博士論文

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**)**

ZHANG XIANG

張 翔

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 Thesis submitted to

Graduate School of Agriculture and Life Sciences

The University of Tokyo

in partial fulfillment of the requirement for the degree of

Doctor of Philosophy

In

Department of Aquatic Bioscience

By

ZHANG XIANG

張 翔

March 2018

DECLARATION

I, Zhang Xiang, hereby declare that the thesis entitled “Construction of ultrahigh-density linkage map by low-coverage whole genome sequencing of doubled haploid population: a case study in torafugu (*Takifugu rubripes*)” is an authentic record of the work done by me and that no part thereof has been presented for the award of any degree, diploma, associateship, fellowship or any other similar title.

March 2018

Zhang Xiang

Laboratory of Aquatic Molecular Biology and Biotechnology,

Department of Aquatic Bioscience,

Graduate School of Agricultural and Life Sciences,

The University of Tokyo

ACKNOWLEGEMENT

Doing Doctor Course in The University of Tokyo is such a delight experience in my life. During the past three years, I have met so many wonderful persons and I have learned so much about doing research and critical thinking. I am so pleased and surprised in the way I become.

I would like to thank Professor Asakawa Shuichi for offering me the opportunity to carry out this interesting PhD project under his supervision. I sincerely appreciate for all his patience, guidance, and full support as always.

I would like to thank all current members of Laboratory of Aquatic Molecular Biology and Biotechnology for all their kind helps. More specifically, I would like to express my gratitude to Dr. Misako Nakaya, for her helps with ordering regents and experimental materials. I would also like to show my greatest appreciation to Ms. Takase Mai who is very kind, patient, willing to offer help in both life and academics.

Finally, my deepest gratitude goes to my parents and my soulmate Wen Ya who are supportive in my whole life. I could barely finish my Doctor Course in UT without their caring and financial support. I would like to dedicate this dissertation to them for their endless love.

CONTENTS

ABBREVIATIONS

AFLPs: amplified fragment length polymorphisms

DH: doubled haploid

CHR: chromosome

cM: CentiMorgan

GBS: genotyping-by-sequencing

H: haploid

LG: linkage group

LOD: logarithm of the odds

NGS: next-generation sequencing

QTL: quantitative trait locus

RAD-seq: restriction-site-associated DNA sequencing

RF: recombination fraction

RFLPs: restriction fragment length polymorphisms

RRLs: reduced-representation libraries

SLAF-seq: specific-locus amplified fragment sequencing

SNPs: single nucleotide polymorphisms

SSG: short segment genotype

STSs: sequence tagged sites:

List of Figures

List of Tables

Abstract

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

1 General Introduction

1.1 Genetic linkage map and its applications

A genetic linkage map, also known as linkage map and genetic map, is a list of genetic elements in order according to their co-segregation patterns of chromosomal crossover or recombination occurring in meiosis (**Figure 1.1**). In theory, the larger of the physical distance between two genetic elements on a chromosome, the higher of the chances of recombination occurring between them. Those elements can be inferred as genes underlying phenotypic characteristics or genetic markers. Therefore, genetic linkage maps can be considered as ordered sets of genetic markers representing "milestones" or "landmarks" along the whole or part of chromosomes. (Cheema and Dicks, 2009) (Collard *et al.*, 2005). Currently, genetic linkage maps are becoming the powerful tools in genetic and genomic research, helping with our understanding of the structural, functional, and evolutionary characteristics of the genome (Rastas *et al.*, 2016). For more detailly and thoroughly characterizing the complex genomes, more and more genetic markers are required to be contained in linkage maps.

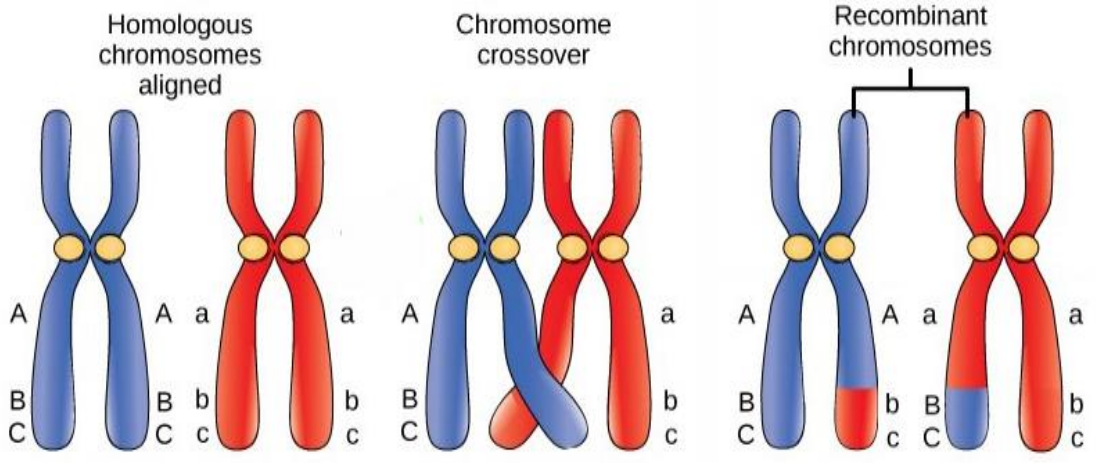


Figure 1.1. The process of crossover or recombination. **The smaller distance between two allelic genetic markers on a chromosome, the smaller is the chance of a crossover occurring between them, and the more likely they are to be inherited together (e.g. marker B and C). Adapted**

from (https://commons.wikimedia.org/wiki/File:Figure_11_01_02.jpg)

Linkage maps are important tools in evolutionary genetics and in the studying field of speciation. Linkage maps lay a strong foundation for comparative genomics and the analysis of synteny across species, providing vital clues toward understanding genome evolution and divergence. For example, a genetic map of the platyfish (*Xiphophorus maculatus*) has been constructed and used to perform comparative analyses with genetic maps or physical maps of other teleosts by Amores et al. These analyses not only help with revealing mechanisms of karyotype evolution among teleost fish, but also facilitating in characterizing chromosome fusion events, distinguishing lineage-independent chromosome fusions, and revealing the stability of syntenies and gene orders in chromosomes of teleosts over hundreds of millions of years (Amores *et al.*, 2014). Likewise, a genetic linkage map of sole (*Solea solea*) has been constructed and suggested itself a tool for evolutionary and comparative analyses among exploited (flat) fishes. Sequences information of the map was aligned with the genomes of four model fish species, medaka (*Oryzias latipes*), Nile tilapia (*Oreochromis niloticus*), three-spined stickleback (*Gasterosteus aculeatus*) and green spotted pufferfish (*Tetraodon nigroviridis*). These comparisons revealed multiple conserved synteny regions with all four species. The map also has been compared to the linkage map of turbot (*Scophthalmus maximus*), another commercially important flatfish species closely related to sole. As the comparisons among related species, the higher degree of synteny has been confirmed between sole and turbot (Diopere *et al.*, 2014).

