Genome analysis of the planarian

Dugesia japonica

APPROVED BY

SUPERVISING COMMITTEE:

Supervisor: Kiyokazu AGATA

Takeshi INOUE

Genome analysis of the planarian *Dugesia japonica*

by

Yang AN

Thesis

Presented to the Faculty of the Graduate School of Science

Kyoto University in Fulfillment

of the Requirements

for the Degree of

Doctor of Science

Kyoto University

January 2015

Table of Contents

CHAPTER 1 ..... **1**

1.1. PLANARIANS **1**

1.2. THE ORIGIN OF PLANARIAN RESEARCH BEFORE THE MOLECULAR BIOLOGY AGE – A LONG HISTORY **3**

1.3. *D. JAPONICA* IS AN EXCELLENT MODEL PLANARIAN IN THE AGE OF MOLECULAR BIOLOGY **7**

1.4. GENOME IS INDISPENSABLE FOR FURTHER STUDY OF *D. JAPONICA* IN THE GENOMICS ERA **9**

1.5. PROJECT AIMS **11**

CHAPTER 2 ..... **12**

2.1. INTRODUCTION **12**

2.1.1. Background 12

2.1.2. Mathematics of DNA library screening 14

2.1.3. Issues that must be considered when designing the practical details of the screening method 19

2.2. RESULTS **25**

2.3. DISCUSSION **29**

2.4. MATERIALS AND METHODS **31**

D. japonica planarian genomic DNA library 31

Pooling 31

qPCR kit and conditions 31

Primers for DjPiwiB gene and positive control gene 32

CHAPTER 3 ..... **33**

3.1. INTRODUCTION **33**

3.2. RAW DATA GENERATION AND QUALITY CONTROL **41**

3.3. ESTIMATION OF GENOME CHARACTERISTICS BEFORE DE NOVO ASSEMBLY 43

3.3.1.Kmer frequency for genome survey 43

3.3.2.*D. japonica* has a complicated genome 45

3.4. DE NOVO ASSEMBLY OF *D. JAPONICA* GENOME 49

3.4.1.Algorithms of genome assembly 49

3.4.2.De novo genome assembly with De Bruijn graph-based method.....51

3.4.3.New strategy to improve genome assembly by overlap-lay-out.....53

3.5. *D. JAPONICA* GENOME ANNOTATION 58

3.5.1.Genome Annotation.....58

3.5.2.Repeat Sequences.....58

3.5.3.De novo transcriptome assembly.....60

3.5.4.Gene prediction and functional annotation.....61

3.6. CNES ARE REGULATORY ELEMENTS IN PLANARIANS 65

3.6.1.Conserved non-coding elements (CNEs).....65

3.6.2.CNEs between two planarian genuses.....65

3.6.3.CNE4 is a regulatory element on DjNDK gene.....69

3.7. DISCUSSION.....71

3.7.1.Improvement of The Planaria Genome.....71

3.7.2.CNEs in Planarians 72

CHAPTER 1

General Introduction

1.1. Planarians

Planarian is a common name for species of non-parasitic Platyhelminthes (flatworms) of the turbellaria class. Planarians are among the simplest bilaterally symmetric acoelomates, and they have three germ layers. Thus, they have been thought to hold an important position in Metazoan evolution (Fig 1-1) [1, 2]. In addition, they have significant value for neurobiology research, since they were the first animals that obtained a brain structure during evolution [3, 4]. Most importantly, planarians have astonishing regeneration abilities [5]. If a planarian is cut into pieces, each piece of the planarian can regenerate into a complete organism, even a complete functional brain. So, for the past few hundred years, planarian has been famous among the scientific community for its regeneration ability.

In Asia, the planarian *Dugesia japonica* (*D. japonica*) is the most widely distributed planarian species, and has been used as a model system for biological research for decades. In this thesis, after reviewing the history of research on this planarian, , I describe my performance of a *D. japonica* genome assembly and annotation project

whose purpose was to facilitate the usage of *D. japonica* as a model animal in the new genomics era and assist future research and applications of planarian biology.

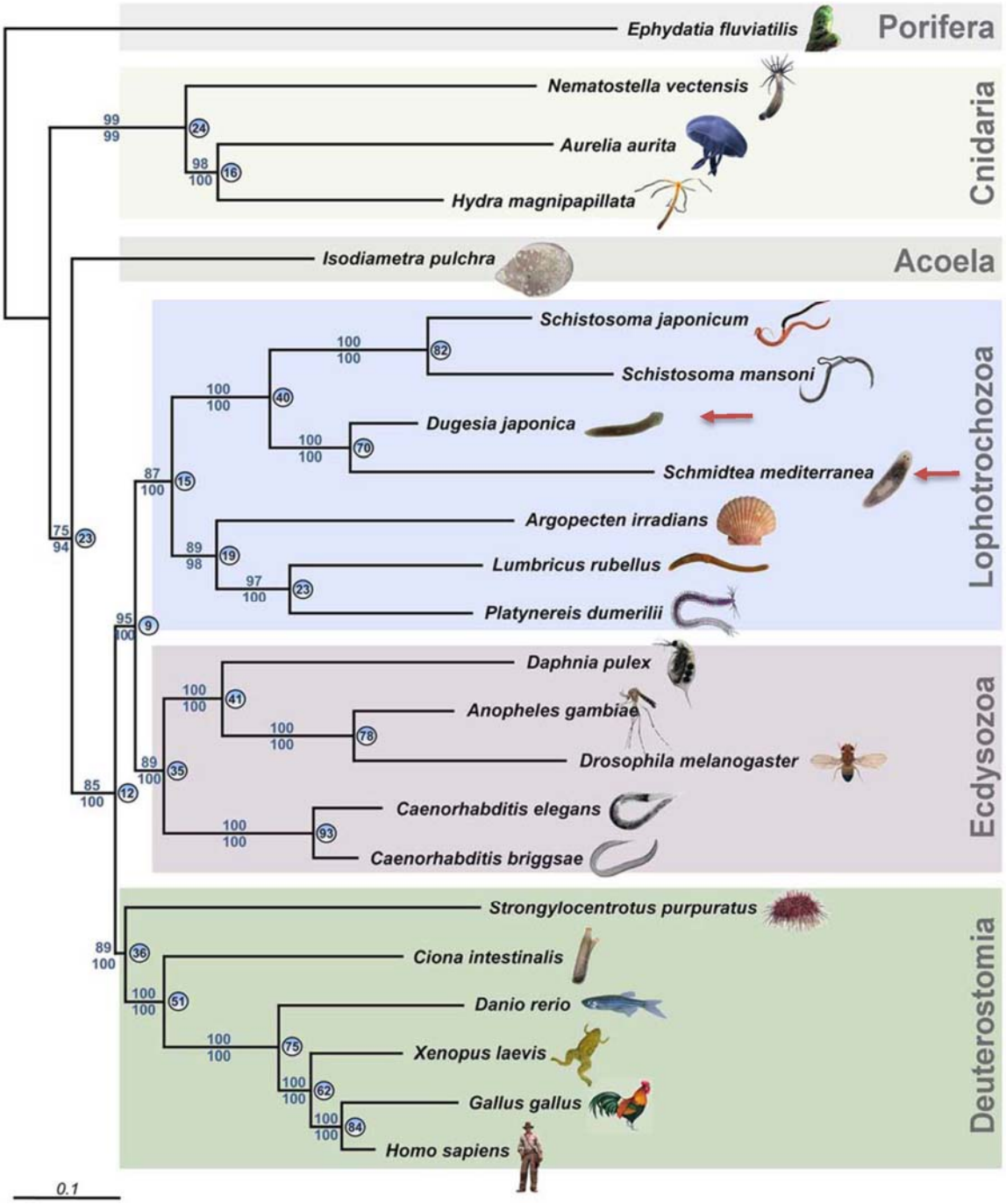
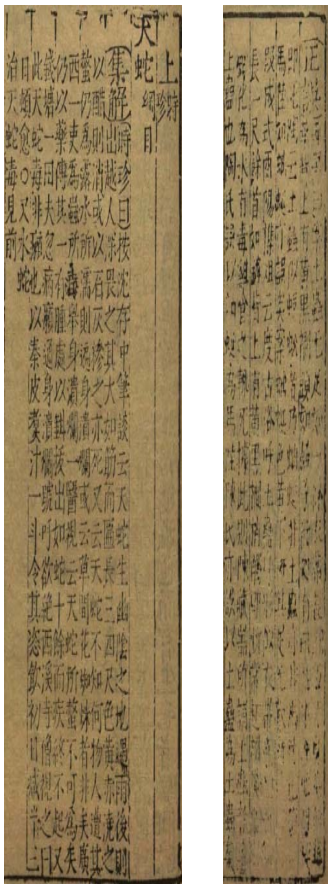


Figure 1-1. The evolutionary position of planarians (taken from [2])

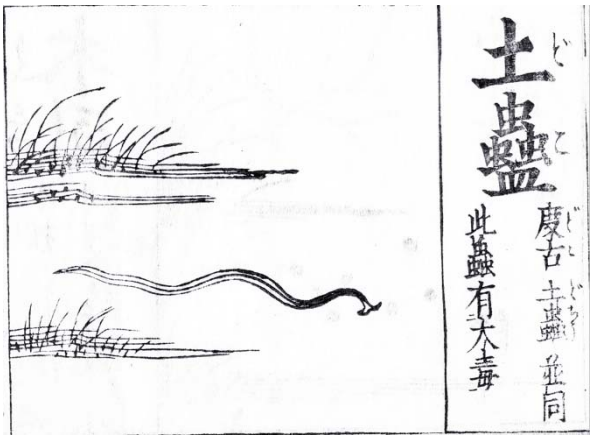
Red arrows point to two commonly used planarians. Although they both belong to the same family, they have long evolution distance between each other.

1.2. The Origin of Planarian Research before the Molecular Biology Age – a long history

In human history, the oldest available written description of planarians can be traced back more than 1200 years ago (739 AD) in China [6], to the book *Supplement to Materia Medica* (written by CHEN Zhangqi at 739 AD). Some extant old books on Chinese medicine directly cite the description about planarians in that book: “The worm has no feet, and looks like a piece of ribbon. It has about 12-15 cm in length and a flat body, which looks like a leek leaf. It has yellow and black pleats on the dorsal surface, and its head is shovel-shaped”[7] (Figure1-2a). This is the oldest record in the world about a planarian. The first known drawing of a planarian appears in a Japanese book, *Kinmo Zui* (written by Tekisai NAKAMURA at 1666, Figure 1-2b)[8]. Its author, for the first time, drew a vivid land-dwelling planarian. In the western world, the observations on planarian lagged 1000 years behind those in Asia. The first reference to a planarian in the western world was recorded in 1766 AD by John Petri Mariae Dana [9-11], and the first description of planarian’s regeneration ability and the first drwing of a freshwater planarian was probably recorded by Peter Simon Pallas in 1774 AD [12, 13]. After Otto Frederik Müller first separated turbellarian species (planarians) from the genus *Hirudo*, and gave a new genus name – “*Planaria*” to these animals[14], planarians’ fascinating regeneration abilities started to attract more and more researchers to dedicate themselves to planarian research, including Charles Darwin [15, 16] and Thomas Hunt Morgan [17, 18], and thus the era of planarian regeneration research started.



a



b

Figure 1-2. The oldest record of planarian (land planarian)

a. The oldest literal description of planarian (Taken from [7]);
b. The oldest figure of planarian (Taken from [8])

Dugesia japonica (*D. japonica*, Figure 1-3) was erroneously known as *Dugesia gonocephala* (a European planarian) until its taxonomic position was determined by Ichikawa & Kawakatsu at 1964 [19]. This species is a typical polymorphic freshwater planarian species widely distributed in the Far East area (including Japan, east part of China, Korea Peninsula, and the Primorsky area of Northeast Siberia in Russia (Figure 1-4)[20, 21]. Many taxonomy, morphology, physiology, ecology and cellular biology studies were performed on this planarian from the 19th century to the end of the 20th

century. However, those studies were hampered by the incomplete understanding of genetics, cellular and molecular biology, and the limitation of experimental methods.

a b c

Figure 1-3. Figures of the planarian *D. japonica*

c. The first figure of *D. japonica* (from [22], 1916)

d. The first colored figure of *D. japonica* (from [23],1922)

e. A photo of *D. japonica* (from [24], 2014)

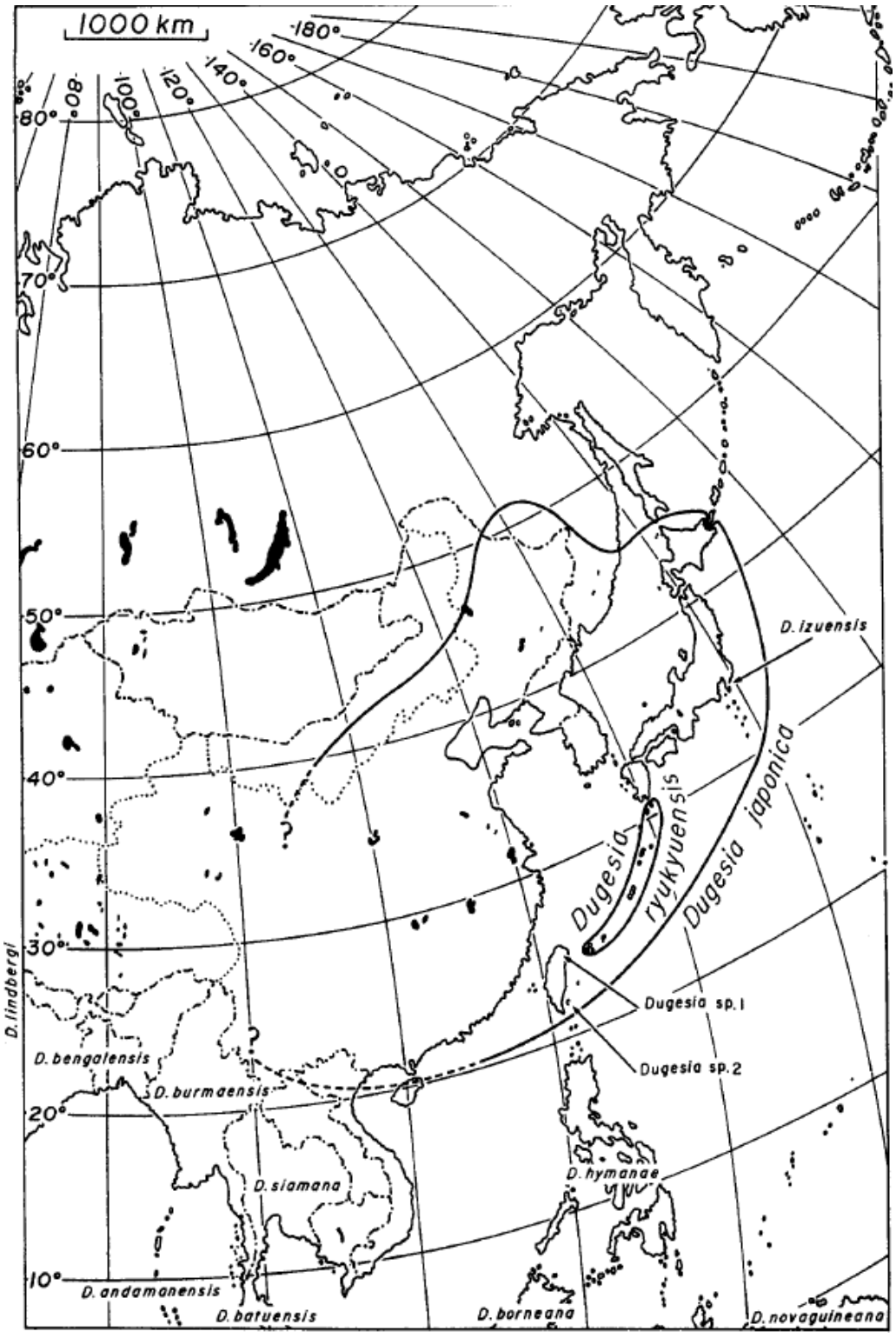


Figure 1-4. Geographical distribution of *D. japonica* (Taken from [21])

1.3. *D. japonica* is an Excellent Model Planarian in the Age of Molecular Biology

Studies on planarian have made great progress aided by the tools and concepts of molecular biology. *D. japonica* was chosen as a model planarian because it has extremely robust regeneration ability, it is the most widely distributed planarian in the world, and is easy for researchers to collect; it can adapt to various circumstances, even to breeding in simple laboratory conditions in autoclaved water; it can reproduce sexually and asexually; it is diploid with a relatively small genome. As *D. japonica* has these favorable characteristics, since 1991, Prof. Kiyokazu Agata's laboratory has endeavored to develop this species as a model system with which to address longstanding problems of biology, such as regeneration, brain formation, and stem cell regulation [4].

Advances in cellular and molecular biology experimental methods as well as nucleic acid sequencing technologies have helped increase our ability to examine the biology of planarians at the molecular level. Using PCR, cDNA libraries, RNA in situ hybridization and immunoscreening, *D. japonica* cell-type-specific genes have been isolated [25]. Highly sensitive *in situ* hybridization methods were developed for identifying mRNA locations and expression in cells [26]. Loss-of-function assays, including RNA interference, were also devised to characterize gene functions [27]. Microarrays have been generated to identify genes important for regeneration, and head-specific genes [28]. Fluorescence Activated Cell Sorting (FACS) was used to isolate discrete cell populations, which could then be used for single-cell gene profiling, functional transplantation studies, and neurobiology study [29-31]. An EST database

was also established for transcriptome level studies [32]. All of these modern research methods and resources have enabled us to link the phenomena of *D. japonica*'s robust regeneration and brain formation to its underlying molecular mechanisms. However, to achieve the complete understanding of those mechanisms, a critical key is still missing, namely, the genome of *D. japonica*.

1.4. Genome Is Indispensable for Further Study of *D. japonica* in the Genomics Era

The essential information of an organism is stored in its genome. With the draft human genome sequences published in 2001, the golden era of genomics arrived. Thus far, thousands of genomes have been sequenced, and have generated exponentially increasing information useful for biological studies (Table 1-1), and to some extent, have changed hypothesis-driven science to discovery-driven (or data-driven) science, which has helped to accelerate the progress of life science research.

In the genomics era, in order to obtain the overall genetic information and perform efficient studies on the planarian *D. japonica*, and to continue to use it as a model system for regeneration, evolution, and development, a *D. japonica* genome project is indispensable. Considering the availability of the powerful cellular and molecular experimental tools mentioned above, determining and analyzing *D. japonica* genome sequences will help to fully exploit the useful characteristics of this species, and facilitate its further usage.

As noted above, *D. japonica* is the most widely distributed planarian in the world, and it is the easiest planarian species to breed in the laboratory. Combined with the already available expressed sequence tag (EST) data[32], a *D. japonica* planarian genome project will be a critical resource for developing this invertebrate model system in the genomics era by facilitating gene identification, RNAi screens, comparative genomics, phylogenetic analysis, microarray experiments, genetic screens, and the identification of promoter and enhancer sequences for further study of the molecular mechanisms of

regeneration and brain formation. With such valuable information available, the planarian *D. japonica* will serve as a powerful model animal.

