METAGENOMİCS SEQUENCİNG METHODOLOGİES: REVİEW

Pritee Chunarkar Patil

Department of Bioinformatics, Bharati Vidyapeeth Deemed University (BVDU),
RGITBT, Pune, INDIA.

preetichunarkar@gmail.com, preeti.chunarkar@bharatividyapeeth.edu

ABSTRACT

Metagenomics can be done by opting different approaches for different samples. Even for the same sample, DNA extraction procedures can be different. Due to these approaches, the concentration and purity of the DNA could be variable. Due to the availability of the different procedures for the metagenomics, the standardization of a specific method cannot be implied directly. There is no specific protocol available which can be followed as a standard protocol for the metagenomics data analysis. Every research problem has been adapted different techniques to solve it. This new branch of science put forth the direct method to gain maximum knowledge of unknown microbial communities. But every time there are chances of getting new diversity as per the sample collection protocol, DNA extraction methodologies, sequencing and data analysis approaches. Metagenomics prove to be a complex science as per the interdisciplinary studies involved. Sanger Sequencing to Big Data Analytics, step by step advancement in the strategies differentiate setup scale and end product analysis for DNA sequencing methodologies. Present data will try to review the available metagenomics methodologies with their applications and drawbacks.

Keywords: Metagenomics, Methodologies, DNA, Protocol

INTRODUCTION

Metagenomics is a proposed field for the identification of the uncultivable micro-organisms from the environmental samples. Metagenomics is becoming familiar day by day to the all respective classes (students, academicians, scientists, researchers etc.) of people. It has put forth the new era considered to be 99% of uncultured micro-organisms (Amman etal., 1995; Schloss and Handelsman, 2003). The word was first coined by Covacci and Rondon (Covacci et al., 1997 andRondon et al., 2000). Conventional methods of microbiology are insufficient for the information flow of uncultivable microorganisms(Hugenholz et al, 1998).

Metagenomics can be applied for any environmental sample like soil (Alvarez et al., 2013) , sea water (Rosario et al., 2009), pond water (Kapardar, 2010), fresh water, sediments (Wang et al., 2012) , saline water (Santos etal., 2010), lake water (Deagle et al., 2012), sewage water (Singh etal., 2010), human guts (Blottière et al., 2013), reclaimed water(Rosario et al., 2009) etc. Different approaches have been undertaken by different group of scientists to reveal the metagenomics library.Few metagenomics approaches will be discussed in this review.

DIFFERENT APPROACHES FOR METAGENOMICS

The basic approach is same for all methods as:

Leena and Luna International, Chikusei, Japan.
(株) リナアンドルナインターナショナル,筑西市,日本

ISSN: 2186-8476, ISSN: 2186-8468 Print
www.ajsc. leena-luna.co.jp

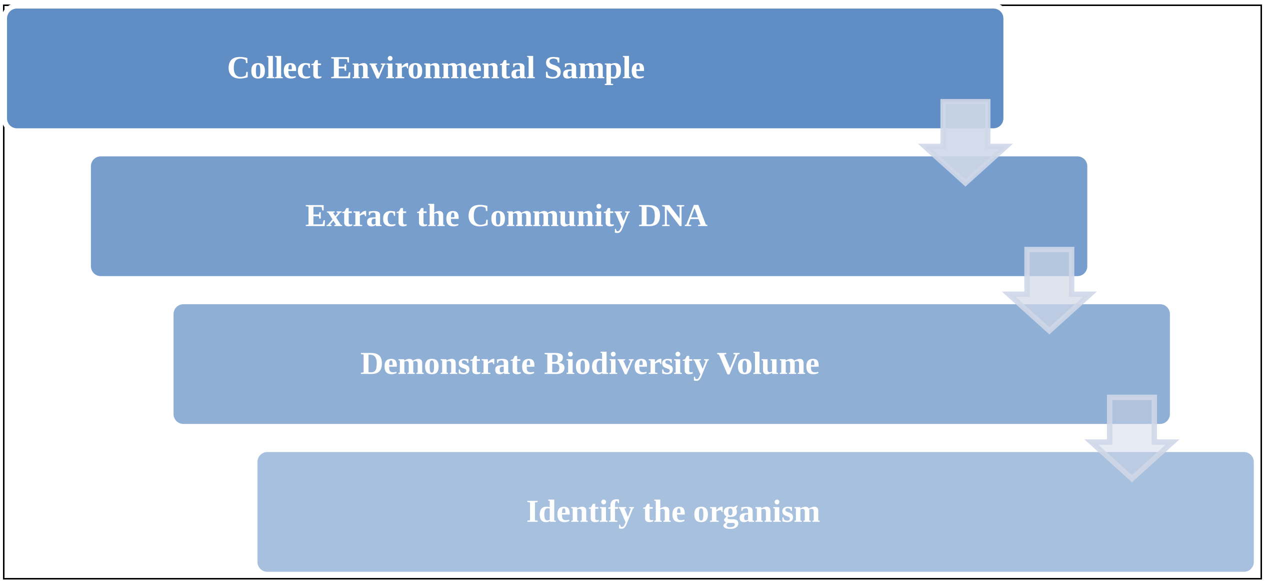
Collect Environmental Sample
Extract the Community DNA
Demonstrate Biodiversity Volume
Identify the organism

