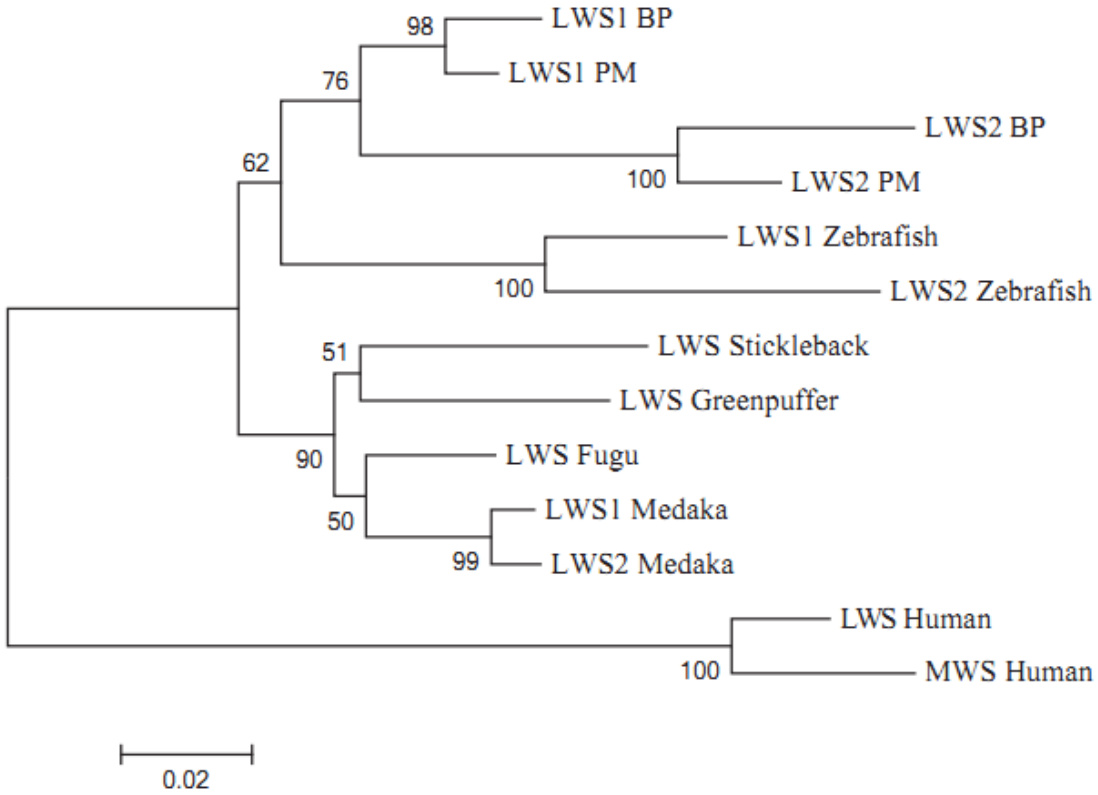
c.


Stickleback group XVII
SWS2A
LWS
Fugu scaffold 79
SWS2B
LWS
Greenpuffer chr.11
SWS2B
LWS
Madeka ch.5
SWS2A
SWS2B
LWS-1
LWS-2
Zebrafish ch.11
SWS2
LWS-1
LWS-2
PM Scaffold213
350481
352471
344672
346252
360595
371345
354441
356563
339561
342649
SWS2A
SWS2B
LWS-1
PM Scaffold2142
40660
47013
LWS-2
BP Scaffold62
435497
437550
446040
447647
450513
453491
419501
431456
SWS2
LWS1
hcfc1a
gn13l
BP Scaffold256
762941
768503
LWS-2

d.



Supplementary Figure 19. Distribution and synteny of four types of opsin genes. Colors represent different opsin genes or gene families: green for RH2, dark blue for RH1, light blue for SWS2, red for LWS, pink for SWS1, yellow for conserved genes, grey for other genes and purple bars for orthology. Numbers near genes show start and end coordinates of genes. In figure **d**, tnpo3 and calua conserved genes were identified in BP and PM genomes but SWS1 opsin gene was absent between them.



Supplementary Figure 20. Phylogenetic analysis of LWS/MWS opsin genes in fish and human. Phylogenetic tree was constructed with NJ method in MEGA4 software and the bootstrap values (1000 replicates) were shown at the nodes. Phylogenetic analysis of LWS from teleosts and human (as outgroup species) demonstrated that they were duplicated independently in zebrafish, medaka and ancestor of mudskippers lineages, separately. Protein sequences of LWS/MWS were downloaded from NCBI (The accession numbers of LWS/MWS: LWS Human: NP_064445.1, MWS Human: NP_000504.1, LWS1 Zebrafish: NP_571250.1, LWS2 Zebrafish: NP_001002443.1, LWS1 Medaka: BAE78645.1, LWS2 Medaka: BAE78646.1, LWS Stickleback: BT027981.1, LWS Greenpuffer: AAT38457.1, LWS Fugu: AAT38456.1).

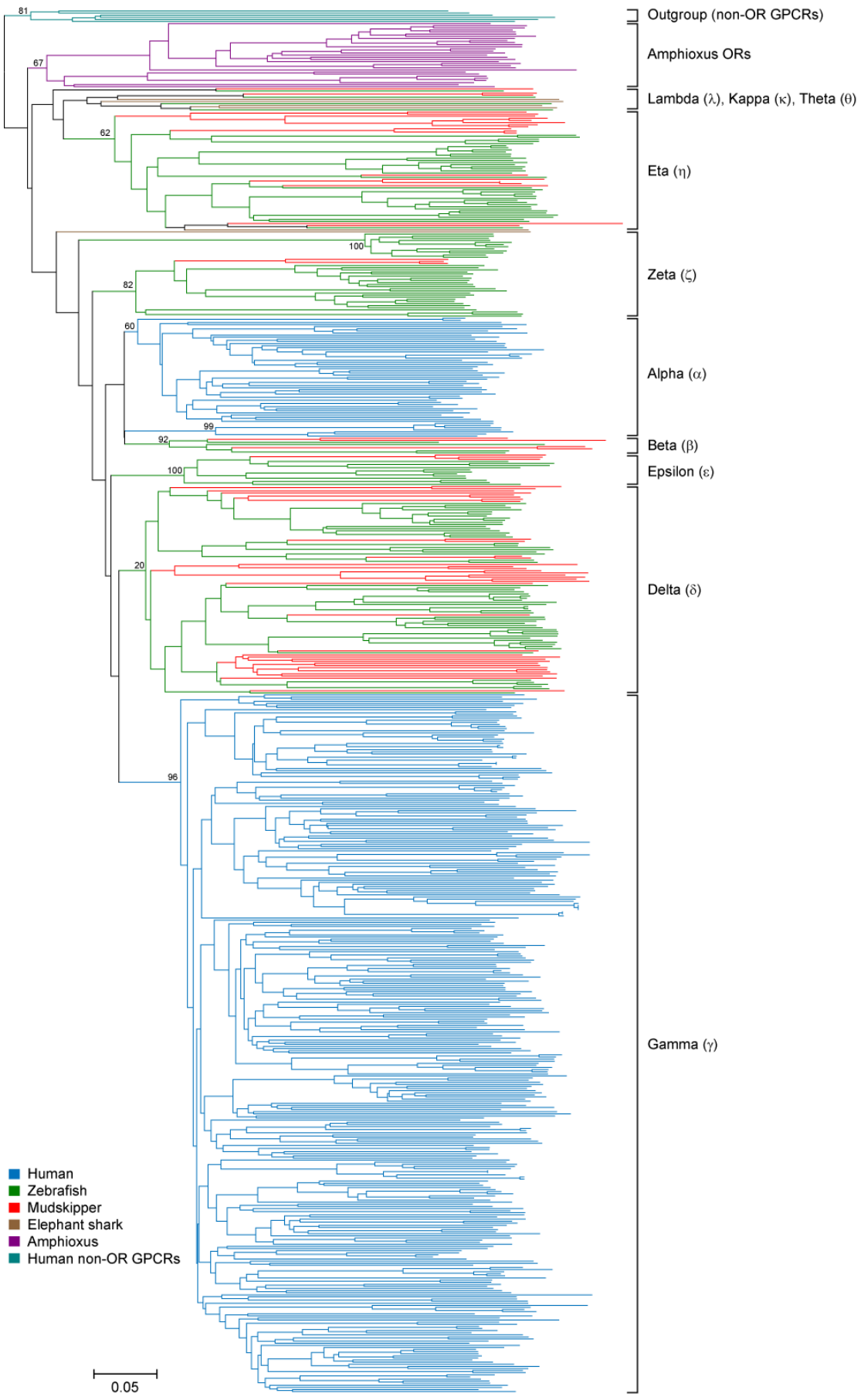
```
LWS Human           MAQQWSLQRLAGRHPQDSYEDSTQSSIFTYTNSNSTRGPFEGPNYHIAPRWVYHLTSVWM
MWS Human           MAQQWSLQRLAGRHPQDSYEDSTQSSIFTYTNSNSTRGPFEGPNYHIAPRWVYHLTSVWM
LWS1 Zebrafish      MAEHWGDAIYAARR---KGDETTREAMFTYTNSNNTKDPFEGPNYHIAPRWVYNVATVWM
LWS2 Zebrafish      MAE-WANGAFAARR---RGDETTRDNAFSYTNSNNTRDPFEGPNYHIAPRWVYNVATVWM
LWS2 BP             MAEEWGKQVYGARR---YHDETTRGSIFTYTNSNNTRDPFEGPNYHIAPRWVYNLSTAWM
LWS2 PM             MAEDWGKQVYAARR---YHEDTTRGAAFTYTNSNNTRDPFEGPNYHIAPRWVYNLSTAWM
LWS Stickleback     MAEEWGKQAFAARR---YNEDTTRGSMFVYTNSNNTKDPFEGPNYHIAPRWVYNLSTLWM
LWS1 BP             MAEEWGKQVFAARR---YHEDTTRGSAFTYTNSNNTRDPFEGPNYHIAPRWVYNLSTVWM
LWS1 PM             MAEEWGKQVFAARR---YHEDTTRGSAFTYTNSNNTRDPFEGPNYHIAPRWVYNVSTVWM
LWS Greenpuffer     MAEEWGKQSFAARR---YHEDSTRGSAFVYTNSNHTRDPFEGPNYHIAPRWVYNLATLWM
LWS Fugu            MAEEWGKQSFAARR---YHEDTTRGSAFVYTNSNHTRDPFEGPNYHIAPRWVYNVATVWM
LWS1 Medaka         MAEEWGKQVFAARR---HNEDTTRGSAFTYTNSNHTRDPFEGPNYHIAPRWVYNVATVWM
LWS2 Medaka         MAEQWGKQVFAARR---QNEDTTRGSAFTYTNSNHTRDPFEGPNYHIAPRWVYNLATLWM
                    **  *       *         *    * ***** *  ***************     **

LWS Human           IFVVTASVFTNGLVLAATMKFKKLRHPLNWILVNLAVADLAETVIASTISIVNQVSGYFV
MWS Human           IFVVIASVFTNGLVLAATMKFKKLRHPLNWILVNLAVADLAETVIASTISVVNQVYGYFV
LWS1 Zebrafish      FFVVVASTFTNGLVLVATAKFKKLRHPLNWILVNLAIADLGETLFASTISVINQFFGYFI
LWS2 Zebrafish      FFVVVASTFTNGLVLVATAKFKKLRHPLNWILVNLAIADLGETLFSSTISVINQVFGYFI
LWS2 BP             LFVVVASVFTNGLVLVATAKFKKLRHPLNWILVNLAIADLGETIFASTISVCNQFFGYFI
LWS2 PM             LFVVVASVFTNGLVLVATAKFKKLRHPLNWILVNLAIADLGETVFASTISVCNQFFGYFI
LWS Stickleback     FIVVALSVFTNGLVLVATAKFKKLQHPLNWILVNLAIADLGETVFASTISVCNQFFGYFI
LWS1 BP             LFVVVASVFTNGLVLVATAKFKKLRHPLNWILVNLAIADLGETVFASTISVCNQFFGYFI
LWS1 PM             LFVVVASVFTNGLVLVATAKFKKLRHPLNWILVNLAIADLGETVFASTISVCNQFFGYFI
LWS Greenpuffer     FFVVVLSVFTNGLVLVATAKFKKLRHPLNWILVNLAVADLGETLFASTISVCNQFFGYFI
LWS Fugu            FIVVVLSVFTNGLVLVATAKFKKLRHPLNWILVNLAIADLGETVFASTISVCNQFFGYFI
LWS1 Medaka         FFVVVLSVFTNGLVLVATAKFKKLRHPLNWILVNLAIADLGETVFASTISVCNQFFGYFI
LWS2 Medaka         FFVVVLSVFTNGLVLVATAKFKKLRHPLNWILVNLAIADLGETVFASTISVCNQFFGYFI
                      **  * ******* ** ***** ********** *** **   ****  **  ***
                                                                              180
LWS Human           LGHPMCVLEGYTVSLCGITGLWSLAIISWERWLVVCKPFGNVRFDAKLAIVGIAFSWIWS
MWS Human           LGHPMCVLEGYTVSLCGITGLWSLAIISWERWMVVCKPFGNVRFDAKLAIVGIAFSWIWA
LWS1 Zebrafish      LGHPMCIFEGYTVSVCGIAALWSLTVISWERWVVVCKPFGNVKFDAKWASAGIIFSWVWA
LWS2 Zebrafish      LGHPMCIFEGYTVSVCGIAGLWSLTVISWERWVVVCKPFGNVKFDGKWASAGIIFSWVWA
LWS2 BP             LGHPMCIFEGYTVSVCGITALWSLAIISWERWVVVCKPFGNVKFDAKWASGGIIFSWSWA
LWS2 PM             LGHPMCIFEGYTVSVCGITALWSLAIISWERWVVVCKPFGNVKFDAKWASGGIIFSWAWA
LWS Stickleback     LGHPMCVFEGYVVSVCGITALWSLTIISWERWIVVCKPFGNVKFDAKWATAGIVFSWIWS
LWS1 BP             LGHPMCVFEGYTVSTCGIAALWSLTIISWERWIVVCKPFGNVKFDAKWASGGIIFSWVWA
LWS1 PM             LGHPMCIFEGYTVSTCGIAALWSLTIISWERWIVVCKPFGNVKFDAKWASAGIIFSWVWS
LWS Greenpuffer     LGHPMCIFEGYVVSVCGIAGLWSLTIISWERWIVVCKPFGNVKFDAKWATAGIVFSWIWP
LWS Fugu            LGHPMCVFEGYTVSTCGIAALWSLTIISWERWVVVCKPFGNVKFDAKWATGGIVFSWVWA
LWS1 Medaka         LGHPMCVFEGYVVSTCGIAALWSLTIISWERWVVVCKPFGNVKFDAKWAIGGIVFSWVWS
LWS2 Medaka         LGHPMCVFEGYVVSTCGIAALWSLTIISWERWVVVCKPFGNVKFDAKWAIGGIVFSWVWS
                    ******  *** ** ***  ****  ****** ********* ** * *  ** *** *
```

