Fish Genome Sequencing: Status, Challenges and Opportunities Ahead

Authors

N. S. Nagpure Arunima Kumar Verma Shivesh Pratap Singh* U. K. Sarkar S. P. Singh

Molecular Biology and Biotechnology Division National Bureau of Fish Genetic Resources Lucknow Uttar Pradesh

*Department of Zoology Govt. PG (Autonomous) College Satna Madhya Pradesh

Focal Points at a Glance

Fish genome sequencing would cause a genomic revolution in fisheries domain which would have an impact on aquaculture enhancement and fish production. This development will pave way for future management of fish genetic resources as a basis for promoting new cost effective and sustainable aquaculture practices in the country.

Abstract

Fishes constitute a highly diversified group of aquatic vertebrates that can play a prominent role in the genomic research and also provide penetrating insight into vertebrate genome evolution. The intrinsic characters of fish genome, its small size, compactness, lack of junk DNA and rounds of tetraploidisation make it an important material for providing answers to many biological questions. Thus, the fish genome sequencing projects became an important area for genomic research. The vast array of data generated from fish genome sequencing projects as that of zebrafish, fugu, tetradon, medaka and stickleback have not only helped in understanding the fish genome organisation but also provided insight to have an understanding of gene evolution and functioning in humans and other vertebrates. This paper reviews the current status of fish genome data and highlights further scope of their utilisation in fish biodiversity research and aquaculture.

Introduction

Fishes are an economically important group of aquatic vertebrates characterised by high distribution, vast diversity in morphology, behaviours and astonishing adaptability. All these phenotypic characters have got embedded into the small, compact and plastic genome all through the ages of evolution. They are potentially a cheap source of protein, rich in nutrients, minerals (iodine and calcium) and vitamins (A and D), contain good concentration of polyunsaturated fatty acids that reduce the risk of cardiovascular diseases and anti-oxidants like selenium and co-enzymes Q-10 that help to fight against diseases. Fishes serve as one of the most important as well as an economical source required for fulfilling the unending food demands of our ever increasing population. Fishes are also an important stress relieving commodity in the form of ornamental fishes and are recognised as a powerful income and employment generator across the world. In India, over 14 million farmers are earning their livelihood from the fisheries sector. With an annual fish production of over 7.6 million tonnes, the sector accounts

for a turnover of over Rs. 300 billion, contributing to over 1 percent of total GDP and five per cent of the agricultural GDP. Besides the export of food fishes, the ornamental fish export has also increased in value from Rs 3.2 crore in 2001-02 to Rs 5.6 crore in 2006-07.

As aquaculture continues to contribute significantly to food availability and income generation of people, there is the requirement of fishes with better genetic constitution as a priority in order to contribute sustainable livelihoods and enhanced social wellbeing. The genetic enhancement of fish species can not only lead to better health of fishes but also amplify the revenue generation and economic yields of our country manyfold. Within last two decades, the fisheries research has been elevated from the level of traditional aquaculture techniques to analytical genetic technologies. The traditional aquaculture techniques employed conventional approaches of breeding and selection, while the genetic technologies now employed use of molecular markers, molecular breeding, genome mapping, microarrays analysis and single nucleotide polymorphisms (SNP) analysis that have shown more promising results in comparatively lesser time. The application of genetic enhancement technology in fishes by development of transgenic fishes (as salmon, trout, tilapia, catfish) and oysters has also led to engineering of traits such as growth, disease resistance and temperature tolerance that contribute to intensification and elevation of aquaculture standards across the globe. In the midst of the application of all these technologies, the Human Genome Project has also boosted research and facilitated multi-directional avenues into genomic studies of numerous organisms as bacteria *Haemophilus influenza* and *E*. coli, yeast Saccharomyces cerevisiae, fruitfly Drosophila melanogaster, silkworm Bombyx mori, nematode Caenorhabditis briggsae and Caenorhabditis elegans, and fishes, elephant shark fish Callorhinchus milii, fugu Takiugu rubripes, tetraodon Tetraodon nigroviridis etc. The onset of whole genome sequencing projects began with shotgun sequencing methodology which was a time consuming process. However, it was replaced recently by an extremely rapid and robust "next generation" sequencing technology that could generate massive amount of genomic and proteomic data in short time. The enormous amount of data generated from whole genome sequencing of the fish species such as Zebrafish, Fugu, tetraodon, and medaka have helped to understand the genomic composition, gene constitution, trait identification and gene expression in fishes. Recently, application of bioinformatics tools and techniques has also escalated fisheries research and the computational analysis of the available data has given fruitful results in a short span of time.

