DNA Research **21**, 85–101, (2014)
Advance Access publication on 9 October 2013

doi:10.1093/dnares/dst043

# High-Resolution Linkage and Quantitative Trait Locus Mapping Aided by Genome Survey Sequencing: Building Up An Integrative Genomic Framework for a Bivalve Mollusc

Wenqian Jiao[1,†], Xiaoteng Fu[1,†], Jinzhuang Dou[1,†], Hengde Li[1,2], Hailin Su[1], Junxia Mao[1], Qian Yu[1], Lingling Zhang[1], Xiaoli Hu[1], Xiaoting Huang[1], Yangfan Wang[1], Shi Wang[1,*], and Zhenmin Bao[1,*]

*Key Laboratory of Marine Genetics and Breeding, College of Marine Life Sciences, Ocean University of China, 5 Yushan Road, Qingdao 266003, China[1] and The Center for Applied Aquatic Genomics, Chinese Academy of Fishery Sciences, 150 Qingta West Road, Beijing 100141, China[2]*

*To whom correspondence should be addressed. Tel. +86 532-82031969. Fax. +86 532-82031969. Email: swang@ouc.edu.cn (S.W.); Tel. +86 532-82031960. Fax. +86 532-82031960. Email: zmbao@ouc.edu.cn (Z.B.).

Edited by Dr Yuji Kohara
(Received 7 May 2013; accepted 9 September 2013)

## Abstract

**Genetic linkage maps are indispensable tools in genetic and genomic studies. Recent development of genotyping-by-sequencing (GBS) methods holds great promise for constructing high-resolution linkage maps in organisms lacking extensive genomic resources. In the present study, linkage mapping was conducted for a bivalve mollusc (*Chlamys farreri*) using a newly developed GBS method—2b-restriction site-associated DNA (2b-RAD). Genome survey sequencing was performed to generate a preliminary reference genome that was utilized to facilitate linkage and quantitative trait locus (QTL) mapping in *C. farreri*. A high-resolution linkage map was constructed with a marker density (3806) that has, to our knowledge, never been achieved in any other molluscs. The linkage map covered nearly the whole genome (99.5%) with a resolution of 0.41 cM. QTL mapping and association analysis congruously revealed two growth-related QTLs and one potential sex-determination region. An important candidate QTL gene named *PROP1*, which functions in the regulation of growth hormone production in vertebrates, was identified from the growth-related QTL region detected on the linkage group LG3. We demonstrate that this linkage map can serve as an important platform for improving genome assembly and unifying multiple genomic resources. Our study, therefore, exemplifies how to build up an integrative genomic framework in a non-model organism.**

**Key words:** bivalve; genome sequencing; 2b-RAD genotyping; linkage mapping; quantitative trait locus mapping

## 1. Introduction

High-resolution linkage maps are exceptionally valuable tools in many genetic and genomic applications, such as fine-scale quantitative trait locus (QTL) mapping, characterization of recombination hotspots, comparative genome analysis and genome scaffolding. A key prerequisite for constructing a high-resolution linkage map is the availability of a large number of genetic markers (preferably, single-nucleotide polymorphisms—SNPs), which has been unfortunately out of reach for most non-model organisms until just a few years ago due to the huge investments in large-scale marker discovery and array-based genotyping. Recent advances in the next-generation sequencing technologies have revolutionized our ability to obtain massive genomic resources in a rapid and cost-effective

[†] These authors contributed equally to this work.

manner and, therefore, greatly stimulate the development of a variety of genotyping-by-sequencing (GBS) methods.[1] Among available GBS methods, restriction site-associated DNA (RAD) has gained popularity for the construction of high-density linkage maps,[1] and several methods with simpler RAD library preparation protocols have been developed.[2,3] For example, Wang *et al.*[2] have recently developed the 2b-RAD method by adopting Type IIB restriction enzymes to simplify RAD library preparation. The streamlined 2b-RAD method features even and tunable genome coverage, providing a flexible genotyping platform to diminish the extensive sequencing effort required for species with large genomes.

Bivalves comprise ~14 000 species worldwide,[4] constituting the second largest group of molluscs. Many bivalves are of economic importance and support both commercial fisheries and mariculture efforts. They make up an important source of food all over the world, with a production of over 11.7 million metric tons in 2008, corresponding to 22% of the global aquaculture production.[5] Elucidation of genetic bases underlying economically important traits, such as growth and reproduction, has been a central task in bivalve genetic and breeding studies for the purpose of genetic improvement. Linkage maps have been constructed for many bivalve species with the aim for QTL mapping of important traits. However, the resolution of these maps is generally low (mostly 10–20 cM), thus limiting their usefulness in determining quantitative trait genes. Great progresses have been recently achieved by transcriptome sequencing for large-scale identification of candidate genes involved in bivalve growth and reproduction.[6–10] Polymorphisms in some candidate genes have also been associated with the variation of bivalve growth traits.[11,12] Bivalves exhibit fascinatingly diverse sexual patterning. Some species are dioecious, whereas some are simultaneous hermaphrodites with a few as protandrous hermaphrodites (male when young, then later become female) or proterogynous hermaphrodites (female when young, then later become male). Currently, the mechanism of sex determination remains poorly understood in bivalves. Some recent studies have associated the sex determination of bivalves with doubly uniparental inheritance (DUI) of mitochondria[10,13] and a hypothetical model has been proposed.[10] However, the generality of the proposed model is constrained because DUI is a peculiar mtDNA inheritance system that exists only in a few bivalve species.

The scallop *Chlamys farreri* (Jones et Preston 1904) naturally distributes along the seacoasts of China, Japan and Korea, and is a commercially important bivalve species in China. With increasing acreage devoted to artificial farming, several issues such as declining production and frequent disease outbreaks have raised great concern in modern scallop aquaculture. Currently, genetic studies focusing on scallop growth, reproduction and immunity represent active research directions. Genetic and genomic resources for *C. farreri* have rapidly expanded, including low-density linkage,[14–17] physical[18] and cytogenetic maps,[19] fosmid and bacterial artificial chromosome (BAC) libraries,[20,21] a large number of expressed sequence tags.[7,22] The direction of the present study was also shaped by our immediate need for building up an integrative genomic platform that unifies a variety of genomic resources, such as genome map, linkage map and physical map, and facilitates future genetic, genomic and breeding studies on *C. farreri*.

## 2. Materials and methods

### 2.1. *Genome survey sequencing and gene annotation*

A two-year-old *C. farreri* individual was chosen for genome survey sequencing. Genomic DNA was extracted from its adductor muscle using the standard phenol/chloroform extraction method.[23] Three paired-end DNA libraries with insert sizes of 175, 500 and 800 bp were constructed by following the Illumina's standard DNA library preparation protocol and were then sequenced using the Illumina HiSeq2000 platform.

Raw reads were first filtered to remove low-quality reads resulting from base-calling duplications or adapter contamination. Clean reads were assembled using the SOAPdenovo software, which applies the *de Bruijn* graph structure to construct contigs. Subsequently, all reads were realigned to the contigs and the paired-end information was used to join unique contigs into scaffolds. Finally, intra-scaffold gaps were filled using paired-end extracted reads that had one read uniquely aligned on a contig and another read located in a gap region. Genomic sequences were archived in the Sequence Read Archive (SRA) database (accession no. SRP018107).

Our group has recently reported the most comprehensive transcriptome resource for *C. farreri* by 454 sequencing of cDNA libraries generated from diverse developmental stages and adult tissues,[7] which provide an excellent basis for gene annotation of the obtained scaffolds. Genomic scaffolds were first annotated by aligning the 20 056 transcriptomic isogroup sequences (i.e. the longest isotig in each isogroup) onto these scaffolds using the GMAP program.[24] To increase the annotation rate, the unannotated scaffolds were compared against the Swiss-Prot and non-redundant (NR) protein databases using BlastX, with an $E$-value threshold of $1E-4$. For scaffolds without significant matches, tBlastX search with an $E$-value threshold of $1E-4$ was conducted against the NR nucleotide database for further annotation. To increase computational speed, all Blast searches were limited to the top 10 significant hits for each query.

### 2.2. *Mapping families*

Twenty full-sib families were established in May 2009 by single-pair mating of 'Penglai-Red' scallops, an elite *C. farreri* strain developed by our group with features of red shell, fast growth and disease resistance. Growth-related traits including shell length, shell width, shell height and body weight were measured for all families at the age of 15 months. One of these families exhibiting high within-family variation of growth traits was chosen for linkage and QTL analysis.

### 2.3. *2b-RAD sequencing*

2b-RAD libraries were prepared for two parents and 96 progenies by following the protocol developed by Wang *et al.*[2] For the parents, standard BsaXI libraries were constructed, whereas for the progenies, reduced representation (RR) libraries were constructed using adaptors with 5′-NNT-3′ overhangs to target a subset of all BsaXI fragments in the *C. farreri* genome. A unique barcode was incorporated into each library during library preparation, and then all libraries were pooled for single-end sequencing (1 × 50 bp) using an Illumina GA-II sequencer. All the 2b-RAD sequences were archived in the SRA database (accession no. SRA065207).

### 2.4. *Sequence data preprocessing and* de novo *genotyping*

Raw reads were first trimmed to remove adaptor sequences. The terminal 3-bp positions were excluded from each read to eliminate artefacts that might have arisen at ligation sites. Reads with no restriction sites or containing ambiguous base calls (N), long homopolymer regions (>10 bp), excessive numbers of low-quality positions (>5 positions with quality of <10) or mitochondrial origins were removed. The remaining trimmed, high-quality reads formed the basis for subsequent analysis.

*De novo* 2b-RAD genotyping was performed using the RADtyping program v1.0 (http://www2.ouc.edu.cn/mollusk/detailen.asp?Id=727) under default parameters. The RADtyping program has recently been developed by our group, and it is an integrated pipeline that enables both accurate *de novo* codominant and dominant genotyping in mapping populations.

