SIP Protocol Series No.2

Genetic Connectivity Survey Manuals

Cross-ministerial Strategic Innovation Promotion Program

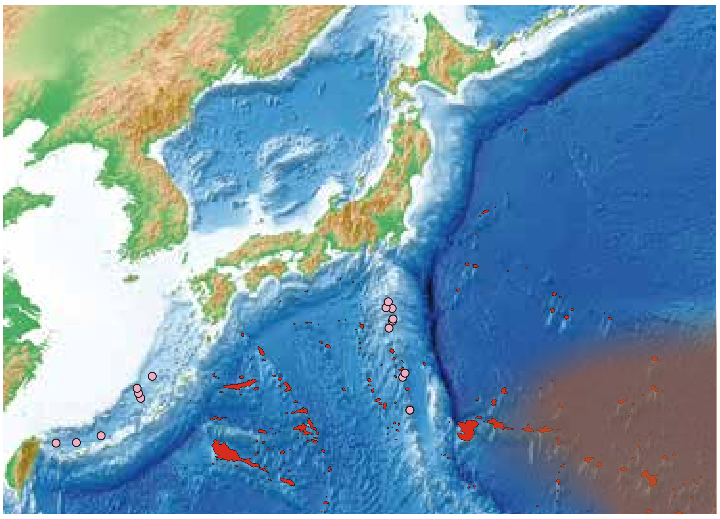
The Strategic Innovation Promotion Program (SIP) was launched by the Council for Science, Technology, and Innovation (CSTI), which oversees projects that target scientific and technological innovation in line with Japanese government directions as stated in the Comprehensive Strategy on Science Technology and Innovation and the Japan Revitalization Strategy. This interdisciplinary program among government agencies, academic institutes and private sectors addresses eleven issues. One of these issues is Next-Generation Technology for Ocean Resources Exploration.

Zipangu in the Ocean Program and Protocols for Environmental Survey Technologies

Zipangu in the Ocean Program is a technical study of the development of submarine mineral deposits that takes into consideration the wise use of these resources.

One research area is the ecological survey of organisms and their long-term monitoring. However, an ecosystem consists of various interrelated factors; thus, in addition to a comprehensive understanding of the system, observation and analysis of each component to its most elemental level are unavoidable. Recently, increased environmental awareness and the necessity of forming a consensus have become key issues in conducting development activities. Growing concern for the environment by the public and the diversification of the use of maritime areas have complicated the interests of stakeholders. To facilitate the formation of a consensus under these conditions, it is important for standardized methods to be implemented. This will ensure that research processes are transparent and that the collection of survey data is objective.

This protocol series aims to introduce more accurate, user-friendly, objective and effective underlying technologies required to understand the environmental impact of submarine mineral resource development. We believe that creating such a technology tool-kit will allow us to develop these resources in a sustainable manner.



○ Thermal activity ● Cobalt-rich crust ● REY-rich mud ※1 based on *Usui et.al(1994)*

Table of Contents

chapter 1

Introduction....2

chapter 2

Genetic Connectivity Surveys....4

2-1 Selection of the Target Species....4

2-2 Selection of Genetic Marker....4

2-3 Sampling....5

1 Sampling Methods

2 Sample Size

3 Sample Preservation

2-4 Taxonomic Identification....6

2-5 DNA Analysis....7

1 DNA Extraction

2 PCR

A.DNA Polymerase

B. PCR Primers (for Cytochrome C Oxidase Subunit I Gene)

C.Preparation of the PCR Solution and Reaction Conditions

3 Verifying PCR Amplification Products (Agarose Gel Electrophoresis)

4 Purification of the PCR Amplification Products

5 Sequencing

2-6 Data Analysis....11

1 Species Identification using DNA Sequences

A.Homology Search using BLAST

B.Molecular Phylogenetic Trees

2 Population Genetic Analysis

A.Haplotype Network

B. Analysis of Molecular Variance (AMOVA)

C.Coalescent Theory

Notes....14

References....15

Memo....16

chapter 1
Introduction

“ Environmental impact assessments (EIAs) are indispensable for the sustainable development of seabed mineral resources. After assessing the impact, it is necessary to consider policies for preserving the environment, but what kind of information is required when considering such policies?

The development of seabed mineral resources is accompanied by a series of activities including mining, processing, lifting, and transportation. Although it may still take some time before each of these methods is established, it goes without saying that any method will exert an impact on the marine environment. The EIAs make prior estimations and evaluations of such impacts and in the actual implementation of developments, consideration may be given to halting the project based on the results of an impact assessment. Moreover, even where an assessment does not result in halting of a project, it is necessary to consider policies for mitigating against environmental impact by establishing protected areas in the adjacent area, and considering sustainable development measures.

Of these, this document deals with surveys for the selection of protected areas. The establishment of a protected areas may have two aims. The first is the prior protection of an equivalent area, as compensation for the seafloor that may be lost or disturbed by development. This case demands some areas that has the size, environmental conditions, and similarity of characteristic distribution of organisms in the protected areas. The main focus of the second aim is promoting the regeneration of the ecosystem of the developed region by leaving a fixed area as a source for the reintroduction of communities. In this case, rather than the similarity of size, environmental conditions, or distribution of organisms, there must be a two-way movement of organisms between the developed area and the protected area, and the protected area must be the upstream supplier of organisms. The connection of areas by this two-way movement of organisms is called ecosystem connectivity.

However, be it for equivalent ecosystems or the supply of a reintroduced biome, selecting a protected area based on scientific foundations is not simple, even with an understanding of the basics. With regards to supplying a source for re-introducing a biome, attempts to assess ecosystem connectivity from the dispersion of planktonic larvae have included: physical ocean model simulations, breeding experiments to understand the behavioral characteristics of larvae, particle tracking methods for marker particles with neutral buoyancy that resemble larvae, and chemical fingerprint analysis, which predicts the diffusion path of larvae from stable isotopes and trace elements.

However, none of these can be readily called well-established methods, but because investigation into the main factors that control ecosystem connectivity has been narrowed to larvae dispersion, there is scope for the consideration of new methods. A method whereby inter-ecosystem connectivity is extrapolated from genetic markers has been proposed and is gaining attention. This method makes indirect estimations from the genetic information of each ecosystem and is not limited to the dispersion process.

This document describes a method for estimating the population connectivity using genetic markers (Fig 1).