India is still in a very infant stage in respect of fish bioinformatics research. Therefore, the level of Indian fisheries research on the subject needs to be elevated by performing data analysis of available genomic data as well as by initiating genome sequencing projects for Indian fishes. In the context of the present requirement, the Indian Council of Agricultural Research (ICAR) has initiated a project funded by World Bank under National Agricultural Innovation Project (NAIP) entitled 'Establishment of National Agricultural Bioinformatics Grid in ICAR'. The domain institute of National Bureau of Fish Genetic Resources (NBFGR) is working in bioinformatics as part of its fisheries research. The objectives of the project are development of agricultural

bioinformatics grid which stores compiled data about molecular information of the fishes, creation of local databases and bioinformatics data warehouse (BinDW) for fish genomic, transcriptomic and proteomic resources across species and create and promote inter-disciplinary research groups with focus on fish computational research. Two databases have been designed under the project; one for fish barcode data (Fish Barcode Information System) and other for chromosomal information on fishes (Fish Karyome). The biggest challenge that Indian fisheries sector is facing today is the deficit of ample number of sequenced whole genomes of Indian fishes. Thus, whole genome sequencing of Indian fishes has now become an imperative requirement so that exhaustive data generated from analysed sequencing can be computationally for in-depth knowledge of molecular mechanism involved in

Apart from this, there is an urgent need to amalgamate the modern biotechnology and bioinformatics techniques with the traditional aquaculture practices in order to get better results in relation to fish health, and its improved productivity and its disease resistant varieties. This contribution reviews the work done and genomic data generated as part of the massive fish genome sequencing projects, their applications and further scope of their utilisation in order to benefit and improve the aquaculture technologies and fisheries research.

Fish genomics research

The small size and compact nature of genome having high gene density and scarce amount of junk DNA in it has made fish a good material for genomic research. Though the length of fish genome is smaller in comparison to other vertebrates, yet it provides a similar repertoire of gene constitution like other vertebrates. Hence, it is considered as a good model for understanding comparative vertebrate genome biology. Besides this, there is a better chance of finding an exon in fishes rather than in other vertebrates, especially in mammals, by sequencing the same quantity of DNA. Thus the small and high quality genome of fish serves to be a good candidate for rapid, economical and highly informative genome sequencing. The sequencing of fish genome is important for determining the genetic information present in the genomes as

well as deducing relationships between various genes and providing insight into fish genome organisation and its evolution. The genome projects can illuminate genes responsible for the traits such as growth, reproductive potential, stress tolerance, and disease resistance in fishes and at the same time it can make a lucid picture of relationship of fish with its parasites. The pathogenicity in fishes can also be better understood by comparison of fish and parasite genome that might facilitate in understanding the disease mechanism in fishes at molecular level. The major fish genome sequencing projects undertaken globally were in respect of zebrafish (Danio rerio), fugu (Takifugu rubripes), medaka (Oryzias latipes), tetraodon (Tetraodon nigroviridis) and elephant shark (Callorhinchus milii). The ongoing genome sequencing projects being carried out are of Atlantic salmon (Salmo salar), Common carp (Cyprinus carpio) and some cichlid species.

Global Fish Genome Sequencing Programmes

1. Zebrafish (Danio rerio) Genome Project: The zebrafish genome project was initiated in 2000 by collaboration between the Wellcome Trust Sanger Institute, United Kingdom and the Zebrafish community in which the genome of 'Tuebingen' strain of zebrafish was sequenced. Zebrafish is a tropical freshwater fish (Fig.1) belonging to the minnow family (Cyprinidae) of order Cypriniformes which serves as a popular aguarium fish and is utilised in research for experimental studies in oncology, toxicology, reproductive, developmental biology regenerative medicine, evolutionary theory and many other fields. Zv9 is the latest integrated whole genome shotgun assembly of the zebrafish genome and it is presently being maintained by Genome Reference Consortium (GRC) at NCBI, USA. The zebra fish genome has been sequenced and analysed at the Sanger Institute and its automatic annotation is available at Ensembl (European Bioinformatics Institute, Europe), while the manual annotation is provided by the Human and Vertebrate Analysis and Annotation (HAVANA) group which is released at regular intervals into the Vertebrate Genome Annotation (Vega). Various genomic resources from which Zebrafish genomic information can be retrieved are: Ensembl Zebrafish browser (Curwen et al., 2004), Vega Zebrafish browser (Wilming et al., 2008), ZFIN (zebrafish model organism database) (Bradford et al., 2010), FishMap (Bhartiya et al; 2010), @