### 2.5. *Linkage map construction*

Segregating markers that could be genotyped in at least 80% of the individuals were considered as high-quality markers and were retained for further analysis. For each segregating marker, goodness of fit of the observed with the expected Mendelian ratio was assessed using the $\chi^2$ test, and those markers conforming to the expected Mendelian ratios ($P \geq 0.05$) were included in map construction. Linkage analysis was performed using the JoinMap 4.0 software.[25] Sex-specific maps were first constructed for each parent. Maternal and paternal datasets were created using the function 'Create Maternal and Paternal Population Nodes' in the JoinMap program, which was also used to partition 1:2:1-type data into 1:1 female-type and 1:1 male-type data. A logarithm of odds (LOD) threshold of at least 6.0 was used to assign markers into linkage groups. The regression mapping algorithm was selected for map construction. The Kosambi mapping function was used to convert the recombination frequencies into map distances (centiMorgans). A consensus map was established by integrating the sex-specific maps through shared markers using the MergeMap software.[26]

### 2.6. *QTL mapping and association analyses of growth traits and sex*

The distributions of growth traits were first assessed using the univariate procedure of MATLAB software to check whether they followed normal distributions. Pearson correlations among four growth traits were calculated using pairwise comparison.

The regression method was used for QTL mapping of both growth traits and sex. Here, sex was used as a binary trait, supposing the haplotypes of the male and female parents are $H_{pp}$, $H_{pm}$, $H_{mp}$ and $H_{mm}$, that $a_f$, $-a_f$, $a_m$ and $-a_m$ are their respective effects on the trait of interest, and $P_{pp}$, $P_{pm}$, $P_{mp}$ and $P_{mm}$ are the respective individual inheritance probabilities at putative positions in the corresponding parental haplotypes. Following the above assumption, $P_{pp} + P_{pm} = P_{mp} + P_{mm} = 1$. Therefore, the hypothesis to detect each QTL is:

$$y = \mu + s + (P_{pp} - P_{pm})a_f + (P_{mp} + P_{mm})a_m + e \quad (1)$$

and its null hypothesis is:

$$y = \mu + s + e \quad (2)$$

where $y$ is the observed value of the trait of interest; $\mu$, the population mean; $s$, the fixed effect of sex and $e$, the random error.

For each chromosome, midpoints of all marker brackets along the chromosome were considered to be putative QTL positions, and the inheritance probabilities for individuals at each position were calculated using the IBD program[27] and the corresponding LOD value was determined by:

$$\text{LOD} = \frac{1}{2}(\log_{10} e)\text{LRT} = 0.217\text{LRT} \quad (3)$$

where the LRT statistic was calculated through the comparison of Equations (5) and (6). Chromosome-wide and genome-wide critical threshold values for QTL detection were approximated using Piepho's method.[28] If a QTL was detected, the confidence interval was calculated using Li's method.[29]

Association analysis was performed as a complement approach to QTL mapping, which was implemented using an R package called 'GWAF'[30] that tests genetic association between genotypes and traits by fitting a linear model accounting for within full-sib pedigree correlation. The additive model in GWAF was selected by use of the kinship coefficient matrix in the analysis to avoid false positives owing to unexpected familial correlation. The functions 'lme.batch' and 'gee.lgst.-batch' were used to perform a global test (i.e. Wald $\chi^2$ test) of genotypic effects with growth traits and sex, respectively. Bonferroni correction was carried out to account for multiple comparisons.

### 2.7. *Integration of the linkage map, genomic scaffolds and a physical map*

To associate genomic scaffolds with the linkage map, all markers in the linkage map were mapped onto the scaffolds using the SOAP2 software[31] (parameters −M 4, -v 2). Marker-associated scaffolds were further used as linkers to integrate the linkage map with a BAC-based physical map previously constructed for *C. farreri*.[18] These scaffold sequences were compared against available BAC-end sequences (BESs) using BlastN with an *E*-value cut-off of 1*E*−10. A reciprocal comparison was performed to ensure one-to-one matches between the scaffolds and BAC contigs.

## 3. Results

### 3.1. *Genome survey sequencing of* C. farreri

The genome size of *C. farreri* is relatively large and was previously estimated as ∼1.2 Gb.[20] Genome survey sequencing was performed for a single *C. farreri* individual by sequencing three paired-end DNA libraries with insert sizes of 175, 500 and 800 bp based on the Illumina HiSeq2000 platform. In total, 62.4 Gb of sequencing data, equivalent to ∼52× genome coverage, were produced (Table 1). *K*-mer analysis revealed remarkably high heterozygosity (∼1.4%) in the *C. farreri* genome (Fig. 1). Owing to the high genome heterozygosity, assembly of these data produced a large number of small scaffolds (N50 = 1.5 kb), with the longest one only reached to 46 kb (Table 2). All the scaffold sequences are accessible at http://ipl.ouc.edu.cn/fuxiaoteng/cf_SRA_data. Of 20 056 isogroups in the *C. farreri* transcriptome,[7] 18 316 (91.3%) could match with 16 489 genomic scaffolds, suggesting that the majority of genes were present in our reference genome dataset. Of the mapped transcriptomic isogroups, 88.4% had been functionally annotated through the homology-based comparison against public databases. Summing up all scaffolds resulted in a total length of ∼870 Mb, corresponding to ∼70% of genome coverage; though not complete, it constitutes the most informative reference genome currently available for *C. farreri*.

**Table 1.** Summary of genome survey sequencing data for *C. farreri*

| Paired-end libraries | Read length (bp) | Insert size (bp) | Total data (G) | Sequencing depth (X) |
|---|---|---|---|---|
| L1 | 100 | 180 | 32.7 | 27.24 |
| L2 | 100 | 500 | 21.4 | 17.82 |
| L3 | 100 | 800 | 8.4 | 6.98 |
| Total | – | – | 62.5 | 52.04 |

### 3.2. *Optimization of a 2b-RAD sequencing plan for linkage mapping*

Because of the large genome size (∼1.2 Gb) of *C. farreri*, it is necessary to devise a cost-effective sequencing plan that can balance the total sequencing cost and genotyping accuracy. In total, 210 570 BsaXI sites were identified from the reference genome dataset. Based on the genome coverage (∼70%) of this dataset, the total number of BsaXI sites in the *C. farreri* genome was estimated as ∼0.3 million. Sequencing all of the BsaXI sites to 20× for a large mapping population (e.g. hundreds of individuals) would require a large sequencing investment. A notable feature of the 2b-RAD method is its ability to fine tune marker density as needed by constructing RR libraries using less-degenerate adaptors.[2] For example, an RR library constructed using adaptors with 5′-NNT-3′ overhangs only targets ∼1/10th of all

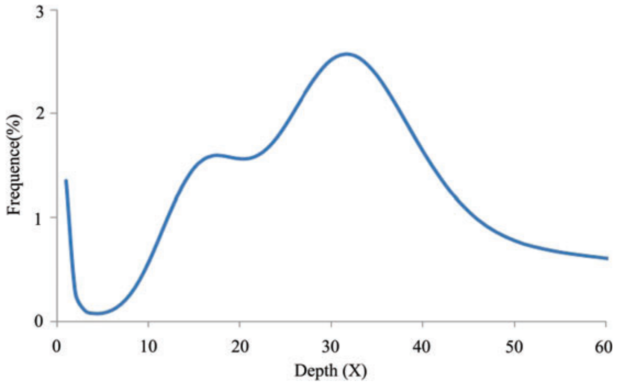

**Figure 1.** 17-mer frequency distribution of sequencing reads. All 17-mer sequences were extracted from high-quality, paired-end reads generated by sequencing three genomic libraries with different insert sizes. The frequency of each 17-mer was calculated and plotted. Two peaks rather than one were observed, indicating high genome heterozygosity.

**Table 2.** Statistics of genome assembly for *C. farreri*

| *K*-mer = 37 | Contig | | Scaffold | |
|---|---|---|---|---|
| | Size (bp) | Number | Size (bp) | Number |
| N90 | 330 | 708,971 | 395 | 595,809 |
| N80 | 503 | 497,805 | 603 | 420,096 |
| N70 | 713 | 353,715 | 851 | 298,664 |
| N60 | 960 | 249,175 | 1142 | 210,001 |
| N50 | 1267 | 170,707 | 1517 | 143,665 |
| Average length | 807 | – | 958 | – |
| Longest | 36,486 | – | 46,176 | – |
| Total length | 863,450,239 | – | 870,802,763 | – |
| Total number | 1,069,270 | – | 908,703 | – |
| Length >100 bp | 1,069,270 | – | 908,703 | – |
| Length >2 kb | 82,556 | – | 94,956 | – |

the BsaXI sites in the *C. farreri* genome, and sequencing of such a library to 20× would require only 0.6 million reads per individual (in contrast to 6 million reads per individual for a standard BsaXI library), thus dramatically reducing the total sequencing cost.

### 3.3. *2b-RAD sequencing of a* C. farreri *mapping population*

Based on the calculations above, RR libraries were constructed for a *C. farreri* mapping population, sequencing of which produced 0.6–1.2 million high-quality reads per progeny. For the parents, standard libraries were constructed and sequenced to a sufficient depth (70–80×). As shown in Fig. 2, target sites of the RR libraries could be reproducibly detected among different progenies and their average sequencing coverage was ~6-fold higher than that of standard libraries. The sequencing depth for progenies ranged from 13.4 to 23.8 with an average of 16.75. Clustering parental reads resulted in 181 625 representative reference sites consisting of 138 141 parent-shared sites and 43 484 parent-specific sites. After filtering low-quality sites, 117 113 parent-shared and 35 799 parent-specific sites remained. These reference sites contained 92% of the unique BsaXI sites inferred from the scaffolds, suggesting that unique sites were well represented in the high-quality reference sites.