Linkage maps are the primary approach in genotype–phenotype association studies, further enabling the investigation of the genetics of complex phenotypic traits. So far, lots of genes or quantitative trait locus (QTLs) related to complex and economic trait have been identified, such as grain shape and size of wheat (Wu *et al.*, 2015), male size and color variation of guppy (Tripathi *et al.*, 2009), fruit-related traits of pear (Wu *et al.*, 2014), cold tolerance of common carp (Sun and Liang, 2004), and sexual and growth-related traits of half-smooth tongue sole (Song *et al.*, 2012). Linkage studies

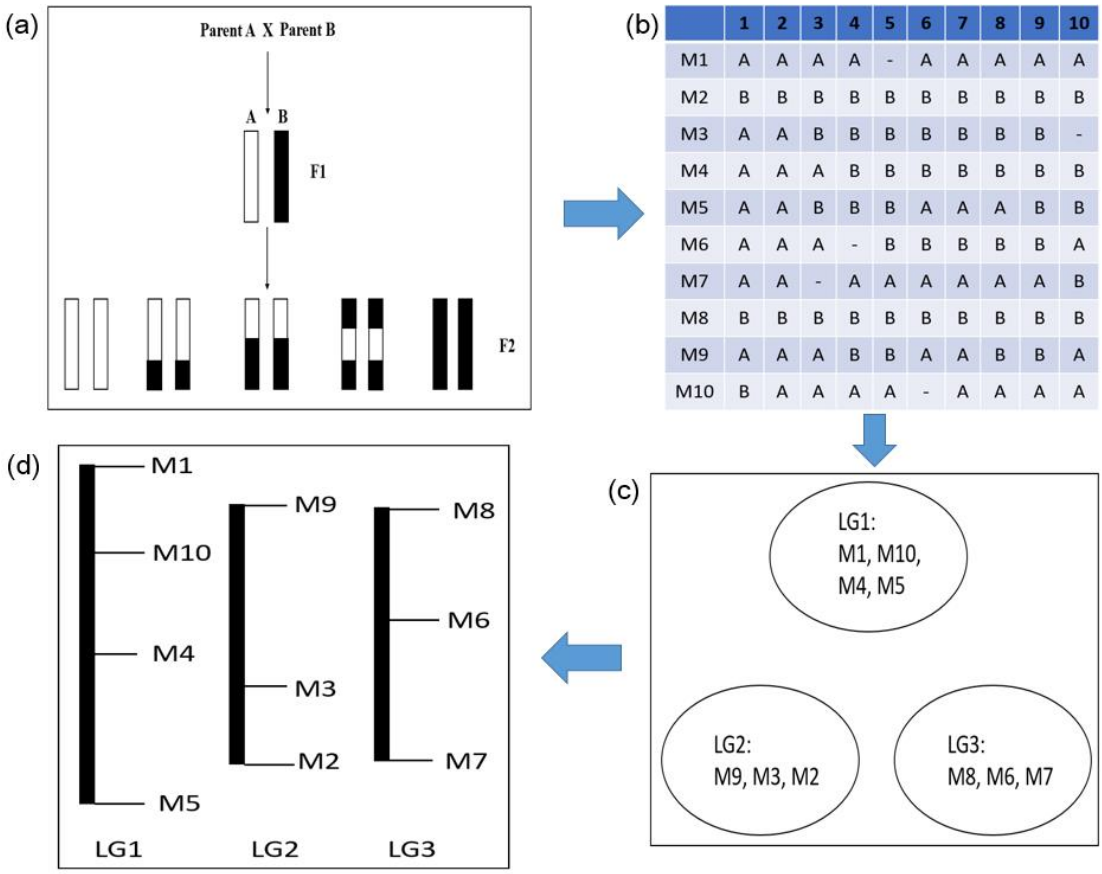
also have led to the identification of genes causing or substantially increasing the risk of many diseases and birth defects. For example, genes that cause many Mendelian disorders such as Huntington disease (Gusella *et al.*, 1983) and cystic fibrosis (Tsui *et al.*, 1985) have been identified via linkage analyses, so as genes with risk alleles that cause increases in the risk for complex diseases, including the BRCA1 and BRCA2 genes with breast cancer (Marcus *et al.*, 1996), and NOD2 gene with Crohn's disease (Hugot *et al.*, 2001).

Linkage maps also contribute toward characterization of genome structure and in turn, serving as the backbones for anchoring unplaced/misplaced scaffolds for chromosome-scale assembly. Current methods of DNA sequence *de novo* assembly could produce contigs and scaffolds of limited length, which results in that many of these assemblies contain thousands of individual sequences without information of how these pieces are assembled into chromosomes (Fierst, 2015). A high-quality linkage map of densely spaced markers is valuable for affirming the correct placement of scaffolds on the chromosomes and for "proofing" the order of markers within the assemblies (Lewin *et al.*, 2009). Therefore, currently, almost all chromosome-scale assemblies are constructed after integration with a linkage map, such as the latest published chromosome-scale genome sequences of zebrafish (Howe *et al.*, 2013), barley (International Barley Genome Sequencing *et al.*, 2012), grass carp (Wang *et al.*, 2015b), Atlantic salmon (Lien *et al.*, 2016) and Japanese flounder (Shao *et al.*, 2017).

1.2 Workflow of linkage map construction

In general, there are four main steps of linkage map construction **(Figure 1.2)** including: (1) crossing experiment, (2) identification of genetic polymorphisms, (3) genetic markers grouping and (4) genetic markers ordering (Cheema and Dicks, 2009). Normally, the first step is a time-consuming and laborious process to produce a mapping population. The second step is responsible for generating the dataset for linkage map construction. The possible errors and missing data of the dataset would have large impact on the accuracy of the constructed linkage map. Therefore, the second step is a

critical step. Nowadays, the third and fourth steps can be carried out by several software tools, such as JoinMap (Stam, 1993), MSTmap (Wu *et al.*, 2008), ASMap (Taylor and Butler, 2017) and Lep-MAP (Rastas *et al.*, 2013). Except JoinMap, other tools also can handle large SNPs dataset produced by high-throughput sequencing. These softwares also are integrated with the map modification processes.



	1	2	3	4	5	6	7	8	9	10
M1	A	A	A	A	-	A	A	A	A	A
M2	B	B	B	B	B	B	B	B	B	B
M3	A	A	B	B	B	B	B	B	B	-
M4	A	A	A	B	B	B	B	B	B	B
M5	A	A	B	B	B	A	A	A	B	B
M6	A	A	A	-	B	B	B	B	B	A
M7	A	A	-	A	A	A	A	A	A	B
M8	B	B	B	B	B	B	B	B	B	B
M9	A	A	A	B	B	A	A	B	B	A
M10	B	A	A	A	A	-	A	A	A	A

Figure 1.2 The general process of genetic map construction. (a) Crossing experiment; (b) Identification of polymorphism; (c) Genetic markers grouping; (d) Genetic markers ordering.

1.2.1 Production of mapping populations via crossing experiment

A genetic linkage map is based on the principle of loci co-segregation pattern during meiosis. However, in most situations, the recombined gametes produced by meiosis are not directly observable. Therefore, designed and labor-consuming mating schemes or crossing experiments for the production of mapping populations and keeping inbred lines lasting several months even to several years are normally required

to infer the number of recombined gametes, which is the key to success in linkage mapping. The population size used in genetic mapping studies varied from 50 to 250 individuals (Mohan *et al.*, 1997), and a larger population size is in need for higher resolution mapping.

Different designs of crossing experiment would generate different types of mapping populations possessing certain characteristics, which must take into account along with the linkage estimation procedure. Several commonly used populations can be utilized for linkage mapping (Collard *et al.*, 2005). F_2 populations derived from F_1 hybrids is one of the simplest types of mapping populations by selfing and sibling mating. The main advantage of F_2 populations is the easily generating of them. Recombinant inbred lines (RILs) can be constructed after inbreeding from F_2 individuals (Broman, 2005). The major disadvantage of producing RILs is the long period requiring six to eight generations in general, whereas the major advantage of RILs is the "true-breeding" lines that can be multiplied and reproduced for eternal propagating in different laboratories for further linkage analyses and the addition of genetic markers to existing maps, which ensures the identical material to all investigators (Broman, 2005). Backcrossing (BC) population, also derived from crossing the F_1 hybrid to one of the parents, is another useful mapping population (Song *et al.*, 2005). Haploid/Doubled haploid (H/DH) populations can be generated via natural or artificial uniparental reproduction, which are found in a wide range of species in several kingdoms. Owing to the well-developed H/DH technologies in plants and teleosts, the H/DH population has become one of ideal mapping populations for genetic linkage map construction (Dunwell, 2010; Komen and Thorgaard, 2007). In addition, there are also other designs of crossing experiment in use despite of that not all software tools of genetic mapping are supportive.