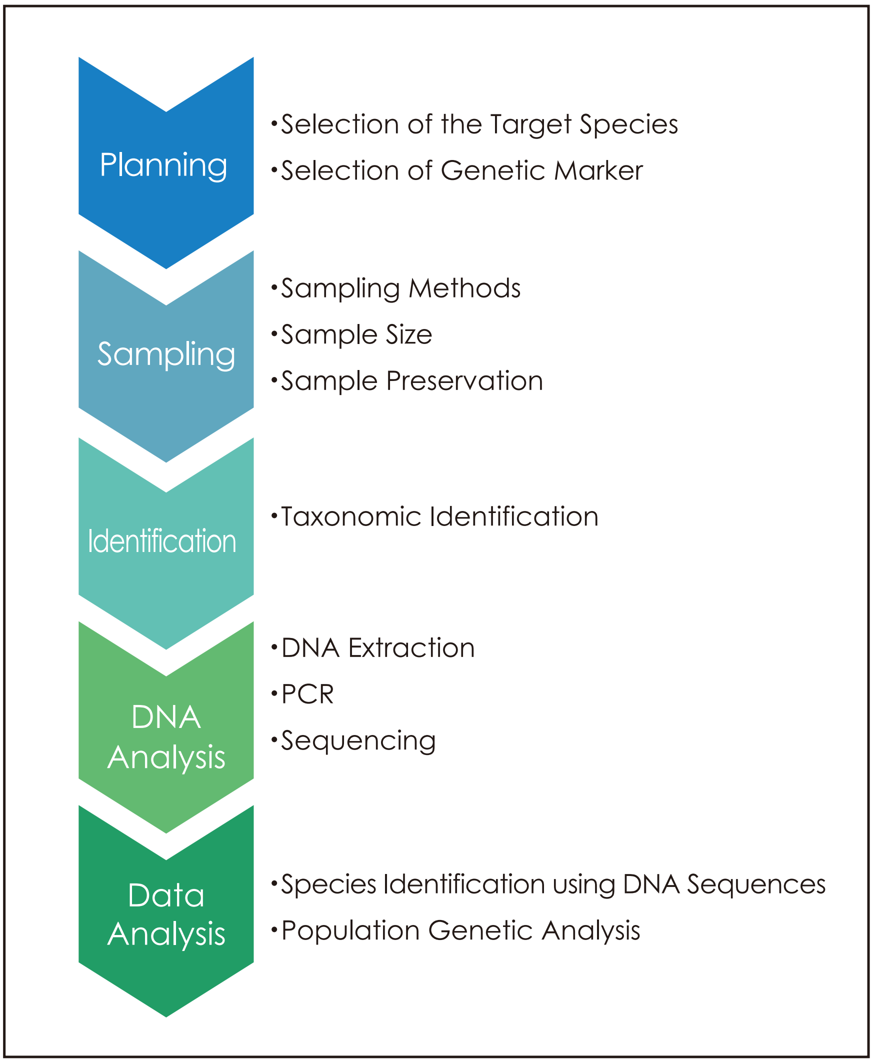

Planning
・Selection of the Target Species
・Selection of Genetic Marker
Sampling
・Sampling Methods
・Sample Size
・Sample Preservation
Identification
・Taxonomic Identification
DNA Analysis
・DNA Extraction
・PCR
・Sequencing
Data Analysis
・Species Identification using DNA Sequences
・Population Genetic Analysis

Fig. 1 The survey process

chapter 2

Genetic Connectivity Surveys

“

2-1. Selection of the Target Species

Most deep-sea benthos lives on hydrothermal vents are sessile or low mobility. These adults are difficult to migrate among distant populations. Instead, they have planktonic larval stage, dispersal of larvae allowed to supply individuals to other populations. Since hydrothermal vents are present in patches in the vast area, it is necessary to disperse larvae more widely in hydrothermal vents species than coastal species.Larvae are usually as small as 0.5 mm or less and are considered as particles carried by ocean currents. The transport distance of larvae is influenced by various factors such as behavior, ocean current, temperature, presence of keys for metamorphosis. The species that flow horizontally while rising to the surface layer or the middle layer tends to increase the travel distance. Ocean currents that determine the direction and distance of larvae transport are not always constant. For example, the ocean current near the bottom layer changes greatly in the direction of flow with a half day cycle. The transport direction and speed of water column differs depending on the distance from the seabed. Furthermore, the transport distance varies depending on planktonic larval duration, and the density of the larva affects the sedimentation pattern.Planktonic larval duration and dispersal pattern varies from species to species. Information on planktonic larval duration is obtained from only some species (Hilário et al., 2015). For example, planktonic larval duration of tubeworms Riftia pachyptila is 38 days on average (Marsh et al., 2001), that of deep-sea clams Bathymodiolus childressi is 240 days on average (Arellano & Young, 2009).

Three higher taxa (Phylum: Mollusca, Arthropoda, and Annelida) are known to be dominant in vent communities (Baker et al., 2001). We recommend to select several species from these taxa that have different ecological characteristics.

2-2. Selection of Genetic Marker

Mitochondrial DNA (mtDNA)[1] , microsatellites[2] , single nucleotide polymorphisms (SNPs)[3], and allozymes[4] have been used as genetic markers for investigating the genetic connectivity of populations (Hellberg et al., 2002, Baco et al., 2016).

In previous studies of deep-sea benthos, the COI gene[5] sequences has been mainly applied to various taxon. Because there are universal PCR primers that can be

ous substitutions[6] are found in the third locus of codons that code for amino acids within the COI gene, making it easier to compare the genetic variation levels among populations within the same species (Watanabe et al., 2010). Furthermore,

arcoding for classification group estimation in many animals, and a large number of nucleotide sequences are registered in Genebank. Deep-sea benthos include species that are difficult to identify morphologically or cryptic species[7]. Based on the determined nucleotide sequence of the COI gene, molecular phylogenetic identification of target species can be performed at the same time. Therefore, the COI gene is optimal as a genetic marker for investigating genetic connectivity. However, the rate of base substitution of

COI gene has been reported as being extremely slow in the Anthozoa and the Porifera (Huang et al., 2008), it should be considered for use.

MtDNA is generally maternally inherited, it has the disadvantage that detection of interspecific hybridization and mixed population cannot be examined. In addition, since the results vary greatly due to the influence of genetic drift, analysis with only a single gene should increase the number of genetic markers used for analysis (Toews & Brelsford, 2012). If possible, it is also recommended to add analyses using co-dominant markers such as microsatellites.

Since the species specificity of microsatellite markers is so high than mtDNA markers, it is best to use markers that have been developed for target species. Depending on the characteristics of the marker, there are cases where it can be applied even to other species in the same genus. The possibility of PCR amplification and the polymorphism level with markers should be checked in advance to evaluate the availability of markers. If markers are not developed even for the target species or closely-related species, it is necessary to develop markers by a method utilizing the next generation sequencing (Nakajima et al., 2014).

