論文の内容の要旨

水圏生物科学専攻

平成 27 年度博士課程入学

氏名 張 翔

指導教員 浅川 修一

論文題目　Construction of ultrahigh density linkage map by low-coverage whole genome sequencing of doubled haploid population: A case study in torafugu (*Takifugu rubripes*)

（ダブルハプロイド集団に対する低カバー率の全ゲノムシーケンシングによる超高分解能連鎖地図の作成：トラフグ（*Takifugu rubripes*）を用いたケーススタディ）

A genetic linkage map is a powerful tool in genetic and genomic research. It lays a strong foundation for comparative genomics and provides vital clues toward understanding genome evolution and divergence. It facilitates genotype–phenotype association mapping and enables investigating the genetics of complex phenotypic traits. It also contributes toward characterization of genome structure and serves as the backbone for anchoring unplaced/misplaced scaffolds for chromosome-scale assembly.

During linkage map construction, or called linkage mapping, producing mapping populations and keeping inbred lines are normally very time-consuming and laborious work lasting several months even to several years. In the meanwhile, more and more genetic markers are also required for more accurately and thoroughly identification of genetic polymorphisms of each individual for linkage mapping.

Owing to the rapid development of next-generation sequencing (NGS) in the last decade, the ability to simultaneously sequence a large number of individuals in a multiplex manner has now become possible, so that an entire population can be rapidly genotyped for linkage mapping. Since 2010, low-coverage whole-genome resequencing has been employed for constructing the genetic linkage maps of rice, shiitake mushroom, and safflower. However, all of these cases relied on the

availability of high-quality reference genome sequences and/or designed inbred pedigree lines to carry out the linkage mapping prior to resequencing of the mapping population. This requirement currently limits the wide application of low-coverage whole-genome resequencing strategy in non-model organisms, especially for those with unexplored genomes.

Torafugu (*Takifugu rubripes*) is a popular species with economic importance in the waters of East Asia, and has emerged as an ideal model in genomic studies owing to its compact genome. In fact, the torafugu genome is considered to be one of the smallest (~400 Mb) among vertebrates and is approximately eight times smaller than the human genome. Another advantage of torafugu as a model for genetic analysis is its similarity to mammals, including a shared body plan and physiological systems. Thus, the compact genome can favor the discovery of genes and gene regulatory regions with clear counterparts in the human genome, and the torafugu genome can further serve as a reference to understand the structure, function, and evolution of vertebrate genomes. In the most recent fifth version of the torafugu genome assembly (FUGU5), 72% of the scaffolds have been located and oriented after integration with the genetic linkage map of torafugu comprising 1,220 microsatellite markers. Therefore, the construction of a higher-density genetic linkage map of torafugu is needed to be able to expand the contiguity and improve the quality of the genome assembly.

Here, we chose torafugu as the test model to develop an effective strategy for ultrahigh-density genetic linkage map construction using low-coverage whole-genome sequencing without requiring a high-quality reference genome and the laborious establishment of inbred lines. An ultrahigh-density genetic linkage map should be also in need to improve the quality of FUGU5.

Preparation of torafugu DH population for low-coverage sequencing were described in **Chapter 2**. In recent years, the H/DH population has been exploited as an ideal population type for genetic linkage map construction, particularly in plants and teleosts due to their well-developed H/DH technologies. A wild female torafugu was subjected to mito-gynogenesis for generating a DH population. In brief, mature oocytes were fertilized with inactive sperm of a male torafugu that had been pretreated with ultraviolet radiation. After fertilization, the eggs were subjected to cold-shock treatment, followed by incubation in aerated tanks with fresh seawater. Several days after artificial insemination, hundreds of eggs were observed to contain embryonic bodies, which were selected for further analysis. We performed low-coverage whole-genome sequencing of the DH population. Genomic DNA was extracted from the selected 192 eggs after homogenization. DNA libraries of these individuals were prepared and barcoded, and were then subjected to next generation sequencing using the Illumina systems. After quality control and removing the 23 samples with very low sequencing coverage, a total of 71.32 of sequencing data, consisting of 2 × 100-bp paired-end reads with an average insert size of 230 bp, were obtained from 169 samples of the generated DH torafugu population.

Subsequently, the details about de novo assembly and SNP discovery were described in **Chapter 3**. After removing 4 samples with partial heterozygosity, a total of 69.58 Gb of sequencing data from 165 samples were obtained. According to the torafugu genome size (approximately 400 Mb), the total sequencing data coverage was estimated at 174, whereas the average coverage for each sample was 1.05 ± 0.76. After performing de novo assembly on the SOAPdenovo2 assembler under a k-mer value of 58 using the sequencing data, a relative low-quality assembly of a total size of 356.59 Mb and N50 size of 22,235 bp was generated, which was composed of 54,127 scaffolds with the length ranging from 200 to 264,568 bp. The sequencing data of the above 165 samples were mapped to the obtained de novo assembly, respectively. SNPs of each sample were called and used for genotype scoring. After SNPs calling from the sequencing data of each DH individual, a total of 1,070,601 SNPs were discovered in the DH population, using the above *de novo* assembly as the reference. Owing to the low-coverage (≈1×) whole genome sequencing of each individual, a low-resolution SNP dataset was obtained.