### 3.4. *High-resolution linkage mapping*

In total, 7458 polymorphic markers were identified, of which 6842 were heterozygous in at least one parent, including 2196 co-dominant and 4646 dominant markers. Dominant markers showing 3 : 1 segregation pattern in progenies were not included in subsequent analysis due to their inability to construct sex-specific maps. Totally, 4881 markers that conformed to the expected Mendelian ratios ($P \geq 0.05$) were included in the linkage analysis. These markers were partitioned into maternal (1 : 1 female type; 2473 markers) and paternal (1 : 1 male type; 2175 markers) datasets for constructing sex-specific linkage maps. At the LOD threshold of 6.0, the markers were grouped into 19 linkage groups, corresponding to the haploid chromosome number of *C. farreri*.[32] The female map contained 2025 markers and spanned 1175.4 cM with an average marker interval of 0.59 cM, whereas the male map contained 1861 markers and spanned 1154.9 cM with an average marker interval of 0.62 cM. Significant differences in recombination rates between the female and male maps were observed for linkage groups LG5, LG8, LG10 and LG15 (Table 3). LG5 and LG10 showed higher recombination rates in females than in males, whereas the contrary was found for LG8 and LG15.

The sex-specific maps were further integrated using anchor markers that were heterozygous in both parents. The consensus map contained 3806 markers (Fig. 3, Table 3) and spanned 1543.4 cM with an average marker interval of 0.41 cM. The length of each linkage group ranged from 54.9 to 99.6 cM, and the marker density varied from 123 to 274 across the linkage groups. Mapping all the markers to the annotated genomic scaffolds revealed 623 markers (18–53 per linkage group; Table 3) that resided in or close to genes. These markers represent a valuable resource for future comparative genome analysis.

Genome length was estimated to be 1544.2 cM ($G_{e1}$) and 1559.5 cM ($G_{e2}$) using two different methods.[33,34] The average of the two estimates (1551.9 cM) was used as the expected genome length. Genome coverage of the consensus map was nearly complete and reached 99.5%. Given the estimated genome size of ~1.2 Gb for *C. farreri*,[20] the average recombination rate across all the linkage groups was ~1.3 cM/Mb.

### 3.5. *QTL mapping and association analyses of growth traits and sex*

Pairwise comparisons among five growth traits using Pearson's correlation revealed that all of them were highly correlated with correlation coefficients ranging from 0.77 to 0.96 (Table 4). The strongest correlation was observed between shell length and shell height ($r = 0.96$, $P = 1E-10$). A genome scan for these growth traits showed that they exhibited quite similar LOD profiles (Fig. 4), indicating that these traits were possibly regulated by the same set of genes. Therefore, a principle component analysis was conducted, which revealed that the first principal component (PC1) explained ~90% of the total phenotypic variation (Fig. 5). To increase the QTL detection

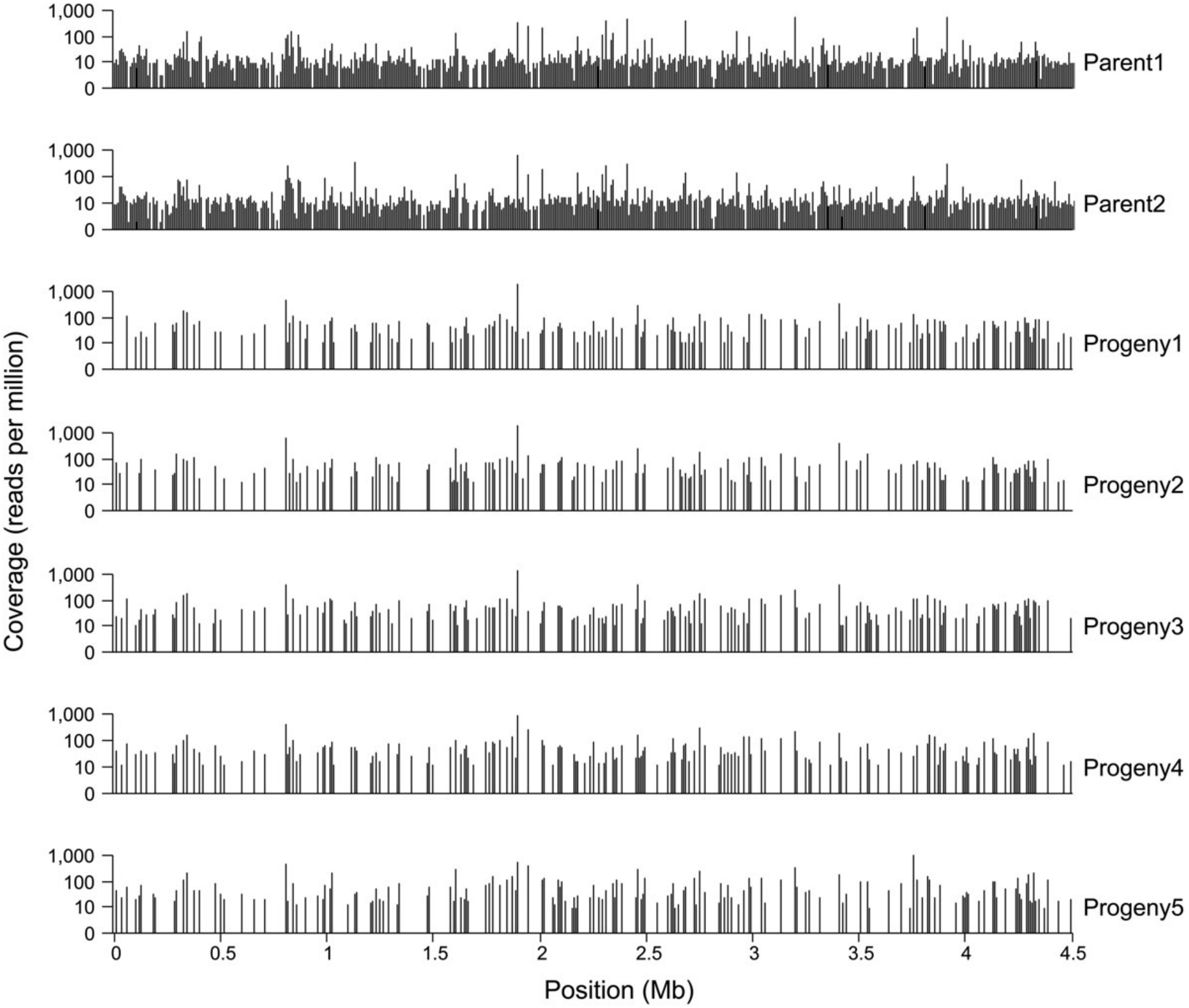

**Figure 2.** Distribution of sequencing coverage for target sites derived from 365 randomly chosen *C. farreri* scaffolds (concatenated). For parents, sequencing data were generated from standard BsaXI libraries, whereas for progenies, RR libraries were sequenced. Target sites of the RR libraries could be reproducibly sequenced among progenies and their sequencing coverage was ~6-fold higher than for standard libraries.

power, PC1 was utilized as a composite trait for QTL mapping. Two significant QTLs were detected on LG1 and LG3 (Fig. 6a and Table 5) and the one on LG3 that passed the genome-wide significance threshold showed a stronger QTL signal. The two QTLs explained 11.4 and 16.9% of the total phenotypic variation, respectively. Given the high marker density in the linkage map, association analysis was also performed to serve as a complementary approach to evaluate the QTL mapping results. Association analysis revealed a similar distribution pattern across all linkage groups as in QTL mapping analysis (Fig. 6a) and supported QTL mapping results in general. Within the confidence interval of the QTL detected on LG3, one marker named f68558 was found to be located in the intron region of a transcription factor gene, *PROP1* (Homeobox protein prophet of Pit-1) (Fig. 7 and Supplementary Table S1). It has been shown that *PROP1* can regulate the production of growth hormone,[35] a peptide that simulates growth, cell reproduction and regeneration in animals. Identification of *PROP1* within the QTL region suggests that this gene may represent an important candidate QTL gene that is worthy of further investigation.

While for sex, a highly significant QTL was finely mapped on LG11 with the confidence interval of 0.37 cM (34.61−34.98 cM; Fig. 6b). Moreover, no global difference (female : male = 1.01) in the recombination rate was observed for this linkage group between the two sex-specific maps (Table 3). Association analysis revealed

**Table 3.** Summary of the consensus linkage map in *C. farreri*

| Linkage group | Number of markers | Length (cM) | Average marker interval (cM) | Female/male recombination rate | Number of annotated markers |
|---|---|---|---|---|---|
| 1 | 274 | 96.09 | 0.35 | 0.97 | 41 |
| 2 | 272 | 88.62 | 0.33 | 0.81 | 43 |
| 3 | 271 | 95.36 | 0.35 | 1.04 | 37 |
| 4 | 241 | 75.95 | 0.32 | 1.00 | 39 |
| 5 | 235 | 71.48 | 0.31 | 1.37 | 44 |
| 6 | 230 | 99.62 | 0.44 | 1.24 | 53 |
| 7 | 223 | 77.77 | 0.35 | 1.00 | 42 |
| 8 | 207 | 68.57 | 0.33 | 0.70 | 35 |
| 9 | 203 | 91.94 | 0.46 | 1.30 | 31 |
| 10 | 199 | 54.99 | 0.28 | 2.25 | 30 |
| 11 | 189 | 96.59 | 0.51 | 0.97 | 28 |
| 12 | 183 | 92.82 | 0.51 | 0.87 | 26 |
| 13 | 172 | 81.04 | 0.47 | 1.14 | 32 |
| 14 | 167 | 65.84 | 0.40 | 1.34 | 18 |
| 15 | 166 | 78.2 | 0.47 | 0.62 | 33 |
| 16 | 162 | 84.01 | 0.52 | 0.94 | 27 |
| 17 | 147 | 81 | 0.55 | 0.82 | 23 |
| 18 | 142 | 64.36 | 0.46 | 1.07 | 21 |
| 19 | 123 | 79.11 | 0.65 | 1.00 | 20 |
| All | 3,806 | 1543.36 | 0.41 | 1.01 | 623 |

27 sex-related markers with high statistical significance ($P < 1E-6$; Supplementary Table S2) and all of them fell into the narrow QTL region identified by the QTL mapping analysis. Interestingly, one sex-related marker named f93422 was located in the coding region of a transcription factor gene, *ZNFX1* (NFX1-type zinc finger-containing 1), which has been shown to be tightly linked with *AMHR2* (anti-Mullerian hormone receptor type II), the sex-determination gene in the tiger pufferfish, *Takifugu rubripes*.[36]

### 3.6. *Integration of the linkage map, genomic scaffolds and a physical map*

A high-resolution linkage map can serve as an important platform for integrating multiple genomic resources. One such application is to improve genome assembly by orienting genomic scaffolds. For *C. farreri*, a majority of larger genomic scaffolds contained at least one BsaXI site (77% for scaffolds of $> 5$kb; 92% for scaffolds of $>10$ kb) and therefore, it could be potentially oriented using a high-density linkage map. Of the 3806 markers in the consensus linkage map, 2174 could be unambiguously anchored to scaffolds. These scaffolds were not only useful for directing a higher-level genome assembly, but also could serve as ‘long’ surrogates for 2b-RAD tags to enhance their utility in unifying genomic resources. To demonstrate this potential, we attempted to integrate the linkage map with a BAC-based physical map previously constructed for *C. farreri*.[18] The physical map contained 2514 BAC contigs that had BESs; however, a large fraction of them had only a few BESs (e.g. 73% with less than three sequenced BAC clones), thus limiting the integration efficiency. Even so, 681 BAC contigs could still be linked with the linkage map through sequence comparison between BESs and marker-associated scaffolds (Supplementary dataset). An example showing the integration of LG1 and BAC contigs is present in Fig. 8.