1.2.2 Identification of genetic polymorphisms via genetic markers

The identification of genetic polymorphisms of each individual from a mapping population is the most important step for linkage map construction, which must base on

genetic markers. Genetic markers can represent genetic differences or polymorphisms among individual organisms or species. Generally, they act as ‘signs’ or ‘flags’ and occupy specific genomic positions within chromosomes referring to “loci” (singular “locus”) (Collard *et al.*, 2005).

Previously, classical genetic markers are morphological and biochemical markers according to the observed phenotypes or differences in enzymes by simple electrophoresis and specific staining. In fact, these phenotypes and differences can represent the associated genes. However, the major disadvantage is their vulnerability to environmental factors (Winter and Kahl, 1995).

Recently, traditional DNA (molecular) markers are the most widely and predominantly used markers due to their abundance. They arise from different classes of DNA mutations including substitution mutations (e.g. point mutations), rearrangements (e.g. insertions or deletions) or errors in replication of tandemly repeated DNA. These mutations can be identified by gel electrophoresis and staining with chemicals (e.g. ethidium bromide or silver) or with radioactive and colorimetric probes (Collard *et al.*, 2005). Several types of DNA markers have been developed: Restriction fragment length polymorphisms (RFLPs) are discovered and identified using restriction enzymes that cleave the DNA only at precise “restriction sites” (Daiger *et al.*, 1982); Microsatellites or Simple Sequence Repeats (SSR) consist of tandemly repeated nucleotide sequences with length of 2 to 6 base-pairs (bp) (Powell *et al.*, 1996); Amplified fragment length polymorphisms (AFLPs are detections of DNA restriction fragments by means of PCR amplification (Dekker and Easteal, 1990); Sequence Tagged Sites (STSs) are DNA sequences occurring only once in a genome in a known position (Ruano *et al.*, 1990).

Currently, with the advent and development of next-generation sequencing (NGS), Single Nucleotide Polymorphisms (SNPs) are becoming the dominated genetic markers. SNPs are variations at the single-nucleotide level without changing the overall length of the DNA sequence in certain region. SNPs are the most highly abundant and largest number of genetic markers occurring throughout the genome. For instance, they are

present at the frequency of every 1000 bp in the human genome (Sachidanandam *et al.*, 2001). Most SNPs are distributed in non-coding regions without direct impacts on the phenotypes of an individual. However, some introduced mutations in expressed sequences or regions would influence gene expression (promoters, enhancers), and may induce changes in protein structure and function. These SNPs have the potential to detect functional genetic variations. The specific technologies of genome-wide genetic markers discovery and genotyping using NGS will be introduced in **section 1.3** of this thesis. Those technologies enable simultaneously and parallelly sequence and genotype a large number of individuals for genetic mapping (Davey *et al.*, 2011).

Genetic differences and polymorphisms of each individual can be more detailly characterize with higher abundance of genetic markers. Using the classical or traditional molecular markers, the linkage map could contain several to hundreds of markers (Bishop *et al.*, 1994; Doniskeller *et al.*, 1987; Gyapay *et al.*, 1994; Sun and Liang, 2004), whereas using SNPs combined with NGS technologies, the linkage map can contain not hundreds but thousands of markers across the genome (Davey *et al.*, 2011; Wang *et al.*, 2015a; Yang *et al.*, 2013). Nevertheless, more and more genetic markers are still required for covering and characterizing the entire genome, considering that normal vertebrate genomes contain million-scale SNPs.

The progenies will inherit a genetic marker from either of the parents. Therefore, after the polymorphism screening and genotyping of each individual of the mapping population, a dataset consisting of the genetic markers $\times$ individuals matrix can be obtained. The quality of genetic linkage map depends on the data that were used to construct it. Therefore, the dataset must be critically checked for all possible errors and missing data. Even a low frequency of typing errors would have a substantial impact on the order and length of a linkage map resulting in the introduction of a double recombination which is not occurring. On the other hand, individuals with too much missing data would result in very little information in the contribution for map estimation. The presence of missing values in the genetic markers dataset suggest the loss of information about the number of true recombination that has taken place along

the chromosomes. A simulation study has investigated the effects of typing errors and missing values on the construction of linkage maps. As it concluded, missing values had less effects than typing errors, but they reduced the number of correctly ordered markers and produced shorter map lengths (Semagn *et al.*, 2006a).

1.2.3 Genetic markers grouping

Genetic markers grouping is the third step dividing the genetic markers dataset into distinct linkage groups. In ideal situation, there is one-to-one correspondence between the linkage groups and chromosomes. So far, this step can be performed through several approaches and software tools according to different strategies. The strategy employed in MAPMAKER (Lander *et al.*, 1987) starts with calculation of the maximum likelihood distances of all two-point markers and their corresponding logarithm of the odds (LOD) scores. The linkage would be built between two markers when the LOD score is more than 3 and the inter-marker distance is less than 80 Haldane cM. In this strategy, the nearest neighbor locus and marker clusters or linkage groups are growing by adding each genetic marker one by one when it represents the lowest recombination value with the current markers of the linkage group.

The approach of MSTMAP (Wu *et al.*, 2008) is derived from graph theory. Firstly, it creates a complete graph including all genetic markers connecting each other with edge values weighted by two-point function of the data. Then, all edge values more than a certain threshold are chopped, leading to several distinct subgraphs which are corresponding to linkage groups. It is notable the output would be influenced by the input parameter specified by the users when using these grouping methods (Cheema and Dicks, 2009).

1.2.4 Generic markers ordering

Markers ordering is the fourth step aiming to find the relative orders and genetic distances (in cM) of the genetic markers within each linkage group. Historically, several scoring functions have been established to find the orders such as minimum sum of

adjacent recombination fractions (Falk, 1989), maximum sum of adjacent LOD scores (Weeks and Lange, 1987), the minimum number of crossovers (Thompson, 1987; Van Os *et al.*, 2005), the product of adjacent recombination fractions (Wilson, 1988), the minimum entropy (Kozik and Michelmore, 2006), the minimum weighted least squares marker order (Stam, 1993), the maximum likelihood (ML) (Jansen *et al.*, 2001; Lander *et al.*, 1987) and the maximum number of fully informative meiosis (Edwards, 1971). Consequently, the relevant software tools would facilitate with this step such as JoinMap (Stam, 1993), MSTmap (Wu *et al.*, 2008), ASMap (Taylor and Butler, 2017) and Lep-MAP (Rastas *et al.*, 2013).

1.3 The discovery of genetic markers using NGS

The ideal molecular approach for population genomics should not only cover hundreds of polymorphic genetic markers but cover the entire genome by a single, simple and reliable experiment (Luikart *et al.*, 2003). Nowadays, with the advent of NGS, several approaches are developed to enable the discovering and genotyping of not hundreds but tens of thousands of genetic markers, which have made substantial contributions to current genetic and genomic studies.

1.3.1 Restriction-enzyme-based methods

For accurately genotyping, it is too expensive to sequence the whole genome of a population at high coverage. There are emerging many NGS-based methods for genetic markers discovery such as reduced-representation libraries (RRLs) (Altshuler *et al.*, 2000) including RRL-based specific-locus amplified fragment sequencing (SLAF-seq) (Sun *et al.*, 2013), restriction-site-associated DNA sequencing (RAD-seq) (Baird *et al.*, 2008), and genotyping-by-sequencing (GBS) (Elshire *et al.*, 2011), which are depending on restriction enzymes to produce a little portion of genome for reduced representation. All the methods are involving the key steps as follow: (1) digesting the

multiple samples of genomic DNA (from individuals of a mapping population) with one or more restriction enzymes; (2) selecting or reducing the restriction fragments after digestion; (3) sequencing of the final set of fragments which should be less than 1 kb in size owing to the read-length limits of current NGS platforms. Polymorphisms in the resulting sequenced fragments can be used as the genetic markers (Davey *et al.*, 2011).