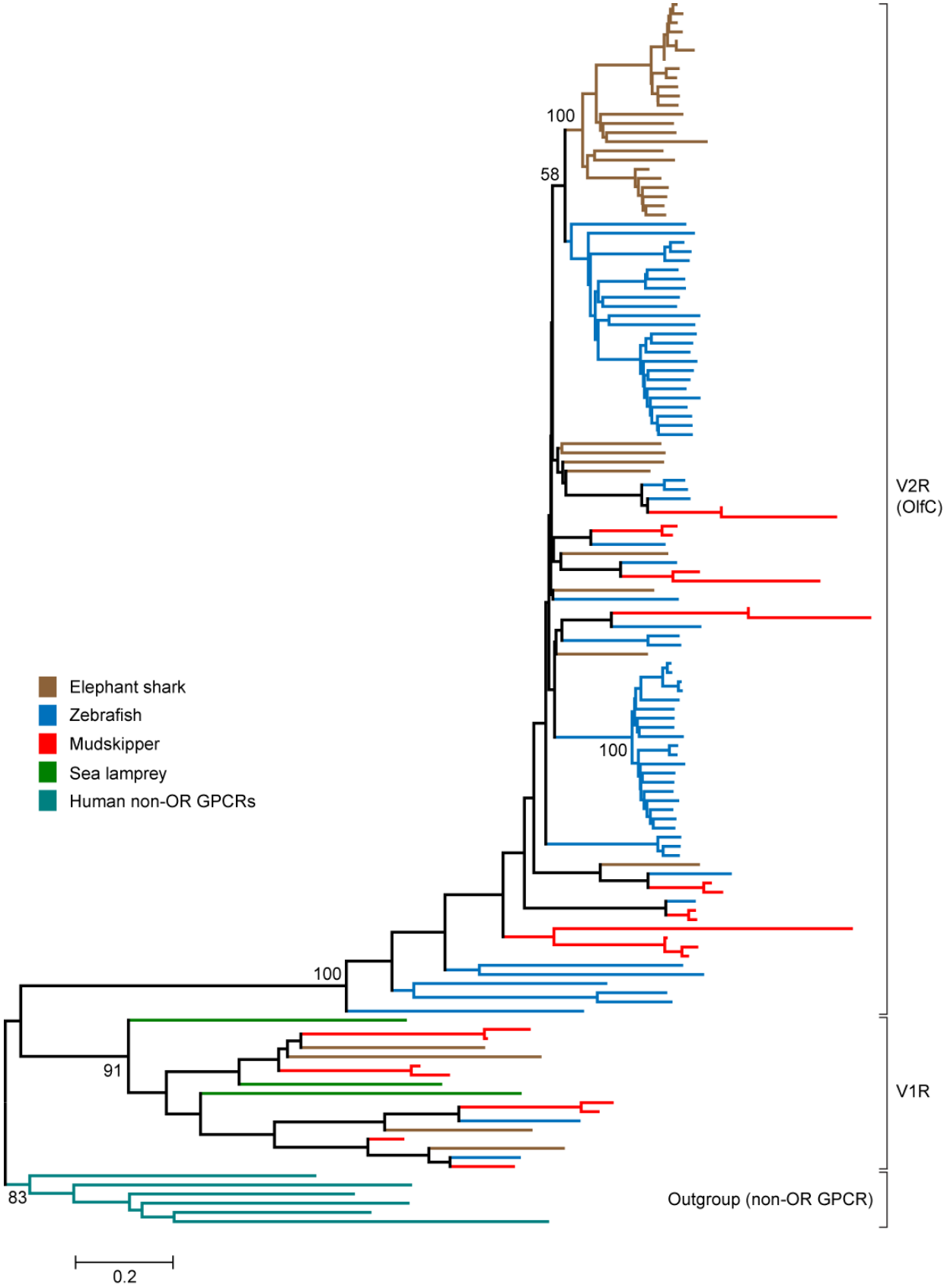
```
                                  197
LWS Human              AVWTAPPIFGWSRYWPHGLKTSCGPDVFSGSSYPGVQSYMIVLMVTCCIIPLAIIMLCYL
MWS Human              AVWTAPPIFGWSRYWPHGLKTSCGPDVFSGSSYPGVQSYMIVLMVTCCITPLSIIVLCYL
LWS1 Zebrafish         AAWCAPPIFGWSRYWPHGLKTSCGPDVFSGSEDPGVQSYMVVLMITCCIIPLAIIILCYI
LWS2 Zebrafish         AVWCAPPIFGWSRYWPHGLKTSCGPDVFGGNEDPGVQSYMLVLMITCCILPLAIIILCYI
LWS2 BP                FFWCAPPIFGWSRYWPHGLKTSCGPDVFSGSEDPGVWSYMVTLMITCCFLPLSIIILCYV
LWS2 PM                SFWCAPPIFGWSRYWPHGLKTSCGPDVFSGSEDPGVWSYMITLMVTCCFLPLSIIILCYI
LWS Stickleback        AVWCAPPIFGWSRYWPHGLKTSCGPDVFSGSEDPGVQSYMIVLMITCCLIPLAIIILCYL
LWS1 BP                AAWCSPPIFGWSRYWPHGLKTSCGPDVFSGSEDPGVQSYMIVLMLTCCILPLTVIILCYL
LWS1 PM                AIWCAPPIFGWSRYWPHGLKTSCGPDVFSGSEDPGVQSYMIVLMLTCCILPLAIIILCYL
LWS Greenpuffer        ICWCAPPIFGWSRYWPHGLKTSCGPDVFSGSEDPGVQSYMIVLMITCCIIPLAIIVLCYL
LWS Fugu               AVWCAPPIFGWSRYWPHGLKTSCGPDVFSGSEDPGVQSYMIVLMITCCIIPLAIIILCYL
LWS1 Medaka            AVWCAPPVFGWSRYWPHGLKTSCGPDVFSGSDDPGVQSYMIVLMITCCIIPLAIIILCYL
LWS2 Medaka            AVWCAPPVFGWSRYWPHGLKTSCGPDVFSGSDDPGVQSYMIVLMITCCIIPLAIIILCYL
                         *  ** ********************** *   *** ***  ** ***  **  * ***
                                                        277     285
LWS Human              QVWLAIRAVAKQQKESESTQKAEKEVTRMVVVMIFAYCVCWGPYTFFACFAAANPGYAFH
MWS Human              QVWLAIRAVAKQQKESESTQKAEKEVTRMVVVMVLAFCFCWGPYAFFACFAAANPGYPFH
LWS1 Zebrafish         AVYLAIHAVAQQQKDSESTQKAEKEVSRMVVVMIFAYCFCWGPYTFFACFAAANPGYAFH
LWS2 Zebrafish         AVFLAIHAVAQQQKDSESTQKAEKEVSRMVVVMVLAFCLCWGPYTAFACFAAANPGYAFH
LWS2 BP                AVWWAIHSVAMQQKESESTQKAEKEVSRMVIVMIMAFCLCWGPYAVFACFAAGNPGYSFH
LWS2 PM                AVWWAIHSVAMQQKESESTQKAEKEVSRMVVVMILAFCLCWGPYTVFACFAAANPGYSFH
LWS Stickleback        AVWLAIRAVAMQQKESESTQKAERDVSRMVVVMIVAYIVCWGPYTTFACFAAANPGYAFH
LWS1 BP                AVWWAIHSVAMQQKESESTQKAEKEVSRMVVVMIFAYCFCWGPYTFFACFAAANPGYSFH
LWS1 PM                AVWWAIHSVAMQQKESESTQKAEKEVSRMVVVMIFAYCFCWGPYTFFACFAAANPGYSFH
LWS Greenpuffer        AVWMAIRAVAMQQKESESTQKAEREVSRMVVVMILAYCVCWGPYTFFACFAAANPGYAFH
LWS Fugu               AVWLAIRSVAMQQKESESTQKAEKEVSRMVVVMIVAYCVCWGPYTFFACFAAANPGYAFH
LWS1 Medaka            AVWLAIRAVAMQQKESESTQKAEKEVSRMVVVMIVAYCVCWGPYTFFACFAAANPGYAFH
LWS2 Medaka            AVWLAIRAVAMQQKESESTQKAEREVSRMVVVMIVAYCVCWGPYTFFACFAAANPGYAFH
                        *  **  ** *** ********  * *** **  *   *****  ****** **** **
                             308
LWS Human              PLMAALPAYFAKSATIYNPVIYVFMNRQFRNCILQLFGKKVDDGSELSSASKTEVSSVSS
MWS Human              PLMAALPAFFAKSATIYNPVIYVFMNRQFRNCILQLFGKKVDDGSELSSASKTEVSSVSS
LWS1 Zebrafish         PLAAAMPAYFAKSATIYNPVIYVFMNRQFRVCIMQLFGKKVDDGSEVST-SKTE---VSS
LWS2 Zebrafish         PLAAAMPAYFAKSATIYNPIIYVFMNRQFRVCIMQLFGKKVDDGSEVST-SKTE---VSS
LWS2 BP                PLAAALPAYFAKSATIYNPIIYVFMNRQFRCCIMQLFGKEVEDSSEVST-SKTE---VSS
LWS2 PM                PLAAAVPAYFAKSATIYNPVIYVFMNRQFRSCIMQLFGKEVEDSSEVST-SKTE---VSS
LWS Stickleback        PLAAAMPAYFAKSATIYNPVIYVFMNRQFRSCIMQLFGKEVDDGSEVST-SKTE---VSS
LWS1 BP                PLAAAMPAYFAKSATIYNPIIYVFMNRQFRTCIMQLFGKEVDDGSEVST-SKTE---VSS
LWS1 PM                PLAAAMPAYFAKSATIYNPIIYVFMNRQFRTCIMQLFGKEVDDGSEVST-SKTE---VSS
LWS Greenpuffer        PLAAAMPAYFAKSATIYNPIIYVFMNRQFRVCIMKLFGKEVDDGSEVST-SKTE---VSS
LWS Fugu               PLAAAMPAYFAKSATIYNPVIYVFMNRQFRVCIMKLFGKEVDDGSEVST-SKTE---VSS
LWS1 Medaka            PLAAAMPAYFAKSATIYNPIIYVFMNRQFRTCIMQLFGKQVDDGSEVST-SKTE---VSS
LWS2 Medaka            PLAAAMPAYFAKSATIYNPVIYVFMNRQFRTCIMQLFGKQVDDGSEVST-SKTE---VSS
                       ** ** ** ********** ********** **  **** * * ** *  ****   ***

LWS Human              VSPA
MWS Human              VSPA
LWS1 Zebrafish         VAPA
LWS2 Zebrafish         VAPA
LWS2 BP                VSPS
LWS2 PM                VSPS
LWS Stickleback        VAPA
LWS1 BP                VAPA
LWS1 PM                VAPA
LWS Greenpuffer        VAPA
LWS Fugu               VAPA
LWS1 Medaka            VAPA
LWS2 Medaka            VAPA
                       * *
```

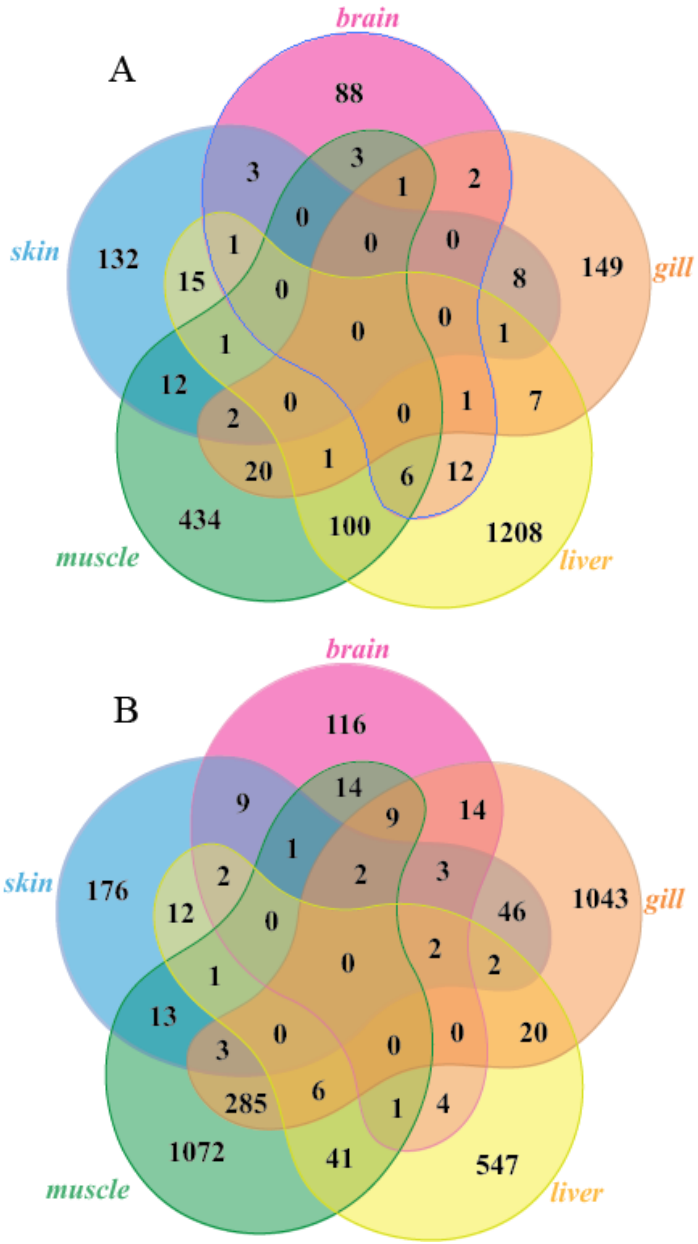
Supplementary Figure 21. Alignment of the protein sequences of LWS/MWS opsin genes from Human, Zebrafish, BP, PM, Stickleback, Pufferfish, Fugu, Medaka. The positions of five critical sites (180, 197, 277, 285, and 308) are marked in blue. Asterisks represent conserved residues in all LWS/MWS opsin genes. The amino acid substitutions of BP's LWS2 genes were same as MWS of Human.


