

# The oyster genome reveals stress adaptation and complexity of shell formation

Guofan Zhang<sup>1\*</sup>, Xiaodong Fang<sup>2\*</sup>, Ximing Guo<sup>3\*</sup>, Li Li<sup>1\*</sup>, Ruibang Luo<sup>2,4\*</sup>, Fei Xu<sup>1\*</sup>, Pengcheng Yang<sup>2\*</sup>, Linlin Zhang<sup>1\*</sup>, Xiaotong Wang<sup>1\*</sup>, Haigang Qi<sup>1</sup>, Zhiqiang Xiong<sup>2</sup>, Huayong Que<sup>1</sup>, Yinlong Xie<sup>2,4</sup>, Peter W. H. Holland<sup>5</sup>, Jordi Paps<sup>5</sup>, Yabing Zhu<sup>2</sup>, Fucun Wu<sup>1</sup>, Yuanxin Chen<sup>2</sup>, Jiafeng Wang<sup>1</sup>, Chunfang Peng<sup>2</sup>, Jie Meng<sup>1</sup>, Lan Yang<sup>2</sup>, Jun Liu<sup>1</sup>, Bo Wen<sup>2</sup>, Na Zhang<sup>1</sup>, Zhiyong Huang<sup>2</sup>, Qihui Zhu<sup>1</sup>, Yue Feng<sup>2</sup>, Andrew Mount<sup>6</sup>, Dennis Hedgecock<sup>7</sup>, Zhe Xu<sup>8</sup>, Yunjie Liu<sup>2</sup>, Tomislav Domazet-Lošo<sup>9</sup>, Yishuai Du<sup>1</sup>, Xiaoqing Sun<sup>2</sup>, Shoudu Zhang<sup>1</sup>, Binghang Liu<sup>2,4</sup>, Peizhou Cheng<sup>1</sup>, Xuanting Jiang<sup>2</sup>, Juan Li<sup>1</sup>, Dingding Fan<sup>2</sup>, Wei Wang<sup>1</sup>, Wenjing Fu<sup>2</sup>, Tong Wang<sup>1</sup>, Bo Wang<sup>2</sup>, Jibiao Zhang<sup>1</sup>, Zhiyu Peng<sup>2</sup>, Yingxiang Li<sup>1</sup>, Na Li<sup>2</sup>, Jinpeng Wang<sup>1</sup>, Maoshan Chen<sup>2</sup>, Yan He<sup>3</sup>, Fengji Tan<sup>2</sup>, Xiaorui Song<sup>1</sup>, Qiumei Zheng<sup>2</sup>, Ronglian Huang<sup>1</sup>, Hailong Yang<sup>2</sup>, Xuedi Du<sup>1</sup>, Li Chen<sup>2</sup>, Mei Yang<sup>1</sup>, Patrick M. Gaffney<sup>10</sup>, Shan Wang<sup>3</sup>, Longhai Luo<sup>2</sup>, Zhicai She<sup>1</sup>, Yao Ming<sup>2</sup>, Wen Huang<sup>1</sup>, Shu Zhang<sup>2</sup>, Baoyu Huang<sup>1</sup>, Yong Zhang<sup>2</sup>, Tao Qu<sup>1</sup>, Peixiang Ni<sup>2</sup>, Guoying Miao<sup>1</sup>, Junyi Wang<sup>2</sup>, Qiang Wang<sup>1</sup>, Christian E. W. Steinberg<sup>11</sup>, Haiyan Wang<sup>1</sup>, Ning Li<sup>2</sup>, Lumin Qian<sup>3</sup>, Guojie Zhang<sup>2</sup>, Yingrui Li<sup>2</sup>, Huanming Yang<sup>2</sup>, Xiao Liu<sup>1</sup>, Jian Wang<sup>2</sup>, Ye Yin<sup>2</sup> & Jun Wang<sup>2,12,13</sup>

**The Pacific oyster *Crassostrea gigas* belongs to one of the most species-rich but genomically poorly explored phyla, the Mollusca. Here we report the sequencing and assembly of the oyster genome using short reads and a fosmid-pooling strategy, along with transcriptomes of development and stress response and the proteome of the shell. The oyster genome is highly polymorphic and rich in repetitive sequences, with some transposable elements still actively shaping variation. Transcriptome studies reveal an extensive set of genes responding to environmental stress. The expansion of genes coding for heat shock protein 70 and inhibitors of apoptosis is probably central to the oyster's adaptation to sessile life in the highly stressful intertidal zone. Our analyses also show that shell formation in molluscs is more complex than currently understood and involves extensive participation of cells and their exosomes. The oyster genome sequence fills a void in our understanding of the Lophotrochozoa.**

Oceans cover approximately 71% of the Earth's surface and harbour most of the phylum diversity of the animal kingdom. Understanding marine biodiversity and its evolution remains a major challenge. The Pacific oyster *C. gigas* (Thunberg, 1793) is a marine bivalve belonging to the phylum Mollusca, which contains the largest number of described marine animal species<sup>1</sup>. Molluscs have vital roles in the functioning of marine, freshwater and terrestrial ecosystems, and have had major effects on humans, primarily as food sources but also as sources of dyes, decorative pearls and shells, vectors of parasites, and biofouling or destructive agents. Many molluscs are important fishery and aquaculture species, as well as models for studying neurobiology, biomineralization, ocean acidification and adaptation to coastal environments under climate change<sup>2,3</sup>. As the most speciose member of the Lophotrochozoa, phylum Mollusca is central to our understanding of the biology and evolution of this superphylum of protostomes.

As sessile marine animals living in estuarine and intertidal regions, oysters must cope with harsh and dynamically changing environments. Abiotic factors such as temperature and salinity fluctuate wildly, and toxic metals and desiccation also pose serious challenges. Filter-feeding oysters face tremendous exposure to microbial pathogens. Oysters do have a notable physical line of defence against predation and desic-

cation in the formation of thick calcified shells, a key evolutionary innovation making molluscs a successful group. However, acidification of the world's oceans by uptake of anthropogenic carbon dioxide poses a potentially serious threat to this ancient adaptation<sup>4</sup>. Understanding biomineralization and molluscan shell formation is, thus, a major area of interest<sup>5</sup>. *Crassostrea gigas* is also an interesting model for developmental biology owing to its mosaic development with typical molluscan stages, including trochophore and veliger larvae and metamorphosis.