University of California Santa Cruz (UCSC) Genome Browser database (Fujita et al., 2011), NCBI Map Viewer. The genome is made up of 25 chromosomes that contain 18,572 known protein coding genes; 192 pseudogenes; 56 immunoglobin genes; 4,444 RNA genes; and 52,873 gene transcripts, according to Ensembl database (refer to http:// asia.ensembl.org/Danio_rerio/Location/ Genome). The annotated genes comprise genes associated with human obesity, major histocompatability complex genes, immunoglobulin genes, aquaglyceroporins for water, glycerol, and trivalent arsenicals transport in zebrafish.

In India, The Institute of Genomics and Integrative Biology (IGIB), New Delhi undertook genome sequencing of wildstrain zebrafish that generated 1,700 million base pairs data having 20x coverage in a period of two months using 'Next Generation' sequencing platform. The genome was sequenced by the Solexa / Illumina sequencing technology. The database created by IGIB for the display of annotated genomic data of zebrafish wild strain is 'FishMap'/which provides facility for submissions of genome-scale data and it currently hosts over 70 datasets (tracks) pertaining to genomescale data on zebrafish.

2. Fugu (Takifugu rubripes) Genome Project: Another important model fish is fugu (Fig.2) which belongs to family Tetraodontidae. It exhibits presence of high quality genome due to small introns and intergenic sequences which make genomic sequencing of fugu a relatively fast and inexpensive process. The Fugu Genome Project was initiated in 1989 by Sydney Brenner and his colleagues Greg Elgar, Sam Aparicio, and Byrappa Venkatesh. In 2000, the International Fugu Genome Consortium was formed, headed by the Joint Genome Institute. (JGI) California, Institute of Molecular and Cell Biology (IMCB), Singapore, and the Human Genome Mapping Resource Centre (HGMP), Cambridge which sequenced and assembled fugu genome (2001) in association with Celera Genomics using the whole genome shotgun method. The genome of fugu can be accessed from IMCB Fugu genome project, University of California, Santa Cruz (UCSC) Genome Browser database, Ensembl Fugu browser. Fugu has the smallest genome size among any vertebrate organisms, about one eighth the size of the human genome (3,000 Mb) and consists of 22 chromosomes. The current version of genome is fifth genome assembly available at IMCB

Browser. It shows 1,138 protein-coding genes, 18,093 novel protein-coding genes, 593 RNA genes, 121 RNA pseudogenes (refer to http://ensembl.fugu-sg.org/Fugu_rubripes_v5/index.html). The annotations include claudins family genes that are associated with tight junctions of epithelial and endothelial cells

3. Medaka (Oryzias latipes) Genome Project: Medaka is a small fish 2-4 cm long (fig.3), commonly found in rice paddies of Southeast Asia and is popular as an aquarium fish because of its hardiness and pleasant colouration. It is extensively used in many areas of biological research, because of short period and gestation prolific reproductivity. The researchers have established sexual behaviour, genetic inheritance of colouration, spawning habits, feeding, pathology, embryological development, and ecological adjustment in the fish. The relatively small genome, short generation time of 7 week and ability to grow in a broad temperature range (6-40°C), has made medaka an attractive model for genomic research. It is a member of sub-family Oryziinae and its genome was sequenced in late 2002 as a collaborative effort of three core laboratories led by H. Takeda (University of Tokyo, Japan), S. Morishita (University of Tokyo, Japan) and Y. Kohara (National Institute of Genetics, Japan). The genome of Medaka can be accessed from National BioResource Project Medaka (NBRP) (Sasado et al., 2010), University of Tokyo Genome Browser (UTGB), University of California, Santa Cruz (UCSC) Genome Browser, Ensembl Genome Browser. The genome comprises 24 chromosomes containing approximately 700 megabases and predicted to have 20,141 genes. The genome statistics depicted at Ensembl show 1,631 known protein-coding genes; 735 RNA genes; 2,20,038 gene exons; 25,397 gene transcripts (refer to http:// asia.ensembl.org/Oryzias_latipes/ Location/Genome).