## 4. Discussion

### 4.1. *Genome sequencing in bivalves*

In spite of species abundance and diverse geographical distribution, little effort has been devoted to decoding bivalve genomes. To date, whole-genome sequencing has been performed for only two bivalve species, i.e. Pearl oyster (*Pinctada fucata*)[37] and Pacific oyster (*Crassostrea gigas*),[38] providing the first glimpse into bivalve genome architecture. Apparently, genome sequencing of a broader range of bivalves would enable a better understanding of their phylogeny, speciation and diversification. Our study represents the first step towards fully decoding a scallop genome.

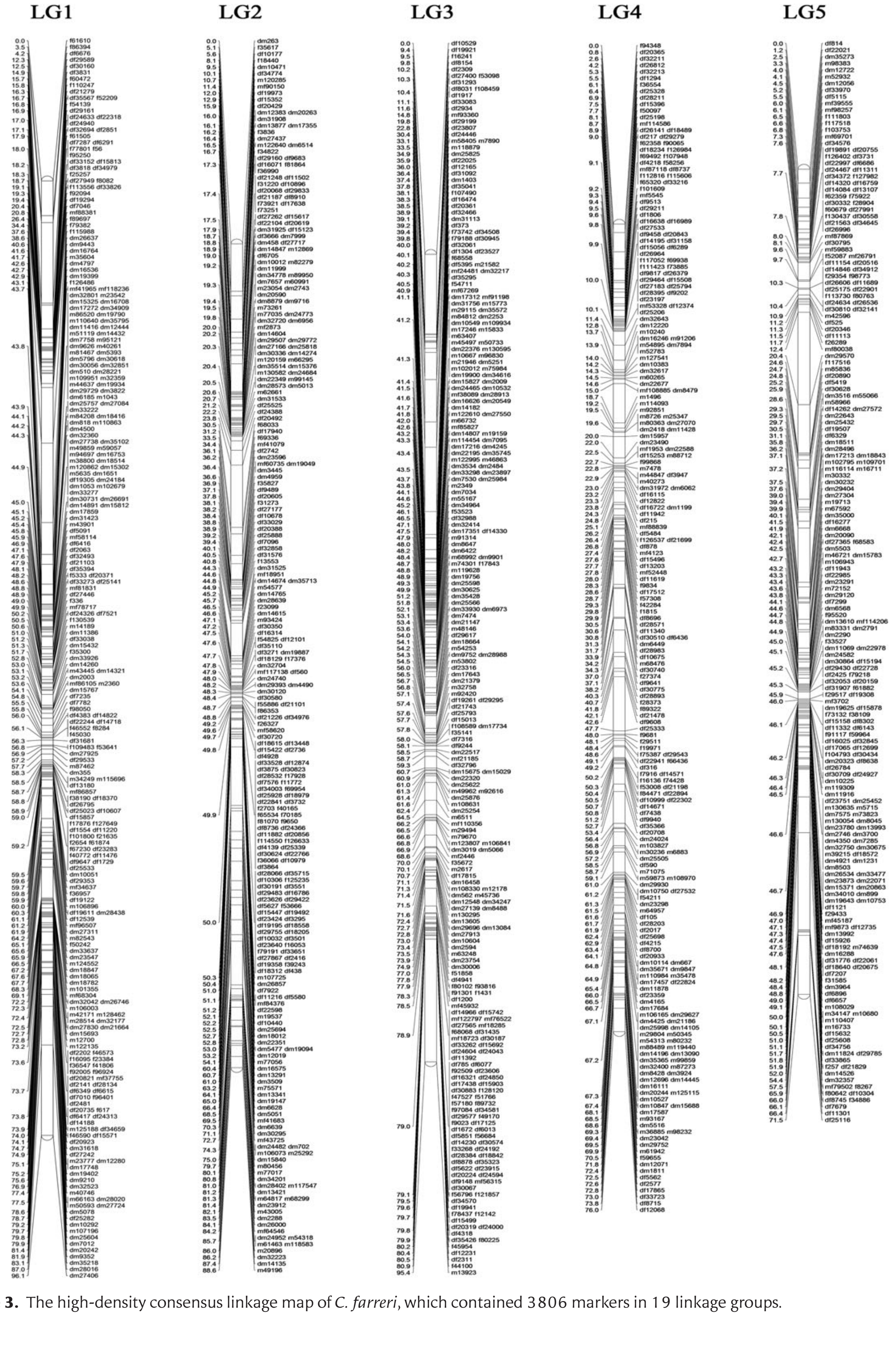

**Figure 3.** The high-density consensus linkage map of *C. farreri*, which contained 3806 markers in 19 linkage groups.

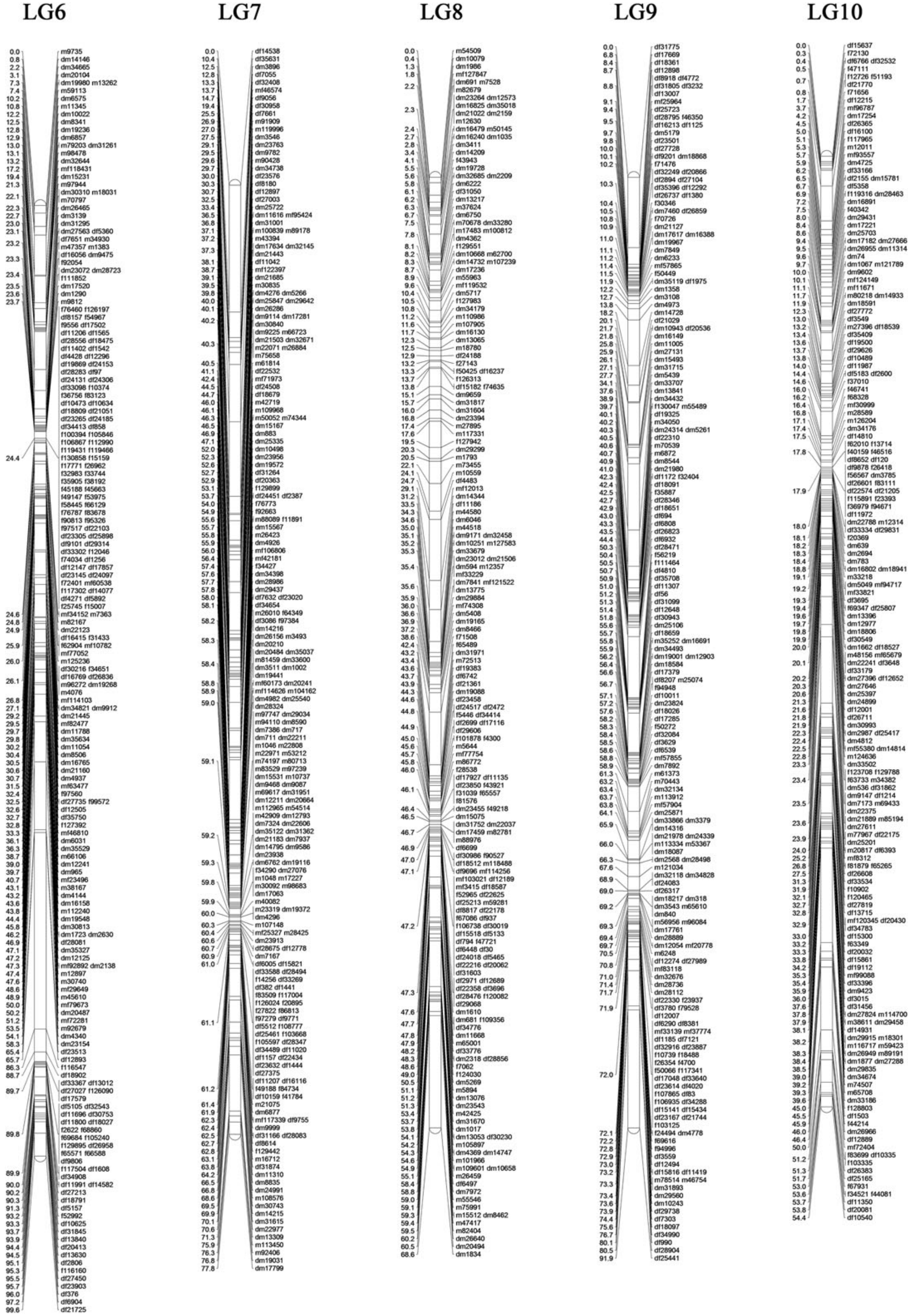

Fig. 3 *Continued*

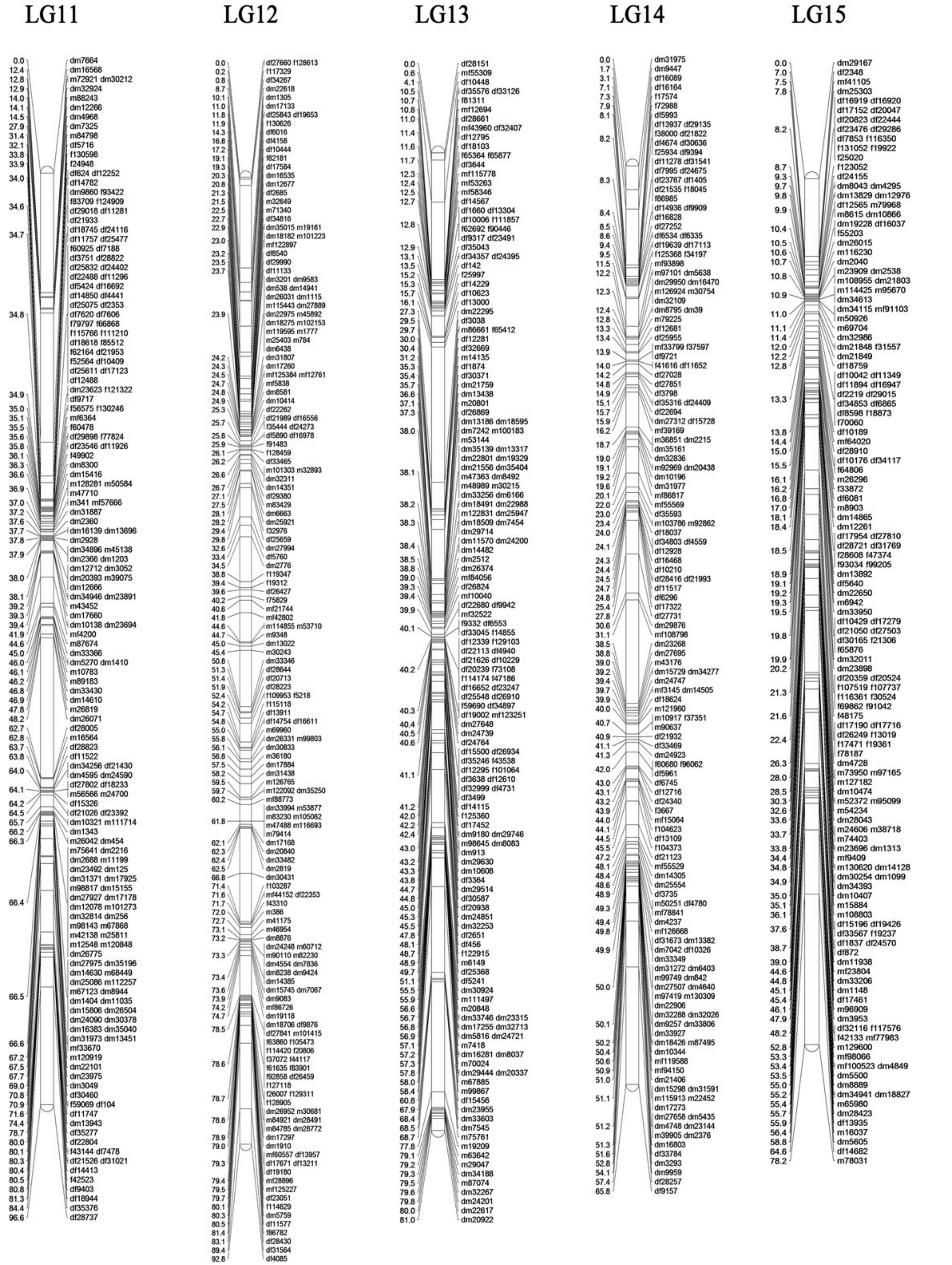

Fig. 3 *Continued*

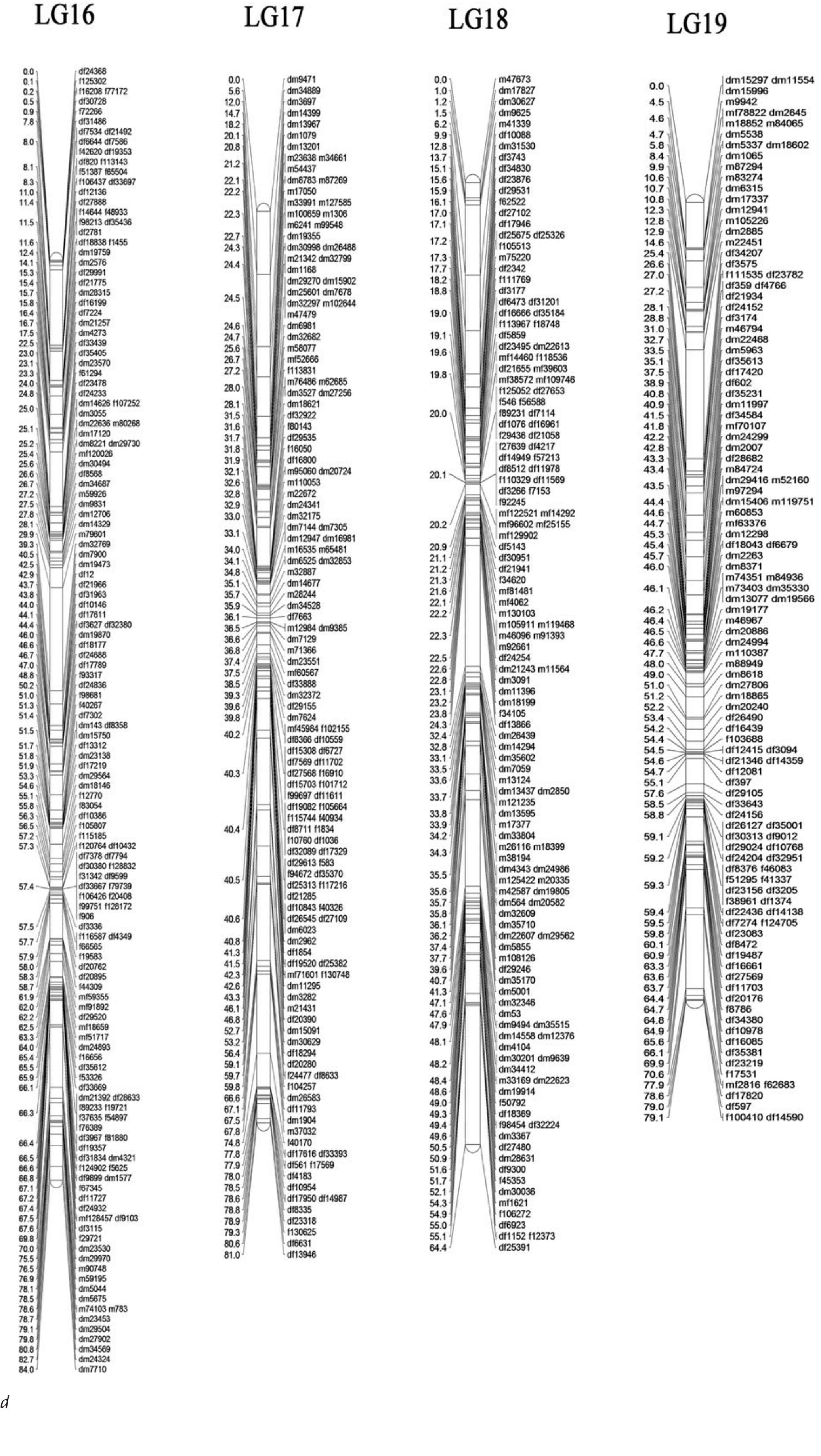

Fig. 3 *Continued*

**Table 4.** Pearson's correlation coefficients among five growth traits in *C. farreri*

| | Shell height | Shell length | Shell width | Body weight | Muscle weight |
|---|---|---|---|---|---|
| Shell height | 1.00 | 0.96 | 0.85 | 0.94 | 0.84 |
| Shell length | – | 1.00 | 0.83 | 0.93 | 0.81 |
| Shell width | – | – | 1.00 | 0.82 | 0.77 |
| Body weight | – | – | – | 1.00 | 0.93 |
| Muscle weight | – | – | – | – | 1.00 |

Genome analysis revealed that *C. farreri* possesses a highly heterozygous genome. This finding is not unusual and has been reported in other bivalve species.[37,38] For example, it has been shown that, for the Pacific oyster, average density of SNPs is one SNP per 60 bp in coding regions and one per 40 bp in non-coding regions.[39] High genome heterozygosity poses the challenge of efficient genome assembly. In such instance, a high-resolution linkage map would be much valuable in that they can orientate the small scaffolds to improve genome assembly.