Table 1-1. Genome sequencing projects statistics

Data from NCBI (http://www.ncbi.nlm.nih.gov/genomes/static/gpstat.html)

Organism	**Complete**	**Draft assembly**	**In progress**	**total**
Prokaryotes	1117	966	595	2678
Archaea	100	5	48	153
Bacteria	1017	961	547	2525
Eukaryotes	36	319	294	649
Animals	6	137	106	249
Mammals	3	41	25	69
Birds		3	13	16
Fishes		16	16	32
Insects	2	38	17	57
Flatworms		3	3	6
Roundworms	1	16	11	28
Amphibians		1		1
Reptiles		2		2
Other animals		20	24	44
Plants	5	33	80	118
Land plants	3	29	73	105
Green Algae	2	4	6	12
Fungi	17	107	59	183
Ascomycetes	13	83	38	134
Basidiomycetes	2	16	11	29
Other fungi	2	8	10	20
Protists	8	39	46	93
Apicomplexans	3	11	16	30
Kinetoplasts	4	3	2	9
Other protists	1	24	28	53
Total:	**1153**	**1285**	**889**	**3327**

1.5. Project aims

There are 5 aims of this project.

1. Assemble a draft genome for *D. japonica* with good quality.
2. Elucidate the genome characteristics of *D. japonica*.
3. Annotate the assembled genome and identify as many coding and noncoding points of interest as possible.
4. Find possible planarian gene regulatory elements.
5. Use the assembly and annotations to investigate the biology of *D. japonica* in the future, including aspects such as wound healing, regeneration, stem cells, tissue homeostasis, and brain formation.

CHAPTER 2.

A colony multiplex quantitative PCR-based 3S3DBC method and variations of it for screening DNA libraries

2.1. Introduction

At the beginning of the *D. japonica* planarian genome project, we constructed a genomic DNA library. Considering the traditional screening methods of DNA libraries are expensive, inefficient, time and labor-consuming, a new method is necessary for researchers to screen out genes of interest for further biological experiments, and survey the genome characters for the later genome project. Here I designed a new high-efficiency screening methods to address this purpose.

2.1.1. Background

Although next-generation sequencing (NGS) is widely used at present, and has been used to assemble many genomes, DNA libraries still have irreplaceable roles. Firstly, by screening a DNA library, researchers can pick desired clones and get very precise sequences of specific genes within those clones. Secondly, even if a genome can be assembled from NGS data, there will still be gaps and uncertain DNA regions that

need to be confirmed; screening a DNA library and sequencing targeted clones can help to achieve gap-closure and to evaluate and correct the assembled genome. Furthermore, assembly of some complicated genomes (with too many repetitive sequences or a high rate of heterozygosity or other variability) is extremely hard to accomplish by NGS alone, and therefore sequencing of DNA libraries is still usually an indispensable method for achieving whole-genome sequencing at present [33, 34]. Thus, a DNA library is still a valuable resource for work such as molecular cloning, physical mapping of genes, and comparative genomics.

To take the best advantage of DNA libraries, a large number of library screening methods have been developed during the past few decades [35, 36]. In early studies, library screening was mainly based on hybridization between clones containing recombinant DNA vectors (bacteriophage, cosmid, plasmid or fosmid), and specific probes (radioactive or synthetic oligonucleotide probes) [37-44]. Later, to avoid the low signal-to-noise ratio and considerable cost of hybridization, PCR-based screening methods were developed. In PCR-based methods, one first isolates DNA from pools of clones, and then uses primers designed to screen the positive pools containing desired clones by PCR, and finally identifies the positive clones by hybridization or further PCR reactions [45-52]. The development of colony PCR [53-55] and quantitative PCR (qPCR) [56] made this method much easier to perform.

However, despite the advantages of the PCR-based screening method, it is still not efficient enough for some requirements of modern genomic research. The arbitrary, inefficient pooling strategy, the culturing of clones, DNA extraction, numerous PCR

steps and electrophoresis procedures generally used for PCR-based screening are time-, money-, and labor-consuming and produce many false-positive results. In order to take better advantage of DNA libraries, here we describe an efficient colony multiplex quantitative PCR-based 3–step, 3-dimension, and binary-code (3S3DBC) method for screening of DNA libraries.

2.1.2. Mathematics of DNA library screening

Screening out one desired clone from a DNA library can be considered a mathematical problem, i.e., how to distinguish one positive sample among a large number of samples. A good library screening method means an optimal solution that needs the least time (i.e., few detection steps, e.g., few PCR rounds in the case of PCR-based screening) and least labor (i.e., simple pooling procedure and small detection number, e.g., a small number of PCR reactions needed in the case of PCR-based screening) to solve this mathematical problem. Usually, screening uses one of three different methods: the dimension-based method, bisection-based method or binary code-based method.

Dimension-based method

Dimension-based methods have been widely used for screening. A one-dimensional method means that all samples are aligned in a one-dimensional line, and the desired sample can be detected by screening them one by one (Figure 2-1). In a two-dimensional method, all samples are arranged into a two-dimensional square (Figure 2-2). After pooling the samples of each row and column, and screening these pools, the desired sample is identified as occupying the intersection of the positive row and column. A three-dimensional method means that all samples are arranged into a three-

dimensional cube (Figure 2-3). After pooling the samples of each layer in the three-dimensional cube and screening them, the desired sample is the sample located at the intersection of the three positive layers. Similarly, in a four-dimensional method, all samples can be arranged into a four-dimensional hypercube, and after pooling the samples of the cubes in each dimension and screening them, the desired sample is located at the intersection of the positive cubes. Furthermore, samples can also be arranged into five-, six-, and so on, dimensions. As it is nearly impossible to depict arrangements in more than three dimensions in our three-dimensional world, some variants of a high-dimensional method can be used for pooling (Figure 2-4 and Figure 2-5). The lowest detection number "n" of the dimension-based method equals $n=D*\sqrt[D]{N}$ (D is the dimension number and N is the total number of samples). A higher dimension method seems better because the higher the dimension number, the lower the detection number required; however, at the same time, it requires more complicated procedures for pooling samples.

Figure 2-1. One dimension-based method

Samples are aligned in a one-dimensional line, and the desired sample can be detected by screening them one by one.

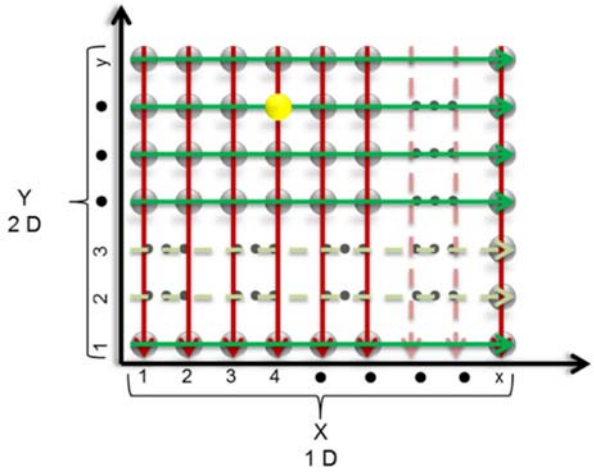


Figure 2-2. Two dimension-based method

Samples are arranged into a two-dimensional square. After pooling the samples of each row and column, and screening these pools, the desired sample is identified as occupying the intersection of the positive row and column.

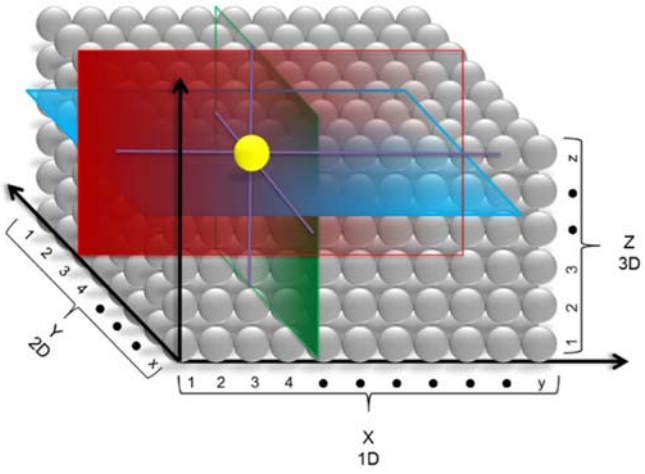


Figure 2-3. Three dimension-based method

A three-dimensional method means that all samples are arranged into a three-dimensional cube. After pooling the samples of each layer in the three-dimensional cube and screening them, the desired sample is the sample located at the intersection of the three positive layers.

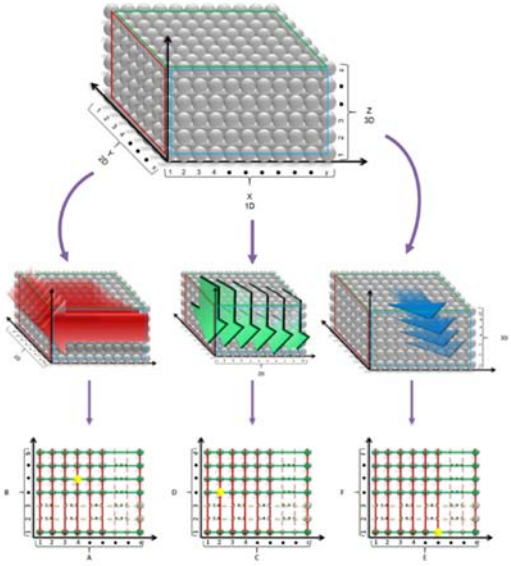


Figure 2-4. Six dimension-based method.

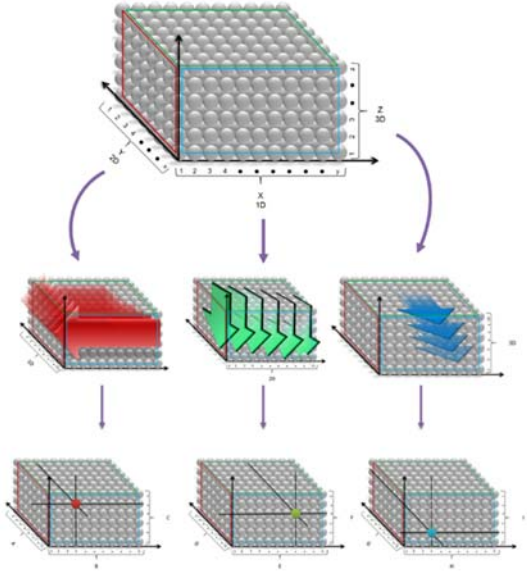


Figure 2-5. Nine dimension-based method.

Bisection-based method

Another simple and commonly used method for screening out one sample from a large sample set is the bisection-based method (Figure 2-6). This method requires several detection steps. In each step, the sample universe is divided into two equal subsets (pools of samples), and the positive subset is detected. The division and detection procedure is repeated at each step until only one sample, which is the desired sample, is left in the final positive subset. The number of detection steps, "n", of this method is $n = 2*\log_2^N-1$ (N is the total number of samples). This exponential bisection method only needs a small number of detections; however, it needs many detection steps (and thus long working time), because the detection at each step is based on the result of the previous detection step.

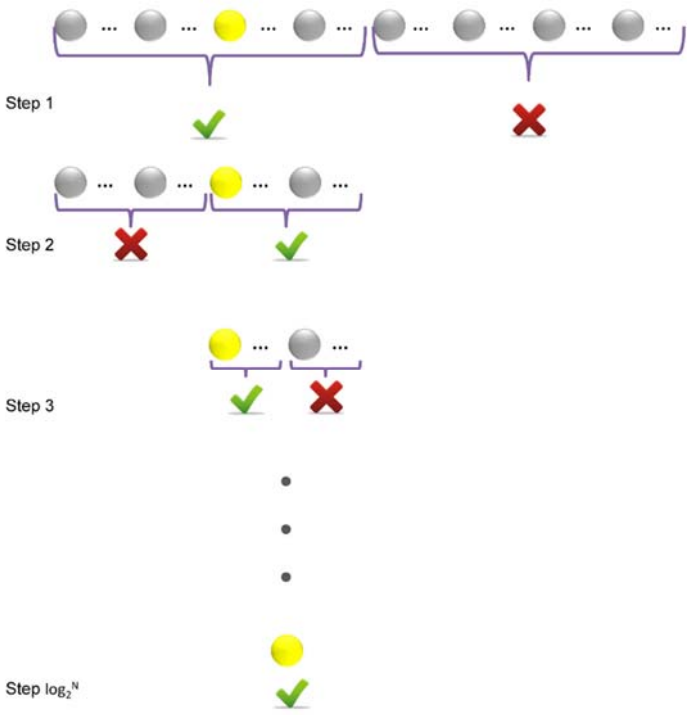


Figure 2-6. Bisection-based method.

Bisection-based method requires several detection steps. In each step, the sample universe is divided into two equal subsets (pools of samples), and the positive subset is detected. The division and detection procedure is repeated at each step until only one sample, which is the desired sample, is left in the final positive subset.

Binary code-based method

In the binary code-based method (Figure 2-7), at first, all samples' assigned decimal code numbers are aligned in a column. Then, all of these decimal code numbers can be converted to binary code numbers, which are arranged in a matrix. In this binary code number matrix, for each column, samples whose assigned binary code numbers include the digit 1 are mixed to form a pool. After detection, each positive pool is marked "1", and each negative pool is marked "0". The final binary code can be converted back into a decimal code that indicates the position of the real positive sample. The detection number (n) equals the total number of pools, $n = \log_2^N$ (N is the total number of samples). This method uses a smaller number of detections compared with the other methods, and all detections can be performed simultaneously and independently. For instance, to screen out one specific sample from 262144 samples requires 193 detections by the 3-dimensional method, 35 detections by the bisection-based method, but only 18 detections by the binary code-based method.

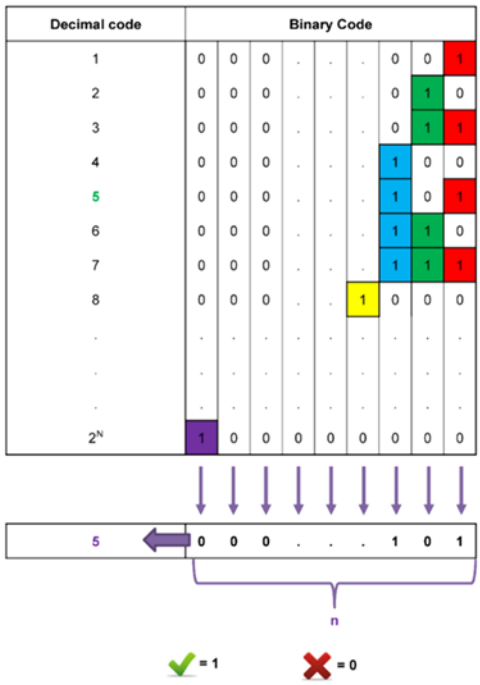


Figure 2-7. Binary code-based method.

In this binary code number matrix, for each column, samples whose assigned binary code numbers include the digit 1 are mixed to form a pool. After detection, each positive pool is marked "1", and each negative pool is marked "0". The final binary code can be converted back into a decimal code that indicates the position of the real positive clone.

2.1.3. Issues that must be considered when designing the practical details of the screening method

Although several mathematical solutions have thus been provided for screening out one desired sample from a sample universe, real cases are much more complicated than the simple mathematical problem. In our actual case, for example, our *D. japonica* planarian DNA fosmid library contained 161,280 clones, which equaled 4-fold coverage of one planarian genome-equivalent. Therefore, a single copy gene could theoretically have 4 possible clones in this library. How to decrease the screening time, labor and cost, but increase the screening sensitivity and accuracy, is always the goal in designing any screening experiment. Accordingly, issues of the detection method and the pooling strategy of the screening design and some other practical considerations are discussed next (Table 2-1).

Detection method

The detection method generally used for DNA library screening has changed over time from hybridization to PCR. Although the PCR method has the advantage of high sensitivity, the clone culturing and DNA extraction before PCR, and the large number of PCR rounds and electrophoresis procedures, are still time- and labor-consuming, which makes the library screening tedious work. In this study, we used a colony multiplex quantitative PCR detection method to make the library screening less tedious. Quantitative PCR (qPCR) can use stored clones directly, and the detection is based only on checking the qPCR dissociation curve without requiring culturing of clones, DNA extraction, or electrophoresis steps. Multiplex primer sets can make it possible to screen several genes simultaneously (different qPCR products can be distinguished by their unique peak position in the dissociation curve). Using more

primer sets can save time, labor and money by reducing the number of PCR reactions and PCR rounds in the screening; however, it also decreases the detection sensitivity and screening accuracy, since too many primer sets will disturb the PCR reaction and the ability to distinguish among PCR products.

Table 2-1. Relationships among issues that must be considered when designing a screening method

Positive and negative relationships of issues (x, y and z) that must be considered when designing a screening method are listed in the table. The functions relating x, y and z (z=f(y); y=f(x)), and their suitable solutions depend on different research laboratories' particular situations.

Issues that must be considered during screening		X	
		Pooling density	Multiplex primers
y	Detection Number	-	-
	Detection Step	-	-
	Detection Sensitivity	-	-
	Detection Accuracy	-	-

Issues that must be considered during screening		Y		
		Number of PCR reactions	Number of PCR rounds	Automation
z	Time	+	+	-
	Labor	+	+	-
	Cost	+	+	-

Pooling density and PCR detection sensitivity

Pooling is used in nearly all PCR-based library screening; and pooling density is one of the most crucial issues that need to be considered because it is related to PCR sensitivity and accuracy, and the number of PCR reactions and rounds, which in turn affect the time, labor and money used for screening. For the dimension-based method, two-dimensional (row/column pools) [49, 57] and three-dimensional

(plate/row/column pools) pooling [50-52, 55, 56, 58-60] strategies have commonly been used. Various higher-dimensional pooling strategies [42, 47, 61, 62] have also been described. Higher-dimensional pooling helps to decrease the number of PCR reactions necessary during screening, and thus to decrease the time, money and labor consumed. However, higher-dimensional pooling is also computationally complicated, and the pooling itself is labor-consuming if appropriate automation machinery is not available. At the same time, higher-dimensional pooling means higher pooling density, which can lead to the concentration of an individual target sample becoming too low to be detected by PCR. Similarly, a bisection-based method may not be suitable for practical screening work, because, in the initial pooling step, this method needs to separate all samples into two super pools, which may result in such high pooling density that the sensitivity of PCR detection is not high enough. To use a section-based method, some studies tried to decrease the pooling density by separating clones into a larger number of super pools [48]. The binary code-based method has a similar problem. This method has been used for detecting protein-protein interaction [63]; however, no such method has been described for DNA library screening. Overly high pooling density is one crucial problem that would impede its application. So, considering the PCR detection sensitivity, a suitable pooling density is important for a PCR-based screening method. To test the PCR detection sensitivity in our planarian DNA library screening, we made and screened some pools with an increasing number of clones. PCR and qPCR experiments showed that PCR was not sufficiently sensitive to detect one sample from a super pool containing a mixture of 10^4 or more clones.