4. Tetraodon (*Tetraodon nigroviridis*) Genome Project: Tetraodon is a small tropical fish (Fig.4) found across Southeast Asia in coastal freshwater and brackishwater habitats. It has been sequenced in collaboration among Genoscope, France and the Broad Institute (MIT), United States, which was supported by the *Consortium National de Recherche en Genomique* and the National Human Genome Research Institute (NHGRI), Maryland. Twenty-one available chromosomes having 385 Mb constitute a small and compact genome

having high gene density. It is eight times more compact than that of a human genome, mostly because intergenic and intronic sequences are reduced in size compared to other vertebrate genomes. The annotation also includes 87 manually curated structures of a number of Hox and Cytokine genes. This species is 20-30 million years distant from T.rubripes, a marine pufferfish from the same family. The genome of Tetraodon can be accessed from T.nigroviridis database (Broad MIT), Ensembl (Tetraodon), Tetraodon genome browser (Genoscope), UCSC Genome Browser database. Ensembl database shows Tetraodon genome statistics as: 1,794 known protein-coding genes, 3,657 novel protein-coding genes, 147 pseudogenes, 813 RNA genes, 2,21,513 gene exons, and 24,078 gene transcripts (refer to http://asia.ensembl.org/ Tetraodon_nigroviridis/Location/ Genome).

5. Elephant Shark (Callorhinchus milii) Genome Project: The elephant shark is a cartilaginous fish (Chondrichthyes) belonging to the subclass Holocephali (Chimaera). It exhibits presence of a model cartilaginous fish genome because of its relatively small genome size (910 Mb long) which is the smallest among all the cartilaginous fishes and one-third the size of the human genome. Its genome is being utilised as a reference genome for understanding the origin and evolution of vertebrates. The gene order (synteny) is more similar between human and elephant shark genomes than between human and teleost fish genomes, even though humans are more closely related to teleost fishes than to the elephant shark. The Elephant Shark Genome Project was launched by the Institute of Molecular and Cell Biology, Singapore, and the Venter Institute. The annotated genome shows presence of Hox genes (Ravi et al., 2009), protocadherin genes and parathyroid hormone gene family (Liu et al., 2010). The putative four Hox clusters indicate that, unlike teleost fish genomes, the elephant shark genome has not experienced additional whole-genome duplication.

Other Genome Sequencing Projects

At present, the Broad Institute, (a premier institute undertaking large number of genome sequencing projects). Cambridge is generating a high quality draft from a Nile tilapia, *Oreochromis niloticus* as well as 4 additional cichlid species representing East African Lake lineages: *Astatotilapia burtoni, Malawi*

(Maylandia) zebra, Pundamila nyererei and Neolamprologus brichardi. Cichlid genome has been serving as an important source for vertebrate evolution as well as addressing important questions of sex determination, behavioural, immunological, and toxicological aspects of tilapia research. The genome research in tilapia can take a step ahead to study social dominance, territoriality, sexual selection, feeding behaviour, and also contribute to the maintenance and improvement of tilapia as an important food source across developing countries.

Very recently, the Centre of Applied Genetics, Heilongjiang Aquaculture Institute and Beijing Institute of Genomics, China, completed the genome sequencing of Cyprinus carpio. The common carp genome has 100 diploid chromosomes with physical size approximately 1,700 Mbp using the whole-genome shotgun sequencing technology. The sequencing generated an assembled scaffold N50 with a length of 404 kb, as well as over 70,000 BAC end sequences, 1.66 million ESTs and a physical map of over 80,000 BAC clones. The draft map so far covers 96.5% BAC end sequences, 96.2% ESTs and 95% physical map of the total carp genome.

The genome sequencing and assembly of freshwater threespine stickleback (Gasterosteus aculeatus) has been accomplished by the Broad Institute, with cDNAs and ESTs being provided by the Stanford Human Genome Center. The assembly has been sequenced by wholegenome shotgun sequencing with a base coverage of approximately 11x. The stickleback genome is approximately 460 Mb in length and comprises 22 pairs of chromosomes pinpointing the important genes for adaptive traits.