A complete genome sequence of *C. gigas* would enable a more thorough understanding of oyster biology and the evolution of Lophotrochozoa. One of the main challenges, however, is the high levels of polymorphism present in oysters and many marine invertebrates<sup>6–8</sup>. To overcome this, an oyster derived from four generations of full-sibling mating (coefficient of inbreeding,  $F = 0.59$ ) was used for genome sequencing and assembly (Supplementary Text B1) through fosmid pooling, next-generation sequencing (NGS) and hierarchical assembling. Combining these genomic data with transcriptomes from different organs, different developmental stages and adults challenged with stressors, in addition to mass spectrometric analysis of shell proteins, allowed us to explore characteristics of the oyster genome

<sup>1</sup>Institute of Oceanology, Chinese Academy of Sciences, Qingdao 266071, China. <sup>2</sup>BGI-Shenzhen, Shenzhen 518083, China. <sup>3</sup>Haskin Shellfish Research Laboratory, Institute of Marine and Coastal Sciences, Rutgers University, Port Norris, New Jersey 08349, USA. <sup>4</sup>HKU-BGI Bioinformatics Algorithms and Core Technology Research Laboratory, Hong Kong. <sup>5</sup>Department of Zoology, University of Oxford, Oxford OX1 3PS, UK. <sup>6</sup>Department of Biological Sciences, Clemson University, South Carolina 29634, USA. <sup>7</sup>Department of Biological Sciences, University of Southern California, Los Angeles, California 90089, USA. <sup>8</sup>Atlantic Cape Community College, Mays Landing, New Jersey 08330, USA. <sup>9</sup>Laboratory of Evolutionary Genetics, Ruđer Bošković Institute, Bijenička cesta 54, P.P. 180, HR-10002, Zagreb, Croatia. <sup>10</sup>School of Marine Science and Policy, University of Delaware, Lewes, Delaware 19958, USA. <sup>11</sup>Institute of Biology, Humboldt Universität zu Berlin Arboretum, Späthstraße 80/81, 12437 Berlin, Germany. <sup>12</sup>Department of Biology, University of Copenhagen, DK-2200 Copenhagen, Denmark. <sup>13</sup>The Novo Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen, DK-2200 Copenhagen, Denmark.

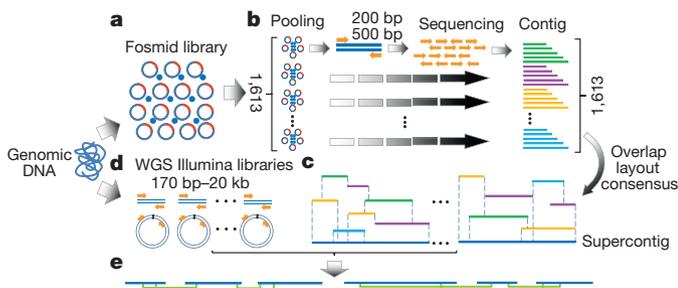
\*These authors contributed equally to this work.

and key aspects of molluscan biology related to stress response and shell formation.

## Sequencing and hierarchical assembly

NGS technology has been successfully applied for *de novo* genome sequencing and assembly using whole-genome shotgun strategies<sup>9–13</sup>. We initially generated 155-fold Illumina whole-genome shotgun reads (Supplementary Table 1), but could not adequately assemble them owing to high levels of polymorphism and abundant repetitive sequences (Supplementary Text B2 and Supplementary Fig. 1). As possible alternative sequencing strategies—such as the addition of longer Roche 454 reads<sup>12,13</sup> or traditional bacterial artificial chromosome (BAC)-to-BAC sequencing—are expensive, we opted instead for a more cost-effective fosmid-pooling strategy. In brief, a fosmid library was constructed, and 145,170 clones (~tenfold genome coverage) were evenly and randomly assigned into 1,613 pools, each of which was sequenced to ~60-fold depth and assembled separately (Fig. 1 and Supplementary Table 1). Contigs from each pool were merged into supercontigs, totalling 1,002 megabases (Mb) (Supplementary Text B4.1–3), which was larger than genome-size estimates of 637 Mb from flow cytometry or 545 Mb from *k*-mer (*k*-base fragment) analysis (Supplementary Text B1, 2,3), owing to failure of some allelic variants to merge (Supplementary Figs 3 and 4). Self-to-self whole-genome alignment with LASTZ<sup>14</sup> and sequencing depth information were used to remove redundancy in the assembly (Supplementary Text B4.4). The resulting 446 Mb of the assembly were retained for further scaffolding using paired-end data (Fig. 1). The final assembly comprised 559 Mb, with a contig N50 size (at which 50% of assembly was covered) of 19.4 kilobases (kb) and a scaffold N50 size of 401 kb (Supplementary Text B4.5 and Supplementary Table 3). Over 90% of the assembly was covered by the longest 1,670 (14%) scaffolds.

To assess the completeness of the assembly, 105-fold coverage of short-insert library reads (<2 kb) that participated in assembly (Supplementary Table 1) were aligned against the assembly. Over 99% of these reads were successfully mapped, using a combination



**Figure 1 | Fosmid-pooling strategy for oyster genome assembly.** Genomic DNA was randomly sheared into fragments. **a, b**, A 40-kb-insert fosmid library was constructed (**a**), and 145,170 fosmid clones were randomly selected and assigned into 1,613 pools, each containing 90 clones covering 0.57% of the diploid genome (**b**). For each pool, three Illumina short-insert barcoded libraries (two 200 bp and one 500 bp) were constructed and ~60-fold coverage of 90-bp reads (20-fold per library) were generated, and assembled using SOAPdenovo with optimizing parameters. Assemblies from each pool were further corrected and reassembled if unexpected connections were detected owing to high similarity sequences from different fosmids, and gaps were filled by local assembly. **c**, Fosmid scaffolds were split into contigs at unfilled regions, leaving no undetermined bases in the sequences. Each base was assigned a Phred-like quality score determined by its coverage and alignment mismatches, and these sequences were merged into supercontigs using the overlap layout consensus method. Redundancy was removed using self-to-self alignment and sequencing depth information. **d**, Whole-genome shotgun Illumina libraries (200-bp to 20-kb inserts) from sheared genomic DNA were constructed for mated-pair Illumina sequencing. **e**, The fosmid supercontigs were linked into scaffolds using (1) the whole-genome shotgun sequences; (2) inferred paired-end information extracted from assembled pool scaffolds with a span size ranging from 50 bp to 37.5 kb; and (3) 225,000 fosmid ends sequenced using Sanger technology.

of Burrows–Wheeler Aligners<sup>15</sup> and the more sensitive LASTZ (Supplementary Fig. 5 and Supplementary Table 4). The integrity of the assembly is further demonstrated by the successful mapping of 99% of the BAC sequenced obtained using the Sanger sequencing technique, and 98% of ~68,000 expressed sequence tags from 454 sequencing (Supplementary Text B5, Supplementary Fig. 6 and Supplementary Tables 5 and 6). Fosmid pooling has been used for re-sequencing<sup>16,17</sup>, and our results show that the combination of fosmid pooling, NGS and hierarchical assembly provides a new, cost-effective alternative for *de novo* sequencing and assembly of complex genomes.