Figure 1. Metagenomics workflow 1

The community DNA can be handled by two methods, wet lab approach andhigh throughput sequencing approach.

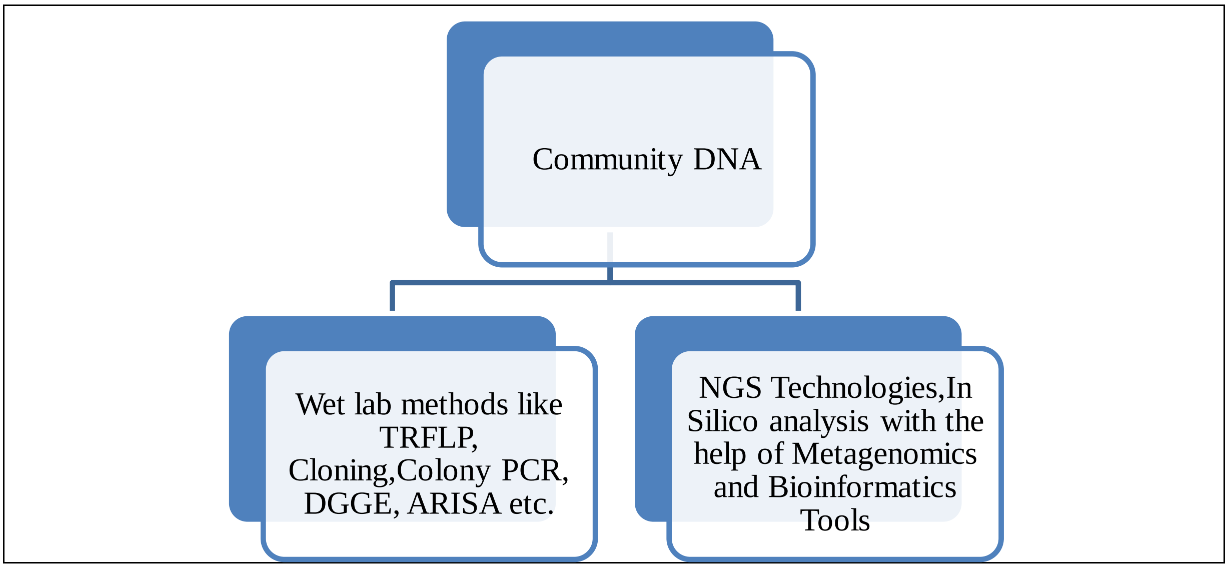
Community DNA
Wet lab methods like TRFLP, Cloning,Colony PCR, DGGE, ARISA etc.
NGS Technologies,In Silico analysis with the help of Metagenomics and Bioinformatics Tools

Figure 2. Metagenomics workflow 2

TRFLP: Terminal Restriction Fragment Length polymorphism, DGGE:Denaturing gradient gel electrophoresis, ARISA:An automated method of ribosomal intergenic spacer analysis, NGS: Next Generation sequencing

Stepwise Description

- ***Collection of the sample:*** Different techniques could be adopted for this as per the sample type. Precaution should be taken for the contamination purpose. So sterilized equipments should be used.
- ***Extraction of the DNA:*** For different samples, different DNA extraction procedures could be opted. The protocol can be standardized by the researchers or ready available kits can be used for fulfilling the goal.

- ***Community DNA Processing:*** Community DNA could be processed for the sequencing (Either conventionally or with advanced technologies). It totally depends on the researcher. Different methodologies for it have been shown in Fig.2.
- ***Biodiversity Identification:*** It could be again done by the conventional 16S rRNA gene (for bacteria)/18S rRNA gene (for fungi) amplification. Then through Advanced methods like NGS, the raw data could be analyzed. Annotation of the data could be handled with the help of Bioinformatics Tools.