Detection accuracy and false positive problem

In a practical case, if a DNA library covers more than one genome-equivalent, the false-positive issue should be taken into account. More positive samples induce more false-positive results in dimension-based and binary code-based methods, and increase the detection number in section-based methods (Figure 2-8). In dimension-based methods, the largest possible number of false-positive results (N) equals n^D-n (where n is the number of true-positive results, and D is the number of dimensions) (Figure 2-9). As the number of dimensions increases, more false-positive results appear, which results in low PCR detection accuracy. The case is much worse for binary code-based methods, because the binary code method can only detect one unique sample from a pool. If there is more than one positive sample in the same pool, the binary code method will give a wrong result that will seriously affect the screening accuracy (Figure 2-10). So, in order to avoid false-positive results, the higher the probability that one super pool contains only one desired clone after pooling, the better. In theory, one single copy gene has 4 positive clones in our planarian DNA library, so the probability (P) that one positive clone will appear in only one super pool equals $P=1*\frac{n-1}{n}*\frac{n-2}{n}*\frac{n-3}{n}$ (n is the number of super pools). The tradeoff between P and n should thus be carefully balanced.

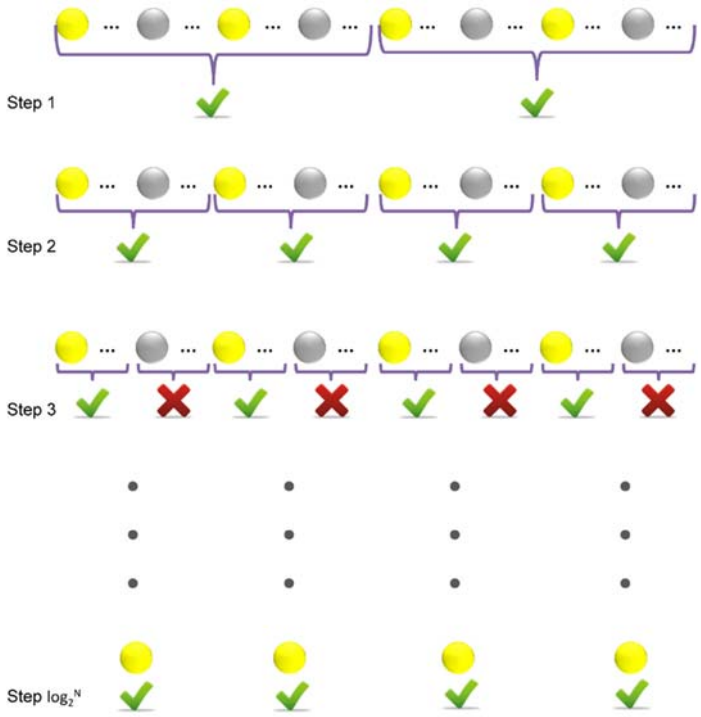


Figure 2-8. Problem of section-based method

When multiple copies of the desired sample are located in the same pool.
Although the section-based method can give the correct result, it requires many more detection numbers

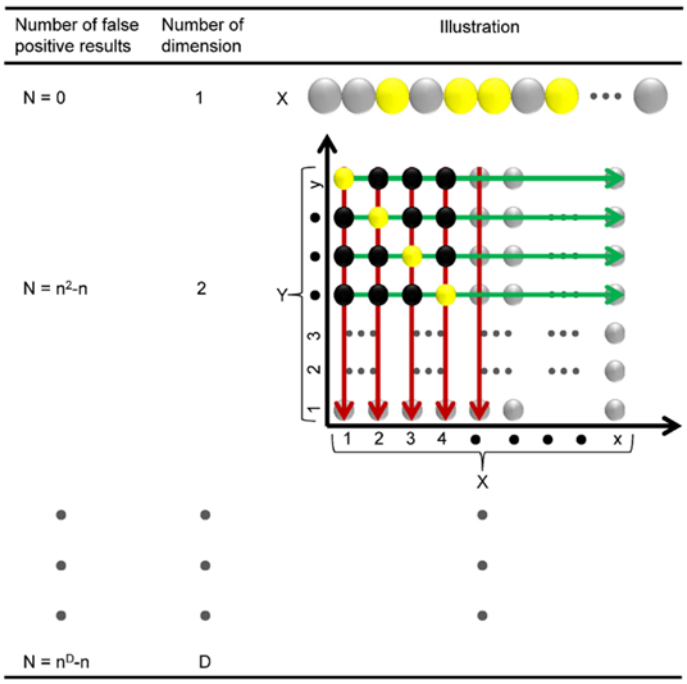


Figure 2-9. Problem of dimension-based method

False-positive results are a common issue when more than one desired sample exists in the same pool. For the dimension-based method, when the dimension number is greater than 1, the largest number of false-positive results “N” equals nD-n (D is the dimension number and n is the number of true positive samples in the library).

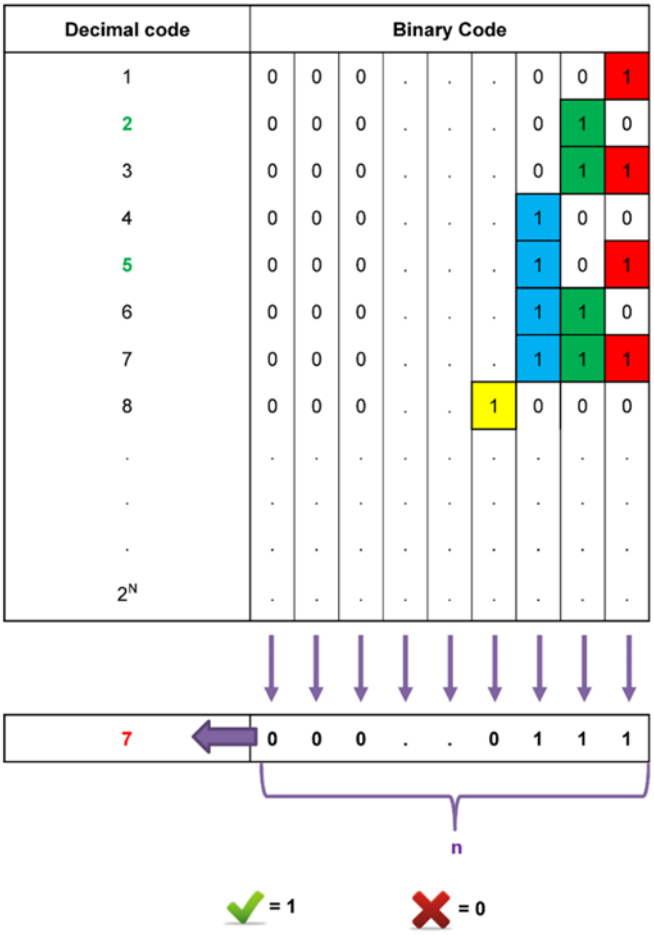


Figure 2-10. Problems of binary code-based method.

The binary-code method even produces a wrong result in this case (for example, if samples No. 2 and No. 5 are both positive samples, this method will give a wrong result: No. 7).

Automation

It is obvious that as the pooling density increases, manual pooling will become a much more complicated challenge, and a labor- and time-consuming task that would probably be impossible to accomplish. This point appears to be one of the main limitations of the high-density pooling strategy. Therefore, the use of a suitable robot should be very important for optimizing the pooling strategy, simplifying screening procedures and reducing cost. In our actual case, our robot could only make one clonal pool from each 384-well plate. So, because of the robot performance limitation, we used two steps to separate a super pool into 3 dimensions (one step for plate dimension and one step for row and column dimensions).

2.2. Results

Screening out a gene by the colony multiplex quantitative PCR-based 3S3DBC DNA screening method

Based on the issues described above, the final solution we devised for our *D. japonica* fosmid library screening work was a “colony multiplex quantitative PCR-based 3S3DBC screening strategy” (Figure 2-11). We used the known sequence from clone DJF-033N19 (this name indicates that the location of this sequence in our *D. japonica* fosmid library was Plate No. 033, Row No. N, Column No. 19) as a positive control, and DjPiwiB was the desired gene we wanted to screen out from this library. Our strategy needed only 3 steps to identify one desired clone from our whole library.

In the first step, in order to make the probability at least 80% that our desired clone appeared no more than once in a super pool, we mixed a total of 161280 clones into 28 super pools. Each super pool was constructed by using fifteen 384-well plates. We chose the number fifteen because it is 2^n-1 (n is a natural number), which could be efficiently used in the following binary code configuration. The total mixed clone number in one super pool was 5760, which was small enough (less than 10^4) for qPCR detection sensitivity. After the detection of 28 simultaneous qPCR reactions, positive super pools were found as verified by qPCR disassociation curves (Figure 2-12); moreover, this experiment verified that multiplex primer sets can be successfully used to screen multiple genes simultaneously with this method. The positive super pools for the positive control sequence DJF-033N19 were super pool No. 1 (384-well plates No. 1 to No. 15) and super pool No. 3 (384-well plates No. 31 to No. 45), and

the positive super pools for the DjPiwiB gene were super pool No. 1 and super pool No. 2 (384-well plates No. 16 to No. 30).

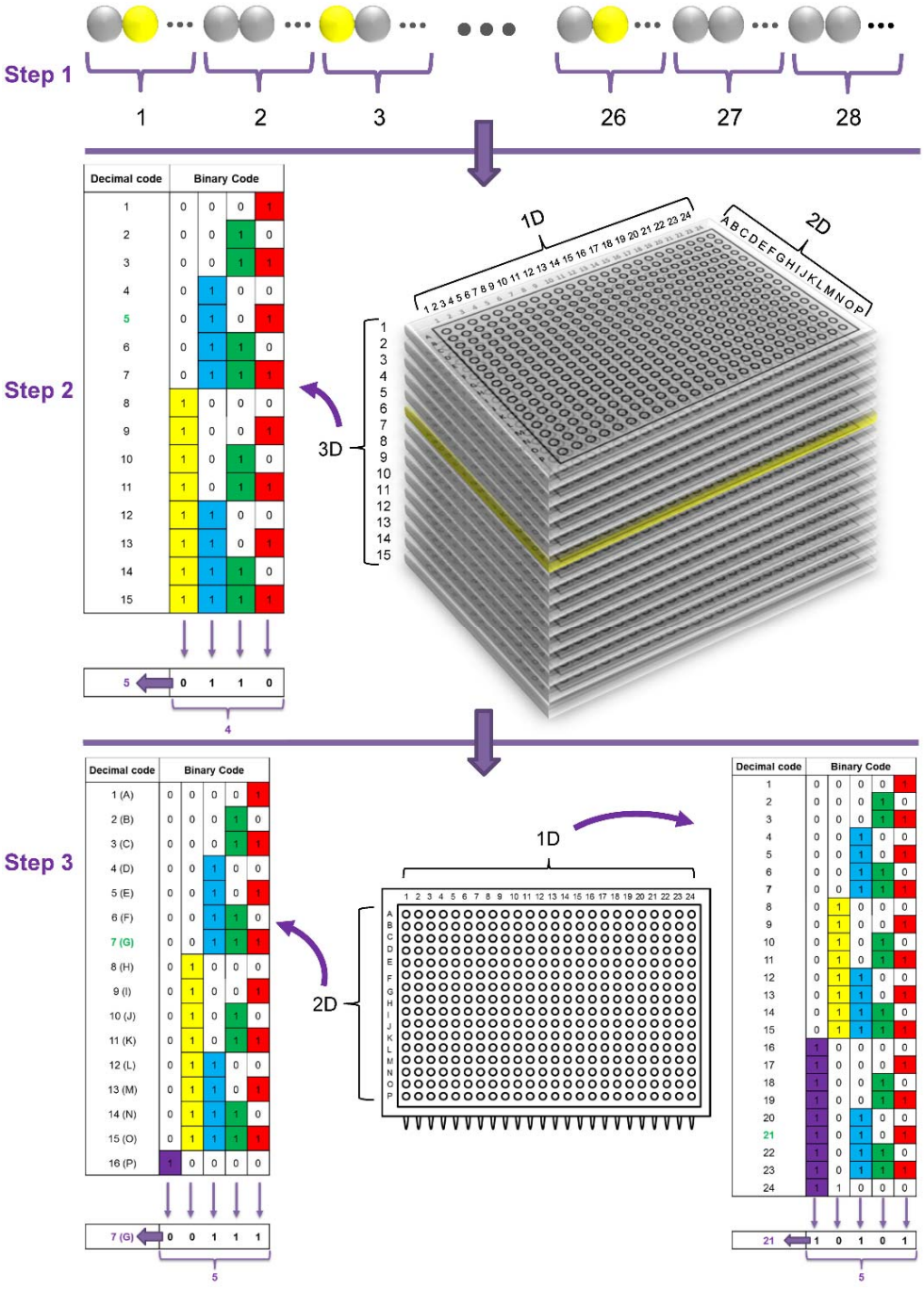
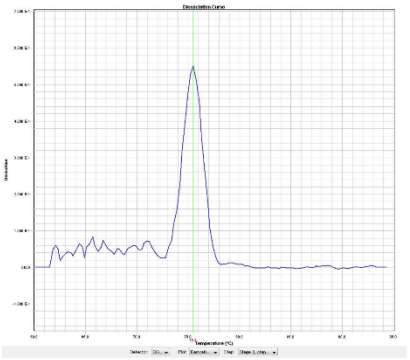
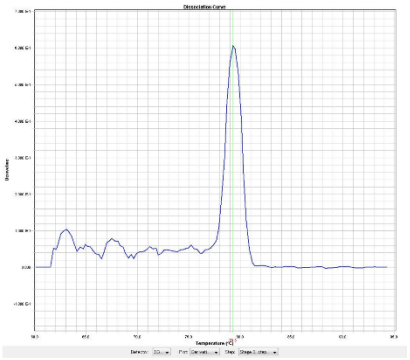


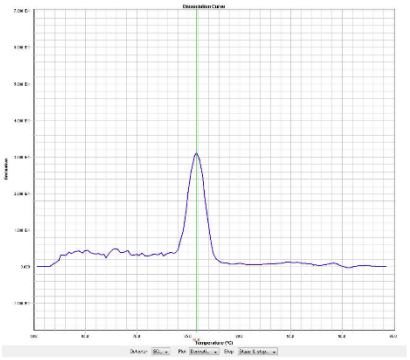
Figure 2-11. A colony multiplex quantitative PCR-based 3S3DBC DNA screening method for planarian DNA library screening.



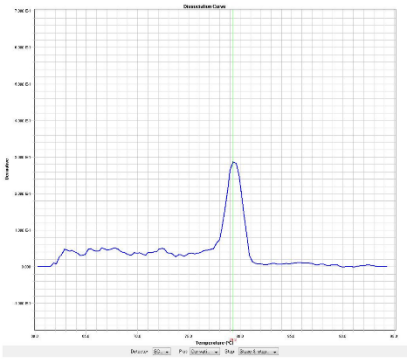
a. DjPiwiB_gDNA



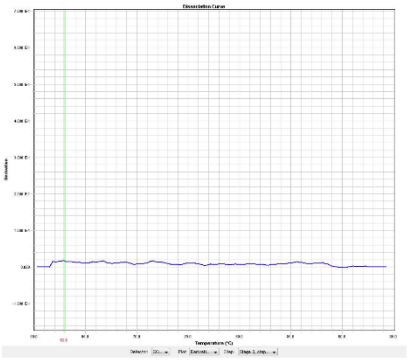
b. DJF-033N19_gDNA



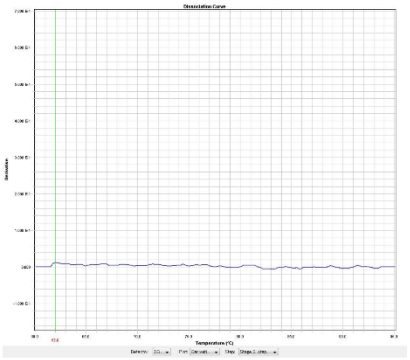
c. DjPiwiB_super pool No. 1



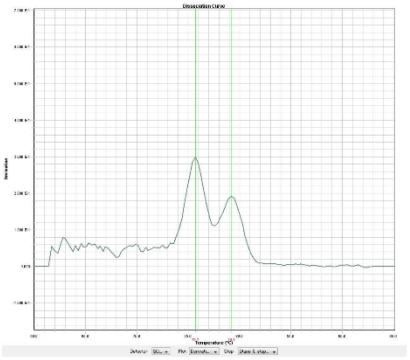
d. DJF-033N19_super pool No. 1



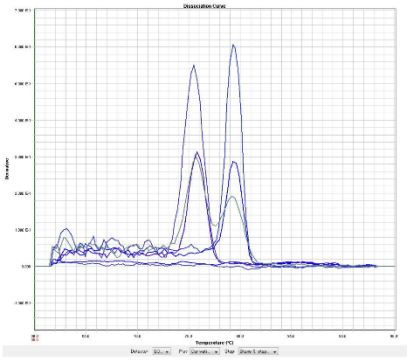
e. DjPiwiB+DJF-033N19_super pool No. 2



f. DjPiwiB+DJF-033N19_H_2O



g. DjPiwiB+DJF-033N19_super pool No. 1



h. Merged

Figure 2-12. Representative results of qPCR disassociation curves during screening.

This figure was snipped from our results obtained using ABI qPCR analysis software SDS 2.4 standalone, and shows representative results of disassociation curves generated from the first step of 3S3DBC DNA screening. Positive DNA template (10 ng of *Dugesia japonica* genomic DNA) was used to detect the qPCR specificity of the respective primer set of DjPiwiB (a) and DJF-033N19 (b). Each primer set could also generate the specific qPCR product after qPCR in a real library super pool (c and d). No nonspecific amplification was detected in the negative super pool or in the H_2O negative template control, using either DjPiwiB or DJF-033N19 multiplex primer sets (e and f). The expected qPCR disassociation curve (double peak shows two qPCR products) was detected in the positive library super pool using DjPiwiB and DJF-033N19 multiplex primer sets (g), and the dissociation temperature values of the two peaks corresponded to the same values obtained in the positive template control and positive super pool results. All figures from (a) to (g) were merged to produce (h).

In the second step, the positive cuboid super pool was pooled by its third dimension - the plate layer dimension. Then, the binary-code pooling method was further used to obtain much higher pooling density. A total of 4 simultaneous qPCR reactions were sufficient to screen out one positive plate from 15 plates. From this step, the positive plates for the control gene clone DJF-033N19 were plate No. 004 and plate No. 033, which indicated that we found the correct positive plate and one extra positive plate. Also for the DjPiwiB gene, we found two positive plates (No. 006 and 025).

In the third step, each positive plate was further pooled into a row dimension pool and a column dimension pool. The binary-code pooling method was again used to make higher pool density of the pools for each dimension. Only 5 simultaneous qPCR reactions were needed to screen out the positive row or column, and to finally distinguish the desired clone from the intersection between the positive row and column. We finally found two positive clones (DJF-033N19 and DJF-004C08) for the positive control DJF-033N19, and two positive clones (DJF-006G21 and DJF-025C08) for the DjPiwiB gene. The positive control demonstrated that we obtained a correct screening result using this method, and later sequencing of the two DjPiwiB gene-positive clones also demonstrated the accuracy of this screening result (Figure 2-12). Accordingly, in the case of our planarian fosmid library, this colony multiplex quantitative PCR-based 3S3DBC screening method only needed 42 qPCR reactions and less than 7 hours to screen out one desired clone from a DNA library containing 161,280 clones. With the same methods, several other genes were also screened out, including DjTh, Dj1020HH, Dj01A, *et al*.