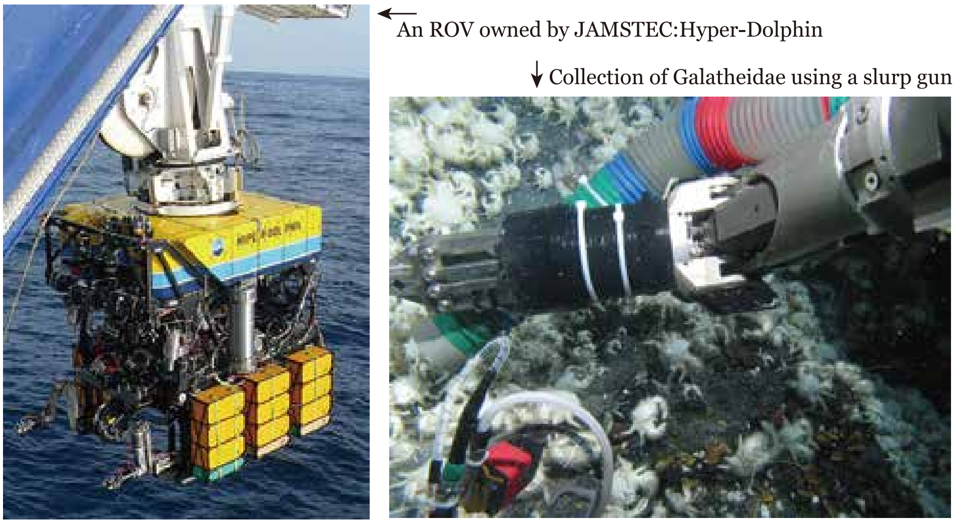
2-3. Sampling

1 Sampling Methods

The equipment used for collecting organisms for a sample differs depending on the environment of the survey location, water depth, and ecological type of the target species. If the seabed is composed of soft sediments, efficient sampling methods include peeling benthos inhabiting on the seabed with dredge type equipment such as beam troll or sledge nets, or digging sediment with organisms by a box corer. Targets for collection include swimming shrimp, tubeworms, sea anemones, starfish, and sea cucumbers for dredge type equipment while box corers are suitable for infaunal animals such as Polychaetes. Additionally, by using ROVs to place incentivized mesh baskets such as bait traps (crab baskets) and collecting them after a fixed time; it is possible to collect carnivorous and scavengerous species.

In the hydrothermal vent area, widely-used techniques include direct collection using a manipulator attached to a remotely-operated underwater vehicle (ROV) and suction collection using a slurp gun (Fig. 2). These methods, which collect target organisms by visual confirmation cause little damage to the samples and pose no risk of collecting non-target organisms, thus making them the most beneficial sampling methods. Such methods are suitable for target organisms that include species which are mainly sessile or limited mobility such as the Gastropoda, Bivalvia, and Galatheidae.

The samples are hauled on deck and after washing with filtered, cooled seawater, classified by species or taxon in trays, and then photographed. The required information, such as number of organisms and organism size, are then recorded.



← An ROV owned by JAMSTEC:Hyper-Dolphin

↓ Collection of Galatheidae using a slurp gun

Fig 2. An example of using an ROV for sampling

2 Sample Size

A common question that accompanies surveys using genetic markers is, how many individuals to sample and analyze for each population to ensure it is representative of the population. If one considers that a sample population is created by random sampling from a relatively large parent set, then, as the sample size becomes larger, the probability that the results correctly approximates the parent set becomes greater. Because the number of haplotypes[8] detected from the population is directly affected by the sample number, if a larger sample is analyzed, it is possible to detect a greater number of rare haplotypes from the population. Since the sampling of seabed organisms itself is extremely time and labor consuming, the amount of sampling that can be obtained at one time may be limited. Considering the time and costs incurred by analyzing large numbers of samples, it is desirable to have the analysis sample numbers as low as possible.

However, since both haplotype diversity[9] and nucleotide diversity[10] have very marked variance in sample sizes of ten individuals or less, it has been stated that sampling requires 25 or more individuals per population (Goodall-Copestake et al., 2012). Additionally, in a simulation analysis that varied the sample number between 2 and 100 specimens, it was reported that the smallest number required to stabilize variations was 20 individuals per population (Luo et al., 2015). From the information above, sample size per population is appropriate to 20 individuals or more.

3 Sample Preservation

After recording (via photographs, etc.) the required morphological information in the on-board laboratory as shown in section 3.1, the samples of collected organisms are preserved either by fixing the specimens in ethanol and storing in a cool dark place, or in a freezer at -20 °C or -80 °C without fixing. The information required for species identification based on the morphological information (which is covered later) is recorded before commencing preservation. It is particularly important, when preserving samples as tissue slices, to record observations carefully to avoid overlooking an individual' s characteristic morphological information.

When fixing with ethanol, containers should be used that can be thoroughly sealed in order to avoid leakage of the fluids. Only high purity fixation ethanol should be used for fixing samples. Additionally, a concentration of 95% ethanol should be maintained after fixation and the amount of sample and seawater stored in the container should be adjusted appropriately[11]. Giving consideration to the sample containers and the size of the storage facilities, small-bodied specimens may be kept as they are, while large specimens should be cut using a disposable scalpel or similar, to produce one or more tissue slices up to 5 mm on a side before storing in a container. To prevent sample contamination, it is desirable for a single container to contain one specimen or tissue slices from one organism. Affix labels with information about the sample, such as species name (scientific or common), collection date, and collection site, to those containers that hold samples. After obtaining the base sequence, record whether genetic information was obtained for classification purposes.

2-4. Taxonomic Identification

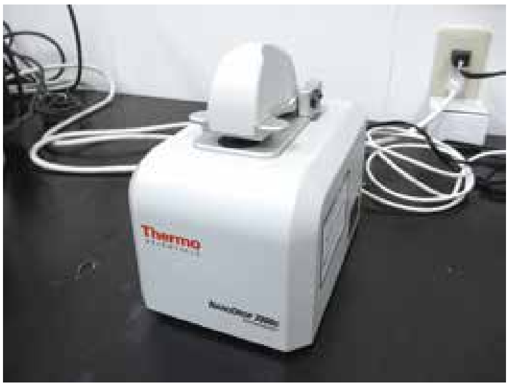
Conduct species identification using morphological and genetic characteristics and record the results for each individual organism used for analysis. Even when only collecting a part of a tissue for storage, obtain morphological information beforehand. Having selected an organism as one that is characteristic to the area of ocean under investigation, conduct identification with genetic base sequences used for the discernment of the taxa. This is important because deep sea organisms which have yet to be researched may be identified. The genetic base sequence provides important information for deciding whether to adopt or reject individuals, such as those of cryptic species (or new species) that cannot be classified from their morphological information, for use in analysis, according to the aims of the later analysis work. Record determined base sequences alongside the morphological information for each individual, and register it in the international DNA base sequence databases. Conducting work based on this kind of morphological and genetic information,

helps to avoid the production of erroneous results due to misidentification, which may be conducive to subsequent taxonomic research. Conduct identification through morphological information with reference to morphological characteristics in previous reports or books, etc. Also refer to the results of previous surveys related to the distribution of species by ocean area (Watanabe et al., 2010, Ishibashi et al., 2015).
Because new species, cryptic species, synonyms, and hybrids, are being found continuously among the deep-sea benthos (Watanabe, 2010), it is highly recommended to always update the information and use it in identification work. Species identification by gene is discussed in section 6.1.