Outgroup (non-OR GPCRs)
Amphioxus ORs
Lambda (λ), Kappa (κ), Theta (θ)
Eta (η)
Zeta (ζ)
Alpha (α)
Beta (β)
Epsilon (ε)
Delta (δ)
Gamma (γ)
81
67
62
100
82
60
99
92
100
20
96
Human
Zebrafish
Mudskipper
Elephant shark
Amphioxus
Human non-OR GPCRs
0.05

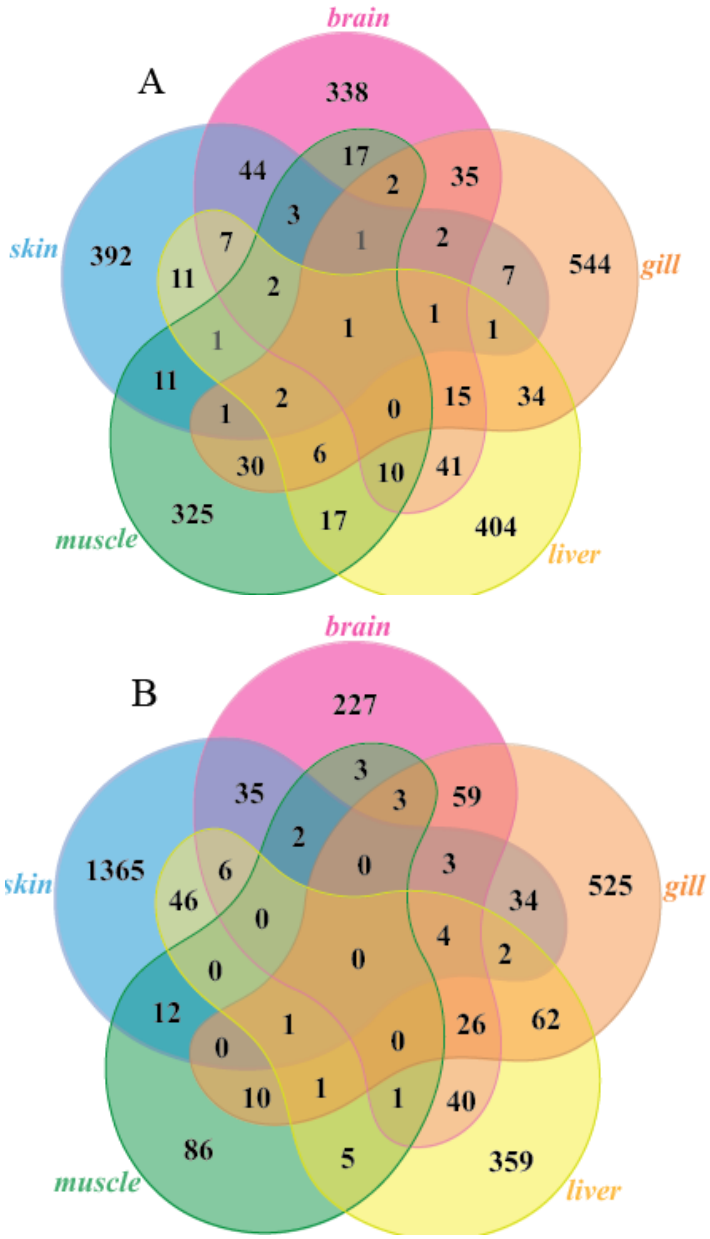
Supplementary Figure 22. Neighbor joining tree of OR-like genes from mudskippers, human, zebrafish, elephant shark, and amphioxus. A total of 657 sequences including six non-OR GPCR (outgroup) sequences were used to generate the tree. Sequences used as outgroup were: human alpha-1Badrenergic receptor (NP_000670.1), human muscarinic acetylcholine receptor M1 (NP_000729.2), human somatostatin receptor type 5 (NP_001044.1), human chemokine-binding protein 2 (NP_001287.2), human G-protein coupled receptor 35 isoform (NP_005292.2), and human G-protein coupled receptor 132 (NP_037477.1). Sequences from different species are color-coded whereas different OR groups are shown as labeled clades. Bootstrap support percentages ≥50 are shown at the main nodes of the tree. Sequences used in the analysis were extracted from the datasets of a previous study[2].


100
58
100
100
91
83
Elephant shark
Zebrafish
Mudskipper
Sea lamprey
Human non-OR GPCRs
V2R
(OlfC)
V1R
Outgroup (non-OR GPCR)
0.2

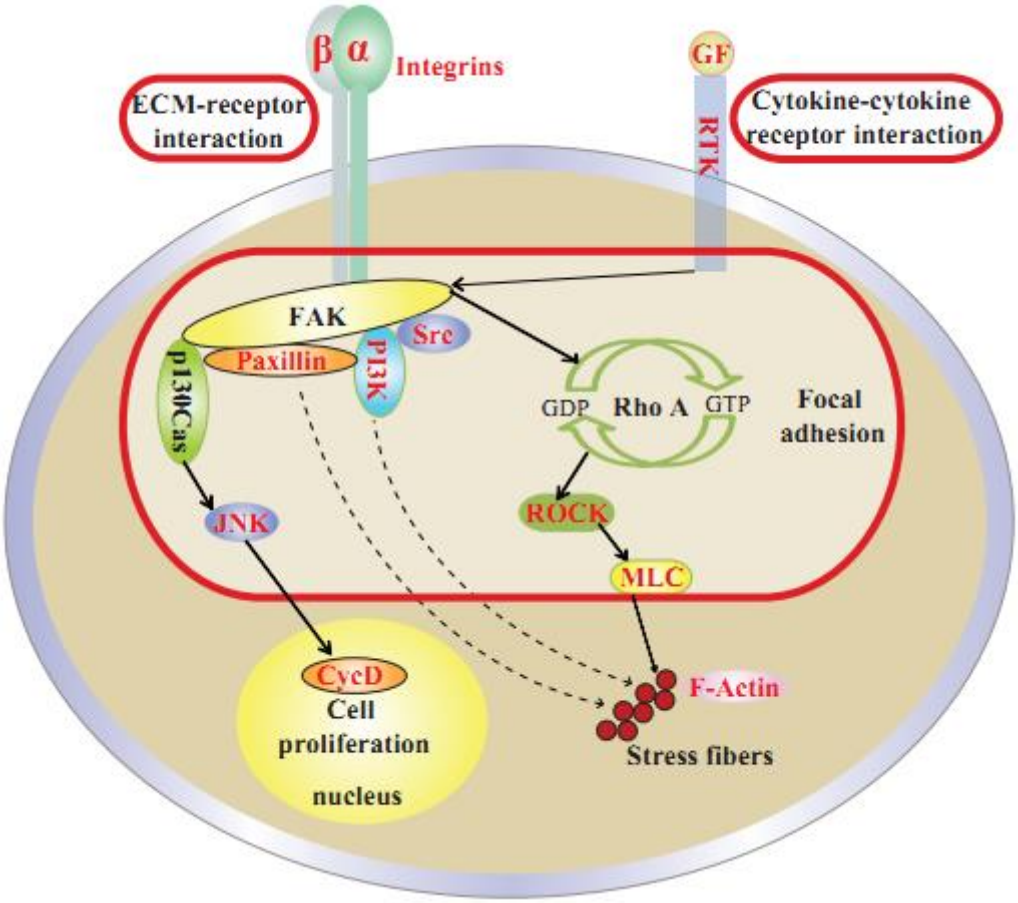
Supplementary Figure 23. Neighbor joining tree of V1R-like and V2R-like (*OlfC*) genes from mudskipper, zebrafish, elephant shark and lamprey. A total of 134 amino acid sequences including six human non-OR GPCRs were used to generate the tree. Non-OR GPCR sequences used were: human alpha-1Badrenergic receptor (NP_000670.1), human muscarinic acetylcholine receptor M1 (NP_000729.2), human somatostatin receptor type 5 (NP_001044.1), human chemokine-binding protein 2 (NP_001287.2), human G-protein coupled receptor 35 isoform (NP_005292.2), and human G-protein coupled receptor 132 (NP_037477.1). Bootstrap support percentages ≥50 are shown at the main nodes of the tree. Red branches denote mudskipper sequences whereas blue, brown and green branches denote sequences from zebrafish, elephant shark and sea lamprey, respectively. Zebrafish V2R sequences were obtained from Alioto's study[3] whereas zebrafish V1R sequences were retrieved from GenBank. Sea lamprey V2R-like sequences from Grus's study[4] were checked against the Petromyzon_marinus_7.0 assembly (www.ensembl.org) to obtain the Ensembl IDs/scaffold number. Non-OR GPCR information is from Niimura's paper [2].



Supplementary Figure 24. Venn diagram of differentially expressed genes in each tissue of BP. The numbers are presented for up-regulated (**A**) and down-regulated (**B**) genes during the air-exposure experiment.



Supplementary Figure 25. Venn diagram of differentially expressed genes in each tissue of PM. The numbers are presented for up-regulated (**A**) and down-regulated (**B**) genes during the air-exposure experiment.



Supplementary Figure 26. Summary of the down-regulated KEGG pathways (map04510) from the air-exposure experiment. Down-regulated genes enriched in KEGG pathways (P-value < 0.05, fold change ≥ 2) of BP and PM were shown in the red boxes. They mainly include 'ECM-receptor interaction', 'Cytokine-cytokine receptor interaction' and 'Focal adhesion'. Representative down-regulated protein-coding genes of 'Focal adhesion' pathway in BP and PM were marked in red. Dot lines represent that more than one step is involved in the process. Abbreviations: **FAK**: focal adhesion kinase, **MLC**: myosin light chain, **Src**: tyrosine-protein kinase Src, **PI3K**: phosphatidylinositol-4,5-bisphosphate 3-kinase, **ROCK**: Rho-associated protein kinase, **CycD**: cyclin D1, **RTK**: receptor tyrosine kinases.

Supplementary Tables

Supplementary Table 1. Summary of sequencing libraries and data production in mudskippers.

Species	Insert Size (bp)	Raw Data			After Filtering		
		Total Data(G)	Reads Length	Sequence Coverage (X)	Total Data(G)	Reads Length	Sequence Coverage (X)
BP	170	44.93	100	45.85	39.05	92	39.85
	250	57.39	150	58.56	44.24	142	45.14
	500	22.15	100	22.60	17.46	92	17.82
	800	17.83	100	18.19	13.51	92	13.79
	2,000	16.92	50	17.27	10.03	44	10.23
	5,000	31.86	50	32.51	9.54	44	9.73
	10,000	28.09	50	28.66	6.44	44	6.57
	20,000	13.55	50	13.83	1.63	44	1.66
Total		232.72		237.47	141.90		144.79
PM	250	45.05	150	57.76	30.28	142	38.82
	800	28.44	100	36.46	21.61	92	27.70
	2,000	20.31	50	26.04	8.69	44	11.14
Total		93.80		120.26	60.58		77.66
SH	250	48.61	150	60.77	29.97	142	37.46
	800	31.13	100	38.92	22.12	92	27.65
Total		79.74		99.69	52.09		65.11
PS	170	25.03	100	33.86	20.28	92	27.44
	500	20.51	100	27.75	16.41	92	22.20
	800	21.11	100	28.57	16.72	92	22.62
Total		66.65		90.19	54.40		72.26

Supplementary Table 2. Summary of mudskipper genome assemblies.

BP	Contig		Scaffold	
	Size(bp)	Number	Size(bp)	Number
N90	4,205	49,051	213,669	631
N80	7,823	33,705	534,171	349
N70	11,501	24,294	1,001,914	218
N60	15,482	17,580	1,562,116	142
N50	20,237	12,508	2,309,662	91
Longest	202,940		17,765,049	
Total Size	896,566,000		966,013,466	
Total	159,400		59,820	
Total Number(>2kb)	65,512		3,365	
PM	**Contig**		**Scaffold**	
	Size(bp)	Number	Size(bp)	Number
N90	4,915	29,048	29,737	3601
N80	9,931	19,292	71,661	2097
N70	15,178	13,626	127,407	1348
N60	20,875	9,696	199,173	897
N50	27,590	6,784	288,532	600
Longest	393,304		3,398,025	
Total Size	699,533,938		715,472,707	
Total	164,733		108,158	
Total Number(>2kb)	39,685		9,327	
SH	**Contig**		**Scaffold**	
	Size(bp)	Number	Size(bp)	Number
N90	1,023	108,534	1,491	70,540
N50	8,413	22,489	14,331	12,423
Longest	94,557		222,328	
Total Size	711,902,481		720,362,804	
Total	384,524		325,765	
Total Number(>2kb)	78,729		222,328	
PS	**Contig**		**Scaffold**	
	Size(bp)	Number	Size(bp)	Number
N90	3,950	42,804	7,574	19,733
N50	16,864	11,553	39,090	4,660
Longest	161,497		429,644	
Total Size	677,746,950		682,959,933	
Total	106,722		65,677	
Total Number(>2kb)	56,637		32,276	

Supplementary Table 3. Assembly statistics from published fish genomes

Genome name	Sequencing platform	Assembled genome size	Scaffold				Contig			
			Total Number	N50 size (kb)	N50 Number	Largest (kb)	Total Number	N50 size (kb)	N50 Number	Largest (kb)
zebrafish	Illumina,Sanger	1.412Gb	4,599	1,551	-	12,372	26,199	25	-	-
coelacanth	Illumina	2.86Gb	-	924	-	-	-	12.7	-	-
greenpuffer	Sanger	342Mb	25,773	100	-	7,612	49,609	16	-	258
medaka	Sanger	700.4Mb	10,357	1,410	-	-	124,126	9.8	-	-
stickleback	Sanger,Illumina	463.3 Mb	-	10,800	-	-	-	83.2	-	-
fugu	Sanger	332.5Mb	12,381	-	-	-	-	16.5	-	-
cod	454	753Mb	157,887	459	344	4,999	284,239	2.8	50,237	77
platyfish	454, Illumina	669Mb	84,533	1,102	8843	7,293	130,963	21	155	203
lamprey	454, Illumina	816Mb	-	173	-	2,400	25,073	-	-	-
lancelets	Illumina	520Mb	3,032	-	-	1,600	81,073	-	-	25
tuna	454, Illumina	800 Mb	16,802	136	-	1,021	192,169	7.6	-	79
BP	Illumina	983Mb	59,820	2,309	91	17,765	159,400	20	12,508	203
PM	Illumina	780Mb	108,158	288	600	3,398	164,733	27	6,784	393

Supplementary Table 4. Nine longest ALLPATH-LG scaffolds versus SOAP*denovo* assembly from the BP genome.

ALLPATH-LG			**SOAP*denovo***				
Scaffold ID	**Scaffold length (bp)**	**Aligned base number (bp)**	**scaffold ID**	**Aligned base number (bp)**	**Start**	**End**	**Identity (%)**
scaffold_1	3,085,967	2,974,442	scaffold28	2,976,313	16,470	3,349,171	99.9
scaffold_37	2,816,217	2,791,995	scaffold7	2,793,677	7,510,584	10,471,089	99.79
scaffold_3	2,836,493	2,812,503	scaffold5	2,814,534	6,443,760	9,443,628	99.78
scaffold_403	3,036,716	2,992,204	scaffold1	2,994,495	7,153,297	10,467,559	99.7
scaffold_524	3,700,222	3,645,705	scaffold3	3,649,317	6,137,145	10,025,468	100
scaffold_537	3,429,106	3,375,691	scaffold10	3,379,201	342,067	4,081,455	99.98
scaffold_572	3,648,356	3,625,189	scaffold17	3,628,374	4,492,581	8,417,319	100
scaffold_752	4,699,871	4,605,704	scaffold6	4,609,190	587,914	5,681,735	99.65
scaffold_87	2,972,370	2,946,872	scaffold27	2,950,228	1,901,586	5,064,086	99.96

Supplementary Table 5. Assessing the sequence coverage of BP and PM genome assemblies by RNA-seq of five tissues.