2.3. Discussion

Here we first described general considerations regarding the detection method and pooling strategy for screening a DNA library, and then detailed our rapid and low-cost new screening method – a colony multiplex quantitative 3S3DBC method, which is superior to the conventional 3D screening method (Table 2-2). By considering the time, labor, cost, detection sensitivity, detection accuracy, automation and other issues in a particular DNA library screening, this method can be modified to produce several other variant methods. For example, if a DNA library has very few clones which altogether cover less than one genome-equivalent, and if a suitable automation robot is available to perform complicated pooling, a 1-step binary code screening method might be adequate. However, usually a DNA library contains hundreds of thousands of clones that cover more than 1 genome-equivalent, so an additional pooling step should be added to make more super pools in order to decrease pooling density and to increase the PCR detection sensitivity and the probability that one positive clone will appear no more than once in a super pool (2-step binary code pooling). Dimension-based pooling is simpler than binary code-based pooling, if a pooling robot is not available. Increasing the number of PCR steps can also help to reduce pooling complications. In addition, since overly high pooling density can reduce PCR detection sensitivity, more pools are required, or a pre-PCR step before pooling is an alternative method to increase PCR detection sensitivity [56]. We also showed here that multiplex qPCR can be used to screen multiple genes simultaneously (we used two primer sets in our experiment). Presumably, more genes could be screened

simultaneously by using more than 2 primer sets, but of course, the number that could be screened would be limited by the same factors (i.e., primer sets design, rate at which primers anneal to their targets, buffer constituents, and annealing temperature, which affect PCR sensitivity and specificity) that limit any multiplex PCR method [64-66]. In conclusion, by making trade-offs and using flexible combinations of approaches to address these issues, our 3S3DBC DNA library screening method can be modified and widely used for screening a variety of DNA libraries, and further employed for other screening-like experiments such as protein-protein interaction and hybridization tests.

Table 2-2**.** Comparison of screening one desired clone from DNA library by 3S3DBC screening method and conventional 3-dimensional method.

Features	S3DBC screening method strategy in our actual case	Conventional PCR screening using 3D pooling strategy
Maximum number of 384-well plates in one super pool	~15	~10
Number of PCR reactions needed to identify a positive super pool	n/15	n/10
Reactions needed to identify the plate ID for one positive super pool	4	10
Reactions needed to identify the clone ID from one positive plate	10	40
Total number of reactions needed to get positive BAC clone ID from whole library	n/15 + 14	n/10 + 50
Multiplexing possibility	Yes	No
Checking on agarose gel No	NO	Needed
Cost	~ $0.34 per qPCR reaction	~ $0.3 per PCR reaction + agarose gel electrophoresis
Procedure duration to screen out one desired clone	~ 6 h	~ 12 h
Data retrieved	Automatically reported from qPCR dissociation curve	Manually check from gel photos

“n” is the total number of plates in the DNA library

2.4. Materials and Methods

D. japonica planarian genomic DNA library

The DNA library was constructed by the National Institute of Genetics in Tokyo Japan. Colonies were picked into freezing solution that contained ampicillin (50 ug/ml), and stored at -80℃

Pooling

Clones (4 ul/clone) from each 384-well plate were mixed into a plate pool using a BioTech EDR-384SⅡ Multi-functional Table Top Pipette Station. Row- and column-pools from each plate were made manually using Eppendorf multichannel pipettes. Each 15-plate pool was further mixed to produce a super pool.

qPCR kit and conditions

The QuantiTect SYBR Green PCR kit was used. According to the manufacturer's qPCR reaction kit instructions, sometimes adding a small amount of extra ExTaq enzyme (0.05ul/10ul reaction) will yield a better result. The qPCR reactions were performed on an ABI PRI明M® 7900HT Sequence Detection System, and the qPCR cycling conditions were: 95°C for 10 mins, [95°C for 30 seconds, 57°C for 30 seconds, 72°C for 50 seconds] (40 cycles), 72°C for 7 mins, followed by dissociation curve analysis.

Primers for DjPiwiB gene and positive control gene

The DjPiwiB primer set was DjPiwiB_Fw (5’-ATGGATCCCATGGCTCCTAATG-3’) and DjPiwiB_Rv (5’-TGCACAGGGACAGGTACACG-3’). The clone DJF-033N19 (plate No.33, row N and column No. 19) was sequenced previously. Its location and known sequence were used for a positive gene control. The primer set for this sequence was DJF-033N19_Fw (5’-AATCGGGAGAACGGGAAGATGTG-3’) and DJF-033N19_Rv (5’-GCCATTCGGAACTTGAGCTTGAC-3’).

CHAPTER 3.

De novo assembly and annotation of *D. japonica* genome

3.1. Introduction

3.1.1. The genome of *D. japonica*

A genome is the complete genetic information contained in an organism's set of haploid chromosomes, and it contains all of the information needed to build and maintain that organism. The genome of *D. japonica* has 2n=16 chromosomes (Figure 3-1a). This is twice the number of chromosomes (Figure 3-1b) in the planarian *Schmidtea mediterranea* (*S. mediterranea*), which is another freshwater planarian species and which possesses 2n=8 chromosomes, and whose sexual strain was estimated to have a 769.5 Mb genome, unpublished data from J. Spencer Johnston).

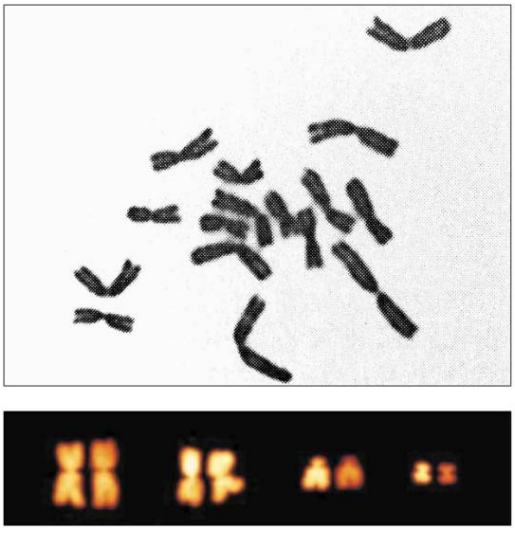


Figure 3-1. The karyotype of two planarian species

a. The karyotype of *D. japonica* (from [67]); b. *Schmidtea mediterranea* (from [68])

The genome of *S. mediterranea* was sequenced using traditional Sanger sequencing, and the data have been available since 2008 [69]. It is a complicated genome - it is A/T rich (69%) and very repetitious (46% of the total genome). Although *D. japonica* is not evolutionarily close to *S. mediterranea* [2] (Figure 1-1), considering that they both belong to the same family, the genome of *D. japonica* may also likely to be complicated.

Before this genome project, the only available omics information of *D. japonica* was its partial transcriptome (EST data)[32], i.e., information about the coding region. To maximize the chances of identifying all of planarian's coding genes from transcriptome data alone, transcriptomes of a large number of planarian individuals at different developmental stages and under various conditions would need to be sequenced and assembled. Although our lab has already performed lots of this work, the fact told us that it was a time- and labor-intensive work, and moreover, transcriptome data alone could not give scientists enough information for further research on molecular mechanisms.

In addition to transcriptome projects, a genome project is required for many applications. The planarian genome sequencing we performed here is the first planarian genome sequencing project in Asia, and the data obtained will help all *D. japonica* researchers; all of the organism's genes will be sequenced, and the information obtained will indicate all alternative splicing transcripts and their exon-intron structures; non-coding regions will be identified, such as those for ncRNAs,

transposons, promoters and enhancers; and genome annotation results will help to guide biological experiments in the future.

Because of the heterozygosity and repeats in the planarian genome, this genome sequencing project was not expected to be able to generate complete chromosome sequences. The aim of this project was rather to get a high-quality draft genome (defined by Chain [70]), with the understanding that some sequence errors and misassemblies are likely to be present in it.

3.1.2. Genome sequencing – three generations

Genome sequencing is a laboratory procedure in which the complete DNA sequence of an organism's genome is determined. The DNA sequence is the most fundamental level of information of a gene or genome, as they contain the instructions for building an organism. No genetic function or evolution could be completely understood without obtaining such information. Sequencing technology to determine DNA sequences has been advancing for more than 40 years, and now there are three generations of sequencing systems.

3.1.2.1. First-generation sequencing:

First generation sequencing, also called Sanger sequencing, was developed during the 1970s [71, 72]. Since the early 1990s, DNA sequence determination has almost exclusively been carried out with capillary-based, semi-automated implementations of Sanger biochemistry (Figure 3-2a). The method is famous for its contribution of all of the DNA sequencing work in the Human Genome Project (1995-2003)[73]. The

advantage of Sanger sequencing is that it can generate long sequencing reads with high quality. However, the greatest limitation of this method is that it is time- and cost-consuming: for example, the human genome took 13 years and 3 billion U.S. dollars.

3.1.2.2. Second-generation sequencing (NGS)

Second-generation sequencing (next-generation sequencing: NGS) techniques provide high speed and high throughput, e.g., genome projects that require several years with Sanger techniques can be finished in a few weeks or even a few days. There are three main NGS platforms from three different companies: Roche, Illumina, and ABI. Although these platforms are quite different in their sequencing biochemistry as well as in how the array is generated, their work flows are conceptually similar [74] (Figure 3-2b).

Roche 454

The start of the NGS revolution was clearly marked by the appearance in 2004 of the Roche 454 sequencing system, which successfully achieved the almost complete genomes of *Mycoplasma genitalium* and *Streptococcus pneumonia* [75]. In the Roche 454 sequencing system, DNA is sheared into small fragments to which adapters are ligated, and the fragments are attached to beads, mixed into an emulsion, amplified by emulsion PCR, and deposited in wells of a picotiter plate, in which sequences are then determined by pyrosequencing [76]. Comparing with other NGS sequencing technology, 454 has can generate longer reads, which are useful for genome projects for repetitious genomes. The cost of this platform is relative high and there are always some homopolymer errors.

Illumina

The Illumina technique was first available for commercial use in 2006. This technique is based on sequencing-by-synthesis chemistry. DNA is sheared into small fragments to which adapters are ligated, then the fragments are attached to the surface of a flow cell, and in situ PCR “bridge” amplification creates clusters that are sequenced by reversible terminator chemistry [76]. Although the sequencing reads are short, the superiority of this technique is obvious – it is cheap and features high throughput. The cost of Illumina sequencing is the cheapest among all NGS sequencing techniques, and the new Illunima platform Hiseq can generate more than 25 Gb genome per day, which makes the Illumina sequencing platforms the most favored at present.

ABI SOLID

ABI released the SOLID sequencing system in 2007. This system is based on sequential ligation of dye-labeled oligonucleotide probes, whereby each probe assays two base positions at a time [76]. The unique point of SOLID is that SOLID sequencing chemistry is based on ligation instead of the DNA polymerase-based chemistry used in Sanger, 454, and Illumina system. Its 2-base encoding system makes this system have an error-correcting function that is suitable for detection of single base mutations. However, this system is not well accepted by researchers because the “color space” output data is not commonly used, and this system does not have obvious advantages over the other systems in price, time or sequencing read length.

Ion Torrent

The Ion Torrent Personal Genome Machine (PGM) was introduced in 2010. It differs from 454, Illumina and ABI SOLID, since this system detects the incorporation of each nucleotide by using a semiconductor pH sensor, rather than by using fluorescent chemistry. Because of this unique technique, Ion Torrent sequencing is cheaper and quicker than that using other NGS systems, although its accuracy and throughput is lower than that of Illumina sequencing.

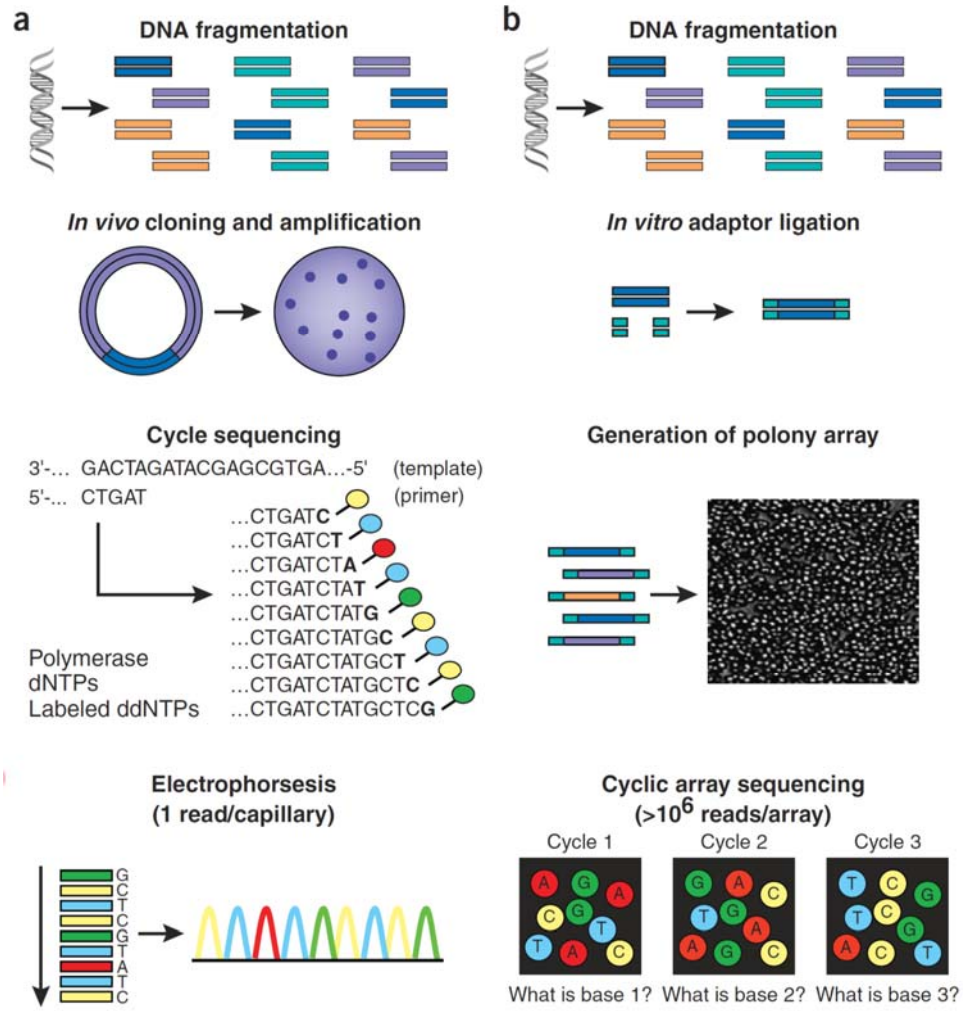


Figure 3-2. Work flow of conventional versus second-generation sequencing (from [74])

a. Conventional Sanger sequencing; b. Second-generation sequencing

3.1.2.3. Third-generation sequencing – SMRT

Third-generation sequencing is a real-time sequencing of single nucleic acid molecules. Typical examples of a 3rd generation sequencer are the PacBio SMRT system (The only commerciallly available 3rd generation sequencer now) and the Oxford Nanopore system. The third-generation sequencing does not need pre-PCR before sequencing, which means there is no PCR bias, and library construction is simple. It is much quicker and can generate much longer reads (10,000 – 40,000 bp) than 2nd generation sequencers. This makes it much more suitable for hyper-variant (high heterozygousity rate and super repetitious) genome projects and whole transcriptome projects. However, the low accuracy rate of the sequencing reads and the high price issues still limit its application.

Considering the advantages and disadvantages (Table 3-1) of each generation of sequencing technology, in this *D. japonica* planarian genome project, we chose both 1st generation (Sanger sequencing for DNA fosmids) and 2nd generation sequencing (Roche 454, Illumina GAIIx and Illumina Hiseq2000 for whole genome shotgun sequencing), in an attempt to generate better assembly results.

Table 3-1 Comparison of sequencing methods

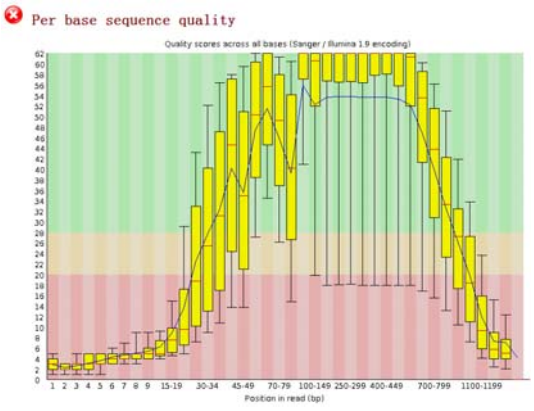
Sequencing Method	Read length	Accuracy	Reads per run	Time per run	Cost per 1 million bases	Advantages	Disadvantages
Sanger(Chain termination)	400 to 900 bp	99.90%	N/A	20 minutes to 3 hours	2400	Long individual reads	More expensive and impractical for larger sequencing projects.
454 (Pyrosequencing)	700 bp	99.90%	1 million	24 hours	10	Long read size. Fast.	Runs are expensive. Homopolymer errors.
Illumina (Sequencing by synthesis)	50 to 300 bp	98.00%	up to 3 billion	1 to 10 days	$0.05 to $0.15	Potential for high sequence yield, depending upon sequencer model and desired application	Equipment can be very expensive. Requires high concentrations of DNA.
SOLiD (Sequencing by ligation)	50+35 or 50+50 bp	99.90%	1.2 to 1.4 billion	1 to 2 weeks	0.13	Low cost per base	Slower than other methods. Has issues sequencing palindromic sequence.
Ion Torrent (Ion semiconductor)	up to 400 bp	98.00%	up to 80 million	2 hours	1	Cheap and fast	Homopolymer errors.
Pacific Bio (Single-molecule real-time sequencing)	10,000 ~40,000bp	99.9999% consensus accuracy; 87% single-read accuracy	50,000 per SMRT cell, or 500–1000 megabases	30 minutes to 4 hours	$0.13–$0.60	Longest read length	Moderate throughput. Equipment can be very expensive.

(Adopted from http://en.wikipedia.org/wiki/DNA_sequencing#Next-generation_methods)

3.2. Raw data generation and quality control

Because errors in sequencing data will cause problems during assembly, data quality estimation and quality control to produce high quality data are required before the next step of de novo assembly. Early quality checks identified significant error rates associated with both data types. For quality control, Cutadapt[77] was used to remove adapter contaminations; SolexaQA[78] was used to trim Illumina sequencing data; PRINSEQ[79] was used to trim Sanger and Roche 454 sequencing data; and then the final sequences quality was estimated using FastQC[80].

