Review

Algae 2016, 31(2): 137-154
http://dx.doi.org/10.4490/algae.2016.31.6.5
Open Access

A novice's guide to analyzing NGS-derived organelle and metagenome data

Hae Jung Song[1], JunMo Lee[1], Louis Graf[1], Mina Rho[2], Huan Qiu[3], Debashish Bhattacharya[3] and Hwan Su Yoon[1,*]

[1]*Department of Biological Sciences, Sungkyunkwan University, Suwon 16419, Korea*
[2]*Division of Computer Science & Engineering, Hanyang University, Seoul 04763, Korea*
[3]*Department of Ecology, Evolution and Natural Resources, Rutgers University, New Brunswick, NJ 08901, USA*

Next generation sequencing (NGS) technologies have revolutionized many areas of biological research due to the sharp reduction in costs that has led to the generation of massive amounts of sequence information. Analysis of large genome data sets is however still a challenging task because it often requires significant computer resources and knowledge of bioinformatics. Here, we provide a guide for an uninitiated who wish to analyze high-throughput NGS data. We focus specifically on the analysis of organelle genome and metagenome data and describe the current bioinformatic pipelines suited for this purpose.

Key Words: bioinformatic; NGS data analysis; organelle genome; metagenome

INTRODUCTION

Following the development of 'first-generation sequencing' by Frederick Sanger (Sanger et al. 1977), a new method was developed in the mid-1990s termed 'second-generation sequencing' or 'next-generation sequencing (NGS)' (Ronaghi et al. 1996). NGS is based on DNA amplification and detects different signals produced by the addition of individual nucleotide to the nascent DNA target (so-called 'sequencing-by-synthesis'; SBS). Compared to Sanger sequencing, NGS technologies are characterized by massively parallel approaches, high throughput, and reduced costs. The rapid progress of NGS technology allowed for a significant increase in the size of datasets that can be used for biological research. Consequently, NGS broadened our understanding of biological phenomena.

There are many kinds of NGS platforms available that have different properties (Table 1). Roche 454 and SOLiD were commercialized early on and have contributed to many research projects (Rothberg and Leamon 2008, Ludwig and Bryant 2011). However, due to the high cost, relatively long running time, and small amount of output, they have been replaced by newer platforms. Illumina (San Diego, CA, USA) is currently the most widely used system because of the large data output (15-1,800 Gbp) with low costs. Furthermore, Illumina provides a large choice of platforms from the benchtop sequencers MiSeq and MiniSeq that are suitable for smaller-scale research, to the HiSeq and HiSeq X Ten for larger-scale genomics, which are applicable for various research purposes. The Ion Torrent is specialized for individual laboratories due to its compact size and relatively low instrument price.

Received April 13, 2016, **Accepted** June 5, 2016
***Corresponding Author**
E-mail: hsyoon2011@skku.edu
Tel: +82-31-290-5915, Fax: +82-31-290-7015

Table 1. Comparison of next-generation sequencing platforms

	Platform	Max read length	Max output	Max read number	Required DNA	Run price	Run time	Single read accuracy (%)	Short description
Ion Torrent	Ion PGM	400 bp	2 Gb	5.5 M	>100 ng	~$350-750	2-7 h	98	Relatively cheap and small instrument. Moderate read length. High error rate. Homopolymer error occurs. Appropriate for targeted and small genome sequencing.
	Ion Proton	200 bp	10 Gb	80 M	>100 ng	~$1k	2-4 h	98	Similar platform to PGM, but produces more output with higher cost.
SOLiD	5500xl	2 × 60 bp	240 Gb	2,400 M	>10 ng	~$5-10k	6-10 days	99.9	Low cost per base, but slow and produces short reads.
	5500	2 × 60 bp	120 Gb	1,200 M	>10 ng	~$2.5-5k	6-10 days	99.9	Similar platform to the 5500xl, but produces lower output at lower cost.
454 (Roche)	GS FLX+	1,000 bp	700 Mb	1 M	>500 ng	~$6k	23 h	99.9	Makes long reads with a rapid run time. Expensive and produces small output.
	GS JR	700 bp	35 Mb	0.1 M	>500 ng	~$1k	10 h	99.9	Smaller version of GS FLX. Cheaper running cost and smaller instrument, but produces smaller output.
Illumina	MiSeq	2 × 300 bp	15 Gb	25 M	>50 ng	~$1.4k	4-55 h	99.9	Produces large output at low cost, but read length is short (for all Illumina platforms). Appropriate for targeted and small genome sequencing.
	NextSeq	2 × 150 bp	120 Gb	400 M	>50 ng	~$4k	12-30 h	99.9	Appropriate for everyday genomics.
	HiSeq	2 × 150 bp	1,500 Gb	5,000 M	>50 ng	~$8-29k	7 h-6 days	99.9	Appropriate for large-scale genomics.
	HiSeq X	2 × 150 bp	1,800 Gb	6,000 M	>100 ng	~$12k	<3 days	99.9	Appropriate for population-scale whole-genome sequencing.
Pacific Biosciences	PacBio	40 kbp (10k-15k bp avg.)	1 Gb	50 k	>10 μg	~$1-1.5k	4 h	86	Makes extremely long reads, but their number is small. Requires a large quantity of DNA. High error rate. Appropriate for research that requires ultra-long reads such as *de novo* assembly of complex genomes.