The main advantage is that a reference genome is not the requisite in these approaches, so they can be appropriate for accurately genotyping in high-density of wild species populations in which reference genomes are not available. However, the disadvantages are that the steps of these approaches are complex and tricky, and are not able to thoroughly discover genetic markers of the entire genome because of just sequencing of a little portion of genome.

1.3.2 Low-coverage whole-genome resequencing

In above methods, little targeted proportions of the genome are subjected to sequencing, so that each genetic marker can be sequenced at high coverage with limited resources, thus enabling the accurately genotyping of genetic markers across many individuals.

An alternative approach is emerging to sequence the whole genome at low coverage per individual. In this strategy, missing genotypes of the genetic markers can be imputed based on their coordinated known positions on a physical map. This strategy has been employed for high-density genetic linkage mapping of rice (Xie *et al.*, 2010), shiitake mushroom (Au *et al.*, 2013), and safflower (Bowers *et al.*, 2016). However, these cases relied on the availability of high-quality reference genome sequences. The parental genotypes of recombinant populations are also either known or can be inferred.

Therefore, the advantages are the simpler and easier procedures of the whole genome sequencing compared to restriction-enzyme-based methods such as SLAF-seq and RAD-seq, and cost-effectively discovering almost all SNPs along with the subsequent thoroughly characterizing the entire genome. However, there are several disadvantages. This strategy remains analytical challenges and no existing software

tools for execution. This strategy also needs large amount of labor sources to produce and keep inbred lines based on which parental genotypes and linkage phase can be known or inferred. Furthermore, this strategy also requires a high-quality reference genome, which limits the application in non-model organisms, especially for those with unexplored genomes.

1.4 Torafugu (Takifugu rubripes)

Torafugu (*Takifugu rubripes*) (**Figure 1.4-1**), also known as tiger puffer and Japanese puffer, is a pufferfish with economic importance. It is distributed in Northwest Pacific Ocean from western part of the Sea of Japan and the East China and Yellow seas northward to Muroran, Hokkaido, Japan (**Figure 1.4-2**). It is not only a popular fish food in Japan, Korea and China, but also is identified as a source of collagen (Nagai *et al.*, 2002). This species also bioaccumulate a kind of valuable neurotoxin compound——tetrodotoxin in liver and ovaries. Tetrodotoxin has been considered as a novel therapeutic agent for pain due to that it blocks voltage-gated sodium channels in nervous system (Nieto *et al.*, 2012).

Figure 1.4-1 Torafugu (*Takifugu rubripes*)

From http://mitofish.aori.u-tokyo.ac.jp/species/detail.html?genus=Takifugu&species=rubripes.

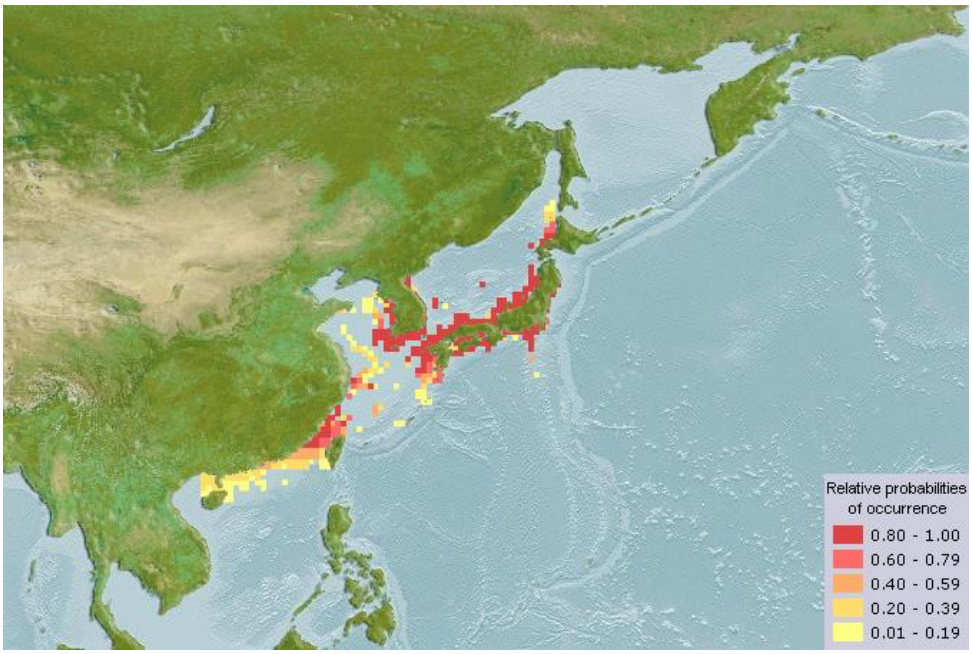


Figure 1.4-2 The distribution of torafugu

Distribution range colours indicate degree of suitability of habitat which can be interpreted as probabilities of occurrence. Adapted from http://www.aquamaps.org/receive.php?type_of_map=regular.

Torafugu has emerged as a model species in genetic and genomic studies. 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 (Brenner *et al.*, 1993). Torafugu has a body plan and physiological systems similar to mammals (Kai *et al.*, 2005). The torafugu genome contains a gene repertoire similar to that in the human genome, so that it can be a model vertebrate genome for discovering genes and gene regulatory elements in the human genome (Kai *et al.*, 2011). The torafugu genome can further serve as a reference to understand the structure, function of vertebrate genomes, due to its compact size (Grutzner *et al.*, 1999; Hedges and Kumar, 2002). Considering that there are only approximately 25 closely related species with limited distributions but remarkable variations in coloration, morphology and behavior, *Takifugu* species also have been strongly suggested as a model system for evolutionary studies of

speciation mechanisms in marine environments and widely used as a reference genome for understanding the evolution of vertebrate genomes and karyotypes (Kai *et al.*, 2011; Yamanoue *et al.*, 2008).

As early as 2002, the torafugu genome has been sequenced to ~6× coverage using the whole-genome shotgun strategy without the aid of any physical or genetic map, yielding an assembly composing of 12,381 scaffolds longer than 2 kb and covering 332 Mb of the genome (Aparicio *et al.*, 2002). In 2005, a genetic map including 200 microsatellite markers has been generated with 10% of the assembly being anchored to the 22 torafugu chromosomes. Before 2011, the torafugu assembly has been expanded to 393 Mb. It consists of 7,213 scaffolds with 511 scaffolds representing 83% of the assembly. Subsequently, in the most recent fifth version of the torafugu genome assembly (FUGU5), 86% of the assembly, including 679 scaffolds, could be anchored to 22 torafugu chromosomes after integration with the genetic linkage map of torafugu containing 1,220 microsatellite markers. In these anchored scaffolds, 72% scaffolds of the whole assembly could be oriented (Kai *et al.*, 2011). Therefore, to expand the contiguity and to improve the quality of the genome assembly, a higher-density genetic linkage map of torafugu is in need.