### 4.2. *Devising an optimal 2b-RAD sequencing plan for linkage mapping*

Large-scale linkage mapping studies generally call for a cost-effective sequencing plan that reasonably balances the sequencing cost and genotyping accuracy. Through simulation analysis, a sequencing depth of 20× for both parents and progenies can provide sufficiently high genotyping accuracy (>96%; unpublished data). Determining the amount of sequencing necessary for a desired sequencing depth would require the knowledge of a total number of restriction sites in a given genome. Without a reference genome, the total number of restriction sites can be simply predicted based on genome size and GC content by assuming sites that are Poisson distributed. However, it has been shown that such estimation departed considerably from the actual numbers in some instances.[1] Our study demonstrates that genome survey sequencing can facilitate devising an optimal 2b-RAD sequencing plan by providing sufficient genome information to determine the minimal amount of sequencing required for a given sequencing depth. Meanwhile, the obtained genome information can also be used to enhance *de novo* 2b-RAD analysis in many other aspects, such as evaluation of the efficiency and reliability of *de novo* reference site reconstruction, identification of genes surrounding 2b-RAD tags and association of 2b-RAD tags with public genomic resources.

For species with very large genomes, sequencing all BsaXI sites at a depth of 20× for all individuals remains a substantial investment. For example, sequencing 200 human individuals with the genome size of ~3 Gb would require 5.2 billion reads, which are approximately equivalent to the number of reads produced from >4 full sequencing runs using the HiSeq2000 platform. One advantage of 2b-RAD methodology is the option to adjust marker density to the desired level using selective adaptors. Sequencing cost can be further reduced by sequencing a subset of BsaXI sites deriving from RR libraries. We have demonstrated that sequencing a *C. farreri* RR library only required about one-sixth of the sequencing cost for a standard library at a given sequencing coverage, thus reducing the total sequencing cost dramatically. Therefore, sequencing RR libraries represents an effective option for large-scale linkage mapping studies dealing with species with large genomes.