Brain (BP)

Dataset	Number	Total length (bp)	Covered by assembly (%)	With >90% sequence in one scaffold		With >50% sequence in one scaffold	
				Number	Percentage	Number	Percentage
>200bp	96,148	78,321,470	95.42	85,668	89.10	93,415	97.16
>500bp	42,947	61,678,206	95.09	37,860	88.16	41,447	96.51
>1000bp	21,528	46,692,094	94.77	18,835	87.49	20,748	96.37

Gill (BP)

Dataset	Number	Total length (bp)	Covered by assembly (%)	With >90% sequence in one scaffold		With >50% sequence in one scaffold	
				Number	Percentage	Number	Percentage
>200bp	78,419	55,094,995	93.58	66,415	84.69	72,566	92.54
>500bp	33,427	40,958,142	94.62	28,907	86.48	31,564	94.43
>1000bp	15,041	28,093,009	95.20	13,180	87.63	14,343	95.36

Liver (BP)

Dataset	Number	Total length (bp)	Covered by assembly (%)	With >90% sequence in one scaffold		With >50% sequence in one scaffold	
				Number	Percentage	Number	Percentage
>200bp	75,138	41,959,532	96.41	66,501	88.51	72,984	97.13
>500bp	24,230	26,780,064	96.87	22,164	91.47	23,640	97.57
>1000bp	9,219	16,399,502	97.06	8,506	92.27	9,033	97.98

Muscle (BP)

Dataset	Number	Total length (bp)	Covered by assembly (%)	With >90% sequence in one scaffold		With >50% sequence in one scaffold	
				Number	Percentage	Number	Percentage
>200bp	54,956	33,093,744	96.56	50,289	91.51	53,633	97.59
>500bp	18,172	22,008,409	96.32	16,309	89.75	17,623	96.98
>1000bp	7,756	14,827,103	96.26	6,903	89.00	7,518	96.93

Skin (BP)

Dataset	Number	Total length (bp)	Covered by assembly (%)	With >90% sequence in one scaffold		With >50% sequence in one scaffold	
				Number	Percentage	Number	Percentage
>200bp	50,530	26,245,000	96.10	45,361	89.77	49,087	97.14
>500bp	15,146	15,333,818	95.82	13,440	88.74	14,589	96.32
>1000bp	4,984	8,376,525	95.94	4,425	88.78	4,813	96.57

Brain (PM)

Dataset	Number	Total length (bp)	Covered by assembly (%)	With >90% sequence in one scaffold		With >50% sequence in one scaffold	
				Number	Percentage	Number	Percentage
>200bp	104,382	98,735,836	94.99	90,944	87.80	101,263	97.45
>500bp	50,273	82,102,874	94.63	42,175	83.89	48,405	96.28
>1000bp	28,316	66,633,609	94.23	22,891	80.81	27,104	95.72

Gill (PM)

Dataset	Number	Total length (bp)	Covered by assembly (%)	With >90% sequence in one scaffold		With >50% sequence in one scaffold	
				Number	Percentage	Number	Percentage
>200bp	83,402	73,859,113	94.31	71,841	86.14	79,041	95.20
>500bp	40,437	60,686,531	94.28	34,115	84.46	38,519	95.26
>1000bp	22,739	48,169,739	94.08	18,894	83.09	21,592	94.96

Liver (PM)

Dataset	Number	Total length (bp)	Covered by assembly (%)	With >90% sequence in one scaffold		With >50% sequence in one scaffold	
				Number	Percentage	Number	Percentage
>200bp	70,863	54,315,518	95.82	62,737	88.53	68,765	97.05
>500bp	30,958	41,996,987	95.54	26,544	85.74	29,836	96.38
>1000bp	15,572	31,163,285	95.28	13,006	83.52	14,934	95.90

Muscle (PM)

Dataset	Number	Total length (bp)	Covered by assembly (%)	With >90% sequence in one scaffold		With >50% sequence in one scaffold	
				Number	Percentage	Number	Percentage
>200bp	56,178	38,841,580	95.63	51,122	90.00	54,758	97.47
>500bp	20,556	28,009,598	95.00	18,137	88.23	19,869	96.66
>1000bp	9,518	20,347,552	94.27	8,158	85.71	9,116	95.78

Skin (PM)

Dataset	Number	Total length (bp)	Covered by assembly (%)	With >90% sequence in one scaffold		With >50% sequence in one scaffold	
				Number	Percentage	Number	Percentage
>200bp	74,237	67,167,717	95.90	65,543	88.29	71,863	96.80
>500bp	34,843	55,009,331	95.78	30,040	86.22	33,566	96.33
>1000bp	19,384	44,102,460	95.62	16,370	84.45	18,598	95.95

Supplementary Table 6. Core eukaryotic genes (CEGs) evaluated for the completeness of genome assemblies

	BP		PM	
	Number	**Completeness (%)**	**Number**	**Completeness (%)**
Complete	206	83.06	215	86.69
Group1	55	83.33	54	81.82
Group2	47	83.93	50	89.29
Group3	52	85.25	52	85.25
Group4	52	80.00	59	90.77
Partial	244	98.39	246	99.19
Group1	66	100.00	65	98.48
Group2	55	98.21	56	100.00
Group3	60	98.36	60	98.36
Group4	63	96.92	65	100.00

Supplementary Table 7. Major subfamilies of transposable elements discovered in mudskipper genomes.

Transposable element	BP			PM			PS			SH		
	Number	Base (kp)	%G	Number	#base (kp)	%G	Number	#base (kp)	%G	Number	Base (kp)	%G
DNA transponson												
other	89,746	16,120	1.67	8,170	1,765	0.25	77,754	12,084	1.77	43,495	6,162	0.86
TcMar	729,333	11,0536	11.45	596,417	78,457	10.97	332,951	45,711	6.69	542,803	60,141	8.35
P	974	93	0.01	1,060	93	0.01	2,141	402	0.06	3,244	373	0.05
Sola	28,451	4,576	0.47	13,449	1,9368	0.27	57,740	8,856	1.30	51,727	5,777	0.80
Harbinger	245	70	0.01	198	48	0.01	44,676	6,046	0.89	127	32	0.004
Kolobok	5483	654	0.07	2,703	314	0.04	2,499	417	0.06	3,485	325	0.05
hAT	684,144	97,841	10.13	213,234	28,069	3.92	364,309	49,624	7.27	268,892	32,621	4.53
LTR Retrotransposon												
other	4,934	1,071	0.11	1,408	267	0.04	1,958	471	0.07	587,974	83,163	11.5
ERVK	10,082	1,037	0.11	13,602	1,512	0.21	2,083	283	0.04	37,745	4,287	0.60
ERV1	6,582	809	0.09	7,140	1,120	0.05	12,854	2,224	0.33	18,636	2,418	0.34
Gypsy	175,064	30,5439	3.16	130,297	17,922	2.50	17,980	2,951	0.43	117,294	16,815	2.33
DIRS	2,116	406	0.04	1,570	268	0.01	1,289	197	0.03	1,402	161	0.02
Non-LTR Retrotransposon												
Penelope	107,209	16,855	1.74	151,562	19,847	2.77	86,770	11,669	1.71	22,011	2,638	0.37
L2	337,900	53,185	5.51	351,414	47,296	6.61	374,547	51,037	7.47	158,230	19,256	2.67
L1	55,206	10,143	1.05	24,240	3,929	0.55	39,543	6,011	0.88	29,894	4,872	0.68
RTE	407,518	63,655	6.59	409,472	54,302	7.59	345,059	45,162	6.61	17,408	2,996	0.37
CR1	2,420	391	0.04	2,248	444	0.06	7,560	1,059	0.16	1,491	163	0.02
Deu	2,959	222	0.02	1,537	118	0.02	37,739	5,391	0.79	3,703	331	0.05
SINE	22,680	2,910	0.30	18,682	2,080	0.29	45,757	4,572	0.67	14,203	1,460	0.20
MIR	13,220	1,584	0.16	31,359	3,655	0.51	22,489	2,987	0.44	52,623	6,072	0.84
tRNA	1,019	164	0.02	674	74	0.01	1,096	130	0.02	188	12	0.001
Simple_repeat	203,807	36,551	3.78	146,226	21,697	3.03	421,541	38,780	5.68	157,860	19,647	2.73
Unclassified	58,553	18,587	1.92	112,482	27,248	3.81	1,688	129	0.02	455479	60,751	8.43
Total (Non-redundance)	3,724,320	453,239	46.92	3,355,963	293,579	41.03	3,951,592	303,212	44.40	3,848,089	297,126	41.25

Supplementary Table 8. Gene annotation summary for the four mudskippers.

BP Gene Set		Number	Average Transcript Length (bp)	Average CDS Length (bp)	Average Exons per Gene	Average Exon Length (bp)	Average Intron Length (bp)
De novo	AUGUSTUS	37,945	11,196.30	1,082.97	5.85	185.10	2,084.96
	Genescan	27,355	23,112.97	1,538.13	8.40	183.02	2,913.82
Homolog	human	17,894	12,397.33	1,377.29	7.90	174.30	1,596.65
	stickleback	25,490	10,695.11	1,232.44	7.10	173.69	1,552.38
	zebrafish	24,710	10,759.89	1,308.21	7.01	186.66	1,573.06
	medaka	28,762	8,960.54	1,117.40	6.20	180.29	1,508.99
	fugu	20,691	12,915.09	1,429.72	8.25	173.32	1,584.39
	greenpuffer	20,220	12,712.73	1,393.94	8.26	168.70	1,558.48
Transcriptome		38,171	15,122.53	2,329.38	8.55	272.41	1,694.21
GLEAN		22,743	17,471.46	1,665.14	9.34	178.33	1,895.86
Final gene set		20,798	18,327.37	1,773.04	10.02	177.03	1,836.24

PM Gene Set		Number	Average Transcript Length (bp)	Average CDS Length (bp)	Average Exons per Gene	Average Exon Length (bp)	Average Intron Length (bp)
De novo	AUGUSTUS	31,106	9,227.87	1,160.30	6.51	178.33	1,465.09
	Genescan	23,654	16,420.92	1,473.87	8.17	180.33	2,083.76
Homolog	human	16,448	9,805.78	1,386.26	8.12	170.64	1,181.89
	stickleback	22,973	8,623.52	1,261.07	7.43	169.61	1,144.13
	zebrafish	22,456	8,632.50	1,314.86	7.33	179.48	1,156.78
	medaka	27,253	6,866.34	1,112.02	6.26	177.66	1,094.10
	fugu	19,944	9,871.53	1,409.42	8.26	170.53	1,164.80
	greenpuffer	19,291	9,731.12	1,379.73	8.28	166.65	1,147.26
Transcriptome		35,824	12,935.97	2,421.91	8.76	276.58	1,355.50
GLEAN		22,588	13,193.95	1,582.04	9.06	174.56	1,440.13
Final gene set		20,927	13,769.21	1,669.25	9.62	173.51	1,403.67

SH Gene Set	Number	Average Transcript Length (bp)	Average CDS Length (bp)	Average Exons per Gene	Average Exon Length (bp)	Average Intron Length (bp)

De novo	AUGUSTUS	19,245	6,658.86	1,048.09	5.75	182.29	1,181.30
	Genescan	19,845	5,967.72	923.70	4.87	189.50	1,301.89
Homolog	human	13,563	4,539.08	1,188.48	6.74	176.34	583.77
	stickleback	19,397	4,423.41	1,128.51	6.46	174.75	603.71
	zebrafish	19,581	4,273.73	1,138.86	6.22	183.02	600.27
	medaka	21,639	3,814.60	1,022.48	5.76	177.60	586.93
	fugu	17,574	4,776.99	1,210.13	6.89	175.56	605.29
	greenpuffer	18,249	4,569.22	1,138.08	6.61	172.11	611.34
GLEAN		17,273	6,624.56	1,199.36	6.37	188.23	1,009.92

PS Gene Set		**Number**	**Average Transcript Length (bp)**	**Average CDS Length (bp)**	**Average Exons per Gene**	**Average Exon Length (bp)**	**Average Intron Length (bp)**
De novo	AUGUSTUS	18,677	10,826.61	1,189.03	6.74	176.35	1,678.35
	Genescan	19,298	9,968.95	1,122.16	6.14	182.69	1,720.38
Homolog	human	15,203	6,150.88	1,245.55	7.15	174.32	798.25
	stickleback	21,470	5,764.18	1,159.13	6.70	173.10	808.42
	zebrafish	21,228	5,674.64	1,183.33	6.53	181.20	812.08
	medaka	24,986	4,774.10	1,033.95	5.78	178.99	782.99
	fugu	19,148	6,343.71	1,259.74	7.23	174.26	816.16
	greenpuffer	19,105	6,235.14	1,218.39	7.13	170.77	817.75
GLEAN		18,156	9,626.33	1,293.45	7.06	183.26	1,375.49

Supplementary Table 9. The statistics of gene structures in mudskippers compared to other typical fish.

Species	Number	Average Transcript Length (bp)	Average CDS Length (bp)	Average Exons per Gene	Average Exon Length (bp)	Average Intron Length (bp)
stickleback	20,772	8,455	1,549.23	10.40	148.90	734.36
zebrafish	26,046	24,122	1,593.15	9.29	171.50	2,717.27
cod	20,084	15,253	1,459.76	12.72	114.79	1,177.23
Latimeria	19,555	36,956	1,563.03	9.95	157.06	3,953.34
Nile tilapia	21,437	14,880	1,714.21	10.90	157.23	1,329.43
medaka	19,671	12,137	1,517.33	10.27	147.80	1,145.94
fugu	18,507	7,497	1,694.35	11.11	152.52	574.06
greenpuffer	19,583	6,065	1,517.44	10.53	144.15	477.44
BP	20,798	18,327.37	1,773.04	10.02	177.03	1,836.24
PM	20,927	13,769.21	1,669.25	9.62	173.51	1,403.67
SH	17,273	6,624.56	1,199.36	6.37	188.23	1,009.92
PS	18,156	9,626.33	1,293.45	7.06	183.26	1,375.49

Supplementary Table 10. The number of genes in each mudskipper with homologs or functional assignment from various databases.