1.5 Objectives of the present study

Linkage maps are essential tools for genetic and genomic studies including evolutionary genetics, phenotype-genotype association studies and chromosome-scale assembly. During the construction of linkage map, crossing experiment normally is a time-consuming and laborious process lasting several months even to several years to produce mapping populations and keep inbred lines. More and more genetic markers are required for more accurate and thorough identification of genetic polymorphisms of each individual. The emergency of the NGS-based technologies may facilitate the achievement of this goal. Among them, the prevalent restriction-enzyme-based methods like SLAF-seq and RAD-seq are complex and tricky, and not able to

thoroughly discover genetic markers of the entire genome because of a little portion of the genome being sequenced. Other strategies, based on low-coverage whole-genome resequencing, can cost-effectively discover almost all SNPs and thoroughly characterize the entire genome, but also with limitations. For instances, these strategies need the information of linkage phase from the inbred lines. They also require a high-quality reference genome, which limits the application in non-model organisms, especially those with unexplored genomes. Therefore, we would like to develop an effective and simple strategy to construct an ultrahigh-density linkage map using low-coverage whole-genome sequencing. This linkage map can contain almost all SNPs thorough the entire genome without requiring inbred lines, linkage phase and a high-quality reference genome, which enables this strategy be applied to non-model and unsequenced species.

On the other hand, torafugu is an important model species in genetic and genomic studies, also with economic importance. After integration with the genetic linkage map of torafugu comprising 1,220 microsatellite markers, the latest version genome FUGU5 has been developed but remaining room for improvement. Therefore, we also are aiming to construct an ultrahigh density linkage map of torafugu to expand the contiguity and improve quality of FUGU5.

6 General discussion

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 sequencing data were used for de novo assembly and further SNP calling to generate a low-call-rate SNP dataset with unknown phase. Based on the relatively low-quality de novo assembly, an SSG was designed as a high-call-rate genetic marker to assign a genotype to a short DNA segment after combing the information of its constituent low-call-rate SNPs. The high-call-rate SSG dataset enabled the construction of an ultrahigh-density genetic linkage map containing most of the information of SNPs (sub-million in this case) of the mapping population.

For testing the accuracy of the present linkage map, we carried out series of analyses. The analyses of recombination fractions and assessment of LOD scores for all marker pairs. The considerably low recombination fractions and the high LOD scores between adjacent markers of each linkage group suggested a low recombination frequency and a strong linkage between them. The 22 linkage groups were distinctly clustered because the higher recombination fractions and lower LOD scores were observed in the marker pairs from different linkage groups than that from the same linkage group. The comparative analyses between the present linkage map and the published high-quality genome FUGU5 also suggested a near-perfect concordance between them.

Torafugu is a popular species with economic importance in the waters of East Asia, and has emerged as an ideal model in genomic studies (Brenner *et al.*, 1993). 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 torafugu genetic linkage map, comprising 1220 microsatellite markers; however, the remaining 14%

have been located but not oriented, and the other 14% have not yet been assigned (Kai *et al.*, 2011). The comparative analyses between the present linkage map and FUGU5 not only indicated that several unassembled regions of FUGU5 can be improved, but also demonstrated potential mis-assemble regions or large segmental polymorphisms in chromosomes. Therefore, after integration with the present linkage map, the contiguity of FUGU5 is expected to be expanded.

Producing mapping populations and keeping inbred lines are normally very time-consuming and laborious work lasting several months even to several years during linkage map construction. In the present study, we did not keep the inbred lines and just rapidly and simply produced DH individuals by well-developed H/DH technologies. 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. Therefore, the present strategy has significant advantage of time and labor saving.

The assembly of a high quality reference genome is also very laborious and time-consuming including the work of de novo assembly and the construction of one or more high-density linkage maps. However, the previous low-coverage resequencing strategy for ultrahigh-density linkage mapping requires a high-quality reference genome, which limits its application on unsequenced and non-model species. 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. In addition, in the whole-genome assembly project, de novo assembly and linkage map construction are independent works contributing to chromosome-scale assembly. However, in our case, both of them can be achieved from the low-coverage whole-genome sequencing of the mapping population. Thus, for a non-sequenced species, de novo assembly, linkage map construction, and further chromosome-scale assembly can be efficiently completed by adopting our strategy.

Our approach has an advantage of simplicity, in that whole-genome sequencing

was applied to each individual sample, whereas existing techniques such as specific-locus amplified fragment sequencing (SLAF-seq) (Sun *et al.*, 2013), reduced-representation libraries (RRLs) (Altshuler *et al.*, 2000), and restriction-site-associated DNA sequencing (RAD-seq) (Baird *et al.*, 2008) demand complicated processes for sequencing a small target portion of the whole genome. We successfully captured most of the SNPs of the population, which would allow for thoroughly characterizing complex genomes, whereas SLAF-seq, RAD-seq, and RRLs are only able to call a small portion of SNPs of the population. The low-coverage sequencing also makes the present strategy cost-effective.

In recent years, the H/DH population has been exploited as an ideal population type for genetic linkage map construction, particularly in plants (Delourme *et al.*, 2013; Lefebvre *et al.*, 1995; Semagn *et al.*, 2006b; Xu *et al.*, 2015; Zhang *et al.*, 2002) and teleosts (Brieuc *et al.*, 2014; Kelly *et al.*, 2000; Sun and Liang, 2004; Young *et al.*, 1998) due to their well-developed H/DH technologies (Dunwell, 2010; Komen and Thorgaard, 2007). The strategy of the present study developed for ultrahigh-density genetic linkage map construction based on an H/DH dataset, would be ideally implemented for various types of sexually reproducing organisms that could be used to generate large numbers of H/DH individuals without requiring other complex crossing schemes and designed inbred pedigree lines to identify the linkage phase. Notably, our strategy could be applied to the construction of an ultrahigh-density genetic linkage map using only single gamete cells, the ubiquitously existing haploids, combined with single-cell sequencing technology, which may further extend the application range and also is our future work.

References

Altshuler D, Pollara VJ, Cowles CR, Van Etten WJ, Baldwin J, Linton L, Lander ES. 2000. An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature 407: 513-516.

Amores A, Catchen J, Nanda I, Warren W, Walter R, Schartl M, Postlethwait JH. 2014. A RAD-Tag Genetic Map for the Platyfish (Xiphophorus maculatus) Reveals Mechanisms of Karyotype Evolution Among Teleost Fish. Genetics 197: 625-U307.

Aparicio S, Chapman J, Stupka E, Putnam N, Chia J-m, Dehal P, Christoffels A, Rash S, Hoon S, Smit A. 2002. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. Science 297: 1301-1310.

Au CH, Cheung MK, Wong MC, Chu AK, Law PT, Kwan HS. 2013. Rapid genotyping by low-coverage resequencing to construct genetic linkage maps of fungi: a case study in Lentinula edodes. BMC Res Notes 6: 307.

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA. 2008. Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. Plos One 3.

Bishop MD, Kappes SM, Keele JW, Stone RT, Sunden SLF, Hawkins GA, Toldo SS, Fries R, Grosz MD, Yoo JY, Beattie CW. 1994. A Genetic-Linkage Map for Cattle. Genetics 136: 619-639.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114-2120.

Bowers JE, Pearl SA, Burke JM. 2016. Genetic Mapping of Millions of SNPs in Safflower (Carthamus tinctorius L.) via Whole-Genome Resequencing. G3-Genes Genomes Genetics 6: 2203-2211.

Brenner S, Elgar G, Sandford R, Macrae A, Venkatesh B, Aparicio S. 1993. Characterization Of the Pufferfish (Fugu) Genome as a Compact Model Vertebrate Genome. Nature 366: 265-268.

Brieuc MS, Waters CD, Seeb JE, Naish KA. 2014. A dense linkage map for Chinook salmon (Oncorhynchus tshawytscha) reveals variable chromosomal divergence after an ancestral whole genome duplication event. G3 (Bethesda) 4: 447-460.

Broman KW. 2005. The genomes of recombinant inbred lines. Genetics 169: 1133-1146.

Cheema J, Dicks J. 2009. Computational approaches and software tools for genetic linkage map estimation in plants. Briefings In Bioinformatics 10: 595-608.