BRIEF METAGENOMICS METHODOLOGIES

1. **Cloning Method -** Collect the community DNA. Then clone the DNA into vectors and prepare the metagenomics library. Identify the sequences with 16S rRNA genes/18S rRNA genes (Pace et al., 1985, Schmidt et al., 1991).

   Advantages **–** The Method israpid, sensitive and robust.

   Disadvantages –

   a. Prior target sequence information is necessary

   b. Target sequence should not be too short in size,

   c. It gives limiting amounts of product and infidelity of DNA replication (Strachan and Read, 1999).

2. **Phylogenetic Marker Method** - Method requires sequencing the DNA and identifying the phylogenetic marker gene to place the organism in taxonomic range (Barns et al., 1999, Buckley et al., 2003).

   Advantages **-** Yield high accuracy for small sequences.

   Disadvantages**-**Only small fragment of the gene can be characterized **(**Teeling et al., 2004, McHardy et al., 2007**).**

3. **High Throughput Sequencing (HTS)**- It consists of three basic steps, DNA Extraction, DNA binding (to different platforms in different technologies) and sequencing (Naiara and Aransay, 2012, Martin and Janet, 2010). The methodology started with large scale sequencing projects like GOS (Global Ocean Sampling) by Venter et al.(2004). This large scale sequencing can be possible due to the time efficient and cheaper NGS technologies like ABI/Life Technologies SOLiD; Helicos HeliScope; Illumina Genome Analyzer; Roche/454 GS FLX Titanium; Sanger capillary sequencing.

Roche/454 GS FLX Titanium

It was first commercially available HTS technology. It was first discovered by 454 Life Sciences and Roche owned it later.

Principal

300-800 bp DNA fragments formed randomly with shearing activity. Two different adapters (+ carries biotin group/- carries nonbiotin group) are ligated to these fragment ends which are attached to single streptavidin beads. Fragments containing biotin are remains attached to beads while nonbiotinylated fragments will be washed out. With the help of limiting dilutions, one to one ratio of the bead and fragments formed. Through emulsion PCR,amplification of the DNA achieved with the help of Picotitre plate platform. GS FLX utilizes Pyrosequencing for the final sequence analysis (ATP>> ADP +PPi >> Light Emits >> Calculate the intensity of Light >> Base Pair Sequence).

Advantages

a. It is first HTS platform.
b. Expected average read length is 400nts.
c. It determines 750mb/day with price of about 20$/Mb.

Disadvantages

a. Average substitution error rate is 10^{3}-10^{4} nt which is higher than Sanger sequencing.
b. Error rate increases due to reduction in enzyme efficiency (Margulies et al., 2005, Quinlan et al., 2008).

Illumina Genome Analyzer

It was first short read technology. It was developed by Soxela and later taken up by Illumina (Bentaley et al., 2006, 2008).

Principal

Special sequencing library created like Roche which will be amplified and immobilized withtwo different adapters for both ends. End specific PCR primers functionalized flow cell surface for amplification. This process is called as Bridge-PCR. Sequencing will be proceeds by using fluorophore labeled reversible terminator and image processing.

Advantages

a. It uses reversible terminator sequencing.
b. It determines 5000Mb/day with price of about 0.50$/Mb.
c. It has lower price per mega base in comparisonwith Roche.

Disadvantages

a. It has higher average error rate of about 10^{-2}-10^{-3}.
b. Crystal, lint and dust particles may be identify as clusters by image analysis.
c. Reversible terminator sequencing increase error rate with increase in the position in determined sequence due to bi –directional phasing. This also results in incorrect termination (Kircher et al., 2009, Dohm et al., 2008).

ABI/Life Technologies SOLiD

It has been developed by Applied Biosystems. It was third HTS technology launched in the market.

Principal

It follows the principal of sequencing-by-ligation. SS copies if library molecules attached to two different adaptors (P1 and P2) are immobilization on paramagnetic beads. Emulsion PCR uses for amplification. Ligation based Sequencing uses five different sequencing primers(X, X-1, X-2, X-3 and X-4) with four different 5' fluorescently labeled DNA octamers with dinucleotide complement recognition core to hybridize with templates. So 35 base template sequenced twice to improve accuracy. (Smith et al., 2010,Valouev et al., 2008).

Advantages

a. Typical read length is between 25 to 75 nt.

b. It has better sequencing accuracy.

c. It determines ~ 5000Mb/day and Price per million bp is ~ 0.50$/Mb (Myllykangas et al., 2012).

Disadvantages

a. Dust, chemical crystal, lint particles could be misidentified as clusters

b. It has higher background error rate, so higher average error rate (compared to Illumina) (Dimalanta et al., 2009).