2-5. DNA Analysis

1 DNA Extraction

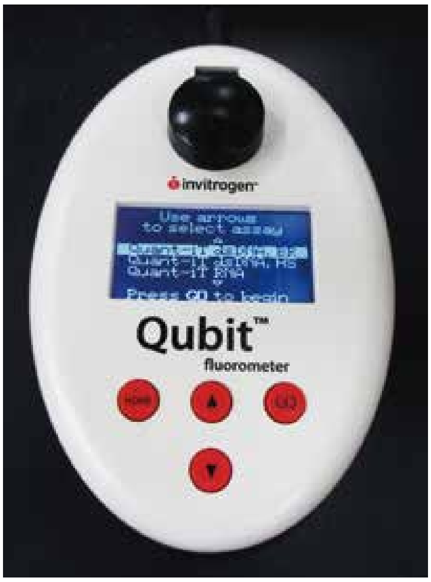
There are several methods for extracting DNA from biological tissues, including the PCI (phenol-chloroform-isoamyl alcohol) method, the CTAB (Cetyl trimethyl ammonium bromide) method, the Chelex 100 resin method, and the silica column method. It is important to choose an appropriate extraction method for the characteristics of the samples used, and the aims of the analysis.
The PCI method is inexpensive and can be scaled up. However, it uses highly toxic substances such as phenol and chloroform, it is necessary to perform the method with sufficient consideration for worker safety and processing of the waste fluids[12]. The CTAB method efficiently removes the PCR inhibitor, polysaccharide. It is a very effective method for handling tissues which contain a large amount of polysaccharides, such as shellfish. (Tel-zur et al., 1999). Methods using the weak positive ion exchange resin, Chelex 100, do not use toxic substances and can extract DNA inexpensively (Walsh et al., 1999), however, the obtained DNA is low purity and yields compared to other methods.
The silica column is a safe, highly efficient, and high purity yielding method, and therefor compensates for the disadvantages of the above extraction methods. Its kits are sold by various manufacturers. When using these kits, it is necessary to choose the kit according to its capabilities and cost. An example is the DNeasy Blood & Tissue Kit by QIAGEN. Ensure that a record is kept regarding the kind of extraction method used, the kit used, and the extraction process, etc.
Measure the concentration of the DNA to verify the purity of the extracted DNA and investigate the amount of extracted DNA to use as a template in the subsequent PCR step. Methods for measuring DNA concentration include those that use a UV spectrophotometer, and those that use a fluorometer. The concentration of DNA is measured with a UV spectrophotometer at a wavelength of 260 nm (A260). Additionally, from the ratio of intensities measured simultaneously at a wavelength of 230 nm (A230) and 280 nm (A280), it is possible to verify the purity of the DNA. If the A260/A280 ratio of the DNA extraction fluid is within the 1.8 to 2.0 range, the purity can be considered as high, while values of 1.8 or less may indicate contamination with protein or phenols, etc. Additionally, if the A260/A230 ratio is 2.0 or less, this may indicate contamination with polysaccharides, salts of Tris or EDTA, or organic solvent. UV spectrophotometers for measuring nucleic acid concentrations include the NanoDrop series by Thermo Fisher Scientific and the NanoVue Plus by GE Healthcare (Fig. 3).



NanoDrop（Thermo Fisher Scientific）

Fig.3 Examples of spectrophotometers

Absorbance measurements by spectrophotometers not only measure the UV absorbance of DNA, but of all substances absorbing at 260 nm, such as contaminating RNAs, proteins, and free nucleotides. Consequently, it is possible that the concentration of DNA will be overestimated. Moreover, if the concentration of DNA in the solution is too low, the UV spectrophotometer cannot make accurate measurements. Fluorometer measurements use fluorescent pigments that only emit light when bound to DNA molecules in a specimen. The concentration of DNA can be calculated from the intensity of fluorescence measured from fluorescent pigments bound to the DNA. This method has the advantages of being able to measure DNA concentrations without being greatly influenced by contaminants, and also is able to quantify low-concentration DNA solutions that would be difficult to measure using UV spectrophotometers. Fluorometers for measuring nucleic acid concentrations include the Qubit series by Thermo Fisher Scientific (Fig.4).



Qubit（Thermo Fisher Scientific）

Fig.4 Example of a fluorometer

2 PCR

A. DNA Polymerase

The DNA polymerases used for PCR can be broadly classified into two categories: the bacteria-derived Taq and Tth polymerases, and the thermophilic archaea-derived Pfu and KOD polymerases. Taq-type polymerases have good amplification efficiency, and a fast elongation rate. The Pfu-type polymerases have proofreading activity (exonuclease activity in the 3'→5' direction), so they can repair bases that were incorporated by mistake during the elongation reaction of the PCR.

Because there are many DNA polymerase products marketed for PCR, it is important to choose one that is appropriate for the characteristics of the samples used, and for the purpose of use. Where it is suspected that PCR inhibitors are present, it may be possible to produce successful PCR reactions using reagents that control the influence of contaminants in the extracted DNA, such as with Ampdirect Plus (Shimadzu). Record the DNA polymerase used.

B. PCR Primers

COI gene of the mitochondrial DNA will be used as an example. It is standard to use a set of universal primers for marine invertebrates, LCO1490/HCO2198 (Folmer et al., 1994). If PCR amplification is not possible using this primer set, it is possible that a base substitution has occurred in the primer annealing region. In this case, refer to primers revised in a taxon-specific manner (Table 1). Ensure that the sequence data and source of the used primer set is recorded.

Table 1 Primer sets for the amplification of CO I gene region

Target Taxa	Primer Name	Sequence (5'→3')	Reference
All	LCO1490	GGTCAACAAATCATAAAGATATTGG	Folmer et al., 1994
	HCO2198	TAAACTTCAGGGTGACCAAAAAATCA	
All	jgLCO1490	TITCIACIAAYCAYAARGAYATTGG	Geller et al., 2013
	jgHCO2198	TAIACYTCIGGRTGICCRAARAAYCA	
Polychaeta	PolyLCO	GAYTATWTTCAACAAATCATAAAGATATTGG	Carr et al., 2011
	PolyHCO	TAMACTTCWGGGTGACCAAARAATCA	
Gastropoda	dgLCO-1490	GGTCAACAAATCATAAAGAYATYGG	Meyer, 2003
	dgHCO-2198	TAAACTTCAGGGTGACCAAARAAYCA	
Decapoda	CrustF1	TTTTCTACAAATCATAAAGACATTGG	Costa et al., 2005
	HCO2198	TAAACTTCAGGGTGACCAAAAAATCA	Folmer et al., 1994
Munidopsis genus	gala_COIF	CATCACTWAGWTTRATYATTCCAGCAGAA	Jones et al., 2007
	gala_COIR	GAAYAGGRTCTCCTCCTCCTAC	
Alviniconcha genus	COI-B	GGATGAACNGTNTAYCCNCC	Kojima et al., 2001
	COI-6	GGRTARTCNSWRTANCGNCGNGGYAT	
Bathymodiolus genus	COIG	GTATTGAATTAGCACGTCCTGGAA	Genio et al., 2008
	COIH	ATACTATTCCAAACCCGGGTAAAAT	

C.Preparation of the PCR Solution and Reaction Conditions

An example PCR solution composition is shown in Table 2. It is beneficial to refer to the manual included with the DNA polymerase being used, and to conduct prior tests to obtain the desired PCR products. This will be helpful in order to adjust the amount of each reagent in accordance with conditions such as the type of DNA polymerase and the concentration of the DNA extracted from the sample (template DNA). When making the PCR solution, also prepare a single negative control that does not contain the template DNA, to ensure that there are no DNA contaminants in the reagents.