## Polymorphism and repetitive sequences

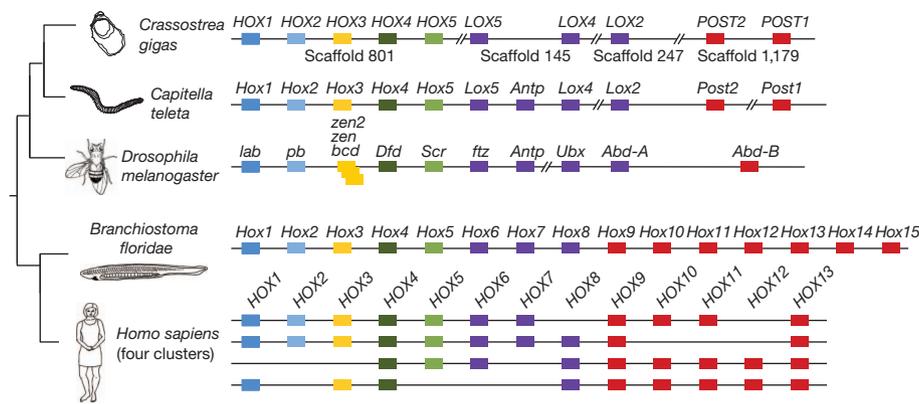
To understand polymorphism in the oyster genome, we analysed allelic variation in the assembled genome (inbred) and one re-sequenced wild oyster (wild) (Supplementary Text C1). The inbred genome contained 3.1 million single-nucleotide polymorphisms and 258,405 short insertion/deletion (indels, 1–40 base pairs (bp)) yielding a sequence polymorphism rate of 0.73%, whereas the wild genome had 3.8 million single-nucleotide polymorphisms and 238,182 indels, or a polymorphism rate of 1.3% (Supplementary Table 7), comparable to previous estimates<sup>18</sup>. This 44% reduction in polymorphism in the inbred genome is smaller than the 59.4% predicted from four generations of brother–sister mating, indicating that selection favouring heterozygotes had occurred<sup>19</sup>. The polymorphism combining inbred and wild (among four haplotypes) was 2.3%, higher than that in most studied animal genomes<sup>20,21</sup> but comparable to that in known high-polymorphism species<sup>7</sup>. In inbred and wild, we found 3,094 short indels located in coding regions inferred to cause frameshift variants in 2,665 genes, providing an important source for recessive lethal mutations.

*k*-mer-based analysis of the oyster genome showed that ~35% of 17-mers had at least two identical copies in the genome, suggesting an abundance of repetitive sequences (Supplementary Fig. 1). Similarly, homology searching and *ab initio* prediction found 202 Mb (36% of the genome) in repetitive sequences (Supplementary Text C2 and Supplementary Table 8). Over 62% of the detected repeats could not be assigned to known categories, reflecting the paucity of genomic information from molluscs<sup>22</sup>. Large numbers of transposase (359) and reverse-transcriptase (779) gene fragments were detected; over 96% of these had detectable transcripts (Supplementary Fig. 8). Alignment of the wild sequence against the assembly identified 20,605 deletions (>100 bp), over 80% of which overlapped with detected transposable elements, suggesting that transposable elements may have an active role in shaping genome variation. Using MITE-hunter<sup>23</sup>, we detected 157,007 copies of miniature inverted-repeat transposable elements (MITEs), accounting for a remarkable 8.82% of the genome (Supplementary Text C2.3 and Supplementary Table 9). Pair-wise comparisons show extremely low sequence divergence in some MITE families (Supplementary Fig. 9), indicating that they may still be active.

## Gene annotation and developmental genomics

A total of 28,027 genes were predicted encoding 50 amino acids or more by combining *de novo* prediction and evidence-based searches using reference genomes, oyster expressed sequence tags and transcriptomes from multiple organs and developmental stages (Supplementary Text D1 and E1 and Supplementary Fig. 11), with 96.1% showing expression (reads per kb per million mapped reads (RPKM) > 1 in at least one transcriptome; Supplementary Text D2). Of the inferred proteins, 21,085 matched entries in the SWISS-PROT, InterPro or TrEMBL databases. These genes plus their transcriptome profile from 7 adult organs and at 38 developmental stages provide valuable resources for comparative genomics analysis (Supplementary Text E2 and 3), functional inference and studies on development and organogenesis (Supplementary Text F2 and Supplementary Fig. 15).

One notable finding of developmental interest is that the oyster Hox gene cluster is broken into four sections (Fig. 2) with flanking non-Hox genes (Supplementary Fig. 16). We did not find a clear



**Figure 2 | Clustering of Hox genes in Pacific oyster *Crassostrea gigas*, polychaete annelid *Capitella teleta*, fruitfly *Drosophila melanogaster*, lancelet *Branchiostoma floridae* and *Homo sapiens*.** Oblique lines indicate