Given the output capability (~2 Gbp) and short running time (2-7 h), the Ion Torrent personal genome machine (PGM) is largely targeted to smaller genomes such as organelle genomes or to prokaryote genome sequencing (Kim et al. 2014*b*, Lee et al. 2015, Yang et al. 2015). The PacBio single molecular real time sequencing (SMRT) platform is referred to as 'third generation sequencing' because the DNA amplification step during library preparation is no longer needed. Consequently, PacBio produces small amounts of output (up to 1 Gbp with 5 Gbp forecast by the end of 2016); however, read length is considerably longer (>10,000 bp) that advantageously differentiates it from other platforms (<400 bp). Its ultra-long read is suitable for *de novo* construction of whole genomes (Tombácz et al. 2014) or for full-length transcriptome sequencing without assembly (ISO-Seq) (Sharon et al. 2013), and is also useful for reducing the re-sequencing step that other platforms require by filling the gaps of complex repeats in the *de novo* assembly (Ferrarini et al. 2013, Loomis et al. 2013, Huddleston et al. 2014).

The development of NGS has been the driving force for major progress in biological research fields. Rapidly generated genome data allow researchers to exploit more information contained in DNA and provides additional opportunities to address profound biological questions. However, handling these high-throughput data is a challenge for beginning investigators. Given this issue, here we describe bioinformatic pipelines that are designed to analyze high-throughput data produced by NGS. Nuclear genome sequencing is not discussed in this paper because of its high complexity. In contrast, organelle genomes are relatively small and easy to handle, thus novices are able to assemble and annotate entire genomes by following relatively simple protocols. This paper introduces detailed pipelines to generate complete eukaryotic organelle genomes, as well as approaches for metagenome analysis, which provides useful information about community structure in natural environments. The methodological pipelines are summarized in Figs 1 and 2.

DNA PREPARATION

The first requirement of any NGS experiment is sufficient, high-quality DNA extracted from organismal tissue. The quantity and quality of DNA largely affects the sequencing results, therefore, this step is of critical importance for NGS. The minimum amount of DNA required varies depending on the platform to be used. Many Illumina protocols require >50 ng of DNA, whereas Ion Torrent platforms require 100 ng or more. PacBio platforms require a larger amount of DNA (15 μg) of high quality (not extensively fragmented) for long-read sequencing. The requirements for each platform are described in Table 1.

The basic process of DNA extraction is composed of two major steps, cell / tissue lysis followed by DNA purification (Csaikl et al. 1998). The lysis step involves cell or tissue disruption to release the DNA. To recover high amount of DNA, proper extraction methods must be used depending on the target organism. For example, in several algal species, high mucus content is a significant hurdle. Because high DNA viscosity may hinder the aggregation of binding buffer and DNA templates, this results in poor DNA yield. Manual extraction tends to leave more mucilage with DNA therefore commercial kits (e.g., DNeasy Plant Mini Kit; Qiagen, Hilden, Germany) are widely used for cells with high polysaccharide content. Furthermore, in order to remove mucilage, the cleaning process after extraction is very helpful even though it reduces DNA yield. Several commercial cleaning kits (e.g., PowerClean DNA Clean-Up Kit; Mo Bio Laboratories, Solana Beach, CA, USA) are available. Another difficulty is rigid cells. Soft tissues are easily broken in liquid nitrogen. However, several organisms with rigid cell walls such as coralline algae are hard to disrupt by grinding. Applying homogenization or bead beating with the appropriate instrument provides a solution to this problem (Lee et al. 2010, Samarasinghe et al. 2012).

Following lysis, detergents, proteins, and any other reagent should be removed. For purification, phenol-chloroform extraction, ethanol precipitation, and spin column-based nucleic acid purification are the most frequently used approaches (Zeugin and Hartley 1985, Boom et al. 1990, Walsh et al. 1991). In many commercial DNA extraction kits, spin column technology is widely used because of its compatibility with standard lab equipment. Manual extraction with the phenol-chloroform method is excellent for maximizing DNA quality. This approach produces high purity and low degraded DNA but with relatively low yield. Therefore, when enough tissue samples for DNA are available the manual method is a good choice for producing high quality of DNA.

After extraction, the quality of DNA needs to be checked using gel electrophoresis. In this step, electrophoresis is performed on 0.8% tris-acetate, ethylenediaminetetra acetic acid agarose gel (50 V, 60 min). High voltage (e.g., >150 V) may heat and melt the gel and result in poor resolution. By observing the band resolution on the gel, the degree of DNA degradation and the size can be estimated. Additionally, the DNA concentration also needs to be