Besides these genome projects, other fishes such as Atlantic salmon, and cichlids have also been sequenced. Atlantic salmon genome is fairly similar to those of warm-blooded vertebrates with respect to size and overall base composition, as it seems to be devoid of isochore structures like other coldwater fish genomes. European subspecies (S. salar europensis), with 29 pairs of chromosomes was sequenced (Hartley, 1987) and approximately 2,495,257 ESTs, 81,000 contigs (i.e. a set of overlapping segments of DNA), 33,709 UniGenes had been identified. The EST databases provide a rich source of material for identifying genetic markers, such as microsatellites and single nucleotide polymorphisms (SNPs).

Opportunities and Challenges

All these major genome projects have generated a vast array of genomic and proteomic data and this baseline data are being utilised as a platform for further research. Genomics draw information from the possible combination of four nucleotides (A, T, G, C) and converts it into biologically relevant information with the help of bioinformatics. The Indian fisheries sector still needs to aggravate its potential by generating a lot of raw genomic data from Indian fishes. Completely assembled genomic sequences and their annotations would provide a clear view of genome organisation of Indian fishes as well as generate fish specific information. The specific genomic peculiarities and their association with traits such as very small body size of zebrafish, large body size and mass of major carps, very high egg laying capacities of carps, stress tolerance in murrels and catfishes can be understood after sequencing their genomes and a gene constructs for the same features can be generated as well as inserted into other fish genomes. Successful introduction of the gene constructs can cause a boom in fisheries research. Thus, sequencing of these genomes would help to identify and even manipulate specific genes for useful traits such as genes for growth rate, colour, disease resistance, cold tolerance and saline tolerance. The data generated at genomic level can be extensively utilised in designing of SNP Chips as well as Expression arrays for detection of traits of economic importance. Association of the SNPs (single nucleotide polymorphism) with particular traits can be outlined on genomic scale and can be utilised for selection of fish strain with superior quality and selective breeding purposes.

Conclusion

As the number of sequenced genomes of fish species will increase, the genomic, transcriptomic, proteomic information would increase exponentially and would serve as a source of potential baseline data for fish computational research. Maximum analysis of such raw data by in silico or computational methodologies can help to interpret gene combinations with superior traits such as body weight, growth rate, and disease resistance. Thus, knowledge generated by genome sequencing projects and use of computational approaches would cause a genomic revolution in fisheries domain which would in turn have considerable impact on aquaculture enhancement and fish production. Enhanced fish production, both in quantity as well as quality will open pathways for the future management of fish genetic

resources in order to develop new cost effective and sustainable aquaculture practices in India and other Asian countries. *Figures on page 63*.

Bibliography

BHARTIYA, D., MAINI, J., SHARMA, M., JOSHI, P., LADDHA, S.V., JALALI, S., PATOWARY, A., PURKANTI, R., LALWANI, M., SINGH, A.R., CHAUHAN, R., SINGH, N., BHARDWAJ, A., SCARIA, V., and SIVASUBBU, S. (2010): FishMap Zv8 updatea genomic regulatory map of zebrafish. *Zebrafish.* 7(2):179-80.

BRADFORD, Y., CONLIN, T., DUNN, N., FASHENA, D., FRAZER, K., HOWE, D.G., KNIGHT, J., MANI, P., MARTIN, R., MOXON, S.A., PADDOCK, H., PICH, C., RAMACHANDRAN, S., RUEF, B.J., RUZICKA, L., BAUER, SCHAPER, H., SCHAPER, K., SHAO, X., SINGER, A., SPRAGUE, J., SPRUNGER, B., VAN, SLYKE, C., and WESTERFIELD, M. (2011): ZFIN: enhancements and updates to the Zebrafish Model Organism Database. *Nucleic Acids Res.* 39: D822-9.

CURWEN, V., EYRAS, E., ANDREWS, T.D., CLARKE, L., MONGIN, E., SEARLE, S.M., and CLAMP, M. (2004): The Ensembl automatic gene annotation system. *Genome Res.* 14(5): 942-50.