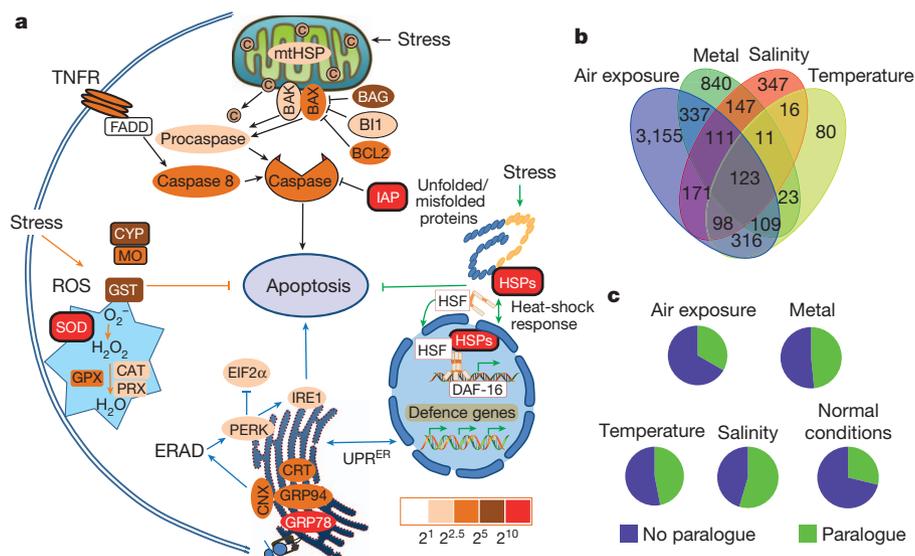
regions of Hox cluster that are non-contiguous or interrupted. Blue denotes anterior Hox genes, yellow denotes paralogy group 3 Hox genes, green and purple denote central Hox genes and red denotes posterior Hox genes.

Antennapedia gene, which is present in other bivalves such as *Pecten* and *Yoldia*<sup>24</sup> (Supplementary Fig. 17). Disruption of the Hox cluster, as also observed in tunicates, nematodes and drosophilids, has been attributed to the loss of temporal co-linearity and modified developmental control<sup>25</sup>. Supporting this model, we find that Hox genes in the oyster are not activated in an order matching their identity or genomic position, with, for example, *HOX4* and *HOX1* peaking before gastrulation, *LOX5* and *POST2* during the trochophore stage and *HOX5* during the pediveliger stage (Supplementary Fig. 18 and Supplementary Table 15).

### Adaptation to environmental stress

Comparison with seven other sequenced genomes identified 8,654 oyster-specific genes (Supplementary Text E3.1) that are probably important in the evolution and adaptation of oysters and other molluscs. With oysters being the only representative, these genes could be shared by other molluscs. Among these genes, gene ontology terms related to ‘protein binding’, ‘apoptosis’, ‘cytokine activity’

and ‘inflammatory response’ are highly enriched ( $P < 0.0001$ ; Supplementary Text E2 and Supplementary Table 17), indicating over-representation of some host-defence genes against biotic and abiotic stress. Manual examination shows that several gene families related to defence pathways, including protein folding, oxidation and anti-oxidation, apoptosis and immune responses, are expanded in *C. gigas* (Fig. 3a and Supplementary Table 18). The oyster genome contains 88 heat shock protein 70 (*HSP70*) genes, which have crucial roles in protecting cells against heat and other stresses, compared with ~17 in humans and 39 in sea urchins. Phylogenetic analysis finds clustering of 71 oyster *HSP70* genes to themselves, suggesting that the expansion is specific to the oyster (Supplementary Fig. 19). Also expanded are cytochrome P450 (Supplementary Fig. 20) and multi-copper oxidase gene families, which are important in the biotransformation of endobiotic and xenobiotic chemicals<sup>26</sup>, and extracellular superoxide dismutases, which are important in defence against oxidative stress. The oyster genome has 48 genes coding for inhibitor of apoptosis proteins (IAPs), compared with 8 in humans and 7 in sea urchins,



**Figure 3 | Expansion, expression and pathway distribution of defence-related genes in *Crassostrea gigas*.** **a**, Expansion and expression of key genes in major stress-response pathways in *C. gigas*. Genes include *HSPs* and *HSF* in the heat-shock response; *GRP78*, *CRT*, *CNX*, *GRP94*, *PERK*, *IRE1* and *EIF2 $\alpha$*  in the endoplasmic reticulum unfolded-protein response (UPR<sup>ER</sup>); *IAPs*, *BCL2* like, *BAG*, *BII*, caspases, *FADD* and *TNFR* in apoptotic pathways; *CYP450* and *MO* in oxidation; and *SOD*, *GPX*, *PRX* and *CAT* in anti-oxidation. Boxes with bold black borders indicate gene families (*HSPs*, *IAPs* and *SODs*) expanded in *C. gigas*, and the filled colours correspond to their degree of upregulation in  $RPKM_{treatment}/RPKM_{control}$  by stress, found in 61 transcriptomes from oysters

challenged with 9 types of stressors (Supplementary Text G2 and Supplementary Table 23). **b**, Venn diagram of common and unique genes expressed in response to temperature, salinity, air exposure and heavy-metal stress (zinc, cadmium, copper, lead and mercury), showing overlap of responses. **c**, Number of genes with and without detectable paralogues differentially expressed under stress and normal conditions, showing that genes responding to stress are more likely to have paralogues ( $P < 1 \times 10^{-10}$ ;  $\chi^2$  test). Green sections of the pie chart represent 1,442, 809, 358, 550 and 7,938 paralogues for air exposure, metal, temperature, salinity and normal conditions, respectively.

indicating a powerful anti-apoptosis system in oysters. Genes encoding lectin-like proteins, including C-type lectin, fibrinogen-related proteins and C1q domain-containing proteins (C1QDCs), are highly over-represented in the oyster genome ( $P < 0.0001$ ; Supplementary Table 18); these genes have important roles in the innate immune response in invertebrates<sup>27–29</sup>. Interestingly, many immune-related genes, including genes coding for Gram-negative bacteria-binding proteins, peptidoglycan-recognition proteins, defensin, C-type-lectin-domain-containing proteins and C1QDCs, are highly expressed in the digestive gland (Supplementary Fig. 21), indicating that the digestive system of this filter feeder is an important first-line defence organ against pathogens.