Table 2 An Example PCR Solution Composition

Reagent Name	Amount (/tube)	Final Concentration
TaKaRa Ex Taq HS (5unit/μL)	0.2 μL	0.05 unit/μL
10×Ex Taq Buffer (Mg2+ plus)	2.0 μL	1×
dNTP mixture	1.6 μL	
Forward primer (10μM)	0.4 μL	0.2 μM
Reverse primer (10μM)	0.4 μL	0.2 μM
Template DNA (50ng/μL)	1.0 μL	Adjusted to 5-200 ng
DNase Free Water (PCR grade)	14.4 μL	
total volume	20.0 μL	

Example PCR reaction conditions are shown in Table 3. Because the PCR reaction conditions must be adjusted according to certain conditions, including the primer set used, the type of DNA polymerase, and the model of thermal cycler, it is best to conduct prior experiments to obtain the desired PCR products. Record the composition of the PCR solution used, the model of the thermal cycler, and the PCR reaction conditions.

Table 3 Example PCR Reaction Conditions

Temperature	Time	Number of Cycles
94℃	2 min	1
94℃	30 sec	30
50℃	30 sec	
72℃	1 min	
72℃	2 min	1
4℃	∞	

3 Verifying PCR Amplification Products (Agarose Gel Electrophoresis)

Agarose gel electrophoresis is used to confirm whether the used primer set obtained DNA fragments of the desired length in the amplification products. The Mupid® series is a submarine type of an electrophoresis device. Use high purity agarose marketed for use in electrophoresis. If the amplification fragments are short, use a concentration of 2% or more. Generally, TBE (Tris-Borate-EDTA) or TAE (Tris-Acetate-EDTA) are used as electrophoresis buffers. Use size markers alongside samples during electrophoresis to find the length of the amplified fragments. For the electrophoresis, also include the negative control that is prepared at the PCR stage to confirm whether there was contamination, and verify that it does not contain amplification products. To visually confirm how the electrophoresis develops, mix the PCR amplification products with a loading dye (a mixture containing dye and glycerin). Electrophoresis is generally performed at 100 V for 20 to 30 minutes.

Following electrophoresis, stain the agarose gel with a nucleic acid staining agent, such as SYBR Gold Nucleic Acid Gel Stain (Invitrogen) or RedSafe (iNtRON Biotechnology), and observe under a blue-LED or UV trans-illuminator. Verify that the desired size of DNA fragment was amplified and that there was no amplification in the negative control. Then, record or save an image of the electrophoresis, either digitally or as a photograph. Record details such as the concentration, buffer solution, voltage, electrophoresis duration, amount of sample used, size marker, and staining method.

4 Purification of the Amplification Products

Before conducting the cycle sequencing described in the next section, remove excess components such as primer and dNTP from (i.e. purification) the PCR amplification products. While there are several methods for purifying the PCR amplification products, purification using enzymes such as exonuclease I are commonly used due to their convenience and the minimal loss of PCR amplification products. ExoSAP-IT® (Affymetrix) is a marketed combination of exonuclease I and shrimp derived alkaline phosphatase. Use reagents such as ExoSAP-IT® according to the procedures designated by the manufacturer. Record the purification method, processed concentration, duration, etc.

5 Sequencing

Use a DNA sequencer to conduct a cycle sequencing reaction, using the purified PCR products as a template, to decode the base sequence. For example, when using an Applied Biosystems capillary DNA sequencer (e.g. the 3130 Genetic Analyzer <Fig.5>) use a BigDye® Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems) for the cycle sequencing reagents. Use either the PCR primers for the sequencing primers or, if the DNA fragment length is greater than 600 bp, design internal primers for appropriate positions, dividing it into two or more regions for sequencing.

fig.5 3130 Genetic Analyzer (Applied Biosystems) DNA sequencer

Retrieve the waveform data output from the DNA sequencer using sequence data analysis software (for example, Sequencing Analysis Software v5.2, which is included as standard in the case of capillary DNA sequencers made by Applied Biosystems, Inc.) and visually confirm that the bases have been assigned correctly. Discard parts, including bases labeled as low confidence (QV values) due to poor isolation of the first few tens of bases from the start of sequence reading. At this stage, if multiple bases are detected consecutively, DNA contamination is likely. Convert the thoroughly examined waveform data into base sequence text data and reconstruct the original sequence from each forward and reverse sequence. If sequencing was conducted by dividing a single DNA fragment into two or more regions using internal primers, assemble (concatenate) sequence data. If the PCR primer sequence was included, remove it. Save the base sequence determined through this processing as text data and use it in the analysis described in section 6.

2-6. Data Analysis

Before investigating the genetic connectivity using the determined base sequences, conduct species identification based on the genetic base sequences of the target organisms using a BLAST search or molecular phylogenetic analysis. Next, investigate the genetic connectivity of the populations via population genetic analysis using the sequences of the gene.

1 Species Identification using COI Sequence

A. Homology Searches using BLAST

Before conducting a genetic population analysis, it is necessary to confirm that the determined base sequences are indeed derived from the species targeted in the survey using international base sequence databases (NCBI/DDBJ/EMBL)[13]. For example, sometimes DNA from an organism other than the target organism, contaminates the extracted DNA. Even for samples that are tested based on morphological characteristics, it is possible that, at the genetic level, the species is a cryptic species or a different species with similar morphology to the target species. Additionally, where the target organism has experienced events such as gene duplication, the target gene region may be duplicated or there may be pseudogenes[14]. Exclude non-homologous base sequences, such as incorrectly decoded sequences and pseudogenes, before conducting a population genetic analysis using the determined base sequences.

More specifically, use the determined sequences with homology search tools that can be used on online, international base sequence databases and BLAST (Altschul et al., 1990). The results of the search will be shown as a list of data on base sequences similar to the determined sequence, with the highest scores listed first. Estimate closely related taxa for the determined sequence based on the displayed scientific names[15]. Save the search results as digital data.

B. Molecular Phylogenetic Trees

If the base sequences obtained from the BLAST searches seem to correspond to the target species, it is beneficial to produce a molecular phylogenetic tree to objectively ground this possibility in the phylogenetic position of the base sequences. To better understand those species that cannot be distinguished from morphological characteristics, such as cryptic species, it is necessary to produce a molecular phylogenetic tree for the determined base sequences and the base sequences of closely-related taxa and verify whether the target organism belongs to a previously known family, or whether it belongs to a new family. Consequently, for comparison, include not only the determined base sequences, but also those sequences shown near the top of the BLAST search in the analysis when producing a phylogenetic tree. Software such as MEGA (Molecular Evolutionary Genetics Analysis), RAxML, and MrBayes, which uses Bayesian estimation, are used for the molecular phylogenetic analysis. Additionally, books on the production of phylogenetic trees (Hall, 2011) may also be a useful reference. If, in the obtained molecular phylogenetic tree, the target species of the analysis forms a monophyletic group[16] that differs from the expected taxon, then it is highly likely that the analysis target is a cryptic species. Consequently, it is necessary to differentiate the expected taxon and the cryptic species taxon in following population genetic analysis. The taxa of determined target organisms are estimated at the genetic level using BLAST homology searches and molecular phylogenetic trees. Record these estimation results as per the species identification. Additionally, save the sequence data (for example, as a text file).