During the quality control, we found that although Sanger sequencing reads are generally thought to be long with high quality, our raw Sanger data quality was not good at either the 3’ or 5’ ends (Figure 3-3a). The Illumina raw data showed a significant decline in quality toward the end of the reads (Figure 3-4a). The same low quality of bases appeared at the 3’ends of 454 sequencing data (Figure 3-5a). After quality control, all sequencing data passed the quality check of FastQC, and quality scores of most of the remaining bases were higher than 30 (Figure 3-3b, Figure 3-4b and Figure 3-5b), which means that those bases had an error rate less than 1/1000. Although only 60% of all the raw data was left, after the quality control, it was still enough for de novo assembly because of the high sequencing depth, and the final data had sufficient quality for use in later procedures.

a 

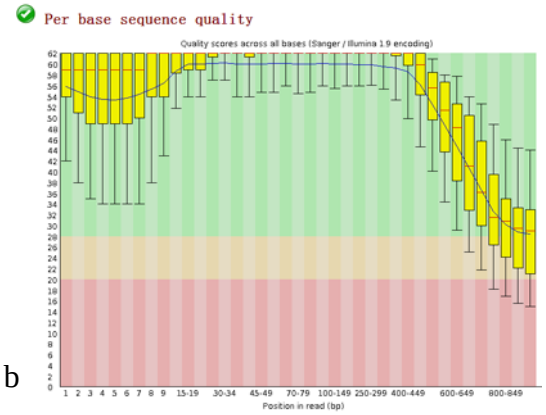
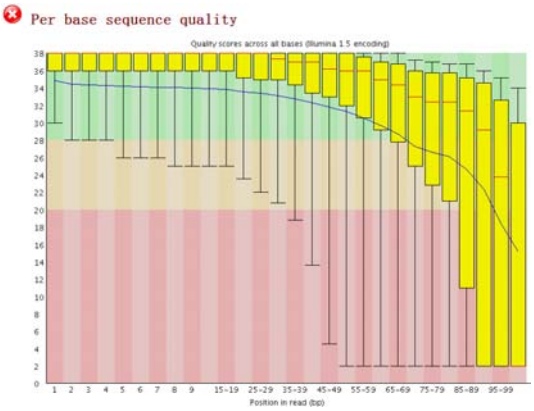
b 

Figure 3-3. Quality control of All Sanger sequencing data
a. The quality of raw data; b. The quality after trimming

a 

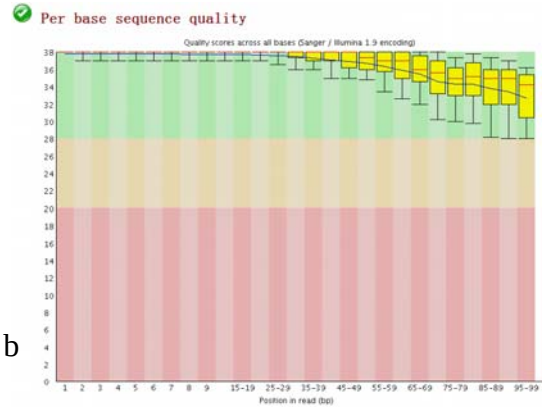
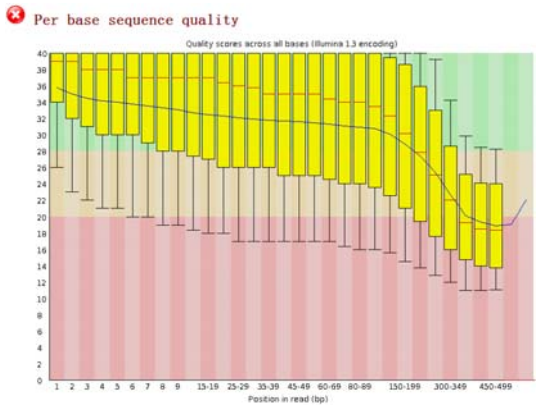
b 

Figure 3-4. Quality control of Illumina Hiseq2000 sequencing data (300bp library)
a. The quality of raw data; b. The quality after trimming

a 

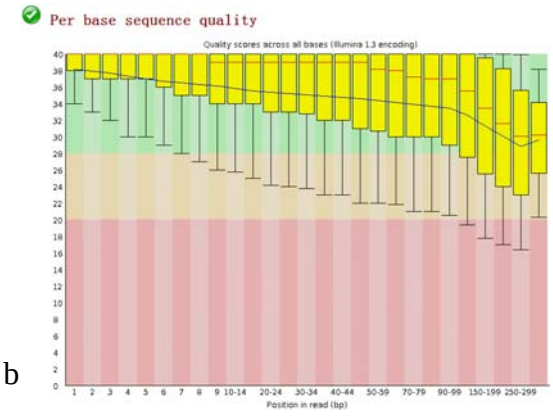
b 

Figure 3-5. Quality control of Roche 454 sequencing data (3kb library)
a. The quality of raw data; b. The quality after trimming

All above figures are generated from FastQC that was used to estimate sequencing data reads quality. Horizontal axis shows base position in reads, and vertical axis shows the quality score. Green, yellow or red color means good, medium or bad quality.

3.3. Estimation of genome characteristics before de novo assembly

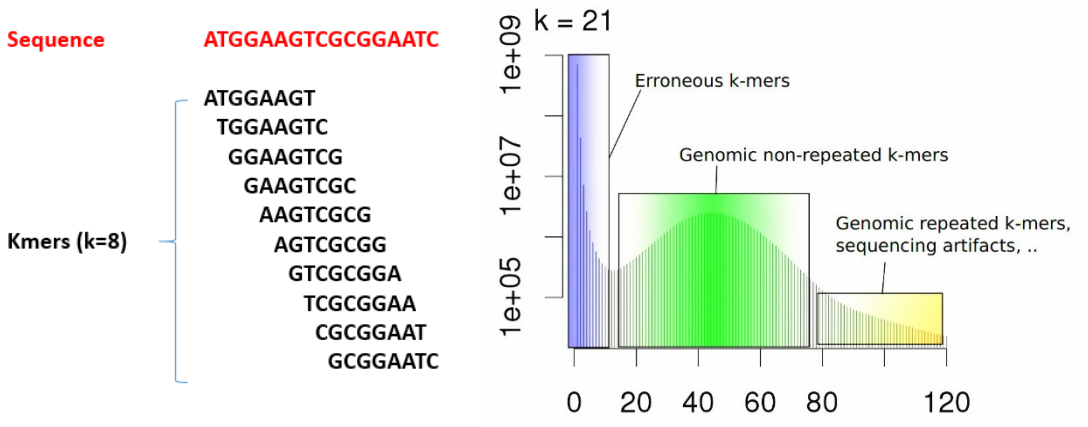
Before de novo assembly was performed, a preliminary genome characteristics survey was needed for estimating the complexity of the genome and the difficulty of the subsequent assembly work.

3.3.1. Kmer frequency for genome survey

In early genome sequencing studies, genomic characteristics were explored by several traditional experimental methods, such as flow cytometry could be used to estimate genome size through C-value [81]; DNA reassociation kinetics (or C_0t analysis) were usually used to measure and classify repetitive sequences in the genome [82]; and molecular markers [83] or DNA microarrays [84] could be used to estimate genome heterozygosity. Nowadays, with the development of new sequencing technology and the application of new assembly algorithms, a kmer frequency method can be used for estimating genome characteristics [85].

Sequencing reads can be consecutively broken into small fragments of sequences with a certain length, called kmers, where the parameter k denotes the length in bases of these sequences. If the read length is “n” and the mer length is “k”, then one read can be broken into n-k+1 fragments (Figure 3-6a). If the k value is large enough, then each kmer should be unique in a given genome, and the distribution histogram of the kmer frequency can normally reflect genome characteristics. For example, if most frequencies are distributed around 50 (the green part of Figure 3-6a), that means that

the sequencing data (the source of those kmers) has 50 times depth of the target genome. In addition, one main peak or two peaks indicates that the genome is haploid or diploid, respectively. Furthermore, kmer frequencies at more than twice the value of the main peak (the yellow part of Figure 3-2b) and smaller than half the value of the main peak (blue part of Figure 3-2b) indicate the repetitiveness of the genome and/or sequencing errors of the data. More relevant discussions and formulas for this application can be found in some papers about algorithms [85, 86].



a b

Figure 3-6. Kmer frequency histogram can be used to estimate genome characteristics

a. Sequencing reads can be consecutively broken into kmers
Read colored sequences is a simulated reads. A kmer means the consecutively broken fragments from the read. In this example, read length n=17 and k=8, so the read can be separated in to n-k+1=16 kmers.

b. Sequencing data quality and genome characteristics can be indicated by kmer frequency histogram (adopted from [86]).
In a kmer frequency histogram, the horizontal axis shows the kmer frequency and the vertical axis shows the number of those kmers. The green main peak shows the genomic non-repeated kmers, and also reflects the data coverage of the genome. The yellow part shows the repeated kemrs, and the blue part shows the error kmers.

3.3.2. *D. japonica* has a complicated genome

Although we tried to perform genome size and heterozygosity estimation using the k-mer frequency histogram derived from trimmed DNA sequencing data, the method that has commonly been used in recent years, the sequencing data of *D. japonica* did not show a typical graph that could be used for genome characteristics estimation by the k-mer frequency model (Figure 3-7). Most of the kmers had low frequencies which might be because of high sequencing error, a highly mutated genome structure or high heterozygosity of the planarian genome. However, in section 3.2, we already showed that our data had high sequencing quality after data quality control, so it seems likely that this strange kmer frequency histogram might have been derived from the genome structure itself. In addition, the large fractions of high-frequency k-mers in the histograms of the *D. japonica* and *S. mediterranea* genomes indicated that planarians have a significant number of repetitive sequences.

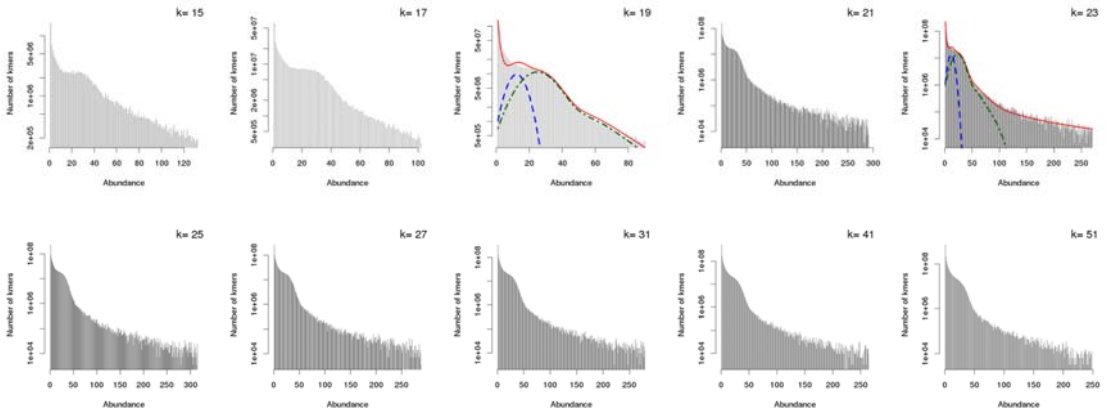


Figure 3-7. Kmer frequency histogram of *D. japonica* genome sequencing data
The 10 figures show kmer frequency of 10 different kmer sizes (15, 17, 19, 21, 23, 25, 27, 31, 41 and 45); however none of them gave a typical kmer model like that shown in Figure 3-6b.The red curves simulated the total kmer distribution, and the green and blue curves simulated homozygous and heterozygous kmers' distributions; however, because the overall distribution didnot match the typical model, those simulations are highly suspicious.

Therefore, in addition, as the above results suggested that the *D. japonica* genome characteristics might be too complicated to be quantified by the k-mer frequency method, we screened out several DNA fosmid clones by a colony multiplex qPCR-based 3S3DBC method, described in Chapter 2, from the *D. japonica* DNA library, and aligned second-generation sequencing genome reads and mRNA reads (downloaded from DDBJ, accession number: DRA002722) to those sequences, with the aim of depicting the genome complexity visually.

By Sanger sequencing, we obtained the full DNA sequences of three DjPiwiB gene fosmids screened from the DNA library (fosmid number: DJF-006G21, DJF-009B05 and DJF-025C08). Mauve alignment [87] among the three sequences showed several Indels (colored blocks expect for the red one in Figure 3-8) which indicated the presence of some retrotransposons around the DjPiwiB genes. DNA alignment of sequencing reads to the fosmid DJF-025C08 showed a large number of SNPs, Indels, and repetitive sequences along it (Figure 3-9a), even in the exon regions of the gene. In addition, alignment of mRNA sequencing reads on the DJF-025C08 sequence (Figure 3-9b), and the MAKER annotation of the sequence (Figure 3-9c) directly showed the repeats (especially retrotransposons) and coding region of this sequence, which further supported our conclusion that the planarian *D. japonica* has a complicated genome.

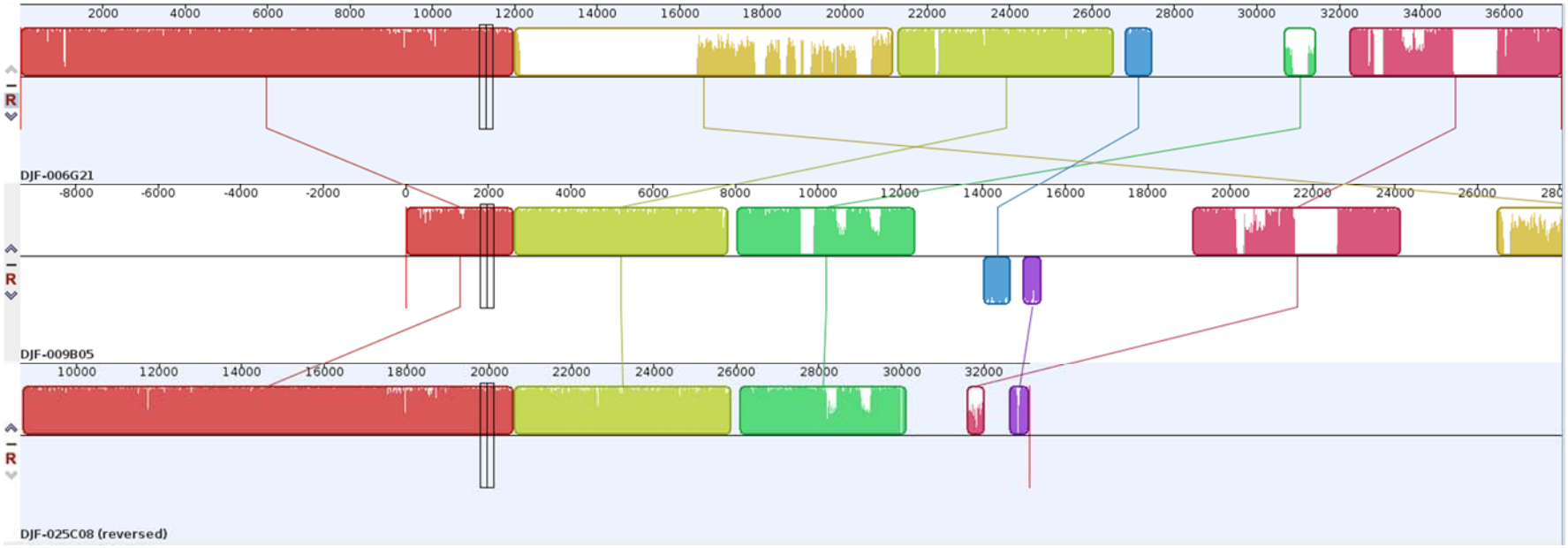


Figure 3-8. Alignment among the three DjPiwiB sequences

This figure shows gene alignment by Mauve software, and indicates the complexity of three structures DjPiwiB genes (from upper to lower: DJF-006G21, DJF-009B05 and DJF-025C08). Each colored block shows similar gene region. The black bars on the red blocks point to the coding region of DjPiwiB.

a

b

c

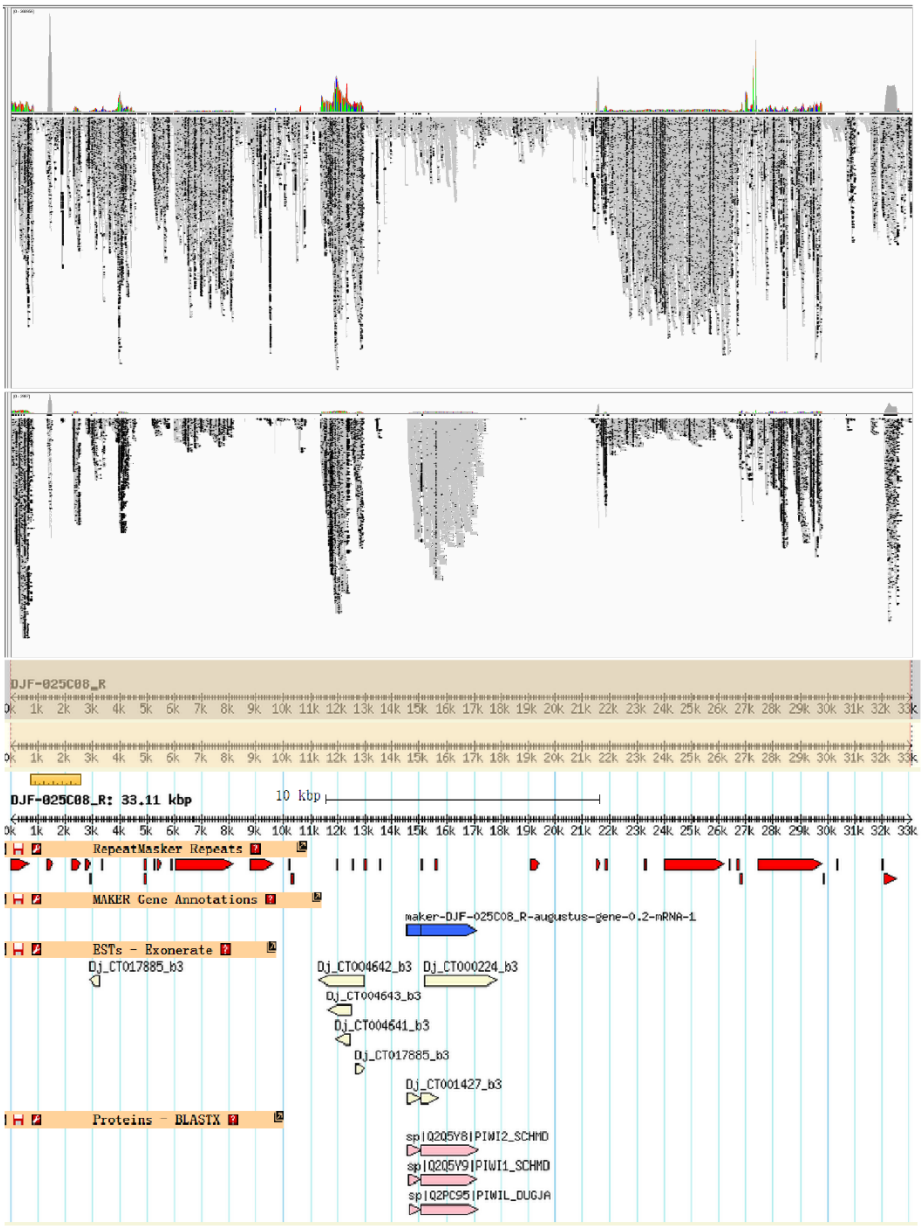


Figure 3-9. Annotation of one DjPiwiB sequence (fosmid clone number DJF-025C08)

The horizontal line in each figure shows the position of bases in the DjPiwiB sequence (~33kb).