BP	Number	Percentage (%)
Total	20,798	
InterPro	16,766	80.61
GO	14,143	68.00
KEGG	15,857	76.24
Swissprot	18,973	91.23
TrEMBL	19,704	94.74
Annotated	19,870	95.54
Unanotated	928	4.89

PM	Number	Percentage (%)
Total	20,927	
InterPro	17,301	82.67
GO	14,533	69.45
KEGG	16,111	76.99
Swissprot	19,316	92.30
TrEMBL	20,020	95.67
Annotated	20,082	95.96
Unanotated	845	4.04

SH	Number	Percentage (%)
Total	17,273	
InterPro	12,340	71.44
GO	10,300	59.63
KEGG	11,473	66.42
Swissprot	14,275	82.64
TrEMBL	15,155	87.74
Annotated	15,221	88.12
Unanotated	2,052	11.88

PS	Number	Percentage (%)
Total	18,156	
InterPro	13,525	74.49
GO	11,388	62.72
KEGG	12,398	68.29
Swissprot	15,483	85.28
TrEMBL	16,154	88.97
Annotated	16,262	89.57
Unanotated	1,894	10.43

Supplementary Table 11. Statistics of reads and mapping ratio of RNA-seq.

Samples	#Total	%Mappable	Samples	#Total	%Mappable	Samples description
BP_BC	45.03	77.82	PM_BC	53.56	82.64	Brain, control
BP_DB3	43.18	80.85	PM_DB3	39.26	74.5	Dry treatment, Brain, 3
BP_DB6	50.92	80.81	PM_DB6	56.13	66.47	Dry treatment, Brain,
BP_SC	46.79	79.02	PM_SC	59.83	83.6	Skin control
BP_DS3	38.07	79.83	PM_DS3	46.66	74.7	Dry treatment, Skin, 3
BP_DS6	52.15	81.86	PM_DS6	53.55	60.04	Dry treatment, Skin, 6
BP_MC	40.06	79.34	PM_MC	55.67	85.7	Muscle, control
BP_DM3	44.20	67.84	PM_DM3	37.93	83.01	Dry treatment,
BP_DM6	50.12	83.16	PM_DM6	49.74	82.1	Dry treatment,
BP_LC	64.11	81.01	PM_LC	58.42	84.13	Liver, control
BP_DL3	58.28	81.65	PM_DL3	39.18	80.11	Dry treatment, Liver,
BP_DL6	49.90	85.47	PM_DL6	43.94	81.8	Dry treatment, Liver,
BP_GC	41.48	75.51	PM_GC	76.62	79.71	Gill, control
BP_DG3	34.83	66.25	PM_DG3	52.40	77.33	Dry treatment, Gill, 3
BP_DG6	44.53	59.95	PM_DG6	40.40	58.81	Dry treatment, Gill, 6

Supplementary Table 12. **Validation of single heterozygous variations (SNVs) by Sanger sequencing.**

Scaffold	Position	Ref. Allele	Heterozygous Allele	Forward Primer	Reverse Primer	SNV status
scaffold6	424789	C	T	AAATGCAGTGGCACCACAGTAG	TGATCACAATCCCAGTGAGGG	valid
scaffold6	4970555	A	C	CTTATCACTGTTAGACCTCCCT	AGAGTGCGCCTTTAACTTTA	valid
scaffold1	8922312	C	T	TCCTGATCCTCGTCAGTAATC	ATGAGGCTGGCTCCAAAA	valid
scaffold1	10802226	T	A	GTGTCACAATTTAAAGGGCCTAT	TGGACATTACCAATTCCAGGT	valid
scaffold1	10986186	C	T	ACGGGCTACAATTTCCAGATA	TTTGGGCAAAGAATAGTGTGT	valid
scaffold7	11210741	T	A	TTACAGGGCAGGTCTATGGA	TTAACCCAACTGTGTAGAAACA	valid
scaffold8	2189800	C	T	ATGGCTGTTCAAATGTTTCTCA	TCCAGATTACTGTAGTTCCCAGA	valid
scaffold8	7285792	G	A	ATCGCATAAACGACTGAAGAC	GCCCCATCACAGCTAACCAT	valid
scaffold8	8834890	C	T	ACATCCCTGACGCTCCATCT	CATACAGTACAACTATCACGCAAC	valid
scaffold4	2184913	C	G	AATGCGACCCGAGTCTGATA	GACAATGCTGGGCGAGTTAT	valid
scaffold4	5747981	A	G	CCACAGGTACACCATGGATC	GACAATGCTGGGCGAGTTAT	valid
scaffold3	1480752	A	T	GACAATGCTGGGCGAGTTAT	ACAATGACGATATGGGGAAA	valid
scaffold3	1560514	C	G	AAGGCTCCACCCCAAACA	TAAAAAGGAACATACATGCAATATA	valid
scaffold3	3990828	A	T	CGTGTTGTGATTGGTTCAGC	ATTGCCAGGTGCCTATTTG	valid
scaffold3	4291996	T	C	GATTTGCAGAAAGATAGAAAGAGA	CTATGCTTGGAATGTAACACAGTA	valid
scaffold3	4404609	G	C	AAACCTCTATAATGGCTCCACT	ATACACATTTGTGCTTCGGTAC	valid
scaffold18	1104863	G	T	TTAGCGTTTGTGATTCTCAAA	CTCAACATAGGCAACATTTTG	valid
scaffold18	1211487	G	A	GCTGTTGAAACTGAATGGATGA	TTTTGCCCTTTGACACTGCTAT	valid
scaffold18	3101063	G	A	AGACACAACATTCGCACTCG	AGCATATCAGATCACATCCCC	valid
scaffold16	1057057	T	G	CTTCATCACAGTGTCGTATGAGG	AACAGGTATTATCCCACCAAAT	valid
scaffold16	1680806	T	C	TGATAAGATAGAAGAGCAATGGC	GCTAGTCTGAGAGAAATCATGTTACA	valid
scaffold16	6583122	T	G	ACTGGACATCCTGGGAACAT	ACCCTGCCAACAGTATGAGA	valid
scaffold41	4335161	T	C	CATTTTTATTATCTACCTATCTAAACAGC	CAAAATAAGCAAAAATAGCAAGAGA	valid
scaffold41	4283440	G	T	GTTACGGGATCTGATCTGTAACTT	GCAAGGAGCATTGTGATCTATT	valid
scaffold38	696976	T	C	AAACCTGCTTAGTATCAGGGAA	TTTAAATGGTTCAGTATGGTGC	valid
scaffold38	61975	G	A	TGGTGTGGGGCCACAGA	GGATGAGAGACAGCAGAGCAT	valid
scaffold45	4052750	G	A	CTACATAAGCATACTTTGCAGCTG	TTTGATAAATACAAGCCTTCGTG	valid
scaffold404	323565	G	C	TCATCTGAATGTAGGCTTTTGTG	TGTGAAACTTTCCTCATGTTCTT	valid
scaffold56	1494565	C	T	GAAGTGCAGCTCTTTGAAGGT	TATGCGATTTAGGATTGGATG	valid
scaffold56	816646	T	C	ATTGTTCTAATTGCGTGAAGG	GGCTGTTATCCGTTTTTTCCA	valid
scaffold56	828401	T	C	CTCTTTGTCTGCTTTACCATTTC	GTACAGCCTTTTCCTTTATTACACT	valid
scaffold41	1641291	G	A	GAGAATACTTCGGCCACATTA	GAACCACAACAGGTCCAAAAC	valid
scaffold38	1729973	A	C	GTTTCCCGTGCCTCTAAACC	ATGGCTTATGGGAGCAGATG	valid
scaffold38	1792393	T	A	AACACTACCTCATGTTATTATGAAATG	TCACCTGTCAAACTATTCGGA	valid
scaffold7	3448654	C	T	CAGAAAGGAGCAGTTTGGC	AAACCGCTCTAACAAGAACATA	invalid
scaffold7	4682320	C	T	CCCCACTCACTGTACGTAAGTAG	AGACTCCAGGAAATGCAACAT	invalid
scaffold3052	1695	T	A	CTGTCTAGGGCCAGCAGAG	TGGATGGCACTCACATGATA	invalid

Supplementary Table 13. TRL13 copy number in ten vertebrates.

Species	complete	patial
BP	**7**	**4**
PM	**6**	**5**
zebrafish	**3**	**4**
medaka	**2**	**0**
stickleback	**3**	**0**
greenpuffer	**2**	**1**
fugu	**3**	**0**
frog	**3**	**0**
lizard	**3**	**0**
mouse	**1**	**0**

Supplementary Table 14. Positive selection analysis on three core genes from ammonia excretion pathway of BP, PM and PS.

Species	Symbol	Description	ω2(whole average)	ω1(other average)	ω0(target)	P-value
BP	CA15	Carbonic anhydrases 15	0.1982	0.1822	0.6615	0.01088035
	NHE3	Na^+/H^+ exchanger 3	0.0919	0.0860	0.2264	0.01620595
	Rhcg1	Rh glycoprotein cg1	0.0659	0.0679	0.0209	0.1130190
PM	CA15	Carbonic anhydrases 15	0.1982	0.1786	0.5112	0.006232303
	NHE3	Na^+/H^+ exchanger 3	0.0919	0.0878	0.1362	0.105934
	Rhcg1	Rh glycoprotein cg1	0.0659	0.0626	0.1353	0.03681123
PS	Rhcg1	Rh glycoprotein cg1	0.0659	0.0631	0.1469	0.04924776

Positively selected genes were marked in red.

Supplementary Table 15. Concentrations (μmol mL-1) of NH_3, NH_4^+, and ammonia (NH_3 & NH_4^+) in the external medium (50% seawater with 8mM NH_4Cl) at 0 h and 24 h after exposure [5].

	B.boddaerti		*P. schlosseri* (PS)	
	0 h	24 h	0 h	24 h
pH 7:				
NH_3	0.036 ± 0.005	0.035 ± 0.003	0.036 ± 0.005	0.049 ± 0.003 [a]
NH_4^+	7.91 ± 0.15	9.04 ± 0.63[a]	7.98 ± 0.12	10.84 ± 0.63 [a]
Ammonia	7.95 ± 0.16	9.17 ± 0.18 [a]	8.12 ± 0.15	10.99 ± 0.69 [a]
pH 8:				
NH_3	0.34 ± 0.02	0.085 ± 0.017 [a]	0.34 ± 0.005	0.16 ± 0.017 [a]
NH_4^+	7.68 ± 0.13	7.53 ± 0.55	7.47 ± 0.15	10.53 ± 0.52 [a]
Ammonia	8.10 ± 0.12	7.59 ± 0.43	7.79 ± 0.18	10.77 ± 0.15 [a]

Note. Results represented means ± SD of five specimens.

[a] Significantly different from the value at 0 h respectively.

Supplementary Table 16. Olfactory receptor gene repertoire in teleost fishes.
Non-OR groups of lambda (λ), kappa (κ) and theta (θ) are not shown.

species	Type 1						Type 2
	Air		Water			Air/Water	Water
	Alpha (α)	Gamma (γ)	Delta (δ)	Epsilon (ε)	Zeta (ζ)	Beta (β)	Eta (η)
zebrafish	0	1	62	12	37	4	38
medaka	0	0	33	3	9	3	20
stickleback	0	0	71	4	18	1	8
fugu	0	0	30	2	4	1	10
BP	0	0	20	1	1	2	8
PM	0	0	17	2	2	2	10
coelacanth[6]	~12	~20	~80	~3	~60	~2	~15
human	58	329	0	0	0	0	0

Supplementary Table 17. Vomeronasal receptor gene repertoire in vertebrates.

Species	V1R	V2R
zebrafish	2	44
fugu	1	18
greenpuffer	1	4
BP	4	8
PM	4	8
Coelacanth [6]	20	Not determined
frog	21	249
chicken	0	0
opossum	98	79
mouse	187	70
rat	106	59
human	5	0

Supplementary Table 18. Up-regulated and down-regulated gene numbers in five tissues of BP and PM under air exposure.

Tissue	BP		PM	
	Up-regulated gene number	Down-regulated gene number	Up-regulated gene number	Down-regulated gene number
Brain	117	177	519	409
Gill	192	1,435	682	730
Liver	1,353	638	553	553
Muscle	580	1,448	429	124
Skin	175	272	487	1,510
Total	2,207	3,444	2,305	2,917

Supplementary Table 19. Common enriched KEGG pathways of down-regulated genes in five tissues of BP and PM under air exposure.

KEGG pathway enrichments of down-regulated genes	BP (multiple tissues)		PM (multiple tissues)	
	Gene number	P-value	Gene number	P-value
Focal adhesion	127	3.93E-09	117	2.94E-09
ECM-receptor interaction	79	3.36E-13	64	2.96E-08
Cytokine-cytokine receptor interaction	47	7.07E-03	54	1.78E-06
Axon guidance	75	4.12E-02	82	2.21E-04
Antigen processing and presentation	25	1.04E-03	24	1.27E-03
Protein digestion and absorption	68	4.74E-09	45	3.46E-03
Hematopoietic cell lineage	25	2.72E-02	25	1.28E-03
TGF-beta signaling pathway	40	1.47E-03	34	1.82E-02

KEGG pathway enrichments of down-regulated genes	BP (brain)		PM (brain)	
	Gene number	P-value	Gene number	P-value
Cytokine-cytokine receptor interaction	9	6.87E-06	19	7.41E-09
NF-kappa B signaling pathway	4	2.67E-02	7	3.05E-02
Hematopoietic cell lineage	3	3.03E-02	7	4.92E-03
NOD-like receptor signaling pathway	4	3.74E-02	13	3.75E-06
Cell adhesion molecules (CAMs)	5	4.69E-02	19	6.58E-06

KEGG pathway enrichments of down-regulated genes	BP (liver)		PM (liver)	
	Gene number	P-value	Gene number	P-value
Gap junction	11	1.53E-02	12	1.45E-03
Amino sugar and nucleotide sugar metabolism	6	3.23E-02	7	3.36E-03

KEGG pathway enrichments of down-regulated genes	BP (muscle)		PM (muscle)	
	Gene number	P-value	Gene number	P-value
Valine, leucine and isoleucine degradation	16	1.97E-05	3	3.62E-03
Glutathione metabolism	12	2.77E-03	2	4.12E-02

KEGG pathway enrichments of down-regulated genes	BP (skin)		PM (skin)	
	Gene number	P-value	Gene number	P-value
Cell adhesion molecules	10	2.21E-03	33	3.33E-02
Leukocyte transendothelial migration	8	2.28E-03	31	4.77E-02

KEGG pathway enrichments of down-regulated genes	BP (gill)		PM (gill)	
	Gene number	P-value	Gene number	P-value
Protein digestion and absorption	44	1.60E-11	16	1.88E-02
Antigen processing and presentation	16	1.25E-04	13	1.14E-05
Natural killer cell mediated cytotoxicity	20	1.61E-02	11	2.67E-02

Supplementary Table 20. Common enriched KEGG pathways of up-regulated genes in five tissues of BP and PM under air exposure.