2 Population Genetic Analysis

The population genetic analysis featured here is a method for using a single species to statistically examine the genetic connectivity between the populations of several ocean regions. Consequently, it is not possible to conduct the population genetic analysis described below when organisms that do not comprise a monophyletic group are included in the molecular phylogenetic tree analysis. Because of this, some caution is required.

When performing a population genetic analysis, it is necessary to use gene loci that are neutral to natural selection. Thus, it is common to use the third position of the codon which does not undergo amino acid substitution, and base substitution occurring in intron (which is said to comprise 90% or more of the metazoan nuclear genome). Additionally, because genetic diversity within a population for neutral gene loci is known to correspond to the extent of adaptive mutations (Hedrick, 2001), they are considered useful indicators for assessing the health of a population. After calculating indicators of population genetic diversity, such as genetic diversity and nucleotide diversity, which are foundational data for the analysis of biological community composition, examine the difference between the genotype frequency of different populations. Genetic diversity[17] , nucleotide diversity[3], *Tajima's D*[18] estimation, and mismatch distribution analysis[19] can be calculated using the population analysis software, Alrequin. Patterns of haplotype frequency distribution obtained from mismatch distribution analyses on four populations with differing backgrounds are shown in Fig.6.

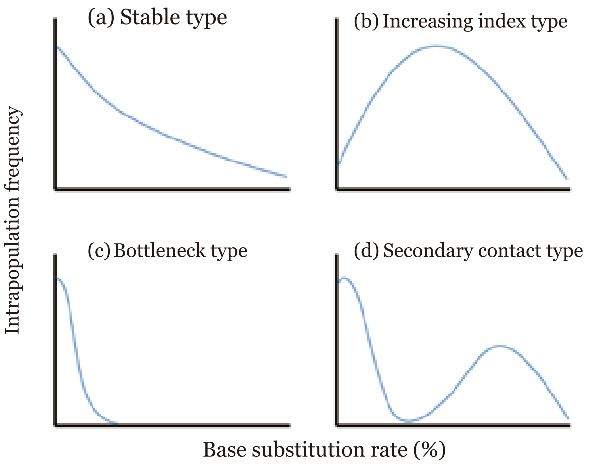


Fig.6 Pattern of frequency distribution of haplotypes in populations with different history
Note: This figure was adapted from Frankham et al. (2007).

A. Haplotype Network

Haplotype network graphs are graphs where connected lines show the relationship of all the haplotypes detected within a population (Frankham et al., 2007). Similarly to the aforementioned mismatch distribution, it is possible to extrapolate the background for populations from their network patterns. Fig.7 shows an example of a haplotype network graph. Haplotype network graphs can be created using the software, TCS.

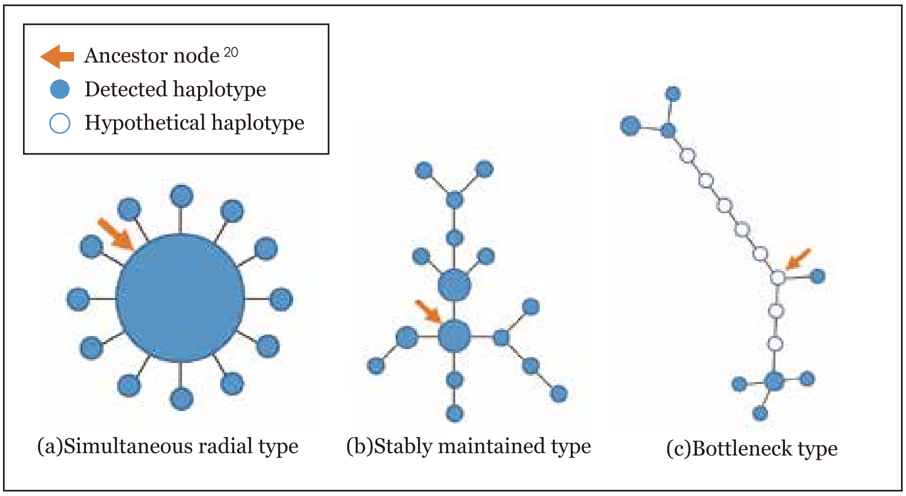


fig.7 Haplotype Network Graph Patterns

B. Analysis of Molecular Variance (AMOVA)

It is possible to objectively assess differences in the frequencies of haplotypes that comprise each population via an F statistic that uses the coefficient of inbreeding[21], or by analysis of molecular variance (AMOVA). An *F* statistic using the coefficient of inbreeding shows the proportion of genetic mutations observed between or within populations of a species, and has been used to examine the differences between populations. There are three *F* statistics: *FIS, FIT*, and *FST,* where I indicates "individual", S indicates "subpopulation" , and T indicates "total population" .

The estimation of these *F* statistics can be implemented using the Alrequin software mentioned in the previous section. AMOVA is the application of the general statistical technique analysis of variance (ANOVA) to genetic information. AMOVA is suited to testing a previously established hypothesis (for example, that obstacles on the seabed split the metapopulation into subpopulations, etc.), with the aim of testing whether the postulated structures exist, through comparison of the *F* statistics.

C. Coalescent Theory

Coalescent theory is a method for retrospectively extrapolating the genetic process that matches the common ancestor population, based on the current genetic information of two different populations. Use the software, MIGRATE-N, if using the maximum likelihood method and Bayes approach to extrapolate the individuals migrating between populations based on this theory. The number of populations for study, and the gene loci, and their base sequence length, as well as the number of individuals in each population, and their base sequence information, are necessary for this analysis. Using this analysis, it is possible to obtain for each population, an estimate for *Ne* (effective population size), and θ (integral component of the base substitution rate) and *Mab,* (the number of migrating individuals from population A to population B).

Notes

1 Mitochondrial DNA (mtDNA) is a circular DNA molecule inside the mitochondria. Excluding some exceptions, these are generally inherited maternally. Because mutations accumulate without the recombination that occurs in nuclear DNA, it is possible to grasp taxonomical phyletic relationships. Moreover, the rate of base substitution is greater than for nuclear DNA, making it suitable for capturing the genetic mutations within and between species. In genetic connectivity surveys, four genetic regions are commonly used: cytochrome b (Cyt-b), NADH dehydrogenase subunit 4 (ND4), cytochrome c oxidase subunit I (COI), and control elements (D-LOOP) (Weersing & Toonen, 2009).