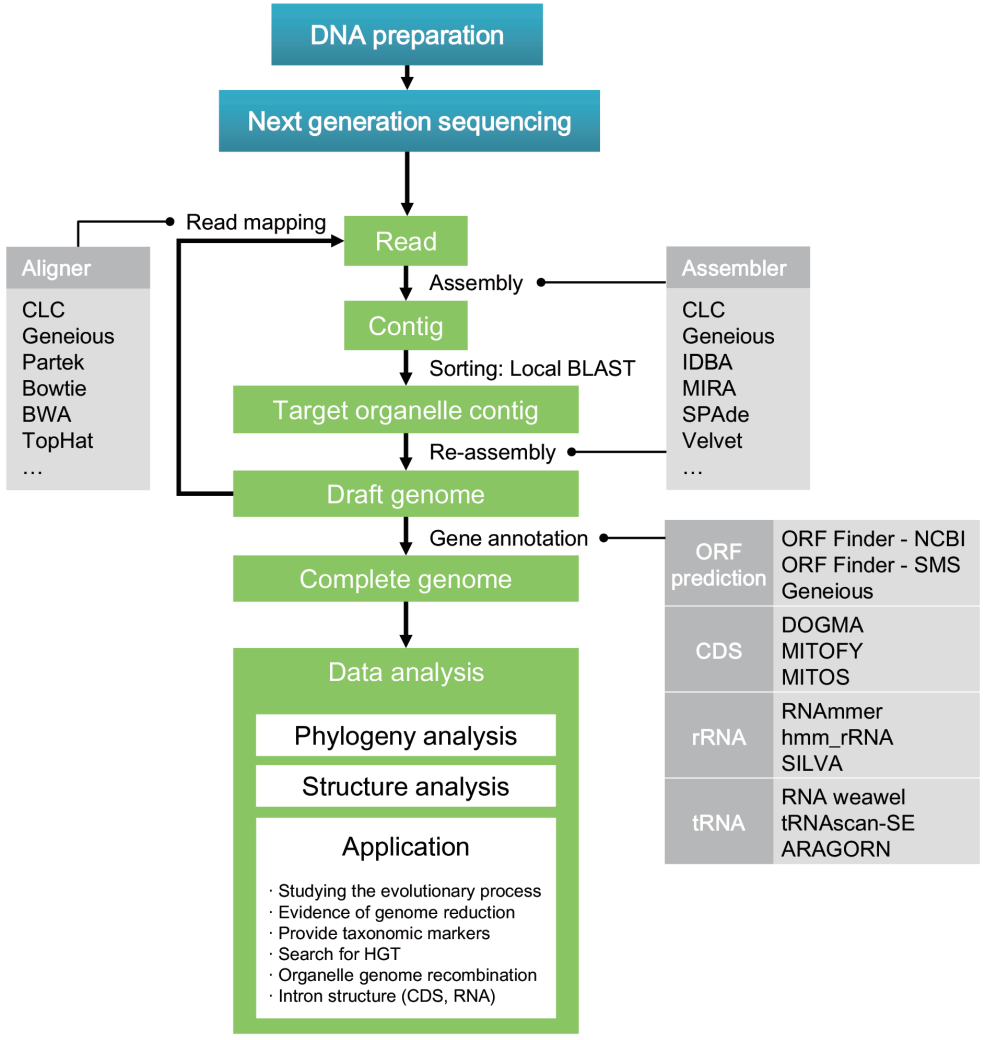

Fig. 1. General strategy for organelle genome reconstruction. Several bioinformatic tools for each process are presented. BWA, Burrows-Wheeler Aligner; CDS, coding sequence; HGT, horizontal gene transfer; IDBA, iterative De Bruijn graph assembler; MIRA, Mimicking Intelligent Read Assembly; ORF, open reading frame.

determined. The most frequently used method for doing this is with a spectrophotometer, which measures the absorbance optical density (OD) of the solution to estimate the DNA concentration.

ORGANELLE GENOME RECONSTRUCTION

Organelle DNA contains valuable genetic information not provided by nuclear DNA such as a conserved gene content that often has a more clearly understood evolutionary history and encodes rapidly diverging sequences suited to studying species-level phylogenetic relationships (Martin and Müller 1998, Vellai et al. 1998, Ingman et al. 2000, McKinnon et al. 2001, Conklin et al. 2009). The conserved gene architecture within smaller, circular genomes makes them easier to use in studying genome level dynamics for phylogeny and evolutionary inferences (Kim et al. 2014*a*). Since the development of NGS technology in 2005, usage of organelle genome in research has accelerated significantly. Numerous organelle genomes have been determined: i.e., 7,644 organelle genomes are available in the NCBI database as of February 2016 (http://www.ncbi.nlm.nih.gov/genome/organelle/).

This chapter provides detailed protocols for reconstructing organelle genomes from NGS high-throughput data using computational tools without the physical isolation of organelles from cells (Fig. 1). Even though different kinds of genome sequencers are available for use, lower-throughput instruments are often better suited to

small-size organelle genomes. This discussion will mainly focus on algal genome construction, although most of these protocols are also applicable to other microorganisms. Building an organelle genome is composed of five steps: 1) contig assembly, 2) identifying organelle genome contigs, 3) generating a draft genome with consensus contigs, 4) gene prediction and annotation, and 5) data analysis.