Cole CJ. 1975. Evolution of parthenogenetic species of reptiles. Intersexuality in the animal kingdom: 340-355.

Collard BCY, Jahufer MZZ, Brouwer JB, Pang ECK. 2005. An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. Euphytica 142: 169-196.

Daiger SP, Wildin RS, Hoffman NS, Su TS. 1982. Restriction Fragment Length Polymorphisms (Rflps) Detected with a Cdna Probe To Argininosuccinate Synthetase (as). American Journal Of Human Genetics 34: A180-A180.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Reviews Genetics 12: 499-510.

Dekker JW, Easteal S. 1990. HLA-DP typing by amplified fragment length polymorphisms (AFLPs). Immunogenetics 32: 56-59.

Delourme R, Falentin C, Fomeju BF, Boillot M, Lassalle G, Andre I, Duarte J, Gauthier V, Lucante N, Marty A, Pauchon M, Pichon JP, Ribiere N, Trotoux G, Blanchard P, Riviere N, Martinant JP, Pauquet J. 2013. High-density SNP-based genetic map development and linkage disequilibrium assessment in Brassica napus L. BMC Genomics 14: 120.

Diopere E, Maes GE, Komen H, Volckaert FAM, Groenen MAM. 2014. A Genetic Linkage Map of Sole (Solea solea): A Tool for Evolutionary and Comparative Analyses of Exploited (Flat)Fishes. Plos One 9.

Doniskeller H, Green P, Helms C, Cartinhour S, Weiffenbach B, Stephens K, Keith TP, Bowden DW, Smith DR, Lander ES, Botstein D, Akots G, Rediker KS, Gravius T, Brown VA, Rising MB, Parker C, Powers JA, Watt DE, Kauffman ER, Bricker A, Phipps P, Mullerkahle H, Fulton TR, Ng S, Schumm JW, Braman JC, Knowlton RG, Barker DF, Crooks SM, Lincoln SE, Daly MJ, Abrahamson J. 1987. A Genetic-Linkage Map Of the Human Genome. Cell 51: 319-337.

Dunwell JM. 2010. Haploids in flowering plants: origins and exploitation. Plant Biotechnology Journal 8: 377-424.

Edwards J. 1971. The analysis of X‐linkage. Annals of human genetics 34: 229-250.

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. 2011. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. Plos One 6.

Falk CT. 1989. A simple scheme for preliminary ordering of multiple loci: application to 45 CF families. Prog Clin Biol Res 329: 17-22.

Felip A, Zanuy S, Carrillo M, Piferrer F. 2001. Induction of triploidy and gynogenesis in teleost fish with emphasis on marine species. Genetica 111: 175-195.

Fierst JL. 2015. Using linkage maps to correct and scaffold de novo genome assemblies: methods, challenges, and computational tools. Frontiers in Genetics 6.

Gadau J. 2009. Phase-unknown linkage mapping in ants. Cold Spring Harb Protoc 2009: pdb prot5251.

Glesener RR, Tilman D. 1978. Sexuality and the components of environmental uncertainty: clues from geographic parthenogenesis in terrestrial animals. The American Naturalist 112: 659-673.

Grell KG. 1973. Alternation of Generations. In: Protozoology: Springer. p 229-231.

Grutzner F, Lutjens G, Rovira C, Barnes DW, Ropers HH, Haaf T. 1999. Classical and molecular cytogenetics of the pufferfish tetraodon nigroviridis. Chromosome Research 7: 655-662.

Gusella JF, Wexler NS, Conneally PM, Naylor SL, Anderson MA, Tanzi RE, Watkins PC, Ottina K, Wallace MR, Sakaguchi AY, et al. 1983. A polymorphic DNA marker genetically linked to Huntington's disease. Nature 306: 234-238.

Gyapay G, Morissette J, Vignal A, Dib C, Fizames C, Millasseau P, Marc S, Bernardi G, Lathrop M, Weissenbach J. 1994. The 1993-94 Genethon Human Genetic-Linkage Map. Nature Genetics 7: 246-339.

Haldane JB, Smith CA. 1947. A new estimate of the linkage between the genes for colourblindness and haemophilia in man. Ann Eugen 14: 10-31.

Hedges SB, Kumar S. 2002. Vertebrate genomes compared. Science 297: 1283-1285.

Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, Collins JE, Humphray S, McLaren K, Matthews L, McLaren S, Sealy I, Caccamo M, Churcher C, Scott C, Barrett JC, Koch R, Rauch GJ, White S, Chow W, Kilian B, Quintais LT, Guerra-Assuncao JA, Zhou Y, Gu Y, Yen J, Vogel JH, Eyre T, Redmond S, Banerjee R, Chi J, Fu B, Langley E, Maguire SF, Laird GK, Lloyd D, Kenyon E, Donaldson S, Sehra H, Almeida-King J, Loveland J, Trevanion S, Jones M, Quail M, Willey D, Hunt A, Burton J, Sims S, McLay K, Plumb B, Davis J, Clee C, Oliver K, Clark R, Riddle C, Elliot D, Threadgold G, Harden G, Ware D, Begum S, Mortimore B, Kerry G, Heath P, Phillimore B, Tracey A, Corby N, Dunn M, Johnson C, Wood J, Clark S, Pelan S, Griffiths G, Smith M, Glithero R, Howden P, Barker N, Lloyd C, Stevens C, Harley J, Holt K, Panagiotidis G, Lovell J, Beasley H, Henderson C, Gordon D, Auger K, Wright D, Collins J, Raisen C, Dyer L, Leung K, Robertson L, Ambridge K,

Leongamornlert D, McGuire S, Gilderthorp R, Griffiths C, Manthravadi D, Nichol S, Barker G, Whitehead S, Kay M, Brown J, Murnane C, Gray E, Humphries M, Sycamore N, Barker D, Saunders D, Wallis J, Babbage A, Hammond S, Mashreghi-Mohammadi M, Barr L, Martin S, Wray P, Ellington A, Matthews N, Ellwood M, Woodmansey R, Clark G, Cooper J, Tromans A, Grafham D, Skuce C, Pandian R, Andrews R, Harrison E, Kimberley A, Garnett J, Fosker N, Hall R, Garner P, Kelly D, Bird C, Palmer S, Gehring I, Berger A, Dooley CM, Ersan-Urun Z, Eser C, Geiger H, Geisler M, Karotki L, Kirn A, Konantz J, Konantz M, Oberlander M, Rudolph-Geiger S, Teucke M, Lanz C, Raddatz G, Osoegawa K, Zhu B, Rapp A, Widaa S, Langford C, Yang F, Schuster SC, Carter NP, Harrow J, Ning Z, Herrero J, Searle SM, Enright A, Geisler R, Plasterk RH, Lee C, Westerfield M, de Jong PJ, Zon LI, Postlethwait JH, Nusslein-Volhard C, Hubbard TJ, Roest Crollius H, Rogers J, Stemple DL. 2013. The zebrafish reference genome sequence and its relationship to the human genome. Nature 496: 498-503.

Hubbs CL, Hubbs LC. 1932. Apparent parthenogenesis in nature, in a form of fish of hybrid origin. Science 76: 628-630.

Hugot JP, Chamaillard M, Zouali H, Lesage S, Cezard JP, Belaiche J, Almer S, Tysk C, O'Morain CA, Gassull M, Binder V, Finkel Y, Cortot A, Modigliani R, Laurent-Puig P, Gower-Rousseau C, Macry J, Colombel JF, Sahbatou M, Thomas G. 2001. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. Nature 411: 599-603.

Huigens M, De Almeida R, Boons P, Luck R, Stouthamer R. 2004. Natural interspecific and intraspecific horizontal transfer of parthenogenesis–inducing Wolbachia in Trichogramma wasps. Proceedings of the Royal Society of London B: Biological Sciences 271: 509-515.