FUJITA, P.A., RHEAD, B., ZWEIG, A.S., HINRICHS, A.S., KAROLCHIK, D., CLINE, M.S., GOLDMAN, M., BARBER, G.P., CLAWSON, H., COELHO, A., DIEKHANS, M., DRESZER, T.R., GIARDINE, B.M., HARTE, R.A., HILLMAN-JACKSON, J., HSU, F., KIRKUP, V., KUHN, R.M., LEARNED, K., LI, C.H., MEYER, L.R., POHL, A., RANEY, B.J., ROSENBLOOM, K.R., SMITH, K.E., HAUSSLER, D., and KENT, W.J. (2011): The UCSC Genome Browser database: update 2011. Nucleic Acids Res. 39: D876-82.

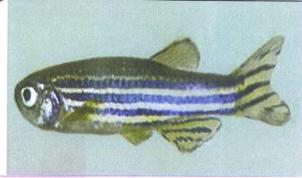
HARTLEY, S.E. (1987): The chromosomes of salmonid fishes. *Biol Rev.* **62**:197-214.

WILMING, L. G., GILBERT, J. G. R., HOWE, K., TREVANION S., HUBBARD, T. and HARROW, J. L (2008): The vertebrate genome annotation (Vega) database. *Nucleic Acids Res.* 36 (Database issue): D753–D760.

LIU, Y., IBRAHIM, A.S., TAY, B.H., RICHARDSON, S.J., BELL, J., WALKER, T.I., BRENNER, S., VENKATESH, B., and DANKS, J.A. (2010): Parathyroid hormone gene family in a cartilaginous fish, the elephant shark (*Callorhinchus milii*). *J Bone Miner Res.* 25(12): 2613-23.

RAVI, V., LAM, K., TAY, B.H., TAY, A., BRENNER, S., and VENKATESH, B. (2009): Elephant shark (*Callorhinchus milii*) provides insights into the evolution of Hox gene clusters in gnathostomes. *Proc Natl Acad Sci U S A*. 106 (38):16327-32.

SASADO, T., TANAKA, M., KOBAYASHI, K., SATO, T., SAKAIZUMI, M., and NARUSE, K. (2010): The National BioResource Project Medaka (NBRP Medaka): an integrated bioresource for biological and biomedical sciences. *Exp Anim.* 59(1):13-23.



Zebrafish (source: www.focusonnature.be)



Fugu (source: www.lbl.gov)



Medaka (source: www.pwri.go.jp)



Tetraodon (source: www.aquahobby.com)





International Conference SUSTAINABLE ORNAMEN

WAY FORWARD March 23-25, 2012 KOCHI KERALA **Gokulam Convention Centre**

Conference Sessions

- Sustainability Issues in Ornamental Fisheries
- 2. Way forward for Sustainable Ornamental Fisheries

Training Sessions

- 2. Re Circulatory and Raceway Systems for Ornamental Fish

- 6. Total Quality Management / BMP

Speakers and Trainers from India, Norway, USA,

* Early Bird Registration Fee (Last Date 28th February, 2012) Rs 2000 (Indian Delegates) & Rs 1000 (Indian Students)

SOFI-WF: 2012 School of Industrial Fisheries, Cochin University of Science and Technology, Lake Side Campus, Fine Arts Avenue, Köchi, Kerala T : +91 484 2354711, F:+91 484 2365952, M : +91 9895070310, +91 9744344477, e : sofi@ornamentalfishes.org, w : www.ornamentalfishes.org

Visit our Website for more details- www.ornamentalfishes.org

J.V.H. Dixitulu National Award for Outstanding work

The Association of Aquaculturists (based in Central Institute of Freshwater Aquaculture, Bhubaneswar), Marine Biological Association of India (MBAI - based in Central Marine Fisheries Research Institute, Kochi) and the Inland Fisheries Society of India (IFSI - based in Central Inland Fisheries Research Institute, Barrackpore), have instituted in 2010, Mr. J.V.H. Dixitulu National Award to recognise outstanding contributions in the field of fisheries extension/ communication in India. The first National Award after its institution by AFSIB was given during the triennial Indian Fisheries Forum organised by the AFSIB in December 2011.

This was presented to Mr. V. Edwin Joseph, Officer-in-Charge, Library and Documentation and the Librarian of CMFRI by Mr. J. V. H. Dixitulu during the 9th Indian Fisheries Forum at Chennai. The award carried a memento, citation and cash award of Rs.15,000/-.



Mr. V. Edwin Joseph receiving the First Mr. J.V.H. Dixitulu National Award for Outstanding work on Fisheries Communication from Mr. Dixitulu