To investigate genome-wide responses to stress, we sequenced 61 transcriptomes from *C. gigas* subjected to nine stressors, including temperature, salinity, air exposure and heavy metals (Supplementary Text G1 and Supplementary Tables 19 and 20). We found that 5,844 genes were differentially expressed under at least one stressor, and genes responding to different stressors showed significant overlap (Fig. 3b and Supplementary Fig. 23a). Air exposure induced a response from the largest number of genes (4,420), indicating that air exposure is a major stressor and that oysters have evolved an extensive gene set in defence. Genes differentially expressed in response to stress are more likely to have paralogues (Fig. 3c), suggesting that expansion and selective retention of duplicated defence-related genes are probably important to oyster adaptation. Under most stressors, genes coding for HSPs, histones, IAPs and protein biogenesis were upregulated, and those for protein degradation downregulated, pointing to concerted responses to maintain cellular homeostasis<sup>30</sup> (Supplementary Text G3 and Supplementary Table 21). Genes involved in the unfolded protein response to cellular stress in the endoplasmic reticulum (coding for calreticulin, calnexin, 78- and 94-kDa glucose-regulated proteins) were upregulated, indicating that protein quality control is critical in cellular homeostasis under stress.

Air exposure induced up to 67-fold upregulation of five highly expressed IAPs (Supplementary Fig. 24a). Other inhibitors of apoptosis were also upregulated: *BCL2* up to fourfold and *BAG* up to 12-fold (Supplementary Fig. 24b). These apoptosis inhibitors were also highly upregulated under heat and low salinity stress. These findings, along with the expansion of IAPs, suggest that a powerful anti-apoptosis system exists and may be critical for the amazing endurance of oysters to air exposure and other stresses. The existence of an intrinsic apoptosis pathway in invertebrates has been controversial, and parts of the pathways have only recently been demonstrated for two lophotrochozoans<sup>31,32</sup>. The finding of key genes belonging to both intrinsic (*BAX*, *BAK*, *BAG*, *BCL2*, *BII* and procaspase) and extrinsic (*TNFR* and caspase 8) apoptosis pathways indicates that oysters have advanced apoptosis systems. Powerful inhibition of apoptosis as shown by genomic and transcriptomic analyses may be central to the ability of oysters to tolerate prolonged air exposure and other stresses.

Heat stress induced a ~2,000-fold increase in expression of five highly inducible *HSP70* genes or a 13.9-fold increase in average expression of all *HSP70* genes, amounting to 4.2% of all transcripts (Supplementary Figs 24c and 25). The genomic expansion and massive upregulation of HSP genes help to explain why *C. gigas* can tolerate temperatures as high as 49 °C when exposed to summer sun at low tide<sup>33</sup>. HSP genes were also upregulated under other stressors and may be central to the oyster defence against all stresses (Supplementary Fig. 25). HSP genes may also inhibit apoptosis by binding to effector caspases<sup>34</sup>.

Genes involved in signal transduction, including genes coding for G-protein-coupled receptors and Ras GTPase, were also activated by stressors (Supplementary Fig. 24f) and over-represented in the oyster genome (Supplementary Table 11). These regulators may have a role in orchestrating stress responses, which seem to be well coordinated (Fig. 3a and Supplementary Fig. 25). The expansion of key defence genes and the strong, complex transcriptomic response to

stress highlight the sophisticated genomic adaptations of the oyster to sessile life in a highly stressful environment.

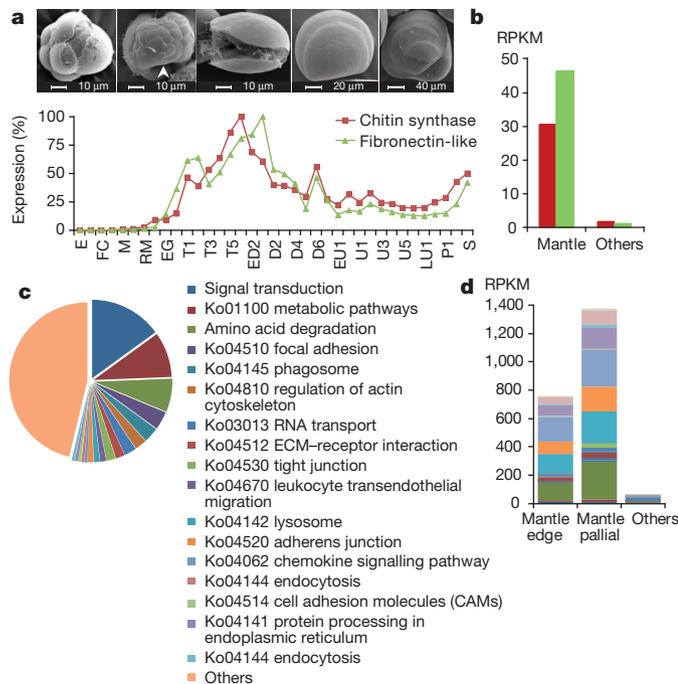
## Shell formation

Calcified shells provide critical protection against predation and desiccation in sessile marine animals such as oysters. Molluscan shells consist of calcium carbonate ( $\text{CaCO}_3$ ) crystals of either aragonite or calcite embedded in an elaborate organic matrix. Two models have been advanced for molluscan shell formation. The matrix model posits that mineralization occurs in a mantle-secreted matrix of chitin, silk fibroin and acidic proteins<sup>35,36</sup>. Chitin and silk proteins are proposed to provide matrix structure, whereas acidic proteins control the nucleation and growth of  $\text{CaCO}_3$  crystals. The cellular model suggests that biomineralization is cell-mediated; that is, crystals are formed in haemocytes and then deposited at the mineralization front<sup>37</sup>.

We searched the oyster genome for genes implicated in shell formation in previous studies and examined their expression in different tissues and at different stages (Supplementary Text H1, 2). We also sequenced peptides from shells, mapped them to the genome and identified 259 shell proteins (Supplementary Text H3 and Supplementary Table 24). Although our search found evidence for the involvement of chitin, we did not find any silk-like proteins encoded in the oyster genome (Supplementary Text H2) but found, instead, many diverse proteins that may have roles in matrix construction and modification. Notably, a gene coding for a fibronectin-like protein was highly expressed at the early developmental stage, when larval shells are formed, in unison with chitin synthase (Fig. 4a) and was mostly expressed in the adult mantle (40× other organ average; Fig. 4b); the fibronectin-like protein was among the most abundant proteins found in oyster shells. Genes coding for laminin and some collagen proteins were also highly expressed in the mantle (Supplementary Fig. 27a) and found in shells. These are typical extracellular matrix (ECM) proteins, and their presence in shells suggests that the shell matrix has similarities to the ECM of animal connective tissues and basal lamina. Unlike silk fibroins that can self assemble<sup>38</sup>, the formation of fibronectin fibrils in the ECM is cell mediated<sup>39</sup>. Oyster fibronectin-like proteins have five type-III domains for integrin binding and cell adhesion. Genes coding for integrins were highly expressed in haemocytes (4× other organ average, Supplementary Fig. 27b). Thus, haemocytes may organize fibronectin-like fibril formation in the shell matrix as they do in ECM.