International Barley Genome Sequencing C, Mayer KF, Waugh R, Brown JW, Schulman A, Langridge P, Platzer M, Fincher GB, Muehlbauer GJ, Sato K, Close TJ, Wise RP, Stein N. 2012. A physical, genetic and functional sequence assembly of the barley genome. Nature 491: 711-716.

Jansen J, De Jong A, Van Ooijen J. 2001. Constructing dense genetic linkage maps. TAG Theoretical and Applied Genetics 102: 1113-1122.

Kai W, Kikuchi K, Fujita M, Suetake H, Fujiwara A, Yoshiura Y, Ototake M, Venkatesh B, Miyaki K, Suzuki Y. 2005. A genetic linkage map for the tiger pufferfish, Takifugu rubripes. Genetics 171: 227-238.

Kai W, Kikuchi K, Tohari S, Chew AK, Tay A, Fujiwara A, Hosoya S, Suetake H, Naruse K, Brenner S, Suzuki Y, Venkatesh B. 2011. Integration of the Genetic Map and Genome Assembly of Fugu Facilitates Insights into Distinct Features of Genome Evolution in Teleosts and Mammals. Genome Biology And Evolution 3: 424-442.

Kearney M. 2005. Hybridization, glaciation and geographical parthenogenesis. Trends in ecology & evolution 20: 495-502.

Kelly PD, Chu F, Woods IG, Ngo-Hazelett P, Cardozo T, Huang H, Kimm F, Liao LY, Yan

YL, Zhou YY, Johnson SL, Abagyan R, Schier AF, Postlethwait JH, Talbot WS. 2000. Genetic linkage mapping of zebrafish genes and ESTs. Genome Research 10: 558-567.

Komen H, Thorgaard GH. 2007. Androgenesis, gynogenesis and the production of clones in fishes: a review. Aquaculture 269: 150-173.

Kozik A, Michelmore R. 2006. MadMapper and CheckMatrix—python scripts to infer orders of genetic markers and for visualization and validation of genetic maps and haplotypes. In: Proceedings of the Plant and Animal Genome XIV Conference, San Diego.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: An information aesthetic for comparative genomics. Genome Research 19: 1639-1645.

Lander ES, Green P, Abrahamson J, Barlow A, Daly MJ, Lincoln SE, Newberg LA. 1987. MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics 1: 174-181.

Lefebvre V, Palloix A, Caranta C, Pochard E. 1995. Construction Of an Intraspecific Integrated Linkage Map Of Pepper Using Molecular Markers And Doubled Haploid Progenies. Genome 38: 112-121.

Lewin HA, Larkin DM, Pontius J, O'Brien SJ. 2009. Every genome sequence needs a good map. Genome Res 19: 1925-1928.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754-1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079.

Lien S, Koop BF, Sandve SR, Miller JR, Kent MP, Nome T, Hvidsten TR, Leong JS, Minkley DR, Zimin A, Grammes F, Grove H, Gjuvsland A, Walenz B, Hermansen RA, von Schalburg K, Rondeau EB, Di Genova A, Samy JK, Olav Vik J, Vigeland MD, Caler L, Grimholt U, Jentoft S, Vage DI, de Jong P, Moen T, Baranski M, Palti Y, Smith DR, Yorke JA, Nederbragt AJ, Tooming-Klunderud A, Jakobsen KS, Jiang X, Fan D, Hu Y, Liberles DA, Vidal R, Iturra P, Jones SJ, Jonassen I, Maass A, Omholt SW, Davidson WS. 2016. The Atlantic salmon genome provides insights into rediploidization. Nature 533: 200-205.

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P. 2003. The power and promise of population genomics: from genotyping to genome typing. Nat Rev Genet 4: 981-994.

Luo RB, Liu BH, Xie YL, Li ZY, Huang WH, Yuan JY, He GZ, Chen YX, Pan Q, Liu YJ, Tang JB, Wu GX, Zhang H, Shi YJ, Liu Y, Yu C, Wang B, Lu Y, Han CL, Cheung DW, Yiu SM, Peng SL, Zhu XQ, Liu GM, Liao XK, Li YR, Yang HM, Wang J, Lam TW, Wang J. 2012.

SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 1.

Marcus JN, Watson P, Page DL, Narod SA, Lenoir GM, Tonin P, Linder-Stephenson L, Salerno G, Conway TA, Lynch HT. 1996. Hereditary breast cancer: pathobiology, prognosis, and BRCA1 and BRCA2 gene linkage. Cancer 77: 697-709.

Mardis ER. 2008. Next-generation DNA sequencing methods. Annu. Rev. Genomics Hum. Genet. 9: 387-402.

Metzker ML. 2010. Sequencing technologies—the next generation. Nature reviews genetics 11: 31-46.

Mohan M, Nair S, Bhagwat A, Krishna TG, Yano M, Bhatia CR, Sasaki T. 1997. Genome mapping, molecular markers and marker-assisted selection in crop plants. Molecular Breeding 3: 87-103.

Morton NE. 1955. Sequential tests for the detection of linkage. Am J Hum Genet 7: 277-318.

Murovec J, Bohanec B. 2012. Haploids and doubled haploids in plant breeding. In: Plant breeding: InTech.

Nagai T, Araki Y, Suzuki N. 2002. Collagen of the skin of ocellate puffer fish (Takifugu rubripes). Food chemistry 78: 173-177.

Neaves WB, Baumann P. 2011. Unisexual reproduction among vertebrates. Trends in Genetics 27: 81-88.

Nieto FR, Cobos EJ, Tejada MÁ, Sánchez-Fernández C, González-Cano R, Cendán CM. 2012. Tetrodotoxin (TTX) as a therapeutic agent for pain. Marine drugs 10: 281-305.

Powell W, Machray GC, Provan J. 1996. Polymorphism revealed by simple sequence repeats. Trends In Plant Science 1: 215-222.

Rastas P, Calboli FCF, Guo BC, Shikano T, Merila J. 2016. Construction of Ultradense Linkage Maps with Lep-MAP2: Stickleback F-2 Recombinant Crosses as an Example. Genome Biology And Evolution 8: 78-93.

Rastas P, Paulin L, Hanski I, Lehtonen R, Auvinen P. 2013. Lep-MAP: fast and accurate linkage map construction for large SNP datasets. Bioinformatics 29: 3128-3134.

Ren J, Wu P, Trampe B, Tian X, Chen S, Lübberstedt T. 2017. Novel technologies in doubled haploid line development. Plant biotechnology journal.

Ruano G, Gray MR, Miki T, Ferguson-Smith AC, Ruddle FH, Kidd KK. 1990. Monomorphism in humans and sequence differences among higher primates for a sequence

tagged site (STS) in homeo box cluster 2 as assayed by denaturing gradient electrophoresis. Nucleic Acids Res 18: 1314.

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, Hunt SE, Cole CG, Coggill PC, Rice CM, Ning ZM, Rogers J, Bentley DR, Kwok PY, Mardis ER, Yeh RT, Schultz B, Cook L, Davenport R, Dante M, Fulton L, Hillier L, Waterston RH, McPherson JD, Gilman B, Schaffner S, Van Etten WJ, Reich D, Higgins J, Daly MJ, Blumenstiel B, Baldwin J, Stange-Thomann NS, Zody MC, Linton L, Lander ES, Altshuler D, Grp ISMW. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409: 928-933.

Semagn K, Bjornstad A, Ndjiondjop MN. 2006a. Principles, requirements and prospects of genetic mapping in plants. African Journal Of Biotechnology 5: 2569-2587.

Semagn K, Bjornstad A, Skinnes H, Maroy AG, Tarkegne Y, William M. 2006b. Distribution of DArT, AFLP, and SSR markers in a genetic linkage map of a doubled-haploid hexaploid wheat population. Genome 49: 545-555.