In figure a & b, the Grey and black points show matches and mismatches between sequencing reads and the reference DjPiwiB sequence, respectively.
In figure c, red arrows show repeat sequences annotated from RepBase; blue arrow shows the gene prediction from Augustus (it precisely reflects the location of the DjPiwiB coding region.); cream colored arrows show the *D.japonica* ESTs that aligned with the DjPiwiB sequence; and pink arrows show protein predictions by BlastX.

a. Alignment of genome sequencing reads on DJF-025C08 sequence
b. Alignment of mRNA sequencing reads on DJF-025C08 sequence
c. MAKER genome annotations

3.4. De novo assembly of *D. japonica* genome

3.4.1. Algorithms of genome assembly

There are basically two approaches for sequencing reads assembly – overlap graphs and de Bruijn graphs.

Overlap-layout graph

Most of the tools that have been developed to be used for assembling long reads (Sanger sequencing reads or Roche 454 sequencing reads) are based on the overlap-layout-consensus algorithm. They compute all pair-wise overlaps between sequencing reads and use this information to construct a graph. Each node in the graph corresponds to a read, and an edge denotes an overlap between two reads (Figure 3-10). The overlap graph is used to compute layouts between reads and make a consensus sequence of contigs, which works very well when the number of sequencing reads is limited, and the overlap between reads is significant. The popular assemblers Celera and Newbler are assembly tools based on the overlap-layout graph, and are commonly used to assemble sequencing reads from Sanger and Roche 454, respectively.

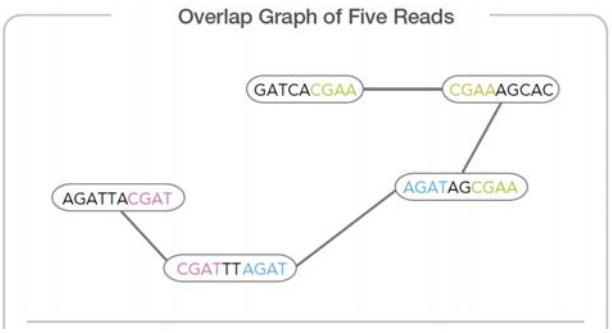


Figure 3-10. Overlap-layout-graph

(adopted from www.illumina.com)

Each circle in the figure shows a sequencing read and also a graph node. Same colored nucleotides indicate overlaps between reads.

De Bruijn Graph

Because the traditional overlap-lay-out graph is computationally intensive, it does not work well with the increasing data produced from second-generation sequencing machines. The most popular algorithm for assembling second-generation sequencing data is the De Bruijn graph. To reduce computational effort, at first, the De Bruijn graph consecutively breaks reads into smaller sequences, called kmers, where the parameter k denotes the length in bases of these sequences; and then it captures overlaps of length k-1 between kmers (Figure 3-11). The De Bruijn graph uses links between neighboring kmers that are derived from reads, so it does not need pairwise reads alignment. Because of this procedure, the redundancy of data is automatically reduced, and therefore high sequencing coverage depth does not matter, and this is why it is more suitable for dealing with second-generation sequencing data. The most popular assemblers based on this algorithm include SOAPdenovo, Allpath-lg, Velvet, and ABySS.

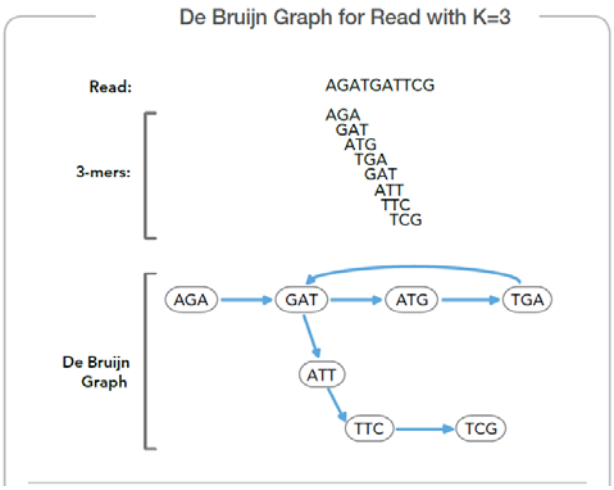


Figure 3-11. De Bruijn Graph

(adopted from www.illumina.com)

This figure shows the kmer generation, and how De Bruijn graph algorithm assembly works by using those kmers. Because k=3, the length of overlaps is k-1=2 in this figure. Blue rows indicate the order of the kmers and their overlaps.

3.4.2. De novo genome assembly with De Bruijn graph-based method

Because most of our genome sequencing data came from second-generation sequencers, it was reasonable to use a De Bruijn graph-based genome assembler. The most popular and best-performng ones are Allpath-lg[88] , SOAPdenovo[89] and Velvet[90]. I used each of them to assemble the genome sequencing reads after quality control (The assembly run of Allpath-lg was terminated because it needed more than 500 Gb memory). However, no matter what assembler I used, what parameters I set, or how strictly I trimmed the input data, the de novo assembly contigs and scaffolds were very short (Table 3-2).

Table 3-2. *D. japonica* genome assembly results from De Bruijn graph assemblers

Software	Pre-overlap	SOAP Corrector	kmer	Contig N50	Longest Contig	Scaffold N50	Longest Scaffold	Total Scaffold Size
Velvet v1.0	X	X	37	157 bp	6 kb	-----	-----	0.2945Gb
SOAPdenovo v1.3	X	X	23	217 bp	7 kb	516 bp	17 kb	0.2736 Gb
SOAPdenovo V1.3	X	X	31	279 bp	19 kb	481 bp	61 kb	0.4047 Gb
SOAPdenovo v1.5	X	X	64	288bp	10 kb	587 bp	12 kb	0.5054 Gb
SOAPdenovo v1.3	O	X	31	811 bp	100 kb	1222 bp	126 kb	0.5719 Gb
SOAPdenovo v1.3	O	O	31	343 bp	180 kb	497 bp	315 kb	0.53789 Gb
SOAPdenovo v1.5	O	O	45	323 bp	7 kb	655 bp	31 kb	0.6042 Gb

This result was not very surprising: the De Bruijn graph algorithm assembles sequencing reads based on the kmers generated from those reads; however, the kmer frequency histogram (Figure 3-7) we had was not a typical good histogram, in contrast to those obtained in most of the reported second-generation sequencing projects. It is very possible that our strange kmers interfered with the De Bruijn graph assemblers. In view of this, we concluded that the De Buijn graph method might not be suitable for our D. japonica genome sequencing data.

As most of our genome sequencing data came from second-generation sequencing (Illumina GAIIx and Hiseq2000), but the best algorithm (De Buijn graph) for assembling second-generation reads did not work well for our data, there was a need to develop a new strategy which could utilize all of our second-generation data but rely on an overlap-layout graph rather than a De Bruijin graph algorithm.

3.4.3. New strategy to improve genome assembly by overlap-layout

Because De Bruijn graph assemblers could not generate good results from our sequencing data, I developed a new strategy to assemble second-generation sequencing reads by using an overlap-layout algorithm to improve the *D. japonica* genome assembly (Figure 3-11).

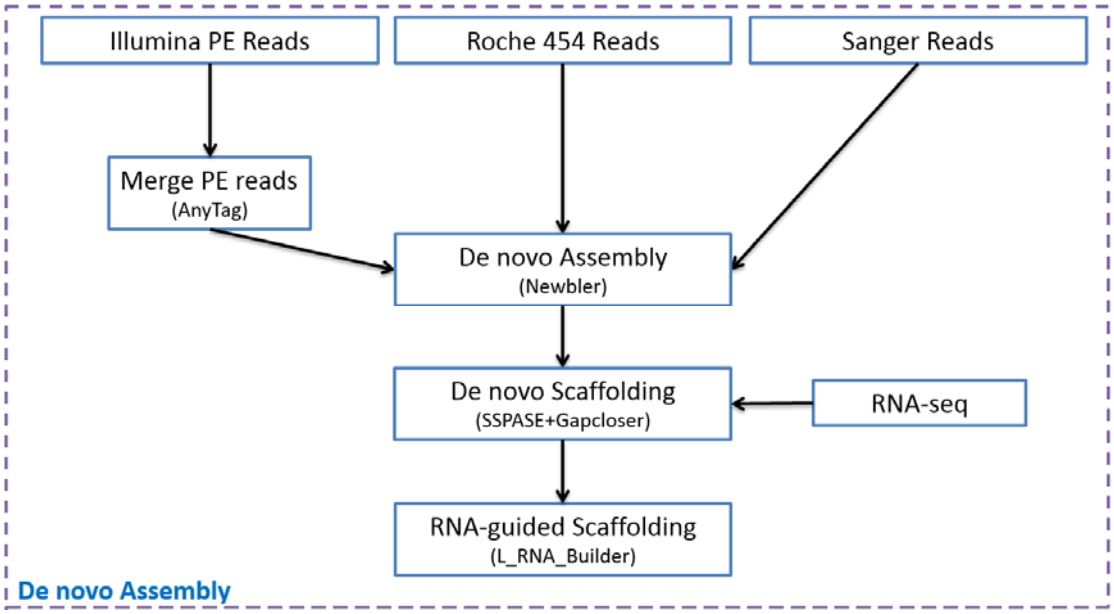


Figure 3-11. De novo assembly strategy of *D. japonica* genome project

Step 1. Generate pseudo-454 long reads from Illumina pair-end short reads

Usually, short assembled contigs are caused by repeated sequences and heterozygosity of a genome. To overcome the repeated sequences problem, long sequencing reads are needed. However, most of our reads were short reads (~100bp) derived from second-generation sequencers. Here I used Anytag[91] to fill the gap between paired-end short reads (300bp, 350bp, 400bp and 500bp insert libraries), and merged them into pseudo-454 long reads. This procedure was based on local assembly using a series of paired-

end libraries of stepwise-decreasing insert sizes (Figure 3-12). In addition, this merging procedure has tolerance to heterozygosity, which to some extant helps solve the heterozygosity problem. Finally, all short pair-end sequences were merged to 44,017,434 pseudo-454 reads with an average length of 485bp.

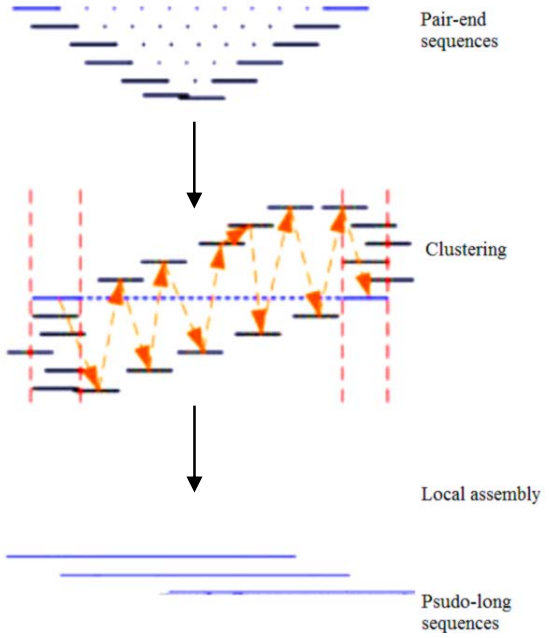


Figure 3-12. Generate pseudo-454 long reads from Illumina paird-end short reads

Short paird-end reads from different insert libraries are clustering, and local clustering in order to generate error-free and long reads.

Step 2. Overlay-layout graph assembly

After merging of short pair-end reads, all of the newly generated pseudo-454 long reads, previous real 454 reads and Sanger long reads were combined together to go through the overlap-layout graph assembler, Newber v2.9, which is popular for de novo assembly of Roche 454 sequencing reads. Accordingly, 951280 contigs were produced with N50 length 1408bp, which showed a great improvement compared with the previous De Bruijn assembler results.

Step 3. Scaffolding and gap closure

After assembly, all un-merged Illumina short pair-end reads and mate-pair reads, Roche 454 mate-pair reads, as well as Sanger end sequencing reads were used for scaffolding by SSPACE[92], followed by GapFiller[93] to fill gaps in scaffolds. After this step, the scaffolds that were generated contained 1.56G bases, which may reflect the draft genome size of the planarian *D. japonica*.

Step 4. Further scaffolding using RNA evidence

For some complex genomes, it is difficult to increase the N50 length even with large mate-pair libraries, which leads to low transcript coverage in its subsequent genome annotation step. In this step, I used L_RNA_scaffolder[94], which takes advantage of long transcriptome reads to order, orient and combine genomic fragments into larger scaffolds in order to facilitate the genome annotation.

Finally, we obtained 202,925 scaffolds with the N50 27,741bp, and contained about 1.56 Gb in total length. The assembly of the *D. japonica* genome was thus improved (Table 3-3).

Table 3-3. Summary of the *D. japonica* genome assembly

	Contigs	Scaffolds
N50 (bp)	1,408	27,741
Longest (bp)	186,265	760,010
Total number (> 1Kb)	286,283	126,524
Total number (> 10Kb)	466	38,208
Total number	951,280	202,925
Total size (bp)	897,448,998	1,565,189,494

Evaluation of assembly results

After de novo assembly, we used two methods to evaluate the quality of the genome assembly: comparisons with previously screened fosmid sequences, and alignment of transcripts on to the genome scaffolds (Figure 3-13).

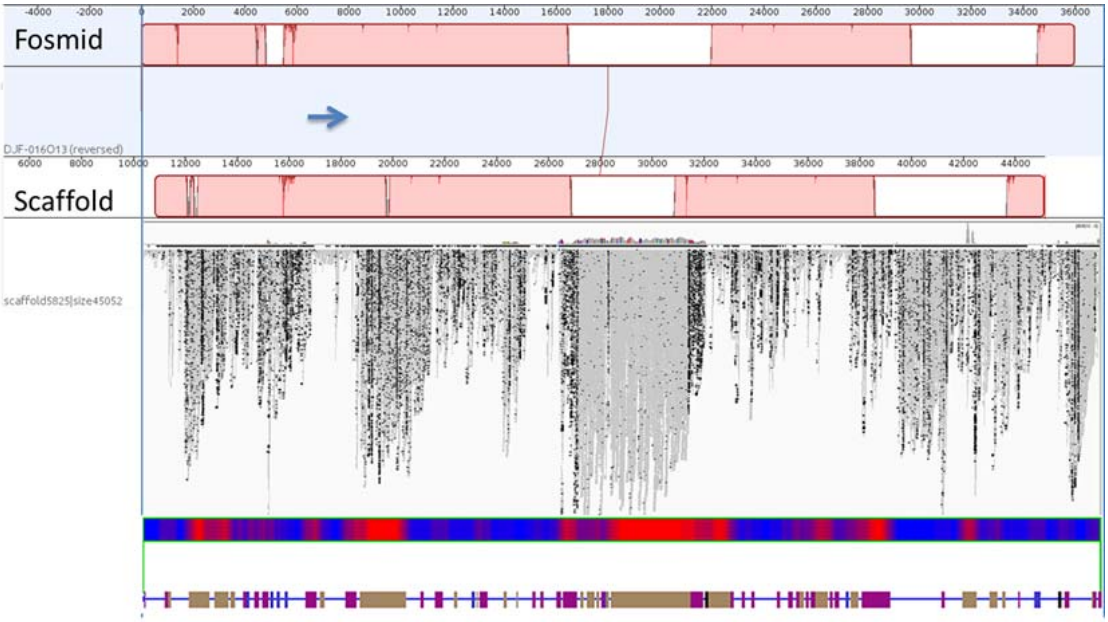


Figure 3-13. Alignment between DjTh fosmid clone and DjTh gene scaffold

Blue arrow shows an exon region of the DjTh gene. White blocks represent un-aligned regions. Pink blocks represent aligned regions. Red blocks show large repeat sequences marked by Repbase. RNA sequencing reads were aligned to the DjTh fosmid sequence, and matched positions were marked by grey points, unmatched positions were marked by black points.

Fosmid sequences (~ 35kb) were aligned to the assembled scaffolds using Mauve. A representative alignment between the DjTh gene's scaffold and its fosmid clone sequence is shown in the upper part of Figure 3-13, which showed good concordance between these two, and indicated good assembly quality of the genome.

In addition, RNA sequencing reads were mapped to the DjTh fosmid sequence (the middle part of Figure 3-13), which together with the RepeatMasker annotation results (the lower part of Figure 3-13), showed that some un-alignment was caused by repetitive sequences, especially retrotransposons (The largest red block in the lower part of Figure 3-13 is annotated as retrotransposon by Repbase, and this region also matched a great number of RNA sequencing reads).

In addition, RNA sequences of D. japonica (ESTs and 454 sequencing data derived from NCBI and DDBJ) were aligned to the assembly scaffolds using BLAT with 95% minimum identity. Counterpart genome scaffolds could be found for more than 98% of all transcripts, and 96% of all the bases in transcripts could be covered by the genome. This indicated the completeness of this genome for protein coding genes. Also, this alignment gives us a way to pick out coding genome regions, namely, the assembled genome scaffolds that match RNA sequences. There were 33924 scaffolds according to the RNA alignment, and they occupied 58.26% of the total genome. The N50 length of those scaffolds is 55,596bp, which is twice as long as the average genome scaffold N50 (27,741). This proved that genome sequences around coding regions are less complicated than other regions, and thus can be assembled more easily.

3.5. *D. japonica* Genome Annotation

3.5.1. Genome Annotation

Genome assembly can only congregate sequencing reads into chromosomes, scaffolds or contigs in a fasta format file full of As, Ts, Cs, and Gs. A draft genome is like a plain map with no labels or legends, on which people won't be able to find places or directions [76]. To annotate a genome is to identify and position each genome element on the draft assembly, like labels on maps.

As Stein pointed out [95], from a technical point of view, genome annotation is actually a process to structure the genome assembly, identify the gene models and genomic features, and make the connection between genome elements and biological meanings.

As described above, I succeeded in assembling the *D. japonica* draft genome, and next I performed analyses to: 1. Identify gene models and genomic features (section 3.5.2 – 3.5.5); and 2. Make the connection between genome elements and biological meanings (section 3.5.6).

3.5.2. Repeat Sequences

Repeat sequences were identified by using both RepeatModeler and RepeatMasker. RepeatModeler was used to build the consensus models of putative interspersed repeats as a new repeat sequence library based on the genome. RepeatMasker was used to search the planarian genome against the combined library Repbase. The two results were integrated to gain a comprehensive analysis of repeats in the *D. japonica* genome.

39.69% of the assembled genome sequences were marked as repetitive elements (Table 3-4), which showed that this planarian genome was highly repetitious. An even higher actual rate of repeat sequences is expected, as the incomplete assembly was interrupted by repeat sequences in the genome, and most 3' and 5' end of assembled contigs were repeat sequences.

In addition, except for unclassified repeats, the majority of repeated elements were retrotransposons and DNA transposons, which matched our observations from the fosmid survey.