KEGG pathway enrichments of up-regulated genes	BP (multiple tissues)		PM (multiple tissues)	
	Gene number	P-value	Gene number	P-value
Metabolic pathways	219	3.63E-04	273	9.52E-16
Arginine and proline metabolism	19	7.01E-04	17	5.17E-03
Fructose and mannose metabolism	15	3.78E-03	13	8.08E-03
Glycerophospholipid metabolism	26	3.93E-03	32	6.08E-05
Complement and coagulation cascades	24	2.94E-02	37	3.88E-06

KEGG pathway enrichments of up-regulated genes	BP (liver)		PM (liver)	
	Gene number	P-value	Gene number	P-value
Arginine and proline metabolism	16	8.40E-05	9	4.41E-04
Metabolic pathways	148	2.04E-04	110	2.36E-19
Peroxisome	18	1.97E-03	8	2.05E-02
Glycerophospholipid metabolism	18	6.95E-03	13	2.33E-04
Propanoate metabolism	8	1.44E-02	4	3.57E-02
beta-Alanine metabolism	7	1.62E-02	4	1.88E-02
Glycine, serine and threonine metabolism	8	2.28E-02	8	1.29E-04
Complement and coagulation cascades	17	2.55E-02	13	5.31E-04
Tryptophan metabolism	8	3.11E-02	6	2.50E-03
PPAR signaling pathway	16	3.25E-02	16	1.30E-06
Starch and sucrose metabolism	9	3.41E-02	7	5.52E-04
Cytosolic DNA-sensing pathway	8	3.77E-02	5	2.07E-02

Glycerolipid metabolism	11	4.28E-02	7	1.12E-02

KEGG pathway enrichments of up-regulated genes	**BP (muscle)**		**PM (muscle)**	
	Gene number	**P-value**	**Gene number**	**P-value**
RNA polymerase	6	2.11E-02	4	1.43E-02
Jak-STAT signaling pathway	10	1.62E-02	7	4.21E-02

Supplementary Note 1

1. Organism background, genome sequencing and assembly

Organism background

Mudskippers are members of four genera from the subfamily Oxudercinae, within the family Gobiidae. The distribution of mudskippers centers mainly in the tropical Indo-Pacific. The known localities of *Boleophthalmus* and *Scartelaos* species are similar, ranging from the Red Sea throughout the East Asian countries, and south to the tropical regions of Australia. The distribution of *Periophthalmodon* species is more restricted to Southeast Asia and mid-northern Australia. *Periophthalmus* species shows the widest distribution among mudskippers, from the west coast of Africa to the Polynesian islands[7].

Sample background and sequencing

Wild individuals of *Boleophthalmu spectinirostris* (BP, female, 1 age), *Periophthalmus magnuspinnatus* (PM, female, 1 age) and *Scartelaos histophorus* (SH, female, 1 age) were collected from Shenzhen Bay at Shenzhen and Qiao Island at Zhuhai, Guangdong Province, China in July of 2012, and *Periophthalmodon schlosseri* (PS, female, 1 age) samples were collected in Malaysia. Genomic DNA was isolated from several mixed tissues by standard molecular biology techniques. Whole genome shotgun sequencing strategy was employed and subsequent short-insert libraries (170-bp, 250-bp, 500-bp & 800-bp for BP, 250-bp & 800-bp for PM and SH, and 170-bp, 500-bp & 800-bp for PS) and long-insert libraries (2-kb, 5-kb, 10-kb & 20-kb for BP and 2-kb for PM) were constructed using the standard protocol provided by Illumina (San Diego, USA). Paired-end sequencing was performed using the Illumina HiSeq 2000 system. In total, we obtained 232.72, 93.80, 79.74 and 66.65 gigabases (Gb) (**Supplementary Table 1**) of raw reads from the libraries of BP, PM, SH and PS, respectively.

All animal experiments in this study were performed in accordance with the

guidelines of the animal ethics committee and were approved by the Institutional Review Board on Bioethics and Biosafety of BGI.

Read processing and genome size estimation by k-mer analysis

Sequencing errors in Illumina reads can disturb short-read assembly algorithms. We therefore performed several highly stringent filtering steps to remove low-quality reads as follows: (1) Reads of short-insert libraries were trimmed of 4 low-quality bases at both ends, and reads of long-insert libraries were trimmed of 3 low-quality bases; (2) For long-insert libraries, duplicated reads were filtered out; (3) We also examined individual reads in all lanes, and discarded reads with 10 or more Ns (no sequenced bases) and low-quality bases. Finally, 141.90 Gb (BP), 60.58 Gb (PM), 52.09 Gb (SH) and 54.40 Gb (PS) of clean reads were obtained (**Supplementary Table 1**) for genome assembly and size estimation.

The estimation of genome size by k-mer frequency distribution analysis is also very sensitive to sequencing errors, so we only chose clean reads from the short-insert libraries (500- or 800-bp). 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 read 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 follow a Poisson distribution in any given dataset. The genome size (G), can be deduced from the formula G=N*(L-17+1)/K_depth, where the N is the total number of reads, and K_depth indicates the frequency occurring more frequently than the others. For BP, N was 337,787,664 and the K_depth was 26 (**Supplementary Figure 1**), therefore the BP genome size was estimated to be 0.983 Gb. Similarly, estimated genome sizes of PM, SH and PS were estimated to be 0.780 Gb, 0.806 Gb and 0.739 Gb, respectively (**Supplementary Figure 1**).

Genome assembly and assessment

For whole-genome shotgun assembly of the four mudskipper species, we employed SOAP*denovo2* (version 2.04.4)[8] with optimized parameters (pregraph -K 27 -p 16 -d 1; contig –M 3; scaff -F -b 1.5 -p 16) to construct contigs and original scaffolds. All reads were mapped onto the contigs for scaffold building by utilizing the paired-end information. This paired-end information was subsequently applied to link contigs into scaffolds in a stepwise manner. Some intra-scaffold gaps were filled by local assembly using the reads in a read-pair where one end uniquely mapped to a contig whereas the other end was located within a gap. Subsequently, SSPACE[9] (version 2.0; using core parameters "-k 6 -T 4 -g 2") was used to link the SOAP*denovo2* scaffolds of BP and PM into super scaffolds with large-insert reads (> 1kb). The final genome assembly of BP is 0.966 Gb with gaps and 0.896 Gb without gaps in total length, which is about 98.2% and 91.1% of the estimated genome size. The contig N50 (the shortest length of sequence contributing more than half of the assembled sequences) is 20.24 kb and the scaffold N50 is 2.31 Mb, which are indicative of a high quality assembly (**Supplementary Tables 2-3**). Similarly, the contig N50s of PM, SH and SV are 27.59 kb, 8.41 kb and 16.86 kb, respectively and the scaffold N50s are 288.53 kb, 14.33 kb and 39.09 kb, respectively (**Supplementary Table 2**).

To assess the quality and accuracy of these assembly results, we applied 5 reliable evaluation methods as follows:

GC content assessment:

GC content for the 4 sequenced mudskippers was calculated to analyze nucleotide distribution and to examine the randomness of sequencing. The GC content analysis was also used to check possible sample contamination because the distribution would be expected to show two clusters. In our GC distribution views of the 4 species, they are normally distributed and we did not discover any other cluster except a core GC one (**Supplementary Figures 2-3**). We subsequently compared the GC content of mudskippers with those of other sequenced teleosts, and found the values in mudskippers are typical among the teleosts; their average GC level is

higher than zebrafish but lower than other fish[10] (such as medaka, fugu and greenpuffer; **Supplementary Figure 3**).

Evaluation of 40-kb fosmid clones:

To check the quality of the assembled BP genome, a 40-kb fosmid library was constructed from BP genomic DNA, and 8 clones were picked randomly and sequenced using Sanger strategy on ABI 3730 sequencer. We used the Nucmer program of MUMmer (version 3.22)[11] software package to align clones and scaffolds at default parameters. All clones could be aligned to their corresponding BP scaffolds with high coverage and identity (**Supplementary Figure 4**). We discovered that almost all gaps in the aligned scaffolds were due to numerous repeats in these regions.

Comparison between ALLPATH-LG scaffolds and SOAP*denovo2* scaffolds:

The same reads from BP were assembled using another de Bruijn graph-based and WGS assembly software named ALLPATH-LG[12], which is designed and maintained by Broad Institute of MIT and Harvard, Cambridge. The assembled results demonstrated that the contig N50 is 5.4 kb and the scaffold N50 is 293 kb. We chose 9 longest scaffolds from the ALLPATH-LG assembly to map against the SOAP*denovo2* scaffolds using Nucmer software. We discovered that each ALLPATH-LG scaffold aligned to only one region of its corresponding SOAP*denovo2* sequence (**Supplementary Figure 5**). The sequences are comparable and the identities are over 99.5% between both assemblies (**Supplementary Table 4**).

Transcriptome evaluation:

By assembling RNA-seq data (details given in Section 3) generated in this study and mapping the assembled fragments to the assemblies using BLAT[13] (E-value=$10e^{-6}$, identity=90%, coverage>90%), we estimated that 92~96% and 94~96% of transcribed regions were covered in the BP and PM genome assemblies, respectively (**Supplementary Table 5**).

CEGMA evaluation:

CEGMA[14] (Core Eukaryotic Genes Mapping Approach)

(http://korflab.ucdavis.edu/Datasets/genome_completeness, version 2.3) was employed to evaluate the gene space completeness of the BP and PM genomes. In total, 248 core eukaryotic genes (CEGs), which were highly conserved and detected in 6 typical genomes (*H.sapiens, D.melanogaster, A.thaliana, C. elegans, S.cerevisiae* and *S. pombe*), were mapped against the mudskipper genome assemblies. The results confirmed that both BP and PM assemblies covered more than 80% of the completed CEGs sequences and more than 98% with partial coverage (**Supplementary Table 6**). In Genis's CEGMA paper[14], he indicated that the CEGMA may have problems with other genes whose structure is interrupted by large insertions of other genes. Hence, the ratios of partial CEGMA coverage (98%) may truly represent the completeness of BP and PM genomes.

Use of these assessment methods confirmed that the assembled genome of BP is of high-quality, and can be considered as a good reference genome. Additionally, the PM genome assembly has long contigs and can be used for comparative genomic analysis. The SH and PS genome assemblies can be utilized for annotation of gene structures and construction of phylogenetic trees (**Figure 2a**).

Supplementary Note 2

2. Genome annotation

Repetitive sequence detection

We constructed a *de novo* repeat library using RepeatModeller (http://www.repeatmasker.org/RepeatModeler.html, version 1.04, default parameter) and LTR_FINDER[15]. To identify known and *de novo* Transposable elements (TEs), we employed RepeatMasker[16] (http://www.repeatmasker.org, version 3.2.9) against the Repbase[17] TE library (version 14.04) and the *de novo* repeat library. In addition, we used RepeatProteinMask (http://www.repeatmasker.org/, Version 3.2.2) implemented in RepeatMasker to detect the TE relevant proteins. We also predicted tandem repeats utilizing Tandem Repeat Finder[18] (http://tandem.bu.edu/trf/trf.html,

version 4.04) with parameters set as “Match=2, Mismatch=7, Delta=7, PM=80, PI=10, Minscore=50, and MaxPerid=2000”.

Protein-coding gene annotation

Homology-based gene prediction: We aligned *H.sapiens* (human), *D.rerio* (zebrafish), *T.rubripes* (fugu), *T.nigroviridis* (greenpuffer), *G.aculeatus* (stickleback) and *O.latipes* (medaka) proteins (Ensembl release 64) to the BP and PM genomes using TblastN with E-value ≤ 1E-5, and then made use of Genewise2.2.0[19] for precise spliced aligning and predicting gene structures. Short genes (less than 150 bp) and premature or frame-shifted genes were removed.

***Ab initio* prediction:** Genome sequences of BP and PM were repeat-masked and 1,500 full-length and random-selected genes from their homology gene sets were used to train the model parameters for AUGUSTUS. Subsequently, we utilized AUGUSTUS2.5[20] and GENSCAN1.0[21] for *de novo* prediction on repeat-masked genome sequences. Short genes were discarded using the same filter threshold as for homology prediction.

Gene structure identification using transcriptome reads: We mapped the mixed RNA reads from liver, muscle, skin, gill and brain samples (details of RNA sample preparation are given in **Supplementary Note 3**) of BP and PM to their genomes respectively using Tophat1.2[22]. Subsequently, we sorted and merged the Tophat mapping results and then applied Cufflink (http://cufflinks.cbcb.umd.edu/)[23] software to identify gene structures to assist gene annotation.

Gene sets integration, optimization and evaluation: All above gene sets were merged to form a comprehensive and non-redundant gene set using GLEAN[24] (**Supplementary Figure 6**). To optimize the consensus gene set, we removed 1) genes shorter than 150 bp, 2) genes annotated as TEs, and 3) *denovo*-only predictions, with highest expression value FPKM (Fragments Per Kilobase of exon per Million fragments mapped) <1 and with no functional assignment. Finally, we obtained the final gene sets containing 20,798 genes for BP and 20,927 for PM

(**Table 1**). Over 70% of the genes are supported by all the three types of evidence, and approximately 95% of the genes were supported by at least two types of evidence (**Supplementary Figure 7**), and around 98% of the genes had FPKM values >0 (**Supplementary Data 1**), all of which confirm that our gene sets of BP and PM are of high quality.

Gene annotation of SH and PS: We adopted the majority of the above-mentioned prediction methods except transcriptomic annotation to identify gene structures on the SH and PS genome sequences; finally, we obtained 17,273 and 18,156 genes, respectively (**Table 1**).

Multiple functional assignments of genes

For the final gene sets of the four examined mudskippers, all proteins were aligned to SwissProt and TrEMBL (Uniprot release 2011.06)[25] by BlastP with an E-value cut-off of 1E-5 to identify the best hit for each protein. Motifs and domains were annotated using InterProScan4.7[26] against publicly available databases including Pfam, PRINTS, ProDom and SMART. Descriptions of the gene products are presented by Gene Ontology[27] (retrieved from the results of the InterProScan). We mapped the final genes to the KEGG[28] pathway maps by detecting the best hit for each gene. In summary (**Supplementary Table 10**), approximately 95% of the genes from BP and PM, and 90% of the genes from SH and PS, are supported by at least one related function from the searched databases (Swiss-Prot, Interpro, TrEMBL and KEGG). Among the remaining, about 5% of the BP and PM genes cannot get any hits, but around 96% of them had transciptome support (RPKM > 0; **Supplementary Data 1**). The InterPro annotation showed that 5,663 and 5,744 domains are present in the BP and PM genomes, and 16,766 and 17,301 genes contain one or more domains, respectively.

Supplementary Note 3

3. Transcriptome analysis

Animals

BP (body weight at 6-9 g) and PM (4-7 g) were collected respectively from mangrove swamps of Shenzhen Bay at Shenzhen and Qiao Island at Zhuhai, Guangdong Province, China in July of 2012.