The involvement of cells in shell formation is further supported by the functional diversity of proteins detected in shells. Many house-keeping proteins, such as elongation factor 1 $\alpha$  and ribosomal proteins, were found in the shell; indeed, most oyster shell proteins are not structural proteins but are distributed in diverse metabolic pathways (Fig. 4c and Supplementary Table 25). This functional diversity of shell proteins mirrors that of cells, which is unexpected under the matrix model. Furthermore, 84% of the 259 shell proteins identified are not classical secreted proteins (Supplementary Text H3.4 and Supplementary Table 24); they may be part of cells or deposited by exosomes<sup>40</sup>. Supporting the presence of exosomes, 61 of the 259 shell proteins matched proteins in the exosome database<sup>41</sup>. Cells and exosome-like vesicles containing calcite crystals have been observed at the mineralization front<sup>37,42</sup>, although their significance in shell formation is debated. This study provides molecular evidence for their presence inside shells and their probable participation in shell formation.

Many shell proteins are enzymes that may be involved in matrix construction or modification. A homologue of penicillin-binding protein is exclusively expressed in mantle (72× other organ average) and highly abundant in shells (Supplementary Fig. 27d). Penicillin-binding protein is a transpeptidase that crosslinks glycopeptides in bacterial cell walls<sup>43</sup> and may have similar functions in the shell matrix. Another notable enzyme found is tyrosinase. The oyster



**Figure 4 | Genes related to shell formation identified from mass spectroscopy analysis of shell proteins and transcriptome data.** **a**, Relative expression (*y* axis) of genes coding for chitin synthase (gene CGI\_10009438) and fibronectin-like (CGI\_10016964) in early development corresponds to the formation of shell gland and first larval shells, as seen in scanning electron microscope photos. White arrow denotes the invagination that forms the shell gland. Developmental stages (*x* axis) and their timeline are defined in Supplementary Table 12. **b**, In adults, chitin synthase and fibronectin-like proteins (same colour as in **a**) are almost exclusively expressed in the mantle compared with other organs. Fibronectin-like is also one of the most abundant proteins found in the shell. **c**, Distribution of shell proteins in diverse Kyoto encyclopedia of genes and genomes (KEGG) pathways indicative of general cellular functions. **d**, Expression of 26 tyrosinase genes in the mantle edge, mantle pallial and other organs. Tyrosinases are abundant in shells and their higher expression in the non-pigmented mantle pallial indicate that their functions are not limited to melanogenesis but are related to shell formation.

genome has an expanded set of 26 genes coding for tyrosinase, compared with one in *Caenorhabditis elegans* and two in humans; most genes coding for tyrosinase are mantle specific (Fig. 4d) and highly enriched among shell proteins ( $P = 8 \times 10^{-6}$ ). Although tyrosinase is a key enzyme in melanogenesis<sup>44,45</sup>, it is most highly expressed in the non-pigmented pallial mantle (Fig. 4d), indicating that it has other functions in the oyster. The mantle secretes tyrosine-rich proteins<sup>46</sup>, and oxidation of tyrosine may be essential for shell matrix maturation. Several proteinases and proteinase inhibitors are highly mantle specific and abundant in shells, and may be involved in matrix formation, modification and protection (Supplementary Table 24). Together, these results indicate that oyster shell matrix is not formed simply by self-assembling silk-like proteins but by diverse proteins through complex assembly and modification processes that may involve haemocytes and exosomes.

### Concluding remarks

We sequenced and assembled the genome of the Pacific oyster using an inbred individual, short-read NGS and a new fosmid-pooling and hierarchical assembly strategy. The draft assembly provided insight into a molluscan genome characterized by high polymorphism, abundant repetitive sequences and active transposable elements. Genomic, transcriptomic and proteomic analyses show unique adaptations of oysters to sessile life in a highly stressful intertidal environment and the complexity of shell formation. The oyster

genome sequence and comprehensive transcriptome data provide valuable resources for studying molluscan biology and lophotrochozoan evolution, and for genetic improvement of oysters and other important aquaculture species.

### METHODS SUMMARY

The sequenced Pacific oyster is an inbred female produced by four generations of brother–sister mating. Genome sequences were produced with Illumina platform using fosmid pooling and assembled with a new hierarchical assembly strategy. Fosmid ends were sequenced by Sanger. Gene models were obtained by integrating results of *de novo* gene prediction and alignment-based methods based on homology and transcriptomic evidence. Transcriptomes were sequenced with Illumina platform. The proteome of the shell was obtained by mass spectrometry. All methods are described in detail in the Supplementary Information.

Received 30 January; accepted 11 July 2012.

Published online 19 September 2012.