2 Microsatellites are repeating sequences with units of around 2-6 bases that are scattered widely throughout a genome, and have advantages such as high diversity and codominance. Because of this, it is possible to use them to detect genetic differences between individuals and populations with a high degree of sensitivity. However, PCR primers for the amplification of regions including microsatellites, must be designed on a per target species basis. A disadvantage of microsatellites is their cost, both financially and of the time it takes to design these primers. Nonetheless, the ease of obtaining large amounts of genomic information from the proliferation of next-generation sequencers is making it possible to significantly reduce the time scales of primer design. The design of primers for microsatellites using next-generation sequencers has been applied to organisms in deep-sea areas of hydrothermal activity (Nakajima et al., 2014) and is expected to advance research into various taxa in the future.

3 SNP is an abbreviation of single nucleotide polymorphism, and denotes a DNA locus where there exist two or more single-base polymorphisms, which are present in DNA sequences with a given frequency (of 1% or more). SNPs usually have two alleles, which have the advantage of being easier to determine than microsatellites. However, since it is not possible to assess the marker error rate of SNPs through family analysis as with microsatellites, they are considered inferior to microsatellites. They have an increased costs that stem from the need for a larger number of alleles to reach a given data threshold (Guichoux et al., 2011). Recently, analysis methods that do not require genomic data, such as the MIG-seq method for decoding with next-generation sequencers following the amplification of fragments bordered by microsatellite regions (Suyama & Matsuki, 2015), are being developed. It is possible that these methods will be widely used in the population analysis of deep-sea organisms in the future.

4 An allozyme is a protein polymorphism detected via electrophoresis, which is derived from different alleles at a single gene locus. Because allozyme analysis can be applied even to organisms for which there is no prior base sequence data at all, it was often used in times when it took time and cost to acquire nucleotide sequence information. However, its use in genetic connectivity surveys is becoming less prevalent recently, given that it has become possible to easily obtain base sequence information, that ethanol fixation of the sample cannot be conducted to maintain enzyme activity, and that the level of detected polymorphisms is relatively low.

5 CO I gene is encoded in the mitochondrial DNA and is one of the genes for cytochrome c oxidase, an enzyme that controls the electron transfer system in the mitochondrial membrane. Its total length is around 1140 bp. A part of CO I gene, approximately 650 bp in length, is being used as a standard barcode region for the animal kingdom and many eukaryotes in an international project into DNA barcoding (a technique to advance the identification of species using short base sequences of specific gene regions).

6 Non-synonymous substitutions are those base substitutions occurring in gene base sequences that cause the mutation of the encoded amino acid.

7 Cryptic species are separate species that were treated as the same species due to an inability to differentiate them morphologically. Species should normally be treated as different species, however in some cases, reproductive isolation is found to have occurred through cross-breeding experiments, or the species is found to belong to different families through molecular phylogenetic analyses using gene base sequences.

8 Haplotype is an abbreviation of haploid genotype.

9 Haplotype diversity is the haploid genotype diversity exhibited by, for example, mitochondrial DNA. It is used as an indicator of genetic diversity.

10 Nucleotide diversity is a measure for observing gene diversity at the base level. Nucleotide diversity is defined as an average proportion of bases that differ between two sequences extracted arbitrarily from a single population. Where a single base substitution haplotype is found in high proportion, such as in populations that exhibit simultaneous radiation, the nucleotide diversity is known to be comparatively low, even where the level of gene diversity is high.

11 When ethanol is mixed with seawater, calcium in the seawater crystallizes, causing a white precipitate. Additionally, change the solution one time, because the permeability of tissues to ethanol is poor. For large specimens, cut tissue slices and then fix and store them.

12 It is best to conduct this work where there is access to draft chambers.

13 International DNA databases include: the DNA data bank of Japan (DDBJ; http://www.ddbj.nig.ac.jp/), the NIH genetic .http://www.ncbi.nlm.nih.gov/genbank/), and the European Molecular Biology Laboratory(EMBL; http://www.ebi.ac.uk/). Each of these is freely available and share data under the International Nucleotide Sequence Database Collaboration (INSDC) framework.Registered data is updated daily, and the number of registrations for

large sequences from metagenome analysis and registrations of whole-genome sequences using next-generation sequencers has been increasing rapidly in recent years.

14 Pseudogenes are copies of genes that have lost their original function due to reasons such as loss of bases or duplication. While pseudogenes cannot produce proteins, the sequence itself remains in the DNA, and since it resembles the original, functional gene, they are easily mistaken as the actual gene. Pseudogenes of CO I gene have also been reported.

15 Even for base sequence data registered in international base sequence databases, the species from which it was derived (the registered scientific name) is generally self-reported by the data registrant, and there is no quality control for species identification. Consequently, the scientific name is not necessarily correct for any given registered data. Confirm the kinds of taxa that are included in any output data. Even for the highest ranked hits, conduct another search using that sequence, if necessary. Additionally, pay attention to the query coverage and E-values. The query coverage value is the proportion of parts of the compared sequence that are similar. If this value is low, it is possible that it is a fragment sequence provided from a sequence meta-analysis, or has a high similarity to a portion with poor or no homology. The E-value is the expectation value that the registered sequence and query sequence (the sequence used for the search) are similar by chance, with lower values indicating that the hit was less likely to be due to chance. Consequently, where the value is zero or exceedingly close to zero, the reliability of the search result is considered to be high.

16 A monophyletic group is a phylogenetic group that consists of a single virtual common ancestor and all its offspring. These are also called clades.

17 Gene diversity is the probability that two arbitrarily chosen alleles from a single population will be different. It is a fundamental measure for the level of genetic diversity of populations.

18 *Tajima' s D* is an indicator for testing the neutrality (the presence or absence of natural selection) of a gene locus based on its allele number and frequency within a population. Additionally, sudden changes in population size can be detected from *Tajima's D*.

19 Mismatch distribution analysis shows the extent to which base substitution occurs between haplotypes for each pair of individuals of a population as a frequency distribution, and can therefore be used to extrapolate the history of the population.

20 Ancestor nodes are the haplotypes in the haplotype network graph that are thought to be the most ancestral (phylogenetically oldest).

21 The inbreeding coefficient expresses the degree of inbreeding. It is also called the fixation index, which is abbreviated as F (f) and therefor also called the F value.

References

1)Altschul, S. F., Gish W., Miller W., Myers E. W., and Lipman D. J. (1990) Basic local alignment search tool. J. Mol. Biol. 215, 403-410.

2)Baco AR, Etter RJ, Ribeiro PA, von der Heyden S, Beerli P, Kinlan BP (2016) A synthesis of genetic connectivity in deep-sea fauna and implications for marine reserve design. Molecular Ecology, 25(14): 3276-3298.

3)Carr CM, Hardy SM, Brown TM, Macdonald TA, Hebert PDN (2011) A tri-oceanic perspective: DNA barcoding reveals geographic structure and cryptic diversity in Canadian polychaetes. PLoS One, 6: e22232.

4)Costa FO, deWaard JR, Boutillier J, Ratnasinghamn S, Dooh RT, Hajibabaei M, Hebert PDN (2007) Biological identifications through DNA barcodes: the case of the Crustacea. Canadian Journal of Fisheries and Aquatic Sciences, 64: 272-295.