The mudskippers were maintained in plastic aquaria with artificial seawater (15‰ salinity) at 27 ± 0.5 ^{o}C. Trizma Base (Sigma Chemical Co., USA) was added (at a final concentration of 10 mM) to the seawater, and pH of the prepared seawater was adjusted to 7 with concentrated HCl. Experiments were performed in the aquarium at BGI, Shenzhen, China. Under the control of air-conditioners, the room temperature was set at 27 ± 0.5 ^{o}C and the humidity was maintained at 75 ± 3%. The seawater was changed every day. The fish were cultured separately in the aquaria for one week before experiments. No attempt was made to separate the genders. During the adaptation period, PM individuals were fed with frozen bloodworms and BP individuals were fed with a local commercial feed. All fish were fasted for 24 h prior to the experiments and were not fed during the subsequent tests.

All animal experiments in this study were performed in accordance with the guidelines of the animal ethics committee and were approved by the Institutional Review Board on Bioethics and Biosafety of BGI.

Air-exposure experiment

Six individuals were placed in Tris (pH 7.0)-15‰ artificial seawater as controls. Ten individuals were placed in plastic aquaria without seawater for air exposure; the room temperature and humidity were maintained at 27 ± 0.5^{o}C and 75 ± 3%, respectively. Samples from the controls were collected at time point zero, whereas tissues from the exposure groups were collected at 3 and 6 hours. For the collection of samples, each fish was killed with a single blow on its head. The gill, brain, liver,

skin and muscle were collected immediately. No attempt was made to separate the red and white muscle. The samples were immediately freeze-clamped in liquid nitrogen with pre-cooled aluminum tongs. All samples were stored at -80^{o}C until use.

Transcriptome sequencing

RNA was extracted from tissues of three randomly picked individuals for a particular time point using TRIzol reagent (Invitrogen, USA) and was then mixed in equal amounts for RNA-seq. We used 90-bp paired-end Illumina reads for transcriptome sequencing on the HiSeq2000 platform. The total raw reads for each sample are given in Supplementary Table 11.

Expression calculation and differentially expressed gene detection

The RNA-seq reads were mapped by Tophat[22] and gene expression levels were measured by FPKM (Fragments Per Kilobase of exon per Million fragments mapped) , which was similar to single-read "RPKM". The Cuffdiff [29] package of Cufflink software (version 2.0.2.Linux_x86_64), with core parameters (–FDR 0.05 --geometric-norm TRUE --compatible-hits-norm TRUE) to reduce certain types of bias caused by differential amounts of RNA reads, was utilized to calculate expression level (FPKM of all genes are shown in **Supplementary Data 1**) and to detect differentially expressed genes with false discovery rate (FDR) <0.05. Combining the results of gene function annotation, we identified GO terms enriched in the differentially expressed genes using EnrichPipeline from Chen's paper[30].

EnrichmentPipeline (http://www.ipm.ioz.ac.cn/kang/webpages/locusttranscriptome.html) for a given gene list was carried out based on the algorithm implemented in GOstat[31], with the whole annotated gene set as the background. GOstat tests for GO terms that are represented by significantly more genes in a given gene set using chi-square test. Fisher's exact test was used when expected counts are below 5, which makes the

chi-square test inaccurate.

Supplementary Note 4

4. Heterozygous SNV calling and population history estimation

Heterozygous single-nucleotide variation identification and distribution

To identify high-quality heterozygous single-nucleotide variations (SNVs) and estimate the level of heterozygosity in the sequenced genome, we mapped short-insert reads (500-800 bp) of BP against its own genome using SOAPALIGNER[32]. After obtaining the read alignment results, we employed SOAPsnp[33] to detect preliminary SNVs. The following candidate SNV set was filtered out: 1) a minimum base quality of twenty, a minimum of 8 (1/4-fold short-insert read depth) and a maximum of 124 (4-fold short-insert read depth) sequencing depth; 2) no other SNVs detected in a 5-bp window on either side of the SNV; 3) each allele with at least three uniquely mapped reads. In total, we discovered 1,683,572 SNVs, indicating that the heterozygosity rate of the BP genome is 0.188%. In all, 11,847 non-synonymous SNVs (these variations code for amino acid substitutions) are located in the coding sequences (CDS). We then identified 6,101 genes impacted by the SNVs. Gene Ontology terms like ‘protein binding’, ‘binding’, ‘ion binding’, ‘ATP binding’, ‘cation binding’, ‘adenyl ribonucleotide binding’ and ‘metal ion binding’, are highly enriched (P-value<0.05) in this gene set. The results of GO enrichment are summarized in Supplementary Data 2.

Additionally, we performed similar SNV detection and filtering (a minimum of 16 and a maximum of 264 sequencing depth, window size of 5) for the PM genome. The number of SNVs is 820,179 and the heterozygosity rate of the PM genome is 0.117%. Similarly, various GO terms relating to binding factors are significantly enriched in the 5,666 heterozygous SNV-affected genes containing 9,512 non-synonymous SNVs (**Supplementary Data 2**).

In order to verify the accuracy of identified heterozygous SNVs, we randomly selected 37 heterozygous SNVs of BP and validated them by PCR amplification and Sanger sequencing (**Supplementary Figure 8 and Table 12**). Thirty-four of the 37 sequenced SNVs were confirmed by double sequence peaks with the heights of each peak being half of other normal sites, confirming these sites as heterozygous SNVs.

Estimation of population history

The distribution of time to the most recent common ancestor (TMRCA) between two alleles in an individual can be related to the history of population size fluctuation.

To estimate the demographic TMRCA history of BP and PM, we performed the Pairwise Sequentially Markovian Coalescent (PSMC) model[34] on heterozygous sites of BP and PM genomes with the generation time (g =1 year)[35] and the mutation rate ($\mu = 3.51*10^{-9}y^{-1}nt^{-1}$)[36]. Finally, we used gnuplot4.4 [37] to draw the reconstructed population history (**Figure 2b**).

Supplementary Note 5

5. Evolutionary analysis of mudskipper genomes

Construction of gene families

Reference protein sequences of *H.sapiens* (human), *D.rerio (*zebrafish*), T.rubripes* (fugu), *X.tropicalis* (African frog), *G.aculeatus* (stickleback), *T.nigroviridis* (greenpuffer), *A.carolinensis* (lizard) and *O.latipes* (medaka) were downloaded from the Ensembl Core database (release 64). The consensus proteome set of the above eight species and our four mudskippers were filtered to remove those protein sequences less than 50 amino acids and resulted in a dataset of 239,304 protein sequences that was submitted to OrthoMCL[38] for protein clustering. A total of 21,149 OrthoMCL groups were built utilizing an effective database size of 239,304 sequences for all-to-all BLASTP strategy with an E-value=1E-5 and a Markov Chain Clustering (MCL) default inflation parameter.

For BP, considered as a representative mudskipper for comparison with four typical vertebrate genomes (human, lizard, frog and zebrafish), we identified 2,215 BP-specific gene families containing 2,394 genes in which 95.8% have expression support (RPKM > 0). We also detected 1,358 BP-specific gene families comprising 1,493 genes in comparison to the four reference teleost species (zebrafish, medaka, stickleback and fugu). GO enrichment analysis (**Supplementary Data 3**) of BP-specific genes was conducted using the Enrich Pipeline as described in the above section.

Phylogenetic tree construction

We extracted 1,913 single-copy (only one gene from each species) families from 12 vertebrate species. Multiple alignments were performed on proteins of each selected family by MUSCLE (version 3.8.31)[39] and we converted protein alignments to their corresponding CDS alignments using an in-house perl script. All the translated CDS sequences were combined into one "supergene" for each species. Four-fold degenerate sites (4D) extracted from the supergenes were then joined into new 4D genes of every species to construct a phylogenetic tree using MrBayes Version 3.2[40] (GTR+gamma model).

In addition, to construct a more precise phylogenetic tree, we removed SH and PS and selected 3,445 one-to-one orthologous from 10 other species. This dataset was also used to generate a phylogenetic tree with the same method (**Supplementary Figure 11**).

Reinterpretation of mudskipper phylogeny

In Figure 2, we observed that the phylogenetic relationship of the four mudskipper species contradicted the morphology-based cladistic tree from a previous study[1] (**Supplementary Figure 12**). To confirm the phylogenetic relationship among the four examined mudskippers, we sought to provide more robust evidence. We focused on the four mudskipper species and used *D. rerio* (zebrafish) as outgroup species, to identify orthologous groups by OrthoMCL. We then concatenated multiple

alignments of the coding nucleotide sequences and the protein sequences (4,306 single copy orthologues) to generate "super coding sequences (CDS) gene" and "super protein sequences (PEP) gene" for each species. Four-fold-degenerate sites and phase1 sites (non-degenerate sites) were extracted separately from the "super CDS gene" for building trees along with protein and CDS sequences. Subsequently, PhyML[41,42] was used to generate a maximum likelihood tree under the HKY85+gamma model for CDS, 4d and phase1 sequences and JTT+gamma model for protein sequences. The NJ[43] method was used for the same set of sequences to build trees under dm and mm model for nucleotide and protein sequences respectively. MrBayes[44] was used to generate Bayesian trees using GTR+gamma and Poisson+gamma model for nucleotide and protein sequences, respectively. All our 10 phylogenetic trees (**Supplementary Figure 13**) show maximal bootstrap support for every branch. We therefore propose that Murdy's morphology-based cladistic revision[7] of these four mudskipper species should be replaced with our new phylogeny presented here.

Divergence time estimation

To estimate the divergence time between mudskippers and other teleosts, as well as among the four mudskipper species, MCMCTree from the PAML package[45] was performed on 4D genes of each species and phylogenetic tree (mentioned in 5.2 Phylogenetic tree construction) using the molecular clock model. We set several reference divergence times (marked by red dots in several branches) from TimeTree database[46] (http://www.timetree.org/) to calibrate the divergence times of other nodes. Furthermore, we discarded SH & PS and calculated the divergence times in 10 species using the same approach.

We obtained a similar divergence time between other teleosts and mudskippers (approximately 140 million years) from the above two divergence trees (**Supplementary Figure 11**). BP and PM diverged about 30 million years ago. In the divergence time trees, we discovered medaka is most closely related to the mudskippers.

Café performance in expansion and contraction of gene families

The Café program[47] is meant for the statistical analysis of the evolution and estimation of the ancestor gene number in gene families on the basis of a birth and death model. We used this to identify gene families that had undergone expansion and contraction based on the topology of the divergence time tree. Orthologous gene families of 10 species (after removing SH and PS) were used for the Café analysis and the results are shown in Supplementary Figure 11.

Detection of positively selected genes

We extracted a total of 4,844 one-to-one orthologous gene families from seven teleosts (fugu, greenpuffer, stickleback, medaka, zebrafish, BP and PM) to identify positively selected genes (PSGs). We generated multiple-protein alignments using MUSCLE version 3.8.31[39] and trimAL version 1.4[48] to remove gaps. These high quality alignments were used to estimate three types of ω (the ratio of the rate of non-synonymous substitutions to the rate of synonymous substitutions) using two models of PAML[45] underlying species tree of these seven teleosts. In detail, branch model[49] (model=2, NSsites=0) was used to detect ω of appointed branch to test (ω0) and average ω of all other branches (ω1), and basic model (model=0, NSsites=0) was used to estimate average of whole branches (ω2), and Orthologs with dS (the rate of synonymous substitution) > 3 or ω0 > 5 were removed[50]. Then Chi-square test was used to check whether ω2 was significantly higher than ω1 and ω0 with threshold p-value < 0.05, which hinted that these genes may be under positive selection or fast evolution. In total, we obtained 722, 705, 64, 724, 1119, 1198 and 1431 positively selected genes from BP, PM, zebrafish, medaka, stickleback, fugu and greenpuffer, respectively.

We also calculated the average dN/dS values for each teleost lineage, and found that the fugu lineage has the highest value (0.184) while the zebrafish lineage has the lowest (0.05). The average values for dN/dS in the BP and PM lineages (0.166 and 0.181) are much higher than that in medaka (0.12; the most closely-related species to

mudskippers), indicating a more rapid evolution or more severe selective pressure in the mudskipper lineage than in medaka lineage. From previous studies[51], human and mouse yielded the highest and lowest dN/dS values (0.163 and 0.116) among mammals. Therefore, to our best knowledge, zebrafish may possess the lowest determined average dN/dS value in the teleost lineage. The dN/dS distributions in lineages of the seven teleosts are summarized in Supplementary Figure 14.

To test whether these PSGs in BP and PM lineages are impacted by their special living environments, we analyzed functional preference of gene ontology using the GO enrichment pipeline (**Supplementary Data 4**). We found many genes enriched in ‘cellular nitrogen compound metabolic process (GO:0034641)’, ‘nitrogen compound metabolic process (GO:0006807)’, ‘response to stress (GO:0006950)’, ‘DNA repair (GO:0006281)’, and ‘metabolic process (GO:0008152)’, which may be closely correlated with mudskippers' adaptation on land.

Identification of specific gained genes in mudskipper lineages

We use MCscan[52] to indentify the orthology information and synteny blocks from BP versus other five teleosts (zebrafish, medaka, stickleback, fugu and greenpuffer), as well as PM versus other five teleosts. In the protein synteny blocks, if one BP and PM protein had no ortholog in five teleosts, but they are reciprocal ortholog in BP and PM, and excluding false positive predictions that could be caused by annotation or genome assembly (gap > 5%), this protein could be defined as specific gained genes in the mudskipper lineage. We found 684 specific gained genes in mudskipper lineage and 657 genes have evidence of transcription.

Supplementary Note 6

6. Land adaptation analyses

Ammonia excretion in the gill

Nine core protein sequences in the gill ammonia excretory pathway, including Na^+-Cl^--K^+ cotransporter (NKCC, ENSDARG00000055313 downloaded from

NCBI)[53] , Na^+K^+-ATPase (NKA, AF286374)[54], carbonic anhydrases (CA, ENSDARG00000015654)[55], cystic fibrosis transmembrane conductance regulator (CFTR, NP_001038348.1)[56], Na^+/H^+ exchanger 3 (NHE, EF591984), H^+-ATPase-V-type-B-subunit (H-ATPase, BC055130)[57], glycosylated Rhesus glycoprotein b (Rhbg, AB218980) and c (Rhcg1, AB218981; Rhcg2, AB218982)[58]. We utilized these key proteins to align against BP, PM and five typical teleost genomes (medaka, zebrafish, fugu, greenpuffer, stickleback; downloaded from Ensembl release 64) by TBLASTN (E-value < 1E-5). Identified sequences were extended 5 kb on both ends of the alignment. GeneWise2.2.0 was used to predict the gene structures and open reading frames (ORFs) in them. A sequence was discarded if there was at least one premature stop codon or frame shift. We then extracted Reciprocal Best Hits (RBHs) between core proteins and new identified protein-code sequences from each genome.