Helicos Heliscope

It is the first sequencer to sequence individual molecule. It is very accurate but costlier. Technology applied in this sequencing is called as asynchronous virtual terminator chemistry.

Principal

At first DS DNA gets fragmented, melted and poly adenylated with last added element as fluorescently labeled Adenine which remains attached to the flowcell. Immobilized poly-T oligonucleotides washed over with this SS DNA for hybridization. After coordinate's identification, 3' adenine is removed. One type of fluorescently labeled nucleotides (A, T, G and C) at a time washed with the polymerases which extends reverse strand. Nucleotide incorporation is slowed down in a way to incorporate one nucleotide at a time by fluorescentlabeling before polymerase is washed away.Continuation of the reaction goes on with flow cell imaging and fluorescent dyes removal (Harris et al., 2008).

Advantages

a. The process is not affected by the biases and errors encounters during library preparation or amplification step.

b. It sequences individual molecule.

c. It could be able to identify DNA modification lost in amplification.

d. Read length is between 24 to 70 nt.

Disadvantages

a. It has a specific limitation due to single molecule sequencing facilitate the sequencing of small amount of DNA.

b. Instrument price is very high (~ 1 million dollars).

c. Template molecule can be lost in wash.

d. It has higher average error rate than any other instrument (Kircher and Kelso, 2010).

Sanger Capillary Sequencing:

It was GE Healthcare MegaBACE instrument, working on the same principal as applied in 1977.

Principal

Fragmented DNA cloned into vectors for amplification. Conventional Chain termination technique is used. Nucleotide Separated with the help of capillary electrophoresis based on their molecular weights. Terminated ddNTPs (fluorescently labeled) read sequentially (Sanger 1977, Gilbert and Maxmum, 1973).

Advantages

a. 384 sequences of the range 600-1000nt can be sequenced in parallel.

b. Cost effective

Disadvantages

a. Low rate amplification,

b. Miss the termination variation

c. End trimming with an error of every 10000-100000nt (Emrich et al 2002, Shibata et al 2000, Hert et al 2008, Shendure and Porreca 2008).

CONCLUSION

Different sequencing technologies seem to facilitate the developing metagenomics domain. It has totally changed the scenario of the research field. From conventional to HTS and HTS to third generation sequencing are opening the new era of the science. Still for developed countries, it is cost economic but for developing countries, doing sequence analysis through NGS is still a challenge due to money constraint. There should be cost economic and convenient method for every researcher from every part of the globe to persue and continue his/her research.

REFERENCES:

[1]. Amann et al., (1995). Phylogenetic identification and in situ detection of individual microbial cells without cultivation.*Microbiol Rev.*, *59*(1):143-69.

[2]. Schloss, P. D. &Handelsman, J. (2003).Biotechnological prospects from metagenomics.*Curr Opin Biotechnol.* , *14*(3):303-10.

[3]. Covacci et al., (1997). From microbial genomics to meta-genomics. *Drug Dev. Res., 41:* 180-192.

[4]. Rondon et al., (2000). Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl. Environ. Microbiol., 66*: 2541-2547.

[5]. Hugenholz et al., (1998). Impact of Culture-Independent Studies on the Emerging Phylogenetic View of Bacterial Diversity.*J. Bacteriol.,180* (18): 4765–74.

[6]. Alvarez et al.,(2013).Development and Biotechnological Application of a Novel Endoxylanase Family GH10 Identified from Sugarcane Soil Metagenome.*PLoS One, Jul 29:8*(7):e70014.

[7]. Rosarioet al., (2009).Metagenomic analysis of viruses in reclaimed water. *Environmental Microbiology,11*(11), 2806–2820.

[8]. Kapardar et al., (2010).Identification and characterization of genes conferring salt tolerance to Escherichia coli from pond water metagenome.*Bioresour Technol., Jun: 101*(11):3917-24.

[9]. Wang et al.,(2012).Comparison of the levels of bacterial diversity in freshwater, intertidal wetland, and marine sediments by using millions of illumina tags.*Appl Environ Microbiol.,Dec:78*(23):8264-71.

[10]. Santos et al.,(2010).The metavirome of a hypersaline environment.*Environ Microbiol., Nov: 12*(11):2965-76.

[11]. Deagle et al., (2012).Population genomics of parallel phenotypic evolution in stickleback across stream-lake ecological transitions.*Proc Biol Sci., Apr: 7*:279(1732):1277-86.

[12]. Singh et al., (2010).Identification of two flavin monooxygenases from an effluent treatment plant sludge metagenomic library.*Bioresour Technol., Nov: 101*(21):8481-4.