5)Folmer O, Black M, Hoeh W, Lutz R, Vrijenhoek R (1994) DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. Molecular Marine Biology and Biotechnology, 3(5): 294-299.

6)Frankham R, Ballou JD, Briscoe DA (2007) Introduction to Conservation Genetics. Cambridge University Press, 644pp.

7)Geller J, Meyer C, Parker M, Hawk H (2013) Redesign of PCR primers for mitochondrial cytochrome c oxidase subunit I for marine invertebrates and application in all-taxa biotic surveys. Molecular Ecology Resources, 13(5): 851-861.

8)Génio L, Johnson SB, Vrijenhoek RC, Cunha MR, Tyler PA, Kiel S, Little CTS (2008) New Record of "Bathymodiolus" Mauritanicus Cosel 2002 from the Gulf of Cadiz (NE Atlantic) Mud Volcanoes. Journal of Shellfish Research 27(1): 53-61.

9)Goodall-Copestake WP, Tarling GA, Murphy EJ (2012) On the comparison of population-level estimates of haplotype and nucleotide diversity: a case study using the gene cox1 in animals. Heredity, 109: 50-56.

10)Guichoux E, Lagache L, Wagner S, Chaumeil P, Léger P, Lepais O, Lepoittevin C, Malausa T, Revardel E, Salin F, Petit RJ. (2011) Current trends in microsatellite genotyping. Molecular Ecology Resources, 11(4): 591-611.

11)Hall B (2011) Phylogenetic Trees Made Easy: A How-To Manual. 4th ed. Sinauer Associates, Sunderland, 282pp.

12)Hedrick PW (2001) Conservation genetics: where are we now? Trends in Ecology and Evolution, 16: 629-636.

13)Hellberg ME, Burton RS, Neigel JE, Palumbi SR (2002) Genetic assessment of connectivity among marine populations. Bulletin of Marine Science, 70(1): 273-290.

14)Huang D, Meier R, Todd PA, Chou LM (2008) Slow mitochondrial COI sequence evolution at the base of the metazoan tree and its implications for DNA barcoding. Journal of Molecular Evolution, 66(2):167-74.

15)Ishibashi J, Okino K, Sunamura M(eds.) (2015) Subseafloor Biosphere Linked to Hydrothermal Systems, TAIGA Concept. Springer Japan, Tokyo, 666pp.

16)Jones WJ, Macpherson E (2007) Molecular phylogeny of the East Pacific squat lobsters of the genus Munidopsis (Decapoda Galatheidae) with the descriptions of seven new species. Journal of Crustacean Biology, 27(3): 477-501.

17)Kojima S, Segawa R, Fijiwara Y, Fujikura K, Ohta S, Hashimoto J (2001) Phylogeny of hydrothermal-vent-endemic gastropods Alviniconcha spp. from the western Pacific revealed by mitochondrial DNA sequences. The Biological Bulletin, 200(3): 298–304.

18)Luo A, Lan H, Ling C, Zhang A, Shi L, Ho SY, Zhu C. (2015) A simulation study of sample size for DNA barcoding. Ecology and Evolution, 5(24): 5869-5879.

19)Meyer CP (2003) Molecular systematics of cowries (Gastropoda: Cypraeidae) and diversification patterns in the tropics. Biological Journal of the Linnean Society, 79: 401-459.

20)Nakajima Y, Shinzato C, Khalturina M, Watanabe H, Inagaki F, Satoh N, Mitarai S (2014) Cross-Species, Amplifiable Microsatellite Markers for Neoverrucid Barnacles from Deep-Sea Hydrothermal Vents Developed Using Next-Generation Sequencing. International Journal of Molecular Sciences, 15(8): 14364-14371.

21)Riehl T, Brenke N, Brix S, Driskell A, Kaiser S, Brandt A (2014) Field and laboratory methods for DNA studies on deep?sea isopod crustaceans. Polish Polar Research, 35(2): 203-224.

22)Suyama Y, Matsuki Y (2015) MIG-seq: an effective PCR-based method for genome-wide single-nucleotide polymorphism genotyping using the next-generation sequencing platform. Scientific Reports, 5: 16963.

23)Tel-zur N, Abbo S, Myslabodski D, Mizrahi Y (1999) Modified CTAB Procedure for DNA Isolation from Epiphytic Cacti of the Genera Hylocereus and Selenicereus (Cactaceae).Plant Molecular Biology Reporter 17(3):249-254.

24)Toews DP, Brelsford A (2012) The biogeography of mitochondrial and nuclear discordance in animals. Molecular Ecology, 21: 3907-3930.

25)Walsh PS, Metzger DA, Higuchi R (1999) Chelex 100 as a Medium for Simple Extraction of DNA for PCR-Based Typing from Forensic Material. Biotechniques, 10 (4):506-513.

26)Watanabe H (2010) Indicator of biodiversity of deep-sea hydrothermal venting area specific animals (in Japanese). AQUABIOLOGY, 32(2): 150-155.

27)Watanabe H, Fujikura K, Kojima S, Miyazaki J, Fujiwara Y (2010) Japan: vent and seep in close proximity. Kiel S (ed) The Vent and Seep Biota Aspects from Microbes to Ecosystems. Topics in Geobiology 33: 379-402.

28)Watanabe H, Kojima S, Fujikura K (2010) Estimation of population dynamics of deep-sea hydrothermal vent community using genetic techniques (in Japanese). AQUABIOLOGY, 32: 561-566.

29)Weersing K, Toonen RJ (2009) Population genetics, larval dispersal, and connectivity in marine systems. Marine Ecology Progress Series, 393: 1-12.

Memo

....

....

....

....

....

....

....

....

....

....

....

....

....

....

....

....

....

....

....

....

....

Memo

....

....

....

....

....

....

....

....

....

....

....

....

....

....

....

....

....

....

....

....

....

Memo

SIP Protocol Series

SIP Protocol No.1

Application of environmental metagenomic analyses for environmental impact assessments

SIP Protocol No.2

Genetic Connectivity Survey Manuals

SIP Protocol No.3

A rapid method to analyze meiofaunal assemblages using an Imaging Flow Cytometer

SIP Protocol No.4

Acquisition of Long-Term Monitoring Images Near the Deep Seafloor by Edokko Mark I

SIP Protocol No.5

Microstructure Measurements aroud Deep Sea floor

-Direct Measurements of the Deep Sea Turbulence flow-

Contact

JAMSTEC (Japan Agency for Marine-Earth Science and Technology)
2-15, Natsushima-cho, Yokosuka-city, Kanagawa, 237-0061, Japan

TEL +81-46-866-3811 FAX +81-46-867-9755
E-mail sip-pc@jamstec.go.jp

ISBN 978-4-901833-28-8

Version.1
May.1.2017

JAMSTEC 国立研究開発法人 海洋研究開発機構
JAPAN AGENCY FOR MARINE-EARTH SCIENCE AND TECHNOLOGY