Contig assembly

NGS produces FASTQ files that contain numerous short sequences called 'reads' and their associated sequencing quality data. The information stored in an individual read is however limited due to its short length. Therefore, reads need to be assembled into contiguous sequences (contigs) using bioinformatic programs. There are two different approaches for assembling reads. The first is *de novo* assembly, whereby short reads are connected into longer sequences by overlapping reads (Paszkiewicz and Studholme 2010). This method uses an assembly algorithm that compares every possible pair of reads, therefore it is a slow process that requires high computing power. The second is reference-guided assembly, which aligns short reads to reference sequences (Gordon et al. 1998). This is faster than *de novo* assembly and can be performed with a smaller number of reads along with the reference sequence that should be similar to the target organism in terms of genome structure. CLC Genomics Workbench (CLC Bio, Aarhus, Denmark) and Geneious (http://geneious.com/) are commercially available and widely used programs that contain both assembly and read mapping with user-friendly interfaces. Other freely available assembly / read mapping programs are listed in Table 2.

Identifying organelle genome contigs

Assembled contig data contains a mixture of sequences encoding nuclear, organelle, and potentially contaminating DNAs. Therefore, contigs of the target organelle need to be identified. To identify all of the potential contigs of the targeted organelle, the sorting process has two steps: 1) to build a reference database with sequences of genetically close taxa and 2) to compare every contig to the reference sequences and select similar contigs based on similarity to the reference.

To build a reference database, sequence data from phylogenetically closely related taxa should be selected. Instead of downloading genome data from a single species, multiple genome data from closely related species is recommended for the reference, because genome data are still sparse from a phylogenetic point of view, and genome structure can be different even between closely related species. Hence, we use all the available algal plastid genomes as reference when we assemble a particular algal plastid genome. Reference data can be collected from the NCBI genome database (http://www.ncbi.nlm.nih.gov/genome/).

Along with the reference database, a tool for comparison between contigs and reference sequences is needed. The BLAST algorithm is useful in this respect (Altschul et al. 1997). Web-based BLAST searches are very convenient, but it is impossible to search through millions of contigs using the web interface. Thus, an automated pipeline with local BLAST is recommended. Local BLAST is a stand-alone software, which can be run on a local computer. To install the local BLAST tool, source codes and installers are available on the NCBI web site (http://www.ncbi.nlm.nih.gov/guide/howto/run-blast-local/). Any computer is capable of running local BLAST; however, for a large amount of genome data, relatively high computing power is needed (in this study, a computer with 64 cores was used to run local BLAST).

Like web BLAST, local BLAST also provides five search algorithms (blastn, blastx, blastp, tblastn, and tblastx). For contig sorting, translated amino acid sequences are more useful because nucleotide sequence comparison can only recognize homologs from very closely related species. Therefore, blastn is not recommended for this purpose. To use translated sequences, reference data should be downloaded in protein format, and contigs need to be used after translation. Among the other four algorithms, blastp and tblastx cannot be used because they compare protein-protein sequences and translated nucleotide-translated nucleotide sequences, respectively. Thus blastx and tblastn are needed to sort contigs. Blastx, however, generally makes underestimates with large genomes, thus using tblastn is generally recommended. As stated above, the blastn algorithm is not useful for identifying protein coding genes, but for rRNA sorting (because RNA does not encode amino acids), blastn is the only applicable algorithm. Local blast is operated using command lines, and detailed commands are included in the source code data (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/). The following setting for tblastn is a good starting point for beginners.

– evalue e-06 (E-value e^{-06} for blast alignment)
– num_threads 1 (use for the single-core calculating, change the number for the different number of cpus)