In **Chapter 4**, genetic marker genotyping was the focus. The main advantage of H/DH individuals for genotyping is that a relatively low sequencing coverage is sufficient without loss of accuracy compared to the coverage necessary for sequencing more common diploid individuals owing to the presence of heterozygous single nucleotide polymorphisms (SNPs). Although the homozygosity of DH individual enables more accurately SNP genotyping, most of the SNPs in each sample were detected once or less, which led to a large quantity of missed genotypes and insufficient data for genotyping calibration because of the low-coverage (≈1×) sequencing of each sample. The obtained low-call-rate SNP dataset with unknown linkage phase was not suitable for linkage map construction. However, based on the above *de novo* assembly, the genotype of adjacent SNPs could be testified by each other. Therefore, the information of SNPs located on each segment was combined and a genotype was assigned to each segment as a new genetic marker termed the short segment genotype (SSG). The maximum segment length of the SSGs was set to 8 kb, so that the SSGs (>0.9 call rate) could harbor as much genome-wide SNPs information as possible while maintaining the length of segments as short as possible. This low-call-rate SNPs dataset was then converted into a high-call-rate SSGs dataset despite of the unknown linkage phase. After 0.9-call rate filtering, 37,398 SSGs containing information of 833,594 SNPs were retained and used for further linkage map construction.

As depicted in **Chapter 5**, an ultrahigh-density linkage map of torafugu was constructed using above high-call-rate SSGs dataset based on method for phase-unknown linkage mapping for DH population. The map consists of 37,343 SSGs in 3,090 unique positions, containing the information of 802,277 SNPs (74.9% of total SNPs). The genetic linkage map contained 22 linkage groups, consistent with the number of chromosomes of the torafugu haploid genome. The genetic distances ranged from 62.75 cM of linkage group (LG)10 to 198.25 cM of LG1, with a total length of 2,319.65 cM. Based on the unique marker positions, the estimated marker intervals ranged from 0.70 cM/marker in LG22 to 0.79 cM/marker in LG1, with an average marker interval of 0.75 cM/marker on the genetic linkage map. The accuracy of the present linkage map was verified by the following analyses. The recombination fractions were considerably low between adjacent markers of each linkage group, indicating a low recombination frequency between them, whereas the logarithm of the odds (LOD) scores between adjacent markers of each linkage group were high, indicating strong linkage between them. The 22 linkage groups appeared to be distinctly clustered. Comparative analyses between linkage map and the latest published genome FUGU5 were carried out. The sequence information of the 37,343 SSGs of the linkage map obtained with the proposed strategy was subjected to BLASTN analyses against the latest published genome FUGU5. Overall, 31,822 SSGs could be mapped to the 22 chromosomes of FUGU5. The results indicated near-perfect concordance between the genetic and physical position of each matched SSG, and 5,521 SSGs could be mapped to the 1,583 (65.97 Mb) unassembled scaffolds, which suggested the regions where FUGU5 can be improved. Furthermore, 180 of these scaffolds (28.2 Mb) contained more than one unique genetic position, suggesting that they might be located on the chromosomes with direction. The flanking sequences of the SNPs contained in this genetic linkage map were also well aligned to

FUGU5, and 532,424 SNPs mapped to the 22 chromosomes of the genome. The results reflected more detailed collinearity between the orders of the SNPs of each linkage group and each chromosome. The results also indicated possible mis-assembled regions or large segmental polymorphisms in chromosome 2, 3, 4, 5, 6, 7, 11, 12, 17, 19, and 20 of FUGU5.

In general, we successfully developed an effective strategy for the construction of an ultrahigh-density genetic linkage map of torafugu based on low-coverage (~1 ×) whole-genome sequencing of each individual of a DH population generated through mito-gynogenesis. The accuracy of the present linkage map was verified by subsequent analyses of recombination fractions and assessment of LOD scores for all marker pairs, along with comparative analyses between the linkage map and FUGU5. In addition, integration with the present linkage map allowed for validation and further refinement of FUGU5. Based on these indicators, an improved genome assembly of torafugu will be achieved in our future work.

The present strategy has significant advantage of time and labor saving because we do not have to keep inbred lines which are time-consuming and laborious. Although the lacking of inbred lines leads to unknown linkage phases which would bring obstacles for the subsequent analyses and strategy design, we successfully designed a strategy for linkage mapping for a low-call-rate SNP dataset with linkage phase-unknown format.

The previous low-coverage resequencing strategy for ultrahigh-density linkage mapping requires a high-quality reference genome, which limits its application. In contrary, the lack of a requirement of a high-quality reference genome for low-coverage whole-genome sequencing expands the application of our proposed strategy to a wide range of non-model and unsequenced species. Furthermore, based on our strategy, de novo assembly, linkage map construction, and further chromosome-scale assembly can be efficiently completed for unsequenced species.

Our approach has an advantage of simplicity compared to the current techniques such as restriction-site-associated DNA sequencing (RAD-seq) which demands complicated processes for sequencing a small target portion of the whole genome. Our strategy can capture most of the SNPs of the population, which would allow for thoroughly characterizing complex genomes, whereas RAD-seq is only able to call a small portion of SNPs.

The strategy of the present study was developed for ultrahigh-density genetic linkage map construction based on an H/DH dataset. Therefore, it can be applied on species which can produce large quantity of H/DH individuals. In our future work, the application range of our strategy may be further extended using only single gamete cells with the combination of single-cell-sequencing technology platform.

Keywords: ultrahigh-density genetic linkage map construction; effective strategy; low-coverage whole genome sequencing; doubled haploid population; *Takifugu rubripes*; unknown linkage phase