- Ponder, W. F. & Lindberg, D. R. in *Phylogeny and Evolution of the Mollusca* (eds Ponder, W. & Lindberg, D. R.) Ch. 1, 1–18 (Univ. of California Press, 2008).
- Walters, E. T. & Moroz, L. L. Molluscan memory of injury: evolutionary insights into chronic pain and neurological disorders. *Brain Behav. Evol.* **74**, 206–218 (2009).
- Talmage, S. C. & Gobler, C. J. Effects of past, present, and future ocean carbon dioxide concentrations on the growth and survival of larval shellfish. *Proc. Natl Acad. Sci. USA* **107**, 17246–17251 (2010).
- Caldeira, K. & Wickett, M. E. Anthropogenic carbon and ocean pH. *Nature* **425**, 365–365 (2003).
- Marin, F., Luquet, G., Marie, B. & Medakovic, D. Molluscan shell proteins: primary structure, origin, and evolution. *Curr. Top. Dev. Biol.* **80**, 209–276 (2008).
- Hedgecock, D. *et al.* The case for sequencing the Pacific oyster genome. *J. Shellfish Res.* **24**, 429–441 (2005).
- Sodergren, E. *et al.* The genome of the sea urchin *Strongylocentrotus purpuratus*. *Science* **314**, 941–952 (2006).
- Small, K. S., Brudno, M., Hill, M. M. & Sidow, A. Extreme genomic variation in a natural population. *Proc. Natl Acad. Sci. USA* **104**, 5698–5703 (2007).
- Li, R. *et al.* The sequence and *de novo* assembly of the giant panda genome. *Nature* **463**, 311–317 (2010).
- Xu, X. *et al.* Genome sequence and analysis of the tuber crop potato. *Nature* **475**, 189–195 (2011).
- Bonasio, R. *et al.* Genomic comparison of the ants *Camponotus floridanus* and *Harpegnathos saltator*. *Science* **329**, 1068–1071 (2010).
- Dalloul, R. A. *et al.* Multi-platform next-generation sequencing of the domestic turkey (*Meleagris gallopavo*): genome assembly and analysis. *PLoS Biol.* **8**, e1000475 (2010).
- Star, B. *et al.* The genome sequence of Atlantic cod reveals a unique immune system. *Nature* **477**, 207–210 (2011).
- Harris, R. S. *Improved Pairwise Alignment of Genomic DNA*. PhD thesis, Pennsylvania State Univ. (2007).
- Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
- Kitzman, J. O. *et al.* Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nature Biotechnol.* **29**, 59–63 (2010).
- Suk, E. K. *et al.* A comprehensively molecular haplotype-resolved genome of a European individual. *Genome Res.* **21**, 1672–1685 (2011).
- Sauvage, C., Bierne, N., Lapègue, S. & Boudry, P. Single nucleotide polymorphisms and their relationship to codon usage bias in the Pacific oyster *Crassostrea gigas*. *Gene* **406**, 13–22 (2007).
- McGoldrick, D. J. & Hedgecock, D. Fixation, segregation and linkage of allozyme loci in inbred families of the Pacific oyster *Crassostrea gigas* (Thunberg): implications for the causes of inbreeding depression. *Genetics* **146**, 321–334 (1997).
- Hillier, L. W. *et al.* Whole-genome sequencing and variant discovery in *C. elegans*. *Nature Methods* **5**, 183–188 (2008).
- Li, R. *et al.* *De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**, 265–272 (2010).
- Gaffney, P. M., Pierce, J. C., Mackinley, A. G., Titchen, D. A. & Glenn, W. K. *Pearl*, a novel family of putative transposable elements in bivalve mollusks. *J. Mol. Evol.* **56**, 308–316 (2003).
- Han, Y. & Wessler, S. R. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res.* **38**, e199 (2010).
- Barucca, M., Olmo, E. & Canapa, A. Hox and paraHox genes in bivalve molluscs. *Gene* **317**, 97–102 (2003).
- Ferrier, D. E. K. & Holland, P. W. H. *Ciona intestinalis* ParaHox genes: evolution of Hox/ParaHox cluster integrity, developmental mode, and temporal colinearity. *Mol. Phylogenet. Evol.* **24**, 412–417 (2002).
- Goldstone, J. V. *et al.* The chemical defenseome: environmental sensing and response genes in the *Strongylocentrotus purpuratus* genome. *Dev. Biol.* **300**, 366–384 (2006).
- Carland, T. M. & Gerwick, L. The C1q domain containing proteins: where do they come from and what do they do? *Dev. Comp. Immunol.* **34**, 785–790 (2010).
- Hanington, P. C. & Zhang, S. M. The primary role of fibrinogen-related proteins in invertebrates is defense, not coagulation. *J. Innate Immun.* **3**, 17–27 (2011).

29. Zhang, S. M., Adema, C. M., Kepler, T. B. & Loker, E. S. Diversification of Ig superfamily genes in an invertebrate. *Science* **305**, 251–254 (2004).
30. Kourtis, N. & Tavernarakis, N. Cellular stress response pathways and ageing: intricate molecular relationships. *EMBO J.* **30**, 2520–2531 (2011).
31. Lee, E. F. *et al.* Discovery and molecular characterization of a Bcl-2-regulated cell death pathway in schistosomes. *Proc. Natl Acad. Sci. USA* **108**, 6999–7003 (2011).
32. Bender, C. E. *et al.* Mitochondrial pathway of apoptosis is ancestral in metazoans. *Proc. Natl Acad. Sci. USA* **109**, 4904–4909 (2012).
33. Galtsoff, P. S. The American oyster *Crassostrea virginica* Gmelin. *Fishery Bull.*, **64**, 1–480 (United States Govt Printing Office, 1964).
34. Mosser, D. D., Caron, A. W., Bourget, L., Denis-Larose, C. & Massie, B. Role of the human heat shock protein hsp70 in protection against stress-induced apoptosis. *Mol. Cell. Biol.* **17**, 5317–5327 (1997).
35. Weiner, S., Traub, W. & Parker, S. Macromolecules in mollusc shells and their functions in biomineralization. *Phil. Trans. R. Soc. Lond. B* **304**, 425–434 (1984).
36. Furuhashi, T., Schwarzinger, C., Miksik, I., Smrz, M. & Beran, A. Molluscan shell evolution with review of shell calcification hypothesis. *Comp. Biochem. Physiol. B Biochem. Mol. Biol.* **154**, 351–371 (2009).
37. Mount, A. S., Wheeler, A. P., Paradar, R. P. & Snider, D. Hemocyte-mediated shell mineralization in the eastern oyster. *Science* **304**, 297–300 (2004).
38. Stark, M. *et al.* Macroscopic fibers self-assembled from recombinant miniature spider silk proteins. *Biomacromolecules* **8**, 1695–1701 (2007).
39. Lemmon, C. A., Chen, C. S. & Romer, L. H. Cell traction forces direct fibronectin matrix assembly. *Biophys. J.* **96**, 729–738 (2009).
40. Keller, S., Sanderson, M. P., Stoeck, A. & Altevogt, P. Exosomes: from biogenesis and secretion to biological function. *Immunol. Lett.* **107**, 102–108 (2006).
41. Mathivanan, S., Fahner, C. J., Reid, G. E. & Simpson, R. J. ExoCarta 2012: database of exosomal proteins, RNA and lipids. *Nucleic Acids Res.* **40**, D1241–D1244 (2012).
42. Mount, A. S., Gohad, N. V., Hansen, D. C., Mueller, K. & Johnstone, M. B. Deposition of nanocrystalline calcite on surfaces by a tissue and cellular biomineralization. US patent 2010/0150982 A1 (2010).
43. Sauvage, E., Kerff, F., Terrak, M., Ayala, J. A. & Charlier, P. The penicillin-binding proteins: structure and role in peptidoglycan biosynthesis. *FEMS Microbiol. Rev.* **32**, 234–258 (2008).
44. Nagai, K., Yano, M., Morimoto, K. & Miyamoto, H. Tyrosinase localization in mollusc shells. *Comp. Biochem. Physiol. B* **146**, 207–214 (2007).
45. Chang, T. S. An updated review of tyrosinase inhibitors. *Int. J. Mol. Sci.* **10**, 2440–2475 (2009).
46. Waite, J. H. in *The Mollusca* Vol. I (eds Hochachka, P. & Wilbur, K. M.) Ch. 11, 467–504 (Academic, 1983).
47. Zhang, G. *et al.* Genomic data from the Pacific oyster (*Crassostrea gigas*). GigaScience. <http://dx.doi.org/10.5524/100030> (2012).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We acknowledge H. Wu, F. Zhang, Q. Tang, Z. Zhu, X. Xu, H. Lin, J. Lei, Z. Xiang, N. Li, J. Xiang and J. Jia for their support of the oyster genome project. We thank F. Han, X. Liu, R. Wu, L. Wang, Y. Wu, L. Yan, H. Niu, H. Li, Y. Wang, J. Liang, Z. Jia, J. Davis and Taylor Shellfish Farms for assistance with DNA, RNA and protein extraction, data analysis and oyster culture, and Y. Lu, C. Lin, H. Peng, Y. Ren, X. Xu, R. Chen and D. Zhang for library construction and sequencing. We thank L. Song, B. Z. Liu, Q. Li, Z. Yu, C. Ke, J. Yu, B. Liu, X. Sun, R. W. Chapman, Y. Han, S. R. Wessler, D. Arendt, E. H. Davidson,