Table 2. Bioinformatic tools provide assembly / alignment function

Name	Web site	Assembly	Alignment (mapping)	Interface	OS	Platform	Licence	Reference
CLC Genomics Workbench	http://www.clcbio.com/products/product-overview/	○	○	GUI	Unix / Linux, Mac OS, Windows	Illumina, Life Technologies, Roche, Sanger	Commercial	Qiagen (2015)
DNA Baser Assembler	http://www.dnabaser.com/	○	○	GUI	Unix / Linux, Windows	Roche, Sanger	Commercial	Biosoft (2012)
Geneious	http://geneious.com/	○	○	GUI	Unix / Linux, Mac OS, Windows	Illumina, Life Technologies, PacBio, Roche	Commercial	Drummond et al. (2011)
SeqMan NGen	http://www.dnastar.com/t-products-seqman-ngen.aspx	○	○	GUI	Unix / Linux, Mac OS, Windows	Illumina, Ion Torrent, PacBio, Roche, Sanger	Commercial	Swindell and Plasterer (1997)
ABySS	http://www.bcgsc.ca/platform/bioinfo/software/abyss	○	×	CL	Unix / Linux	Illumina, Roche, SOLiD, Sanger	NC/A	Simpson et al. (2009)
Euler-sr	http://cseweb.ucsd.edu/~ppevzner/software.html#EULER-short	○	×	CL	Unix / Linux	Illumina, Roche, Sanger	NC/A	Chaisson and Pevzner (2008)
Edena	http://www.genomic.ch/edena.php	○	×	CL	Unix / Linux	Illumina	Open source	Hernandez et al. (2008)
IDBA	http://www.cs.hku.hk/~alse/idba/	○	×	CL	Unix / Linux	Illumina, Roche, Sanger	Open source	Peng et al. (2010)
MIRA	http://sourceforge.net/apps/mediawiki/mira-assembler/	○	×	CL	Unix / Linux, Mac OS	Illumina, Life Technologies, PacBio, Roche	Open source	Chevreux et al. (1999)
PASHA	http://sites.google.com/site/yongchaosoftware/pasha	○	×	CL	Unix / Linux	Illumina	Open source	Liu et al. (2011)
Ray	http://denovoassembler.sf.net/	○	×	CL	Unix / Linux	Illumina, Roche	Open source	Boisvert et al. (2010)
SOAPdenovo	http://soap.genomics.org.cn/soapdenovo.html	○	×	CL	Unix / Linux, Mac OS	Illumina	Open source	Li (2009)
SPAdes	http://bioinf.spbau.ru/en/spades	○	×	CL	Unix / Linux, Mac OS	Illumina, Ion Torrent, PacBio, Roche, Sanger	Open source	Bankevich et al. (2012)
Taipan	http://sourceforge.net/projects/taipan/	○	×	CL	Unix / Linux	Illumina	Open source	Schmidt et al. (2009)
VCAKE	http://sourceforge.net/projects/vcake/	○	×	CL	Unix / Linux, Mac OS	Illumina	Open source	Jeck et al. (2007)
Velvet	http://www.ebi.ac.uk/~zerbino/velvet/	○	×	CL	Unix / Linux	Illumina, Life Technologies, Roche	Open source	Zerbino and Birney (2008)
Partek	http://www.partek.com/star-align-and-quantify	×	○	GUI	Unix / Linux, Mac OS, Windows	Illumina, Life Technologies, Roche, Sanger	Oommercial	Partek Inc. (1998)
Novoalign	http://www.novocraft.com/products/novoalign/	×	○	CL	Unix / Linux, Mac OS	Illumina, SOLiD	NC/A	Li (2013)
Bowtie	http://bowtie-bio.sourceforge.net/index.shtml	×	○	CL	Unix / Linux, Mac OS, Windows	Illumina, SOLiD	Open source	Langmead and Salzberg (2012)
BWA	http://bio-bwa.sourceforge.net/	×	○	CL	Unix / Linux	Illumina, Roche, SOLiD	Open source	Li and Durbin (2009)
Stampy	http://www.well.ox.ac.uk/project-stampy	×	○	CL	Unix / Linux	Illumina	Open source	Lunter and Goodson (2011)
SHRiMP2	http://compbio.cs.toronto.edu/shrimp/	×	○	CL	Unix / Linux, Mac OS	Illumina	Open source	David et al. (2011)
TopHat	https://ccb.jhu.edu/software/tophat/	×	○	CL	Unix / Linux	Illumina	Open source	Trapnell et al. (2009)

GUI, graphic user interface; CL, command line interface; NC/A, free for non-commercial and academics; IDBA, iterative De Bruijn graph assembler; MIRA, Mimicking Intelligent Read Assembly; BWA, Burrows-Wheeler Aligner.

After sorting, the collected contigs need to be manually checked. Due to sequence similarity between organelle DNA with genomes of contaminant bacteria, bacterial contigs need to be identified and separated. Therefore, if the sample is highly contaminated, bacterial contigs should be filtered out using blastn or blastx.

Draft genome with consensus contig

The assembled contigs may not be the full-length organelle genome and the contigs are linear rather than the (typically) expected circular form. This indicates that the contigs are partial genome (circular / linear form can be determined by checking the end to end connection). To assemble a complete genome, the 're-assembly' step using sorted contigs is required. Re-assembly can be done using *de novo* assemblers such as CLC or Geneious, and additional programs are listed in Table 2. *De novo* assembly normally works for many algal organelle genomes, but in some cases different methods should be considered. For instance, the read mapping method is more suitable for genomes with low variation like those found in the chloroplast genomes of land plant (Doorduin et al. 2011).

Once the consensus genome is assembled, several confirmation steps are needed. The first is genome size comparison to sister taxa using a sequence homology check. This step can be performed using the BLAST method and will confirm completion of the target genome. Another step is read-mapping to the consensus contig. This step will reveal the regions of the genome where more sequence data may be needed to ensure accuracy. Low read coverage (less than 50×) indicates insufficient read number or assembly error that needs to be used to inform re-sequencing or re-assembly strategies. For instance, in the PGM platform using the 318-chip we generally produce ca. 50× genome coverage for organelles from 1 Gbp of sequence data that results in a reasonable assembly. If the reads are too limited in number (i.e., less than 50× coverage), then additional sequencing should be done. The specific issues to be considered for this step, however, vary between different NGS platforms. In the case of assembly error, confirmation by highly accurate sequencing (i.e., Sanger method) or mapping with ultra-long read (i.e., PacBio) can provide solutions. Specifically, due to the high frequency of repeated sequences (e.g., inverted repeats or duplicated rRNA operons), which usually results in assembly error, plastid genomes demand the utmost care. In addition, for circular genomes, connection of both ends needs to be checked by additional read mapping. If both ends are not connected, this gap must be filled using Sanger sequencing or an additional NGS run. Read mapping can be performed using available programs (e.g., CLC, Geneious, Partek, and Bowtie) including the aligning function (Table 2, Fig. 1).