### 4.3. *Linkage map construction and QTL analysis*

Thus far, linkage maps have been constructed for many bivalve species including *C. farreri*.[14–17] However, these maps were usually built using hundreds of markers and the resolution was generally low (mostly 10–20 cM), limiting their use in fine-scale QTL mapping and many other applications. In this study, the linkage map constructed for *C. farreri* contained 3806 markers, a marker density that has, to our knowledge, never been reached for any other bivalve species. This linkage map covered nearly the whole genome (99.5%) with a resolution of 0.41 cM. In addition, ~10% of the markers in the linkage map residing in or close to genes as revealed by mapping them to the annotated genomic scaffolds. Our map, therefore, provides sufficient resolution for fine-scale QTL mapping and facilitates the discovery of quantitative trait genes.

Growth traits are of particular interest to scallop researchers due to their high commercial significance in scallop aquaculture. QTL mapping represents an efficient approach to identify genetic loci underlying these traits for marker-assisted selection in genetic breeding. Two growth-related QTLs were detected on the linkage groups LG1 and LG3, which are also supported by the association analysis, a complementary approach to evaluate the QTL mapping results. The markers located at the confidence intervals of these QTLs constitute a valuable marker set for further evaluation of their utility in marker-assisted selection. In particular, the homeotic gene *PROP1*, which was identified from the QTL region on LG3, drew our great interest. This gene encodes a paired-like homeodomain protein that is necessary for *PIT1* (growth hormone factor 1) expression. In mammals, mutations in this gene can result in impaired production of growth hormone and other

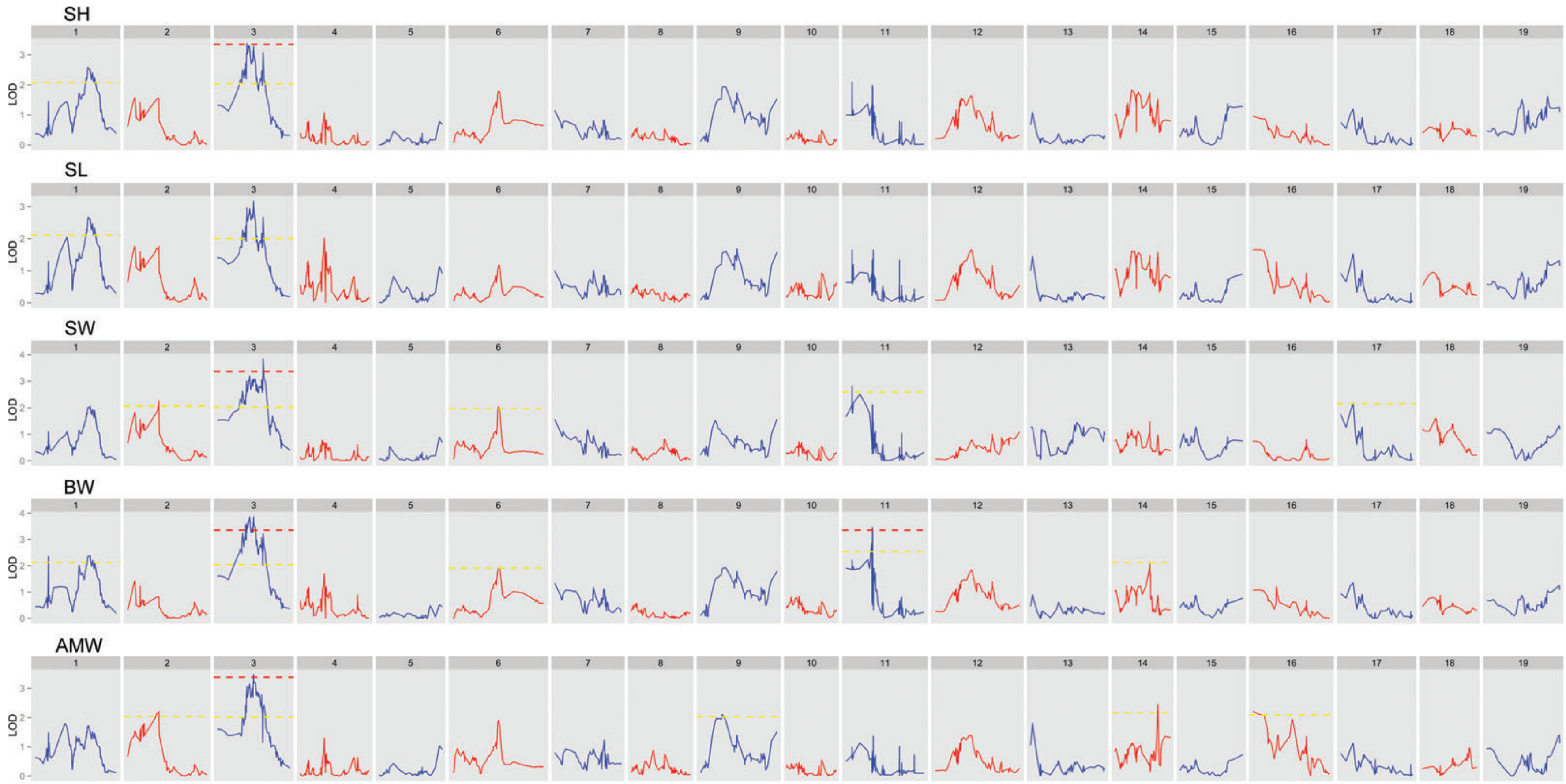

**Figure 4.** A genome scan of LOD profiles for five growth traits in *C. farreri*. All traits exhibited quite similar LOD distributions. SH, shell height; SL, shell length; SW, shell width; BW, body weight; AMW, adductor muscle weight. The dark and light dashed lines indicated the genome-wide and chromosome-wide significance thresholds.

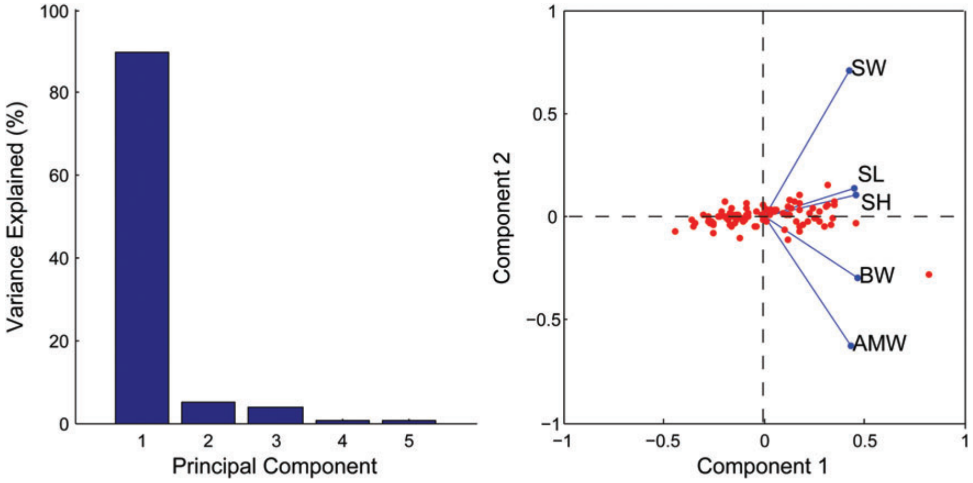

**Figure 5.** Principal component analysis of five growth traits in *C. farreri*. The PC1 explained ~ 90% of the total variance. SH, shell height; SL, shell length; SW, shell width; BW, body weight; AMW, adductor muscle weight.

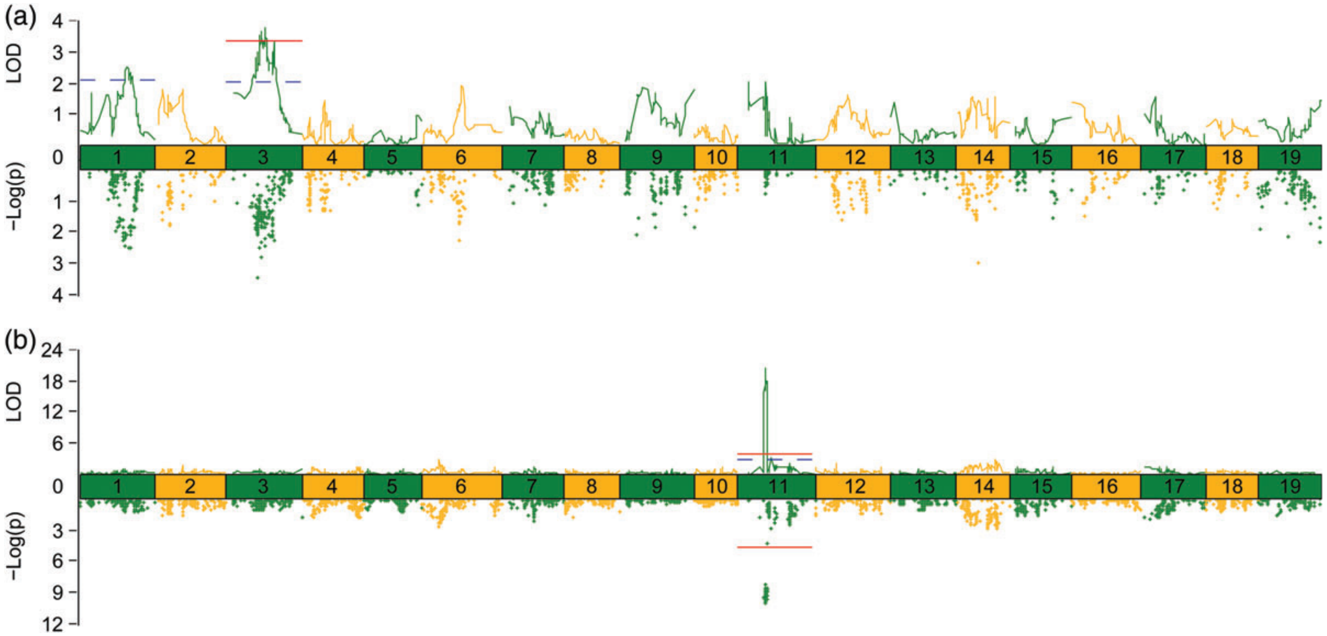

**Figure 6.** QTL mapping and association analysis of growth traits (a) and sex (b) in *C. farreri*. The PC1 for five growth traits was used in QTL mapping and association analyses. The solid and dashed lines indicate the genome-wide and chromosome-wide significance thresholds, respectively.