[13]. Blottière et al.,(2013).Human intestinal metagenomics: state of the art and future.*Curr Opin Microbiol., Jun: 16*(3):232-9.

[14]. Pace et al., (1985). Analyzing natural microbial populations by rRNA sequences. *ASM News, 51*: 4-12.

[15]. Schmidt et al., (1991). Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. *J. Bacteriol., 173*: 4371-4378.

[16]. Strachan, T., & Read, A. P. (1999). Human Molecular Genetics, 2nd edition, New York: Wiley-Liss, Chapter 6:PCR, DNA sequencing and *in vitro* mutagenesis.

[17]. Barns et al., (1999). Wide distribution and diversity of members of the bacterial kingdom Acidobacterium in the environment. *Appl. Environ. Microbiol., 65*: 1731-1737.

[18]. Buckley, D. H. & Schmidt, T. M. (2003) Diversity and dynamics of microbial communities in soils from agro-ecosystems. *Environ. Microbiol., 5*: 441-452.

[19]. Teeling et al., (2004). Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ Microbiol.,* 6: 938–947.

[20]. McHardy et al., (2007). Accurate phylogenetic classification of variable-length DNA fragments. *Nat. Methods., 4*:63–72.

[21]. Naiara, R.-E., & Aransay, A. M. (2012). Introduction. In Rodríguez-Ezpeleta N. et al. (Eds.), Bioinformatics for High Throughput Sequencing (pp 1-9). Springer Science+Business Media, LLC.

[22]. Martin, K,. & Janet, K. (2010). Methods, Models & Techniques, High-throughput DNA sequencing –concepts and limitations. Bioessays, *WILEY Periodicals, Inc., 32*: 524–536.

[23]. Venter et al., (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science, 304*: 66-74.

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

[25]. Quinlan et al., (2008). Pyrobayes: an improved base caller for SNP discovery in pyrosequences. *Nat Methods, 5*: 179–81.

[26]. Bentley, D. R. (2006). Whole-genome re-sequencing.Cu*rr Opin Genet Dev., 16* (6):545–552.

[27]. Bentley et al., (2008). Accurate whole human genome sequencing using reversible terminator chemistry.*Nature,456*:53–59.

[28]. Kircher et al.,(2009).Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome Biol., 10*: R83.

[29]. Dohm et al.,(2008). Substantial biases in ultra-short read datasets from high-throughput DNA sequencing. *Nucleic Acids Res., 36*: e105.

[30]. Smith et al., (2010). Highly-multiplexed barcode sequencing: an efficient method for parallel analysis of pooled samples.*Nucleic Acids Res.,38* (13):e142.

[31]. Valouev et al., (2008). A highresolution, nucleosome position map of C. elegans reveals a lack of universal sequencedictated positioning.*Genome Res.,18* (7):1051–1063.

[32]. Myllaykangas et al., (2012). Overview of Sequencing Technology Platforms. In Rodríguez-Ezpeleta N. et al. (Eds.), Bioinformatics for High Throughput Sequencing (pp 11- 25). Springer Science+Business Media, LLC.

[33]. Dimalanta et al., (2009). Increased Read Length on the SOLiDTM Sequencing Platform. Poster, SOLiDTM System.

[34]. Harris et al., (2008). Single-molecule DNA sequencing of a viral genome. *Science, 320*: 106–9.

[35]. Sanger et al., (1977). Nucleotide sequence of bacteriophage phiX174 DNA. *Nature 265*: 687–95.

[36]. Gilbert, W.& Maxam, A. (1973). The nucleotide sequence of the lac operator. *Proc Natl Acad Sci USA, 70*: 3581–4.

[37]. Emrich et al., (2002). Microfabricated 384-lane capillary array electrophoresis bioanalyzer for ultrahigh-throughput genetic analysis. *Anal Chem., 74*: 5076–83.

[38]. Shibata et al., (2000). RIKEN integrated sequence analysis (RISA) system – 384-format sequencing pipeline with 384 multicapillary sequencer. *Genome Res., 10*: 1757–71.

[39]. Hert et al., (2008). Advantages and limitations of next-generation sequencing technologies: a comparison of electrophoresis and non-electrophoresis methods. *Electrophoresis, 29*: 4618–26.

[40]. Shendure et al., (2011). Overview of DNA sequencing strategies. In edited by Frederick M. Ausubel et al., (Eds.), *Curr Protoc Mol Biol:* (Chapter 7: Unit 7.1.17.1.23). Wiley online library.