Gene prediction and annotation

Once a draft genome is constructed, its constituent genes need to be identified and annotated. Before annotation, gene prediction should occur. Gene prediction is the process of identifying the regions of encoded genes that are likely to occur. This process entails translating nucleotide sequence and finding open reading frames (ORFs). Gene prediction can be performed with some computational programs such as Geneious Pro or ORF Finder (Table 3). During the prediction process, the genetic code setting must be carefully considered. Several species use different translation codons. In particular, alternative start codons significantly change the structure of predicted genes. Translation in several organelles can be initiated from codons other than ATG. For example, translation codon number 11, which is usually used for chloroplast genome, also uses TTG, CTG, ATT, ATC, ATA, and GTG as initiation codons. Many organelle genomes use altered translation codons (e.g., many algal mitochondria use 4, some green algal mitochondria use 22, many plant mitochondria use 1, and so forth), therefore, proper genetic code must be used in the genetic code data in NCBI (http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi). Originally, gene prediction requires cDNA data to identify noncoding exons. Therefore, when transcriptome data or expressed sequence tag (EST) data are available, annotations tend to be more accurate, but producing cDNA data is an expensive and time-consuming process. Furthermore, organelle genomes are generally highly conserved, thus reference-based prediction is sufficient for organelle genome annotation. Predicted ORFs are verified using a BLAST similarity search. For coding gene annotation, blastx is recommended. The blastx program compares the six-frame conceptual translated products of a nucleotide query against a protein sequence database to provide more accurate models and to detect unknown ORF sequence. Moreover, there are some automated annotation tools that are available for use (e.g., DOGMA, MITOFY, and CpGAVAS, see more in Table 3).

After annotation, two things should be checked: 1) whether the lengths of the annotated genes are similar to that of the reference, and 2) whether the proper start codon was used. If nucleotide insertions / deletions exist,

Table 3. Bioinformatic tools used in gene annotation

Category	Name	Web site	Short description	Reference
ORF prediction	ORF Finder - NCBI	http://www.ncbi.nlm.nih.gov/gorf/gorf.html	ORF finding service that provided by NCBI. Finds ORFs in a user's sequence or in the database.	Rombel et al. (2002)
	ORF Finder - SMS	http://www.bioinformatics.org/sms2/orf_find.html	ORF finding service that provided by Sequence Manipulation Suite (SMS). Predicts the range of each ORF.	Rombel et al. (2002)
	Geneious	http://geneious.com/	Commercial annotation program. Provides graphical ORF prediction, and easy-to-use annotation method.	Drummond et al. (2011)
Protein coding gene annotation	DOGMA	http://dogma.ccbb.utexas.edu/	Annotation tool for plant chloroplast and animal mitochondrial genome.	Wyman et al. (2004)
	MITOFY	http://dogma.ccbb.utexas.edu/mitofy/	Annotation tool for plant mitochondrial genome.	Wyman et al. (2004)
	MITOS	http://mitos.bioinf.uni-leipzig.de/index.py	Annotation tool for metazoan mitochondrial genome.	Bernt et al. (2013)
	CpGAVAS	http://www.herbalgenomics.org/0506/cpgavas/analyzer/home	Annotation tool for chloroplast genome. Provides genome map drawing, and analysis results of annotated genome. The result can be submitted to GenBank directly.	Liu et al. (2012)
	Mfannot	http://megasun.bch.umontreal.ca/cgi-bin/mfannot/mfannotInterface.pl	Annotation tool for mitochondrial and plastid genome. Helpful with organelle genomes that contain many introns.	Beck and Lang (2010)
rRNA annotation	RNAmmer	http://www.cbs.dtu.dk/services/RNAmmer/	Predicts 5S/8S, 16S/18S, and 23S/28S.	Lagesen et al. (2007)
	blastn_rRNA	http://weizhong-lab.ucsd.edu/metagenomic-analysis/server/blastn_rRNA/	rRNA prediction tool using BLASTN algorithm.	Huang et al. (2009)
	hmm_rRNA	http://weizhong-lab.ucsd.edu/metagenomic-analysis/server/hmm_rRNA/	rRNA prediction tool using HMMER 3.0.	Huang et al. (2009)
	SILVA	http://www.arb-silva.de/	rRNA database. Provides comprehensive, quality checked and regularly updated datasets of aligned 16S/18S and 23S/28S for all bacteria, archaea and eukarya.	Quast et al. (2013)
tRNA annotation	RNA weawel	http://megasun.bch.umontreal.ca/cgi-bin/RNAweasel/RNAweaselInterface.pl	Predicts introns, tRNAs, rnpB, 5S and SSU in organelle genome.	Lang et al. (2007)
	tRNAscan-SE	http://lowelab.ucsc.edu/tRNAscan-SE/	Provides secondary structure diagrams of the tRNA molecules.	Lowe and Eddy (1997)
	ARAGORN	http://mbio-serv2.mbioekol.lu.se/ARAGORN/	Identifies tRNA genes without introns or with introns at canonical or non-canonical positions.	Laslett and Canbäck (2004)
	Rfam	http://rfam.xfam.org/search#tabview=tab1	A collection of RNA families, each represented by multiple sequence alignments, consensus secondary structures and covariance models.	Griffiths-Jones et al. (2005)
	ARWEN	http://130.235.46.10/ARWEN/	Predicts tRNA in metazoan mitochondrial genome.	Laslett and Canbäck (2008)