Table 3-4. Summary of repeat elements in *D. japonica* genome assembly

Repeat Elements	Numbers of elements*	Length (bp)	Percentage of genome (%)
Retrotransposon	260,765	122,888,499	7.85%
LTR-Retrotransposon	194,395	101,831,050	6.51%
Non-LTR Retrotransposon	66,370	21,057,449	1.35%
DNA Transposon	323,715	109,645,290	7.01%
Unclassified	1,602,282	355,858,117	22.74%
Small RNA	4,762	993,638	0.06%
Simple repeats	411,512	27,237,668	1.74%
Low complexity	80,856	4,603,818	0.29%
Total count	2,683,892	621,227,030	**39.69%**

* Most repeats fragmented by insertions or deletions were counted as one element.

3.5.3. De novo transcriptome assembly

RNA evidence is very important for genome annotation, and will increase the accuracy of annotation. Therefore, RNA sequencing data derived from all resources (RNA sequences from NCBI, ESTs and sequencing reads from DDBJ, and unpublished NGS data in the Agata laboratory) were combined and de novo assembled by Trinity[96]. The final assembled transcriptome had 25,566 transcripts and a total length of 34.8 Mb (Table 3-5).

Table 3-5 Summary of the *D. japonica* transcriptome assembly

Term	Value
Statistics for isotig length	
Min isotig length (bp)	62
Max isotig length (bp)	17,446
N50 isotig length (bp)	1,792
Statistics for numbers of isotigs	
Number of isotigs	25,566
Number of isotigs >=1kb	13,256
Statistics for bases in the isotigs	
Number of bases in all isotigs	34,777,653
Number of bases in isotigs >=1kb	27,150,697
GC content of isotigs	31.71%

One metric for evaluating the quality of a transcriptome assembly is to examine if the number of transcripts assembled appears to be full-length or nearly full-length. Because there was no high quality annotation available from a closely related organism, we compared the assembled transcripts to all known proteins from the Uniprot database, and determined what percentage of the top-matching proteins were covered by our assembled transcriptome.

3.5.4. Gene prediction and functional annotation

For gene prediction, we used evidence-based prediction followed by de novo prediction. In the evidence-based method, 2,857,787 long RNA sequences (including ESTs, 454 sequences and assembled transcripts) were aligned against the *D. japonica* genome with BLAT [97](identity > 95). The best output item for each RNA sequence was taken as evidence of a coding region in the genome, and the information was further used by Augustus [98] to help de novo gene prediction. Although CH-HIT was used to remove redundancy after de novo prediction, the final predicted protein result was still more than my expected (a total of 108195 proteins were predicted). Taking into account that a large number of retrotransposons exist in the genome, as well as the fact that the assembled genome is relatively fragmented, and that some genes could be truncated because of repetitive sequences located within their intron regions, this predicted number of proteins should be an overestimate.

To achieve a more accurate result, I made the prediction of genes only from scaffolds with RNA alignment evidence, and annotated them by Blast2GO annotation. 15601 genes were endowed with GO classified ontology annotation (Table3-6 and Figure 3-14).

Table 3-6 GO ontology annotation

	Ontology type	Number	Percentage (%)
Cellular Component	extracellular region	895	5.7
	extracellular region part	495	3.2
	cell	13946	89.4
	cell part	13946	89.4
	membrane-enclosed lumen	4308	27.6
	envelope	1083	6.9
	macromolecular complex	4990	32
	organelle	11516	73.8
	organelle part	8326	53.4
	extracellular region part	495	3.2
	organelle part	8326	53.4
	synapse part	641	4.1
	cell part	13946	89.4
	synapse	923	5.9
	synapse part	641	4.1
	Ontology type	Number	Percentage (%)
Biological Process	reproduction	3375	21.6
	cellular component biogenesis	3324	21.3
	developmental process	8268	53
	cellular component organization	7007	44.9
	death	3261	20.9
	reproductive process	2957	19
	immune system process	3042	19.5
	response to stimulus	7601	48.7
	multicellular organismal process	9264	59.4
	anatomical structure formation	3746	24
	multi\-organism process	1397	9
	establishment of localization	5436	34.8
	biological adhesion	921	5.9
	metabolic process	10821	69.4
	viral reproduction	940	6
	rhythmic process	292	1.9
	pigmentation	9645	61.8
	locomotion	2549	16.3
	localization	6865	44
	growth	2222	14.2
	cellular process	13255	85
	cell killing	72	0.5
	biological regulation	10467	67.1

	Ontology type	Number	Percentage (%)
Molecular Function	electron carrier activity	79	0.5
	molecular transducer activity	1210	7.8
	enzyme regulator activity	1105	7.1
	transcription regulator activity	870	5.6
	catalytic activity	7287	46.7
	binding	12328	79
	antioxidant activity	35	0.2
	auxiliary transport protein activity	182	1.2
	metallochaperone activity	4	0
	chemorepellent activity	2	0
	translation regulator activity	221	1.4
	structural molecule activity	758	4.9
	chemoattractant activity	16	0.1
	transporter activity	1522	9.8

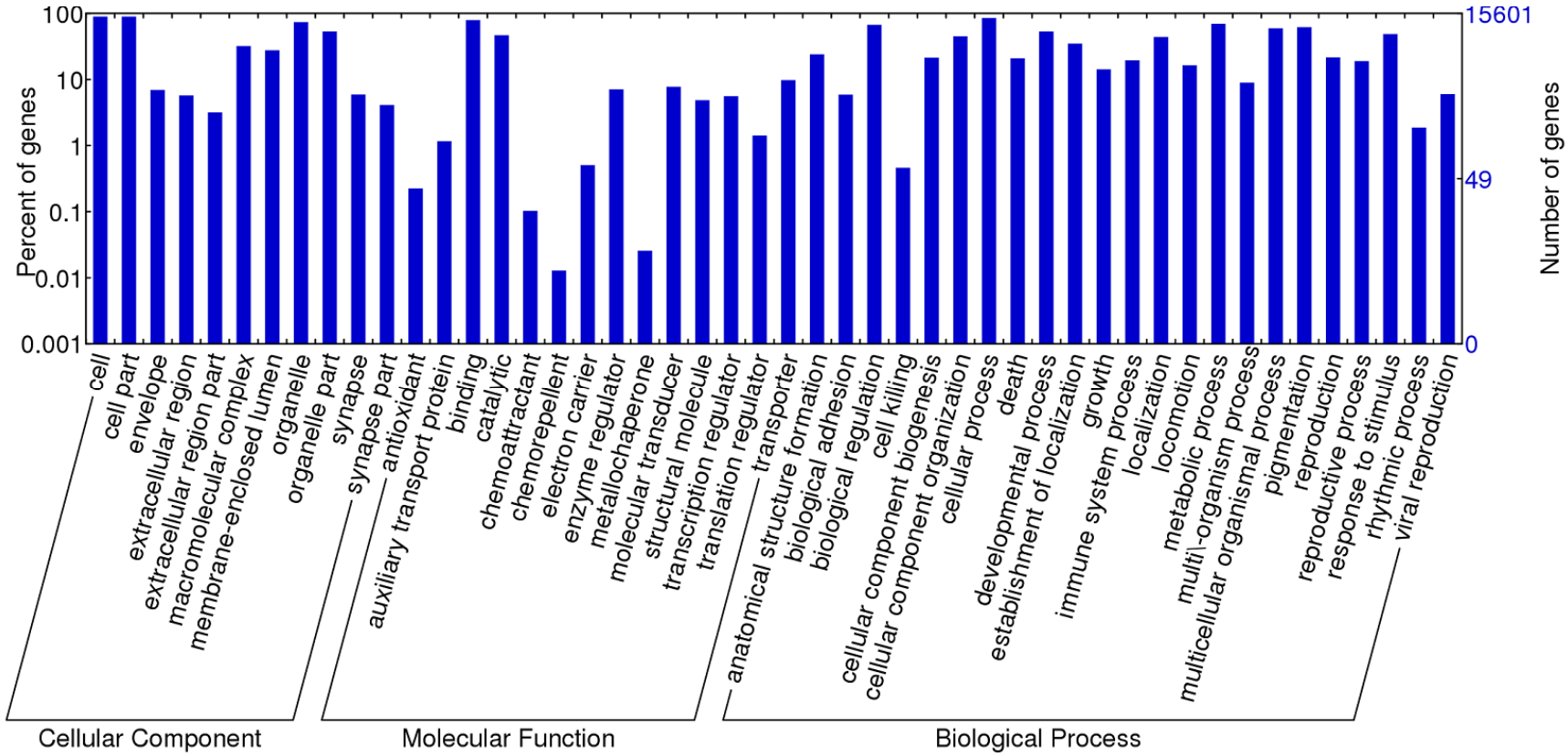


Figure 3-14. Histogram of GO ontology annotation results

3.6. CNEs are regulatory elements in planarians

3.6.1. Conserved non-coding elements (CNEs)

Because of protein functional constraints, coding regions are expected to exhibit sequence conservation between related species. In addition, some non-coding elements also show functional constraints, and such conservation outside of exons can be detected by cross-species comparison. In mammals, conserved non-coding elements (CNEs) were found to contain transcription factor binding sites [99], and they are accepted as beacons of gene regulatory elements [100]. Planarians have been used as a model animal for regeneration and stem cell research for many years. Although numerous functional genes have been identified in planarian, the molecular regulatory networks in planarian are still far from understood, because few gene regulatory elements have been discovered. Only after de novo assembly of the *D. japonica* genome could comparative genomics between the planarians *D. japonica* and *S. mediterranea* be performed. The resultant discovery of CNEs will be of great importance for future planarian research.

3.6.2. CNEs between two planarian genuses

Although D. japonica and S. mediterranea have much morphological, physiological and functional similarity, their evolutionary positons are far from each other, and their gene sequences also have large differences, so finding CNEs between these two planarian

genuses could light the way to find conserved molecular mechanisms in planarians, such as the mechanisms regulating brain formation and regeneration.

To discover CNEs, I performed genome comparison between these planarians (*D. japonica* and *S. mediterranea*) belonging to two different genuses. To increase accuracy, firstly, I chose only the scaffolds with mRNA alignment evidence, because the location of other scaffolds that are separated from coding genes and have no RNA reference is unsure, which may cause many false-positive results. Even if CNEs were predicted from those independent scaffolds, it would be hard to speculate about the relationship between them and functional genes. In order to remove the effect of repeated sequences, all scaffolds were passed through the filter of RepeatModeler and RepeatMasker, and repeat sequences were masked out by the letter “N”. To find conserved regions between these two different genuses, genome comparison was performed between the *D. japonica* genome scaffolds and the scaffolds from the *S. mediterranea* genome, using NUCmer[101] and BlastN. One-to-one matched scaffold pairs were found, and conserved sequences were located between the two genuses. In addition, a further masking procedure was performed masking all coding regions by “N” based on RNA alignment information. Finally, the conserved noncoding elements were generated.

In 2002, a very important gene, DjNDK (a homolog of the vertebrate Fgfrl1 gene), was found to play a crucial role in brain formation during planarian regeneration [102]. However, although more than 10 years have passed since then, we still do not know what regulatory elements restrict the expression of DjNDK to the brain region, or how

this gene functions in brain formation. To test whether the CNEs we found by genome comparison are indeed regulatory elements, and with the hope of finding regulators of the DjNDK gene, we chose to analyze the CNEs on the scaffolds of DjNDK.

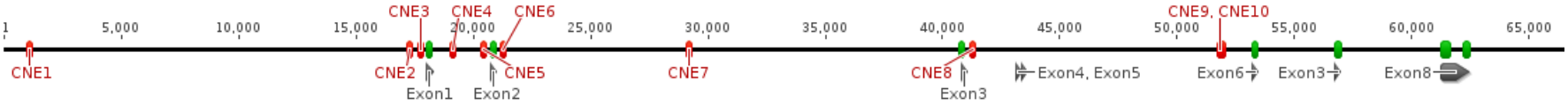
By genome comparison, I found 10 CNEs on the DjNDK scaffolds (Figure 3-16) that are also conserved on the counterpart region of scaffolds of the *S. mediterranea* NDK gene (Figure 3-15).

Figure 3-15. Ten CNEs between the *D. japonica NDK* gene and *S. mediterranea* NDK gene

a. The scaffold of DjNDK. Red blocks show CNEs, Grey arrays indicate exons and green block show conserved region on exons.

b. Ten CNEs alignments between DjNDK and SmedNDK gene. In each alignment, the upper sequence is the noncoding sequence from the *D. japonica NDK* gene and the lower sequence is the noncoding sequence from the *S. mediterranea* NDK gene.

a


1
5,000
10,000
15,000
20,000
25,000
30,000
35,000
40,000
45,000
50,000
55,000
60,000
65,000
CNE1
CNE2
CNE3
CNE4
CNE5
CNE6
CNE7
CNE8
CNE9, CNE10
Exon1
Exon2
Exon3
Exon4, Exon5
Exon6
Exon3
Exon8

b


CNE1
Query: 1 aatgcgattatgaaaatatcaacatttagtcatctg 36
Sbjct: 32183 aatgcgattatgaaaatatcaacatttagtcatctg 32218
CNE2
CNE3
CNE4
CNE5
CNE6
Query: 1 caaggtgtcgtgccaagactttgagtgg 28
Sbjct: 35998 caaggtgtcgtgccaagactttgagtgg 36025
CNE7
CNE8
CNE9
Query: 1 aggctgactcatgaaatgtagagatgtacacgacataaacaatatttacgta 52
Sbjct: 18353 aggctgagtcatgagatgtagagatgtacacgacatagacaatatttacgta 18302
CNE10
Query: 1 tagttgcttttactattgcattgtcattcgg 31
Sbjct: 18201 tagttacttttactattgcattgtcattcgg 18171

3.6.3. CNE4 is a regulatory element on DjNDK gene

To examine whether CNEs in or near the NDK gene have regulatory activity, we firstly chose CNE4 (140bp) as a representative example for transgenic expression experiments (Figure 3-16). The 140bp CNE4 was inserted upstream of the promoter of a beta-actin promoter-driven GFP expression vector to form a reporter construct called CNE-actGFP. When this expression vector was injected into *Xenopus laevis* embryos, the GFP expression pattern was localized in the neural-plate-forming region of the embryo, which was especially evident in the anterior region at the end of gastrulation (Figure 3-16a). This experiment was repeated 132 times to confirm the discovery of this reporter-gene-expression regulation by CNE4.

By aligning the CNE sequence between *D. japonica* and *S. mediterranea*, putative transcription factor-binding motifs were identified. We made 3 point mutations in the expression vector in putative Msx(M), Tcf/lef1(T) and Jun/Fos(J) binding sites, respectively (Figure 13-6b). The bar graph in Figure 3-16c shows the percentage of the embryos that showed GFP expression in the neural plate among the total developed embryos injected with the constructs. Statistical analysis by the chi-square test (http://www.graphpad.com/quickcalcs/chisquared1.cfm) showed that the percentage of positive cases of reporter gene regulation by the wild-type CNE "wt" (140) and by the Jun/Fos mutant constructs were significantly different ($P<0.0001$), which indicated that the 140bp CNE4 is a real regulatory element on the ndk gene, and the transcription factor Jun/Fos might regulate the ndk expression through this region.

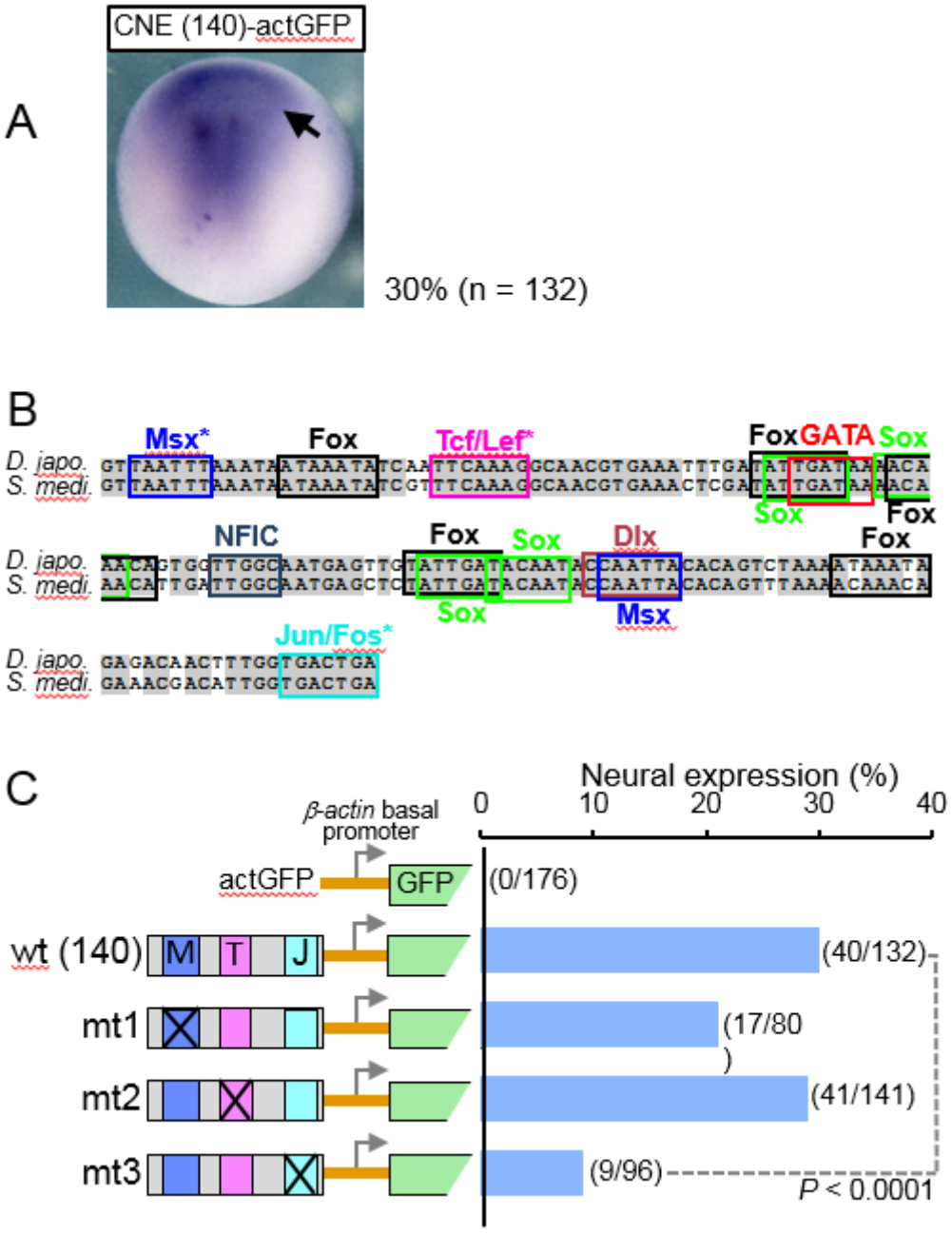


Figure 3-16. Analysis of enhancer activity of non-coding sequences conserved between *D. japonica* and *S. mediterranea* ndk genomic regions

3.7. Discussion

3.7.1. Improvement of the Planarian Genome

In this genome project, we sequenced and de novo assembled the Dugesia japonica planarian genome. Although the new strategies used here helped improve the assembly results, the final genome scaffolds were still short, and such a result was also obtained in the Schmidtea mediterranea genome project. The strange kmer frequency histogram, fosmid sequences analysis, and further genome analysis all showed that planarians have complicated genomes.