J. S. Evans, B. Brown, P. Boudry and B. Lieb for discussions. We thank other faculty and staff at the Institute of Oceanology, Chinese Academy of Science, BGI-Shenzhen and Rutgers who contributed to the oyster genome project. We acknowledge grant support from the National High-Technology Research and Development Program of China (863 program; 2010AA10A110), National Basic Research Program of China (973 Program; 2010CB126401 and 2010CB126402), 863 program (2012AA10A405), Basic Research Program Supported by Shenzhen City (JC2010526019), Shenzhen Key Laboratory of Transomics Biotechnologies (CXB201108250096A), National Natural Science Foundation of China (40730845), Mollusc Research and Development Center, CARS, Shenzhen Key Laboratory of Gene Bank for National Life Science, Taishan Scholar and Scholar Climbing Programs of Shandong. X.G. acknowledges funding from the US Department of Agriculture (2009-35205-05052 and NJ32108) and the Chinese Academy of Science Marine Functional Genomics Oversea Team and Taishan Scholar Fund; P.W.H.H. acknowledges funding from the European Research Council (EU FP7 ERC grant [268513]11); and J.P. acknowledges funding from Beatrice de Pinós of the Generalitat de Catalunya (2009 BP-DGR). We are grateful to Dalian Zhangzhidao Fishery Group Co. Ltd for providing support.

**Author Contributions** G.Z. and X.G. conceived the study and designed scientific objectives. G.Z., Jun W. and X.G. led the project and manuscript preparation. Jun W., X.F. and Y.Y. developed the sequencing strategy. L. Li and X.F. managed the project. R.L. (leader), Yr.L., Z.H., Y.L., Xq.S., B.L., X.J., W.F., Qm.Z., H.Y., L. Luo, B. Wang, Y.M. and P.N. conducted assembly and evaluation; X.F. (leader), P.Y., Zq.X., Y.X., Yb.Z., Y.C., C.P., Y.F., D.F., L.Y., Z.P., Na L., X.W., M.C., L.C., S.Z., Jy.W., Ning L., Gj.Z. and Yr.L. performed genome annotation and data analysis; L. Li, F.X., Hy.Q., F.W., Sd.Z., Jp.W., X.D., J.Z., Q.W. and L.Q. cultured oysters and provided materials; Hg.Q. (leader), L. Li, Jf.W., Z.S. and H.W. performed polymorphism analysis and validation; F.X. (leader), P.W.H.H., J.P., T.D.L., P.Y., J. Liu, X.W., L. Li, N.Z., J. Li, W.W., Yx.L., M.Y. and W.H. conducted developmental biology studies and data analysis; L.Z. (leader), X.G., J.M., Qh.Z., Y.D., C.E.W.S., P.C., B.H., T.Q. and G.M. conducted stress studies and data analysis; X.W. (leader), X.G., T.W., Z. Xu, Y.H., A.M., Xr.S., R.H., B. Wen, F.T. and Y.Z. conducted shell-formation studies and data analysis. Hm.Y. and Jian W. supervised sequencing, assembly and bioinformatics analysis. X.G., S.W. and F.X. performed flow-cytometry analysis. D.H. and P.M.G. provided inbred oysters, BAC sequences and advice. L.Q. and X.L. participated in discussions and provided suggestions. X.G., X.F., L. Li, R.L., F.X., P.Y., L.Z., X.W., Hg.Q. and P.W.H.H. did most of the writing with contributions from all authors.

**Author Information** The oyster genome project has been deposited at DDBJ/EMBL/GenBank under the accession number AFT101000000. All short-read data have been deposited into the Sequence Read Archive (SRA) (<http://www.ncbi.nlm.nih.gov/sra>) under the accession number SRA040229. Short-read data of re-sequencing have been deposited in the SRA under the accession number SRA043580. Raw sequencing data of transcriptomes have been deposited in the Gene Expression Omnibus under the accession number GSE31012. Genomic data are also available at the Comprehensive Library for Modern Biotechnology (CLiMB) repository: doi:10.5524/100030 (ref. 47). Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). This paper is distributed under the terms of the Creative Commons Attribution-Non-Commercial-Share Alike licence, and the online version of the paper is freely available to all readers. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to Jun W. ([wangj@genomics.org.cn](mailto:wangj@genomics.org.cn)), X.G. ([xguo@hsrl.rutgers.edu](mailto:xguo@hsrl.rutgers.edu)), Y.Y. ([yinye@genomics.org.cn](mailto:yinye@genomics.org.cn)) or G.Z. ([gzhang@qdio.ac.cn](mailto:gzhang@qdio.ac.cn)).