ORF, open reading frame.

the stop codon may occur in the middle of a gene resulting in alteration of the length, referred to as a 'pseudogene.' Once all gaps and ambiguous sequences have been identified, polymerase chain reaction (PCR) confirmation is needed to correct these regions. This confirmation step, however, may not be always necessary. Sequence confirmation is generally needed for Ion Torrent or PacBio platforms, because the accuracy of these platforms (98% and 86%, respectively) (Table 1) is lower than that of other platforms such as Illumina (>99.9%). In many cases, the size of the gap is unknown, thus long-range PCR (PCR with long extension time) or primer walking (making additional primers to sequence through the gap) is also useful to fill the gap.

Because rRNA and tRNA do not encode amino acids, the annotation step described above is not applicable for these sequences. Two possible methods are widely used for non-protein coding RNA annotation. The first is using web-based tools, which provide RNA annotation services listed on Table 3. For example, ARAGORN and RNAmmer provide reasonable prediction for tRNAs and rRNAs respectively. However, some genomes are not fit for the listed programs because of the extremely high divergence of RNA, and the presence of introns in some of these genes. In this case, sequences should be manually analyzed by comparison to related species; i.e., a blastn alignment may find the corresponding rRNA or tRNA region between different genomes.

Data analysis

A completed organelle genome provides a rich source of genetic information that can be applied to diverse biological fields including systematics and evolutionary research.

Phylogenomics. Reconstructing phylogenetic tree is one of the major tools used to address taxonomic or evolutionary questions. Compared to the phylogenetic tree of a single gene, phylogeny of multiple concatenated genes from the organelle genome generally provides better resolution (Kim et al. 2014*a*). Conceptually, reconstructing multi-gene trees is identical to methods used for single gene data. However, before phylogenetic analysis, multi-gene sequences need to be combined into a single alignment. For this approach, every gene sequence from the organelle genomes from all target taxa should be extracted into individual files. It is ideal if all of the species contain the same set of genes, otherwise, the gene set should be manually selected. In general, genes that are present in more than 80% of the taxa set are normally chosen. For the gene selection step, the blastn algorithm (search a nucleotide database using a nucleotide query) is appropriate for nucleotide datasets, whereas the blastp algorithm (search a protein database using a protein query) is used for protein datasets. Use of nucleotide alignments might result in phylogenetic 'noise' from saturated silent nucleotide substitutions, thus, using protein dataset is generally recommended. Extracted sequences should be combined into a single sequence file from each species. These steps can be done manually, however, for a large set of genomes, using command lines can be useful. When using protein datasets, the correct genetic code setting for translation must be used (see above). Concatenated gene sets then are aligned into a PHY file using MAFFT (http://mafft.cbrc.jp/alignment/server/) to prepare for phylogenetic analysis. Thereafter, phylogenies can be reconstructed using various standard methods such as RaxML (Stamatakis 2006) or IQtree (Nguyen et al. 2015). These concatenated phylogenetic analysis with organelle genome have been used in many evolutionary biology studies, for example in understanding the evolution of brown algal plastids (Le Corguillé et al. 2009), in finding evolutionary evidence for organelle genome reduction (Qiu et al. 2015), and to identify useful taxonomic markers by comparing the mutation rate of organelle encoded genes (Janouškovec et al. 2013).

Structure analysis. Genome structure analysis may reveal genome-wide differences such as gene gain, loss, duplication, rearrangement or inversion of gene fragments on a genome, and lateral gene transfer. These data also provide additional information about the interrelationships of different taxa. Some bioinformatic methods such as drawing graphical maps or synteny comparison can be used for this purpose. Graphical maps of DNA sequences show the annotation information describing the gene loci, whereas synteny comparison can be used to identify large-scale changes in the genome. Genome maps can be visualized by uploading the genome sequence to web based tools such as OGDraw (http://ogdraw.mpimp-golm.mpg.de) (Lohse et al. 2007) or GenomeVX (http://wolfe.ucd.ie/GenomeVx/) (Conant and Wolfe 2008). For synteny comparison, several multiple genome aligners are available including the two widely used programs MUMmer (http://mummer.sourceforge.net/) (Delcher et al. 1999) and Mauve (http://darlinglab.org/mauve/mauve.html) (Darling et al. 2004). Structural analysis of organelle genome has contributed greatly to our understanding of organelle genome evolution. For example, searching for horizontal gene transfer in red algal genomes (Qiu et al. 2013), the origin of red algal plasmids