We extracted and pooled genomic DNA for sequencing from several hundred planarians, and although all of these planarians came from one individual that has been clonally maintained since 2005, mutations must have accumulated during their nearly 10 years of asexual proliferation via fissioning. Differences of genome types among individuals could thus be a reason for the complexity. Moreover, according to recent observations, even the genomes of cells from one individual planarian are heterogenic, and haploid or polyploid cells are very common in their bodies. So, all of these complicated factors hampered the genome assembly. To achieve a much better result, DNA extracted from a single individual, homotype tissue, or even a single cell could help reduce the problems of genome assembly encountered in this study.
In addition, the genomes of planarians are highly repetitious and possesss low GC content, which are two additional factors that interfere with genome assembly. Longer

sequences (such as reads from third-generation sequencing techniques) will to some extent solve this problem.

3.7.2. CNEs in Planarians

CNEs are well accepted as locations for regulatory elements. In this project, we also proved that one CNE in the ndk gene is a regulatory element which has a binding site for the transcription factor Jun/Fos.

Previous transgenic assays showed that DjNDK was expressed in the anterior part of the neural plate of Xenopus embryo that will form the brain during development. However, in the present transgenic expression experiment, CNE4 could only restrain the reporter gene expression to within the whole neural plate, but not to within the anterior part of the neural plate. This indicated that some other regulatory factor(s) exist in addition to CNE4. We found 10 CNEs, and thus some other CNE(s) could also be regulatory elements for transcription factors, or possibly other regulators such as miRNA may also play a part in the process of regulating ndk gene expression. In addition, CNE4 in planarians has a similar counterpart in human, mouse, and frog NDK genes, and thus we expect that the regulation of the ndk gene and even ndk-related aspects of the brain formation mechanism might be conserved between vertebrates and invertebrates.

Acknowledgements

Planarians could “almost be called immortal under the edge of the knife”, Dalyell (1814). – My interest in the immortality of planarians is the most critical motivation for my work.

I would like to thank my supervisor Kiyokazu AGATA for giving me the opportunity to take on such an interesting and challenging project, and for his continuing support and encouragement.

I thank Alexandre Alié for his fruitful discussions with me during the design of the DNA library screening method, and Elizabeth Nakajima for critically reading and proofreading the manuscript.

Thanks also to the members of the Toyoda and Fujiyama lab for the masses of DNA sequencing data they generated, and all members of Ogino’s lab for their cooperation with the transgenic experiments.

Finally, thanks go to my family, in particular my grandfather, who make everything worthwhile.

References:

[1]. Willmer, P., Invertebrate relationships: patterns in animal evolution. 1990: Cambridge University Press.

[2]. Egger, B., et al., To be or not to be a flatworm: the acoel controversy. PloS one, 2009. 4(5): p. e5502.

[3]. Pagán, O.R., The First Brain The Neuroscience of Planarians. 1 edition ed. 2014: Oxford University Press. 280.

[4]. Agata, K. and Y. Umesono, Evolution of the genetic program controlling brain development. Tanpakushitsu kakusan koso. Protein, nucleic acid, enzyme, 1999. 44(3): p. 245.

[5]. Agata, K., Regeneration and gene regulation in planarians. Current Opinion in Genetics & Development, 2003. 13(5): p. 492 - 496.

[6]. SASAKI, G. and M. Kawakatsu, Bipaliid Land Planarians Recorded in Chinese and Japanese Materia Medica. 2001.

[7]. 李時珍, 本草綱目. 1596, 金陵.

[8]. 中村惕斎, 訓蒙圖彙. 1666: 山形屋.

[9]. Dana, J., De Hirudinis nova specie, noxa, remedi-isque adhibendis. Mélanges Philos. Mathém. Soc. r. Turin (Misc. Taurin.), 1766. 3: p. 199-205

.

[10]. Lue, K.Y. and M. Kawakatsu, History of the study of Turbellaria in China. Part 1: Ages of Materia Medica and of early expeditions by westerners. Hydrobiologia, 1986. 132(1): p. 317--322.

[11]. Artois, T., S. Tyler and J. Dana, World Register of Marine Species (WoRMS). 2014, Society for the Management of Electronic Biodiversity Data (SMEBD).

[12]. Pallas, P.S., Spicilegia zoologica quibus novae imprimis et obscurae animaliu. speciosiconibus atque conamentariis illustratur. Fasc X, Berolini, 1774.

[13]. H. V. Brøndsted, H.V., Planarian regeneration. Vol. 42. 1969: Pergamon.

[14]. Müller, O.F., Zoologiae Danicae prodromus: seu Animalium Daniae et Norvegiae indigenarum characteres, nomina, et synonyma imprimis popularium. 1776: typis Hallageriis.

[15]. Darwin, C., XXIX.—Brief descriptions of several terrestrial Planariae, and of some remarkable marine species, with an account of their habits. Journal of Natural History, 1844. 14(91): p. 241--251.

[16]. Nicholas, F.W. and J. Nicholas, Charles Darwin in Australia. 2008: Cambridge University Press.

[17]. Morgan, T.H., Experimental studies of the regeneration of Planaria maculata. Development Genes and Evolution, 1898. 7(2): p. 364-397.

[18]. Morgan, T.H., Regeneration in planarians. Development Genes and Evolution, 1900. 10(1): p. 58--119.

[19]. ICHIKAWA, A. and M. KAWAKATSU, A New Freshwater Planarian, Dugesia japonica, Commonly but Erroneously Known asDugesia gonocephala (Duges). 日本動物学彙報, 1964. 37(3): p. 185--194.

[20]. 刘德增, 中国的淡水 (三肠目) 涡虫. 动物学杂志, 1989. 24(6): 第38--43 页.

[21]. KAWAKATSU, M., I. OKI and S. TAMURA, Taxonomy And Geographical-Distribution Of Dugesia-Japonica And D-Ryukyuensis In The Far-East. Hydrobiologia, 1995. 305(1-3): p. 55-61.

[22]. IJIMA, I. and T. KABURAKI, Preliminary Descriptions of some Japanese Triclads. Annotationes zoologicae japonenses / Nihon do¯butsugaku iho¯., 1916. v.9 (1915-1920): p. 698.

[23]. KABURAKI, T., On some Japanese Freshwater Triclade; with a Note on the Parallelism in their Distribution in Europe and Japan. The journal of the College of Science, Imperial University of Tokyo, Japan = Tokyo Teikoku Daigaku kiyo. Rika., 1922. v. 44 (1922-1923): p. 618.

[24]. Inoue, T., T. Yamashita and K. Agata, Thermosensory Signaling by TRPM Is Processed by Brain Serotonergic Neurons to Produce Planarian Thermotaxis. Journal of Neuroscience, 2014. 34(47): p. 15701-15714.

[25]. Agata, K. and K. Watanabe, Molecular and cellular aspects of planarian regeneration. 1999. p. 377-383.

[26]. Agata, K., et al., Structure of the Planarian Central Nervous System (CNS) Revealed by Neuronal Cell Markers. Zoological Science, 1998. 15(3): p. 433-440.

[27]. Sanchez, A.A. and P.A. Newmark, Double-stranded RNA specifically disrupts gene expression during planarian regeneration. Proc Natl Acad Sci U S A, 1999. 96(9): p. 5049-54.

[28]. Nakazawa, M., Search for the Evolutionary Origin of a Brain: Planarian Brain Characterized by Microarray. Molecular Biology and Evolution, 2003. 20(5): p. 784-791.

[29]. Hayashi, T., et al., Isolation of planarian X-ray-sensitive stem cells by fluorescence-activated cell sorting. Development, Growth and Differentiation, 2006. 48(6): p. 371-380.

[30]. Inoue, T., et al., Clathrin-mediated endocytic signals are required for the regeneration of, as well as homeostasis in, the planarian CNS. Development, 2007. 134(9): p. 1679-1689.

[31]. Hayashi, T., et al., Single-cell gene profiling of planarian stem cells using fluorescent activated cell sorting and its "index sorting" function for stem cell research. Development, Growth & Differentiation, 2010. 52(1): p. 131-144.

[32]. Nishimura, O., et al., Comparative transcriptome analysis between planarian Dugesia japonica and other platyhelminth species. BMC Genomics, 2012. 13(1): p. 289.

[33]. Zhang, G., et al., The oyster genome reveals stress adaptation and complexity of shell formation. Nature, 2012. 490(7418): p. 49-54.

[34]. You, M., et al., A heterozygous moth genome provides insights into herbivory and detoxification. Nat Genet, 2013. 45(2): p. 220-5.

[35]. Campbell, T.N. and F.Y. Choy, Approaches to library screening. J Mol Microbiol Biotechnol, 2002. 4(6): p. 551-4.

[36]. Seidman, J., Screening of recombinant DNA libraries, in Current protocols in molecular biology. 1994.

[37]. Benton, W.D. and R.W. Davis, Screening lambdagt recombinant clones by hybridization to single plaques in situ. Science, 1977. 196(4286): p. 180-2.

[38]. Hanahan, D. and M. Meselson, Plasmid screening at high colony density. Methods Enzymol, 1983. 100: p. 333-42.

[39]. Suggs, S.V., et al., Use of synthetic oligonucleotides as hybridization probes: isolation of cloned cDNA sequences for human beta 2-microglobulin. Proc Natl Acad Sci U S A, 1981. 78(11): p. 6613-7.

[40]. Traver, C.N., et al., Rapid screening of a human genomic library in yeast artificial chromosomes for single-copy sequences. Proc Natl Acad Sci U S A, 1989. 86(15): p. 5898-902.

[41]. Asakawa, S. and N. Shimizu, High-fidelity digital hybridization screening. Genomics, 1998. 49(2): p. 209-17.

[42]. Han, C.S., et al., Construction of a BAC contig map of chromosome 16q by two-dimensional overgo hybridization. Genome Res, 2000. 10(5): p. 714-21.

[43]. Grunstein, M. and D.S. Hogness, Colony hybridization: a method for the isolation of cloned DNAs that contain a specific gene. Proc Natl Acad Sci U S A, 1975. 72(10): p. 3961-5.

[44]. Sanzey, B., et al., Methods for identification of recombinants of phage lambda. Proc Natl Acad Sci U S A, 1976. 73(10): p. 3394-7.

[45]. Bloem, L.J. and L. Yu, A time-saving method for screening cDNA or genomic libraries. Nucleic Acids Res, 1990. 18(9): p. 2830.

[46]. Liu, J., et al., Large-scale cloning of human chromosome 2-specific yeast artificial chromosomes (YACs) using an interspersed repetitive sequences (IRS)-PCR approach. Genomics, 1995. 26(2): p. 178-91.

[47]. Klein, P.E., et al., A high-throughput AFLP-based method for constructing integrated genetic and physical maps: progress toward a sorghum genome map. Genome Res, 2000. 10(6): p. 789-807.

[48]. Green, E.D. and M.V. Olson, Systematic screening of yeast artificial-chromosome libraries by use of the polymerase chain reaction. Proc Natl Acad Sci U S A, 1990. 87(3): p. 1213-7.

[49]. Israel, D.I., A PCR-based method for high stringency screening of DNA libraries. Nucleic Acids Res, 1993. 21(11): p. 2627-31.

[50]. Kwiatkowski, T.J., et al., Rapid identification of yeast artificial chromosome clones by matrix pooling and crude lysate PCR. Nucleic Acids Res, 1990. 18(23): p. 7191-2.

[51]. Libert, F., et al., Construction of a bovine genomic library of large yeast artificial chromosome clones. Genomics, 1993. 18(2): p. 270-6.

[52]. Crooijmans, R.P., et al., Two-dimensional screening of the Wageningen chicken BAC library. Mamm Genome, 2000. 11(5): p. 360-3.

[53]. Gussow, D. and T. Clackson, Direct clone characterization from plaques and colonies by the polymerase chain reaction. Nucleic Acids Res, 1989. 17(10): p. 4000.

[54]. Pollier, J., et al., An integrated PCR colony hybridization approach to screen cDNA libraries for full-length coding sequences. PLoS One, 2011. 6(9): p. e24978.

[55]. Farrar, K. and I.S. Donnison, Construction and screening of BAC libraries made from Brachypodium genomic DNA. Nat Protoc, 2007. 2(7): p. 1661-74.

[56]. Vu, G.T., P.D. Caligari and M.J. Wilkinson, A simple, high throughput method to locate single copy sequences from Bacterial Artificial Chromosome (BAC) libraries using High Resolution Melt analysis. BMC Genomics, 2010. 11: p. 301.

[57]. Kim, C.G., A. Fujiyama and N. Saitou, Construction of a gorilla fosmid library and its PCR screening system. Genomics, 2003. 82(5): p. 571-4.

[58]. Chumakov, I., et al., Continuum of overlapping clones spanning the entire human chromosome 21q. Nature, 1992. 359(6394): p. 380-7.

[59]. Liu, W., et al., Construction and characterization of a novel 13.34-fold chicken bacterial artificial chromosome library. Anim Biotechnol, 2003. 14(2): p. 145-53.

[60]. Febrer, M., et al., Rapid identification of the three homoeologues of the wheat dwarfing gene Rht using a novel PCR-based screen of three-dimensional BAC pools. Genome, 2009. 52(12): p. 993-1000.

[61]. Bouzidi, M.F., et al., A sunflower BAC library suitable for PCR screening and physical mapping of targeted genomic regions. Theor Appl Genet, 2006. 113(1): p. 81-9.

[62]. Barillot, E., B. Lacroix and D. Cohen, Theoretical analysis of library screening using a N-dimensional pooling strategy. Nucleic Acids Res, 1991. 19(22): p. 6241-7.

[63]. Jin, F., et al., A pooling-deconvolution strategy for biological network elucidation. Nat Methods, 2006. 3(3): p. 183-9.

[64]. Elnifro, E.M., et al., Multiplex PCR: optimization and application in diagnostic virology. Clin Microbiol Rev, 2000. 13(4): p. 559-70.

[65]. Edwards, M.C. and R.A. Gibbs, Multiplex PCR: advantages, development, and applications. PCR Methods Appl, 1994. 3(4): p. S65-75.

[66]. Henegariu, O., et al., Multiplex PCR: critical parameters and step-by-step protocol. Biotechniques, 1997. 23(3): p. 504-11.

[67]. 渉等, プラナリアの形態分化: 基礎から遺伝子まで. 1998.

[68]. S A Nchez Alvarado, A., et al., Proposal for the sequencing of a new target genome: white paper for a planarian genome project. The Schmidtea mediterranea sequencing consortium, 2003.

[69]. Robb, S.M., E. Ross and A.A. Sanchez, SmedGD: the Schmidtea mediterranea genome database. Nucleic Acids Res, 2008. 36(Database issue): p. D599-606.

[70]. Chain, P.S., et al., Genomics. Genome project standards in a new era of sequencing. Science, 2009. 326(5950): p. 236-7.

[71]. Sanger, F. and A.R. Coulson, A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. Journal of molecular biology, 1975. 94(3): p. 441--448.

[72]. Sanger, F., S. Nicklen and A.R. Coulson, DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences, 1977. 74(12): p. 5463--5467.

[73]. Venter, J.C., et al., The sequence of the human genome. science, 2001. 291(5507): p. 1304--1351.

[74]. Shendure, J. and H. Ji, Next-generation DNA sequencing. Nature biotechnology, 2008. 26(10): p. 1135--1145.

[75]. Margulies, M., et al., Genome sequencing in microfabricated high-density picolitre reactors. Nature, 2005. 437(7057): p. 376--380.

[76]. Brown, S.M., Next-generation DNA sequencing informatics. 2013: Cold Spring Harbor Laboratory Press.

[77]. Martin, M., Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. journal, 2011. 17(1): p. pp--10.

[78]. Cox, M.P., D.A. Peterson and P.J. Biggs, SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. BMC bioinformatics, 2010. 11(1): p. 485.

[79]. Schmieder, R. and R. Edwards, Quality control and preprocessing of metagenomic datasets. Bioinformatics {(Oxford,} England), 2011. 27(6): p. 863--864.

[80]. Bioinformatics, B., FASTQC: A quality control tool for high throughput sequence data. 2011, Cambridge, UK: Babraham Institute.

[81]. Dole V Z El, J., J. Greilhuber and J. Suda, Estimation of nuclear DNA content in plants using flow cytometry. Nature protocols, 2007. 2(9): p. 2233--2244.

[82]. Waring, M. and R.J. Britten, Nucleotide sequence repetition: a rapidly reassociating fraction of mouse DNA. Science, 1966. 154(3750): p. 791--794.

[83]. DeWoody, Y.D. and J.A. DeWoody, On the estimation of genome-wide heterozygosity using molecular markers. Journal of Heredity, 2005. 96(2): p. 85--88.

[84]. Gresham, D., et al., Optimized detection of sequence variation in heterozygous genomes using DNA microarrays with isothermal-melting probes. Proceedings of the National Academy of Sciences, 2010. 107(4): p. 1482--1487.

[85]. Liu, B., et al., Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. 2013. p. 47.

[86]. Chikhi, R. and P. Medvedev, Informed and automated k-mer size selection for genome assembly. Bioinformatics, 2013: p. btt310.

[87]. Darling, A.E., B. Mau and N.T. Perna, progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PloS one, 2010. 5(6): p. e11147.

[88]. Gnerre, S., et al., High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci U S A, 2011. 108(4): p. 1513-8.

[89]. Luo, R., et al., SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience, 2012. 1(1): p. 18.

[90]. Zerbino, D.R. and E. Birney, Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res, 2008. 18(5): p. 821-9.

[91]. Ruan, J., et al., Pseudo-Sanger sequencing: massively parallel production of long and near error-free reads using NGS technology. BMC Genomics, 2013. 14: p. 711.

[92]. Boetzer, M., et al., Scaffolding pre-assembled contigs using SSPACE. Bioinformatics, 2011. 27(4): p. 578-9.

[93]. Boetzer, M. and W. Pirovano, Toward almost closed genomes with GapFiller. Genome biology, 2012. 13(6): p. R56.

[94]. Xue, W., et al., L_RNA_scaffolder: scaffolding genomes with transcripts. BMC Genomics, 2013. 14: p. 604.

[95]. Stein, L., Genome annotation: from sequence to biology. Nat Rev Genet, 2001. 2(7): p. 493-503.

[96]. Grabherr, M.G., et al., Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol, 2011. 29(7): p. 644-52.

[97]. Kent, W.J., BLAT--the BLAST-like alignment tool. Genome Res, 2002. 12(4): p. 656-64.

[98]. Stanke, M., et al., AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res, 2006. 34(Web Server issue): p. W435-9.

[99]. Levy, S., S. Hannenhalli and C. Workman, Enrichment of regulatory signals in conserved non-coding genomic sequence. Bioinformatics, 2001. 17(10): p. 871--877.

[100]. Hardison, R.C., Conserved noncoding sequences are reliable guides to regulatory elements. Trends in Genetics, 2000. 16(9): p. 369--372.

[101]. Kurtz, S., et al., Versatile and open software for comparing large genomes. Genome Biol, 2004. 5(2): p. R12.

[102]. Cebrià, F., et al., FGFR-related gene nou-darake restricts brain tissues to the head region of planarians. Nature, 2002. 419(6907): p. 620-624.