pituitary hormones.[35] Polymorphisms in this gene have been associated with production traits in livestock animals, such as cattle,[40] goats [41,42] and sheep.[43] The marker f68558 is probably not the causal locus that is responsible for growth trait variation in *C. farreri*. Instead, we think it is more likely that f68558 is in tight linkage disequilibrium with the causal mutation that is possibly located in the regulatory region (e.g. promoter) of *PROP1*. The expression level of *PROP1* has been checked in major adult tissues (e.g. adductor muscle, digestive gland, gill, mantle, male and female gonads and kidney) of *C. farreri*. It turned out that *PROP1* was not expressed in neither of these tissues (data not shown), suggesting that scallop *PROP1* is probably expressed and functioning during early embryonic development just like its homologues in vertebrates. In vertebrates, *PROP1* is only temporarily expressed during early pituitary development.[44] However, determining the temporal and spatial expression of *PROP1* in bivalves is not straightforward, since the counterpart of pituitary has not been established in bivalves. Further work would be needed to overcome these barriers to evaluate the role(s) of *PROP1* in scallop growth regulation.

Unlike many other animals, scallops do not possess recognizable heteromorphic sex chromosomes,[45] and little is currently known about their sex-determination mechanism. The use of molecular markers in mapping experiments to identify regions involved in sex determination represents an efficient approach to study sex-determination systems in species without heteromorphic sex chromosomes. In this study, both QTL mapping and association analysis revealed a highly significant sex-related QTL on LG11, suggesting the presence of a narrow sex-determination region (34.61–34.98 cM) in *C. farreri*. No global differences in recombination rates were detected for LG11 between the two sexes, which may account for the indistinguishable morphologies of sex chromosomes in *C. farreri*. Interestingly, we found that one sex-related marker (i.e. f93422) was situated at the transcription factor gene, *ZNFX1*, which has been shown to be tightly linked with the sex-determination gene *AMHR2* in the tiger pufferfish.[36] Unfortunately, we do not have direct evidence to support the physical linkage between *ZNFX1* and *AMHR2* based on the available *C. farreri* genome information. Furthermore, searching another two molluscan genomes, i.e. the Pacific oyster *C. gigas* (http://gigadb.org/pacific_oyster) and the owl limpet *Lottia gigantea* (http://metazoa.ensembl.org/Lottia_gigantea), did not support the two genes' linkage either, though this does not necessarily exclude the possibility of their physical linkage because these genome assemblies are still largely fragmented (scaffold N50 is 401 kb for *C. gigas*[38] and 1.87 Mb for *L. gigantean*[46]). Our group has initiated the whole-genome sequencing project for *C. farreri*; therefore, this issue will be settled in the near future.

**Table 5.** QTL mapping of the PC1 of five growth traits in *C. farreri*

| Trait | LG | Confidence interval (cM) | Closest marker (position) | PV (%)[a] | LOD |
|---|---|---|---|---|---|
| PC1 | 1 | 55.13–70.68 | dm28438 (60.34),<br>df12539 (61.10) | 11.38 | 2.52 |
| | 3 | 34.14–61.82 | m7430 (48.70),<br>f17843 (48.72) | 16.89 | 3.76 |

[a]Phenotypic variance (PV) explained by a QTL.

### 4.4. *Integration of the linkage map with multiple genomic resources*

Recent studies have shown that RAD-based linkage maps can greatly improve large-scale genome assembly.[47,48] However, RAD tags are usually short in length (35–100 bp), thus limiting their direct use for integration with other important genomic resources generated from RNA-Seq, ChIP-Seq, exome sequencing or BAC–fosmid-end sequencing. Here, we propose that RAD tags can be effectively extended using the scaffolds obtained from genome survey sequencing. For *C. farreri*, although a majority of the scaffolds were not very long (scaffold N50 = 1.5 kb), they could already serve as 'long' surrogates for 2b-RAD tags to enhance their utility. We have demonstrated that this approach could facilitate the integration of the scallop linkage map with the BAC-based physical map, even though a large fraction of BAC contigs contained only a few BESs (e.g. 73% with less than three sequenced BAC clones). Besides, such approach can also facilitate the identification of genes surrounding a given tag, which can be crucial for the identification of quantitative trait genes in QTL mapping and association studies. Therefore, it can be envisioned that this scallop linkage map would serve as an important platform for unifying genomic resources that have been accumulated for *C. farreri*.

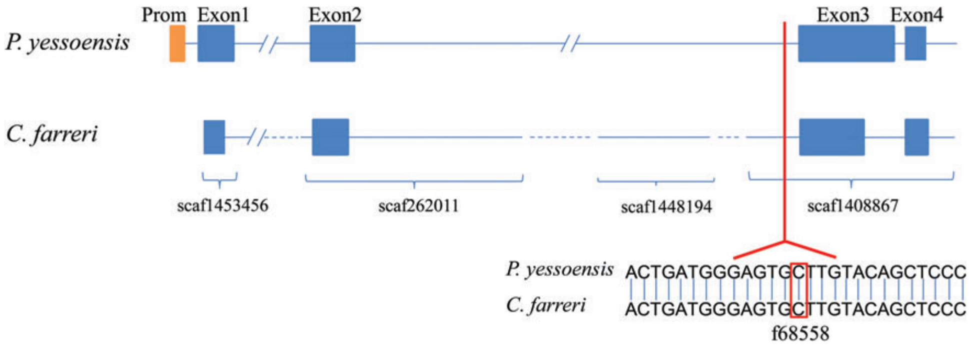

**Figure 7.** The gene structures of *PROP1* (Homeobox protein prophet of Pit-1) in two scallop species. For *P. yessoensis*, the full length of *PROP1* was obtained from an ongoing genome sequencing project. For *C. farreri*, homologous genomic scaffolds were shown. The marker f68558 found in the confidence interval of a growth-related QTL detected on LG3 was located at the intron region of this transcription factor gene. The marker sequences were shown for the two species with the polymorphic locus in *C. farreri* indicated.

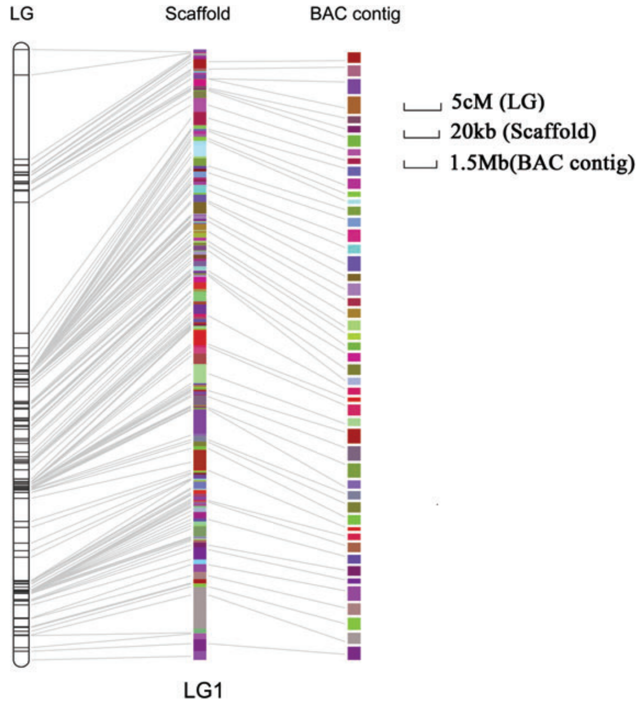

**Figure 8.** Schematic demonstration of the integration of linkage group LG1, genomic scaffolds and a BAC-based physical map.

### 4.5. *Conclusions*

We generated a preliminary reference genome for a bivalve mollusc, expanding the genome resources currently available for bivalves—a group of animals that remain largely unexplored in terms of genome sequencing. A high-resolution linkage map was constructed for *C. farreri* with a marker density that has, to our knowledge, never been achieved in any other molluscs. Two growth-related QTLs and one putative sex-determination region were detected, and candidate genes identified from these QTL regions represent important targets for further evaluation. Integration of the genome map, linkage map and physical map exemplifies how to build up a comprehensive genomic framework in a non-model organism.

**Acknowledgements:** We thank Dr Xiaojun Zhang (Institute of Oceanology, Chinese Academy of Sciences) for kindly providing the physical map data.

**Supplementary Data:** Supplementary Data are available at www.dnaresearch.oxfordjournals.org.

## Funding

This work was supported by the National Natural Science Foundation of China (31130054), National Basic Research Program of China (2010CB126402), National High Technology Research and Development Program of China (2012AA10A402, 2012AA10A 405 and 2012AA092204) and Natural Science Foundation for Distinguished Young Scholars of Shandong Province (2013JQE29001).