PRANK[59] (http://www.ebi.ac.uk/goldman-srv/prank/) was used to align the RBH counterparts of each core protein. Before calculating the ω values, the coding nucleotide sequences were aligned based on protein alignments using in-house perl scripts and alignment gaps were removed using Gblocks[60]. After obtaining high-quality alignments, we estimated average ω across all the branches (ω2), ω of appointed branch to test (ω0), average ω of all other branches (ω1) and used the chi-square test to check whether ω2 is significantly higher than ω1 and ω0 with threshold p-value < 0.05 (**Supplementary Table 14**). Using this method, we observed that CA15 & NHE3 of BP, CA15 & Rhcg1 of PM and Rhcg1 of PS were under positive selection (**Figure 3a**).

Vision alteration

Evolutionary analysis of Arylalkylamine *N*-acetyltransferase (AANAT)

We downloaded AANAT2 (ENSTRUP00000029446), AANAT1a (ENSTRUP00000044871) and AANAT1b (ENSTRUP00000024942) protein sequences of fugu from Ensembl (release 64). These were used for homology searches against the BP and PM genomes using TBLASTN (E-value<1E-5).We

chose alignments with coverage >50% and identity >50% and extended 5 kb on both ends of each alignment. GeneWise2.2.0 was employed to predict the gene structures and open read frames (ORFs) in them. A sequence was discarded if there was at least one premature stop codon or frame shift. We identified three AANATs (bpAANAT1a: Bp_GLEAN_10012923, bpAANAT1b: Bp_GLEAN_10001218, bpAANAT2: Bp_GLEAN_10018012) in BP and 2 AANATs (pmAANAT1b: Pm_GLEAN_10010376, pmAANAT2: Pm_GLEAN_10016636) in PM.

To confirm the loss of AANAT1a from the PM genome, we utilized short-insert reads (from the 250, 500 and 800-bp libraries) of PM to map bpAANAT1a and pmAANAT1b using SOAPALIGNER software. The results showed that there were no reads mapped on the bpAANAT1a gene (**Supplementary Figure 17**).

We also constructed gene trees of all types of AANAT from both teleosts and tetrapods using the NJ method as implemented in MEGA5[1]. All teleost AANAT1as, AANAT1bs and AANAT2s form one clade whereas AANATs of tetrapods form a separate clade (**Supplementary Figure 18**).

Alteration of opsin genes in the mudskipper lineages

We extracted zebrafish opsin genes from NCBI to serve as query sequences (RH2-1, NP_571328.2; RH2-2, NP_878311.1; RH2-3,NP_878312.1; RH2-4, NP_571329.1; LWS-1, NP_571250.1; LWS-2, NP_001002443.1; SWS1, AAH60894.1; SWS2,NP_571267.1). Using the method described in Supplementary Note 6.2.1, we identified four types of opsins (LWS, SWS2, RH1 and RH2) in BP and PM. Like pufferfishes, mudskippers have lost their *SWS1* opsin gene. Since several conserved genes (*slc6a13, synpr, hcfc1a* and *gnl3l*) located at the upstream and downstream of the opsin genes, we identified them in BP and PM genomes. The distribution and synteny of all opsin genes were shown in Supplementary Figure 19.

As the spectral sensitivities of the visual pigments encoded by the opsin genes are related to specific residues (shown to be crucial for spectral tuning), we compared the protein sequences of LWS, SWS2, RH1 and RH2 in BP and PM to those in five other teleosts (downloaded from NCBI). We then constructed gene trees of LWS

(**Supplementary Figure 20**) from several fish (zebrafish, stickleback, fugu, greenpuffer, BP, PM, medaka) and human (as outgroup species) using the NJ method as implemented in MEGA5. In addition, for LWS pigment, the spectral sensitivities are determined mainly by the five tuning sites (180, 197, 277, 285 and 308). Several studies[61-63] have speculated that the ancestral LWS vertebrate opsin contains "SHYTA" in these five sites and suggested that a single mutation at S180A, H197Y, Y277F, T285A or A308S can lead to a -7, -28, -8, -15 or -27 shift, respectively in the absorption spectrum of the LWS opsin. So we followed the "five-sites" rule to predict the absorption spectrum of different species (**Table 2 and Supplementary Figure 21**).

Olfactory and vomeronasal receptor genes in mudskipper genomes

Olfactory receptor-like (OR-like) and vomeronasal receptor (V1R and V2R) genes were searched in the proteomes of the BP and PM by BLASTP using representative OR proteins as queries. Mudskipper proteins showing a match were re-checked by BLAST searches against the NCBI non-redundant database. Only those proteins that gave an 'olfactory receptor' or 'vomeronasal receptor' hit were retained. An OR-dataset was prepared using putative mudskipper OR proteins and those from human, zebrafish, amphioxus and elephant shark obtained from a previous study[2]. Six non-OR G protein-coupled receptors (GPCRs) were also included to serve as outgroups. Similarly, a V1R/V2R-dataset was prepared using proteins from mudskipper and those from zebrafish, elephant shark and sea lamprey[4]. The same six non-OR GPCRs were included as outgroups. We used Clustal-Omega version 1.2.0[64] at default settings to generate multiple alignments for the two datasets. Alignment gaps were removed using Gblocks and neighbor-joining (NJ) trees were generated for the two alignments. The NJ trees were viewed and edited using MEGA5.

Identification of tissue-specific regulated genes

Refer to previous methods[65-67], we identified the tissue-specific regulated genes with this strategy: 1) up-regulated gene identification by choosing the common

genes which expressed high level (rpkm > 1) in 3-h and 6-h air exposure but lower in control and the RPKM ratio of 6-h group vs control and 3-h group vs control must be over two folds. 2) down-regulated gene identification by selecting the common genes which expressed high level (rpkm > 1) in control but lower in 3-h and 6-h groups and the RPKM ratio of control vs 3-h group and control vs 6-h group must be over two folds. We obtained tissue-specific regulated gene numbers (**Supplementary Table 18**), which was similar as that from a medaka hypoxia paper[65]. Venn diagrams of common and unique genes in each tissue were shown in Supplementary Figures 24-25. The common tissue specific enriched regulated pathways of KEGG in BP and PM were shown in Supplementary Tables 19-20.

Supplementary References

1. Tamura, K. *et al*. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* **28**, 2731-2739 (2011).
2. Niimura, Y. On the origin and evolution of vertebrate olfactory receptor genes: comparative genome analysis among 23 chordate species. *Genome Biol. Evol.* **1**, 34-44 (2009).
3. Alioto, T.S. & Ngai, J. The repertoire of olfactory C family G protein-coupled receptors in zebrafish: candidate chemosensory receptors for amino acids. *BMC Genomics* **7**, 309 (2006).
4. Grus, W.E. & Zhang, J. Origin of the genetic components of the vomeronasal system in the common ancestor of all extant vertebrates. *Mol. Biol. Evol*. **26**, 407-419 (2009).
5. Chew, S.F., Hong, L.N., Wilson, J.M., Randall, D.J. & Ip, Y.K. Alkaline environmental pH has no effect on ammonia excretion in the mudskipper *Periophthalmodon schlosseri* but inhibits ammonia excretion in the related species *Boleophthalmus boddaerti*. *Physiol. Biochem. Zool.* **76**, 204-214 (2003).
6. Nikaido, M. *et al*. Coelacanth genomes reveal signatures for evolutionary transition from water to land. *Genome Res.* **23**, 1740-1748 (2013).
7. Murdy, E.O. *A taxonomic revision and cladistic analysis of the oxudercine gobies (Gobiidae: Oxudercinae)*, **Suppl. 11**, 1 (Australian Museum Scientific Publications, Sydeny, 1989).
8. Luo, R. *et al*. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1**, 18 (2012).
9. Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578-579 (2011).
10. Mitchell, K.A., Markham, K.R. & Bayly, M.J. Flavonoid characters contributing to the taxonomic revision of the Hebe parviflora complex. *Phytochemistry* **56**, 453-461 (2001).
11. Kurtz, S. *et al*. Versatile and open software for comparing large genomes. *Genome Biol.* **5**, 12 (2004).
12. Gnerre, S. *et al*. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci. USA* **108**, 1513-1518 (2011).
13. Kent, W.J. BLAT--the BLAST-like alignment tool. *Genome Res.* **12**, 656-664 (2002).
14. Parra, G., Bradnam, K., Ning, Z., Keane, T. & Korf, I. Assessing the gene space in draft genomes. *Nucleic Acids Res.* **37**, 289-297 (2009).
15. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, 265-268 (2007).
16. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* **Chapter 4**, Unit 4.10 (2009).
17. Jurka, J. et al. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462-467 (2005).
18. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573-580 (1999).

19. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14**, 988-995 (2004).
20. Stanke, M. *et al*. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, 435-439 (2006).
21. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol*. **268**, 78-94 (1997).
22. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111 (2009).
23. 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).
24. Elsik, C.G. *et al*. Creating a honey bee consensus gene set. *Genome Biol.* **8**, 13 (2007).
25. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28**, 45-48 (2000).
26. Zdobnov, E.M. & Apweiler, R. InterProScan--an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847-848 (2001).
27. Ashburner, M. *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25-29 (2000).
28. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27-30 (2000).
29. Trapnell, C. *et al*. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.* **31**, 46-53 (2013).
30. Chen, S. *et al*. De novo analysis of transcriptome dynamics in the migratory locust during the development of phase traits. *PLoS One* **5**, e15633 (2010).
31. Beissbarth, T. & Speed, T.P. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics* **20**, 1464-1465 (2004).
32. Li, R., Li, Y., Kristiansen, K. & Wang, J. SOAP: short oligonucleotide alignment program. *Bioinformatics* **24**, 713-714 (2008).
33. Li, R. *et al*. SNP detection for massively parallel whole-genome resequencing. *Genome Res.* **19**, 1124-1132 (2009).
34. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493-496 (2011).
35. Cai, Z. Population structure and reproductive characteristics of mudskipper *Boleophthalmus Pectinirostris,* in ShenZhen bay, China. *ACTA Ecologica Sinica* **16**, 77-82 (1996).
36. Graur, D. & Li, W.-H. *Fundamentals of molecular evolution*, Vol. 2, ***481*** (Sinauer Associates, Sunderland, 2000).
37. Janert, P.K. *Gnuplot in action: understanding data with graphs*, (Manning Publications, Connecticut, ed. 1, 2009).
38. Li, L., Stoeckert, C.J., Jr. & Roos, D.S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178-2189 (2003).
39. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792-1797 (2004).

40. Ronquist, F. et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61**, 539-542 (2012).
41. Guindon, S. et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307-321 (2010).
42. Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52**, 696-704 (2003).
43. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406-425 (1987).
44. Huelsenbeck, J.P. & Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754-745 (2001).
45. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**, 555-556 (1997).
46. Hedges, S.B., Dudley, J. & Kumar, S. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* **22**, 2971-2972 (2006).
47. De Bie, T., Cristianini, N., Demuth, J.P. & Hahn, M.W. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* **22**, 1269-1271 (2006).
48. Capella-Gutierrez, S., Silla-Martinez, J.M. & Gabaldon, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972-1973 (2009).
49. Zhao, H., Yang, J.R., Xu, H. & Zhang, J. Pseudogenization of the umami taste receptor gene Tas1r1 in the giant panda coincided with its dietary switch to bamboo. *Mol. Biol. Evol.* **27**, 2669-2673 (2010).
50. Castillo-Davis, C.I., Kondrashov, F.A., Hartl, D.L. & Kulathinal, R.J. The functional genomic distribution of protein divergence in two animal phyla: coevolution, genomic conflict, and constraint. *Genome Res.* **14**, 802-811 (2004).
51. Groenen, M.A. *et al*. Analyses of pig genomes provide insight into porcine demography and evolution. *Nature* **491**, 393-398 (2012).
52. Tang, H. *et al*. Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res.* **18**, 1944-1954 (2008).
53. Kato, A. *et al*. Differential expression of Na^+-Cl^- cotransporter and Na^+-K^+-Cl^- cotransporter 2 in the distal nephrons of euryhaline and seawater pufferfishes. *Am. J. Physiol. Regul. Integr. Comp. Physiol*. **300**, 284-297 (2011).
54. Liao, B.K., Chen, R.D. & Hwang, P.P. Expression regulation of Na^+-K^+-ATPase alpha1-subunit subtypes in zebrafish gill ionocytes. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* **296**, 1897-1906 (2009).
55. Lin, T.Y. *et al*. Carbonic anhydrase 2-like a and 15a are involved in acid-base regulation and Na^+ uptake in zebrafish H^+-ATPase-rich cells. *Am. J. Physiol. Cell Physiol.* **294**, 1250-1260 (2008).
56. Ip, Y.K. *et al*. Cystic fibrosis transmembrane conductance regulator in the gills of the climbing perch, *Anabas testudineus*, is involved in both hypoosmotic regulation during seawater acclimation and active ammonia excretion during ammonia exposure. *J. Comp. Physiol. B* **182**, 793-812 (2012).
57. Seo, M., Mekuchi, M., Teranishi, K. & Kaneko, T. Expression of ion transporters in gill mitochondrion-rich cells in Japanese eel acclimated to a wide range of

environmental salinity. *Comp. Biochem. Physiol. A Mol. Integr. Physiol.* **162**, 323-332 (2013).

58. Nawata, C.M., Hirose, S., Nakada, T., Wood, C.M. & Kato, A. Rh glycoprotein expression is modulated in pufferfish (*Takifugu rubripes*) during high environmental ammonia exposure. *J. Exp. Biol.* **213**, 3150-3160 (2010).
59. Loytynoja, A. & Goldman, N. An algorithm for progressive multiple alignment of sequences with insertions. *Proc. Natl. Acad. Sci. USA* **102**, 10557-10562 (2005).
60. Talavera, G. & Castresana, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst..Biol.* **56**, 564-577 (2007).
61. Yokoyama, S. & Radlwimmer, F.B. The molecular genetics and evolution of red and green color vision in vertebrates. *Genetics* **158**, 1697-1710 (2001).
62. Davies, W.L. *et al*. Functional characterization, tuning, and regulation of visual pigment gene expression in an anadromous lamprey. *FASEB J.* **21**, 2713-2724 (2007).
63. Ward, M.N. *et al*. The molecular basis of color vision in colorful fish: four long wave-sensitive (LWS) opsins in guppies (*Poecilia reticulata*) are defined by amino acid substitutions at key functional sites. *BMC Evol. Biol.* **8**, 210 (2008).
64. Sievers, F. *et al*. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).
65. Ju, Z., Wells, M.C., Heater, S.J. & Walter, R.B. Multiple tissue gene expression analyses in Japanese medaka (*Oryzias latipes*) exposed to hypoxia. *Comp. Biochem. Physiol. C Toxicol. Pharmacol.* **145**, 134-144 (2007).
66. Ton, C.D. Stamatiou, C.-C. Liew, Gene expression profile of zebrafish exposed to hypoxia during development. *Physiol. Genomics* **13**, 97-106 (2003).
67. van der Meer D.L. *et al.,* Gene expression profiling of the long-term adaptive response to hypoxia in the gills of adult zebrafish. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* **289**, 1512-1519 (2005).