Shao C, Bao B, Xie Z, Chen X, Li B, Jia X, Yao Q, Orti G, Li W, Li X, Hamre K, Xu J, Wang L, Chen F, Tian Y, Schreiber AM, Wang N, Wei F, Zhang J, Dong Z, Gao L, Gai J, Sakamoto T, Mo S, Chen W, Shi Q, Li H, Xiu Y, Li Y, Xu W, Shi Z, Zhang G, Power DM, Wang Q, Schartl M, Chen S. 2017. The genome and transcriptome of Japanese flounder provide insights into flatfish asymmetry. Nat Genet 49: 119-124.

Song WT, Li YZ, Zhao YW, Liu Y, Niu YZ, Pang RY, Miao GD, Liao XL, Shao CW, Gao FT, Chen SL. 2012. Construction of a High-Density Microsatellite Genetic Linkage Map and Mapping of Sexual and Growth-Related Traits in Half-Smooth Tongue Sole (Cynoglossus semilaevis). Plos One 7.

Song X, Wang K, Guo W, Zhang J, Zhang T. 2005. A comparison of genetic maps constructed from haploid and BC1 mapping populations from the same crossing between Gossypium hirsutum L. and Gossypium barbadense L. Genome 48: 378-390.

Stam P. 1993. Construction of integrated genetic linkage maps by means of a new computer package: Join Map. The plant journal 3: 739-744.

Sturtevant AH. 1913. The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. Journal Of Experimental Zoology 14: 43-59.

Sun X, Liu D, Zhang X, Li W, Liu H, Hong W, Jiang C, Guan N, Ma C, Zeng H, Xu C, Song J, Huang L, Wang C, Shi J, Wang R, Zheng X, Lu C, Wang X, Zheng H. 2013. SLAF-seq: an efficient method of large-scale de novo SNP discovery and genotyping using high-throughput sequencing. PLoS One 8: e58700.

Sun XW, Liang LQ. 2004. A genetic linkage map of common carp (Cyprinus carpio L.) And mapping of a locus associated with cold tolerance. Aquaculture 238: 165-172.

Taylor J, Butler D. 2017. R Package ASMap: Efficient Genetic Linkage Map Construction and Diagnosis. arXiv preprint arXiv:1705.06916.

Thompson EA. 1987. Crossover counts and likelihood in multipoint linkage analysis. IMA J Math Appl Med Biol 4: 93-108.

Tripathi N, Hoffmann M, Willing EM, Lanz C, Weigel D, Dreyer C. 2009. Genetic linkage map of the guppy, Poecilia reticulata, and quantitative trait loci analysis of male size and colour variation. Proceedings Of the Royal Society B-Biological Sciences 276: 2195-2208.

Tsui LC, Buchwald M, Barker D, Braman JC, Knowlton R, Schumm JW, Eiberg H, Mohr J, Kennedy D, Plavsic N, et al. 1985. Cystic fibrosis locus defined by a genetically linked polymorphic DNA marker. Science 230: 1054-1057.

Van Os H, Stam P, Visser RG, Van Eck HJ. 2005. RECORD: a novel method for ordering loci on a genetic linkage map. Theor Appl Genet 112: 30-40.

Wang WJ, Hu YL, Ma Y, Xu LY, Guan JT, Kong J. 2015a. High-Density Genetic Linkage Mapping in Turbot (Scophthalmus maximus L.) Based on SNP Markers and Major Sex- and Growth-Related Regions Detection. Plos One 10.

Wang YP, Lu Y, Zhang Y, Ning ZM, Li Y, Zhao Q, Lu HY, Huang R, Xia XQ, Feng Q, Liang XF, Liu KY, Zhang L, Lu TT, Huang T, Fan DL, Weng QJ, Zhu CR, Lu YQ, Li WJ, Wen ZR, Zhou CC, Tian QL, Kang XJ, Shi MJ, Zhang WT, Jang S, Du FK, He S, Liao LJ, Li YM, Gui B, He HH, Ning Z, Yang C, He LB, Luo LF, Yang R, Luo Q, Liu XC, Li SS, Huang W, Xiao L, Lin HR, Han B, Zhu ZY. 2015b. The draft genome of the grass carp (Ctenopharyngodon idellus) provides insights into its evolution and vegetarian adaptation (vol 47, pg 625, 2015). Nature Genetics 47: 962-962.

Weeks DE, Lange K. 1987. Preliminary ranking procedures for multilocus ordering. Genomics 1: 236-242.

Wilson SR. 1988. A major simplification in the preliminary ordering of linked loci. Genet Epidemiol 5: 75-80.

Winter P, Kahl G. 1995. Molecular Marker Technologies for Plant Improvement. World Journal Of Microbiology & Biotechnology 11: 438-448.

Wu J, Li LT, Li M, Khan MA, Li XG, Chen H, Yin H, Zhang SL. 2014. High-density genetic linkage map construction and identification of fruit-related QTLs in pear using SNP and SSR markers. Journal Of Experimental Botany 65: 5771-5781.

Wu QH, Chen YX, Zhou SH, Fu L, Chen JJ, Xiao Y, Zhang D, Ouyang SH, Zhao XJ, Cui Y, Zhang DY, Liang Y, Wang ZZ, Xie JZ, Qin JX, Wang GX, Li DL, Huang YL, Yu MH, Lu P, Wang LL, Wang L, Wang H, Dang C, Li J, Zhang Y, Peng HR, Yuan CG, You MS, Sun QX, Wang JR, Wang LX, Luo MC, Han J, Liu ZY. 2015. High-Density Genetic Linkage Map

Construction and QTL Mapping of Grain Shape and Size in the Wheat Population Yanda1817 x Beinong6. Plos One 10.

Wu YH, Bhat PR, Close TJ, Lonardi S. 2008. Efficient and Accurate Construction of Genetic Linkage Maps from the Minimum Spanning Tree of a Graph. Plos Genetics 4.

Wyatt R, Odrzykoski IJ, Stoneburner A. 1989. High levels of genetic variability in the haploid moss Plagiomnium ciliare. Evolution 43: 1085-1096.

Xie W, Feng Q, Yu H, Huang X, Zhao Q, Xing Y, Yu S, Han B, Zhang Q. 2010. Parent-independent genotyping for constructing an ultrahigh-density linkage map based on population sequencing. Proc Natl Acad Sci U S A 107: 10578-10583.

Xu Y, Huang L, Ji D, Chen C, Zheng H, Xie C. 2015. Construction of a dense genetic linkage map and mapping quantitative trait loci for economic traits of a doubled haploid population of Pyropia haitanensis (Bangiales, Rhodophyta). BMC Plant Biol 15: 228.

Yamanoue Y, Miya M, Matsuura K, Miyazawa S, Tsukamoto N, Doi H, Takahashi H, Mabuchi K, Nishida M, Sakai H. 2008. Explosive speciation of Takifugu: another use of fugu as a model system for evolutionary biology. Molecular biology and evolution 26: 623-629.

Yang HA, Tao Y, Zheng ZQ, Zhang QS, Zhou GF, Sweetingham MW, Howieson JG, Li CD. 2013. Draft Genome Sequence, and a Sequence-Defined Genetic Linkage Map of the Legume Crop Species Lupinus angustifolius L. Plos One 8.

Young WP, Wheeler PA, Coryell VH, Keim P, Thorgaard GH. 1998. A detailed linkage map of rainbow trout produced using doubled haploids. Genetics 148: 839-850.

Zhang J, Guo W, Zhang T. 2002. Molecular linkage map of allotetraploid cotton (Gossypium hirsutum L. x Gossypium barbadense L.) with a haploid population. Theoretical And Applied Genetics 105: 1166-1174.