## References

1. Davey, J.W. and Blaxter, M.L. 2011, RADSeq: next-generation population genetics, *Brief Funct. Genomics*, **9**, 416–23.
2. Wang, S., Meyer, E., McKay, J. and Matz, M.V. 2012, 2b-RAD: a simple and flexible method for genome-wide genotyping, Nat. Methods, **9**, 808–10.
3. Peterson, B.K., Weber, J.N., Kay, E.H., Fisher, H.S. and Hoekstra, H.E. 2012, Double digest RADseq: an inexpensive method for *De Novo* SNP discovery and genotyping in model and non-model species, *PLoS One*, **7**, e37135.
4. Appeltans, W., Ahyong, S.T., Anderson, G., et al. 2012, The magnitude of global marine species diversity, *Curr. Biol.*, **22**, 2189–202.
5. FAO. 2010, *The State of World Fisheries and Aquaculture 2010*. FAO Fisheries and Aquaculture Department, FAO: Rome.
6. Hou, R., Bao, Z., Wang, S., et al. 2011, Transcriptome sequencing and de novo analysis for Yesso scallop (*Patinopecten yessoensis*) using 454 GS FLX, *PLoS One*, **6**, e21560.
7. Wang, S., Hou, R., Bao, Z., et al. 2013, Transcriptome sequencing of Zhikong scallop (*Chlamys farreri*) and comparative transcriptomic analysis with Yesso scallop (*Patinopecten yessoensis*), *PLoS One*, **8**, e63927.
8. Clark, M.S., Thorne, M.A., Vieira, F.A., Cardoso, J.C., Power, D.M. and Peck, L.S. 2010, Insights into shell deposition in the Antarctic bivalve *Laternula elliptica*: gene discovery in the mantle transcriptome using 454 pyrosequencing, *BMC Genomics*, **11**, 362.
9. Meyer, E. and Manahan, D.T. 2010, Gene expression profiling of genetically determined growth variation in bivalve larvae (*Crassostrea gigas*), *J. Exp. Biol.*, **213**, 749–58.
10. Ghiselli, F., Milani, L., Chang, P.L., et al. 2011, *De novo* assembly of the Manila clam *Ruditapes philippinarum* transcriptome provides new insights into expression bias, mitochondrial doubly uniparental inheritance and sex determination, *Mol. Biol. Evol.*, **29**, 771–86.
11. Guo, H., Bao, Z., Li, J., et al. 2012, Molecular characterization of TGF-β type I receptor gene (Tgfbr1) in *Chlamys farreri*, and the association of allelic variants with growth traits, *PLoS One*, **7**, e51005.
12. Wang, C., You, Y., Wang, H. and Liu, B. 2012, Genetic diversity of the sulfotransferase-like gene and one nonsynonymous SNP associated with growth traits of clam, *Meretrix meretrix*. *Mol. Biol. Rep.*, **39**, 1323–31.
13. Breton, S., Stewart, D.T., Shepardson, S., et al. 2011, Novel protein genes in animal mtDNA: a new sex determination system in freshwater mussels (Bivalvia: Unionoida)? *Mol. Biol. Evol.*, **28**, 1645–59.

14. Wang, S., Bao, Z., Pan, J., et al. 2004, AFLP linkage map of an intraspecific cross in *Chlamys farreri*, *J. Shellfish Res.*, **23**, 491–9.
15. Li, L., Xiang, J.H., Liu, X., Zhang, Y., Dong, B. and Zhang, X.J. 2005, Construction of AFLP-based genetic linkage map for Zhikong scallop, *Chlamys farreri* Jones et Preston and mapping of sex-linked markers, *Aquaculture*, **245**, 63–73.
16. Wang, L., Song, L., Chang, Y., Xu, W., Ni, D. and Guo, X. 2005, A preliminary genetic map of Zhikong scallop (*Chlamys farreri* Jones et Preston 1904), *Aquaculture Res.*, **36**, 643–53.
17. Zhan, A., Hu, J., Hu, X., et al. 2009, Construction of microsatellite-based linkage maps and identification of size-related quantitative trait loci for Zhikong scallop (*Chlamys farreri*), *Anim. Genet.*, **40**, 821–31.
18. Zhang, X., Zhao, C., Huang, C., et al. 2011, A BAC-based physical map of Zhikong scallop (*Chlamys farreri* Jones et Preston), *PLoS One*, **6**, e27612.
19. Zhang, L., Bao, Z., Wang, S., Hu, X. and Hu, J. 2008, FISH mapping and identification of Zhikong scallop (*Chlamys farreri*) chromosomes, *Mar. Biotechnol.*, **10**, 151–7.
20. Zhang, L., Bao, Z., Cheng, J., et al. 2007, Fosmid library construction and initial analysis of end sequences in Zhikong scallop (*Chlamys farreri*), *Mar. Biotechnol.*, **9**, 606–12.
21. Zhang, Y., Zhang, X., Scheuring, C.F., et al. 2008, Construction and characterization of two bacterial artificial chromosome libraries of Zhikong scallop, *Chlamys farreri* Jones et Preston, and identification of BAC clones containing the genes involved in its innate immune system, *Mar. Biotechnol.*, **10**, 358–65.
22. Wang, L., Song, L., Zhao, J., et al. 2009, Expressed sequence tags from the Zhikong scallop (*Chlamys farreri*): discovery and annotation of host-defense genes, *Fish Shellfish Immunol.*, **26**, 744–50.
23. Maniatis, T., Fritsch, E.F. and Sambrook, J. 1989, *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY.
24. Wu, T.D. and Watanabe, C.K. 2005, GMAP: a genomic mapping and alignment program for mRNA and EST sequences, *Bioinformatics*, **14**, 1859–75.
25. Stam, P. 1993, Construction of integrated genetic linkage maps by means of a new computer package: Joinmap, *Plant J.*, **3**, 739–44.
26. Wu, Y., Close, T.J. and Lonardi, S. 2011, Accurate construction of consensus genetic maps via integer linear programming, *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **8**, 381–94.
27. Sorensen, P. and Thomsen, H. 2003, *Using DMU for QTL mapping*. Aarhus University: Foulum, Denmark.
28. Piepho, H.P. 2001, A quick method for computing approximate thresholds for quantitative trait loci detection, *Genetics*, **157**, 425–32.
29. Li, H. 2011, A quick method to calculate QTL confidence interval, *J. Genet.*, **90**, 355–60.
30. Chen, M.H. and Yang, Q. 2010, GWAF: an R package for genome-wide association analyses with family data, *Bioinformatics*, **26**, 580–1.
31. Li, R., Yu, C., Li, Y., et al. 2009, SOAP2: an improved ultrafast tool for short read alignment, *Bioinformatics*, **25**, 1966–7.
32. Wang, M.L., Zheng, J.S. and Yu, H. 1990, The karyotyping of *C. farreri*, *J. Ocean Univ. Qingdao*, **20**, 81–5.
33. Chakravarti, A., Lasher, L.K. and Reefer, J.E. 1991, A maximum likelihood method for estimating genome length using genetic linkage data, *Genetics*, **128**, 175–82.
34. Fishman, L., Kelly, A.J., Morgan, E. and Willis, J.H. 2001, A genetic map in the *Mimulus guttatus* species complex reveals transmission ratio distortion due to heterospecific interactions, *Genetics*, **159**, 1701–16.
35. Wu, W., Cogan, J.D., Pfaffle, R.W., et al. 1998, Mutations in PROP1 cause familial combined pituitary hormone deficiency, *Nat. Genet.*, **18**, 147–9.
36. Kamiya, T., Kai, W., Tasumi, S., et al. 2012, A trans-species missense SNP in *Amhr2* is associated with sex determination in the tiger pufferfish, *Takifugu rubripes* (fugu), *PLoS Genet.*, **8**, e1002798.
37. Takeuchi, T., Kawashima, T., Koyanagi, R., et al. 2012, Draft genome of the pearl oyster *Pinctada fucata*: a platform for understanding bivalve biology, *DNA Res.*, **19**, 117–30.
38. Zhang, G., Fang, X., Guo, X., et al. 2012, The oyster genome reveals stress adaptation and complexity of shell formation, *Nature*, **490**, 49–54.
39. Sauvage, C., Bierne, N., Lapègue, S. and Boudry, P. 2007, Single nucleotide polymorphisms and their relationship to codon usage bias in the Pacific oyster *Crassostrea gigas*, *Gene*, **406**, 13–22.
40. Thomas, M.G., Enns, R.M., Shirley, K.L., Garcia, M.D., Garrett, A.J. and Silver, G.A. 2007, Associations of DNA polymorphisms in growth hormone and its transcriptional regulators with growth and carcass traits in two populations of *Brangus bulls*, *Genet. Mol. Res.*, **6**, 222–37.
41. Lan, X., Pan, C., Zhang, L., Zhao, M., Zhang, C., Lei, C. and Chen, H. 2009, A novel missense (A79V) mutation of goat *PROP1* gene and its association with production traits, *Mol. Biol. Rep.*, **36**, 2069–73.
42. Xu, T.S., Liu, J.B., Yao, D.W., Cai, H.F., Chen, H., Zhou, H.L., et al. 2010, The prophet of PIT1 gene variation and its effect on growth traits in Chinese indigeous goat, *J. Anim. Vet. Adv.*, **9**, 2940–6.
43. Zeng, X.C., Chen, H.Y., Jia, B., Zhao, Z.S., Hui, W.Q., Wang, Z.B., et al. 2011, Identification of SNPs within the sheep *PROP1* gene and their effects on wool traits, *Mol. Biol. Rep.*, **38**, 2723–8.
44. Dasen, J.S. and Rosenfeld, M.G. 2001, Signaling and transcriptional mechanisms in pituitary development, *Annu. Rev. Neurosci.*, **24**, 327–55.
45. Basoa, E., Alfonsi, C., Perez, J.E. and Cequea, H. 2000, Karyotypes on the scallops *Euvola ziczac* and *Nodipecten nodosus* from the Gulf of Cariaco, Sucre State, Venezuela, *Bol. Inst. Oceanogr. Venez.*, **39**, 49–54.
46. Simakov, O., Marletaz, F., Cho, S.J., et al. 2013, Insights into bilaterian evolution from three spiralian genomes, *Nature*, **493**, 526–31.
47. Heliconius Genome Consortium. 2012, Butterfly genome reveals promiscuous exchange of mimicry adaptations among species, *Nature*, **487**, 94–8.
48. Catchen, J.M., Amores, A., Hohenlohe, P., Cresko, W. and Postlethwait, J. 2011, Stacks: building and genotyping loci *de novo* from short-read sequences, *G3 (Bethesda)*, **1**, 171–82.