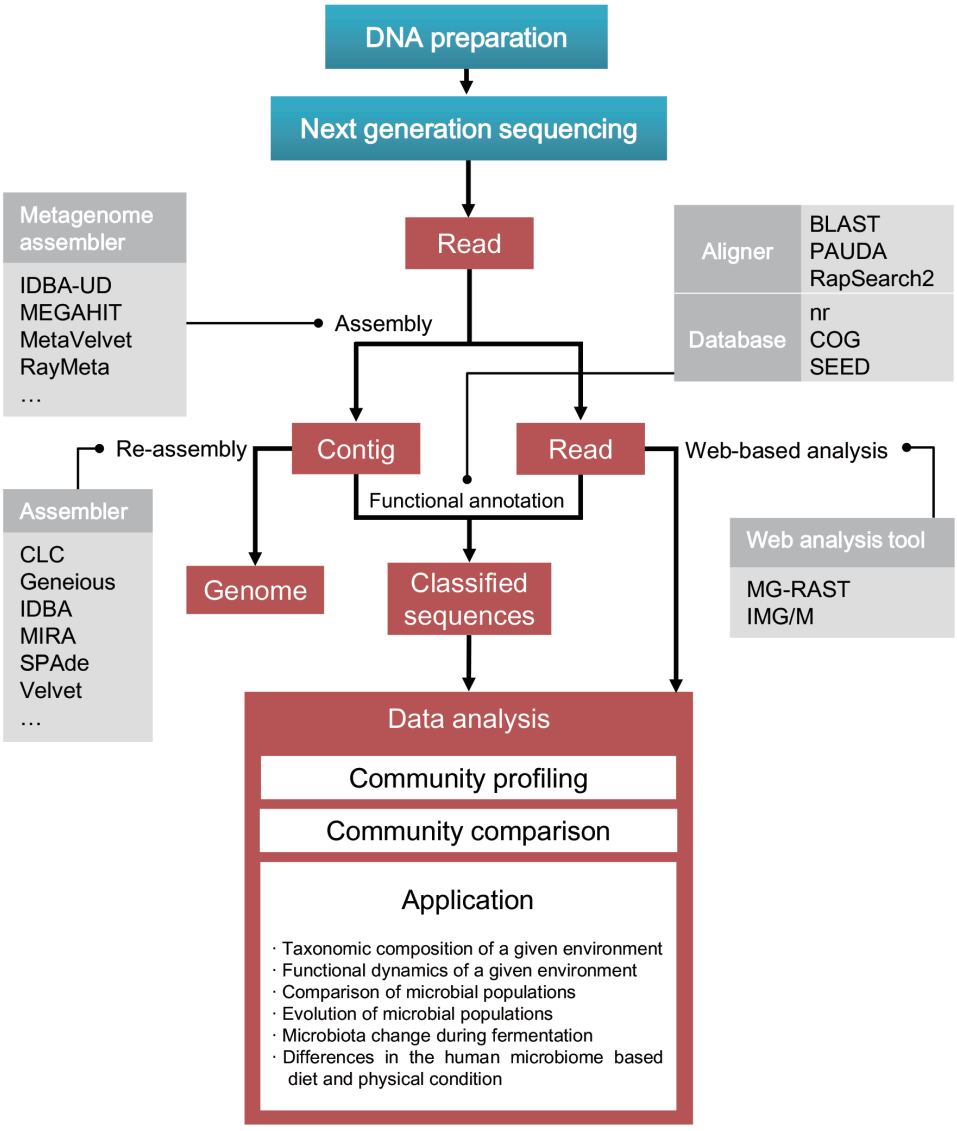

Fig. 2. General strategy for metagenomic approach. Several bioinformatic tools for each process are presented. COG, Clusters of Orthologous Group; IDBA, iterative De Bruijn graph assembler; IMG/M, Integrated Microbial Genomes and Metagenomes; MG-RAST, metagenomics Rapid Annotation using Subsystem Technology; MIRA, Mimicking Intelligent Read Assembly.

(Lee et al. 2016), organelle genome conservation (Yang et al. 2015) and recombination (Maréchal and Brisson 2010), the finding of introns within tRNA, which encodes a plastid intron maturase (Janouškovec et al. 2013) have indicated the utility of organelle genome structure analysis.

METAGENOME ANALYSIS

The term 'metagenome' refers to the 'collective genomes of environmental microflora,' which are directly isolated from an environmental sample (Handelsman et al. 1998). Metagenomics is the study of microbial organisms using genome sequence data derived from environments such as soil, marine water, air, or sediment cores. Essentially, metagenomic analysis focuses on the full characterization of the natural population, which addresses community composition, their functional dynamics and relative abundance among different environments or different time points (Scholz et al. 2012). Technological advances in NGS fueled a revolution in metagenomic sequencing and analysis. Increased throughput and cost-efficiency coupled with additional

technological advances have extended the importance of metagenomics. This technological development allowed more comprehensive investigation of diverse microbial communities of extreme complexity such as human gut (Weinstock 2012), global ocean microbiome (Sunagawa et al. 2015), and palaeomicrobiome (Warinner et al. 2015). Given the enormous sequencing data, the advanced computational methods are required, and recently, several systems and tools have been developed to apply in the analysis of complex metagenome datasets (Mocali and Benedetti 2010).

Here, we describe methodological approaches for high-throughput metagenome sequence analysis (Fig. 2). There are two general types of analysis depending on the research aim. If the research aims at reconstructing the genome from a mixture of multiple organism sequences, and the reads are enough to recover entire genome, 1) contig assembly is needed to construct genomic contigs. Whereas, to profile the community structure, 2) taxonomic / functional assignment to the individual reads or short contigs (functional annotation) is the suitable method. Furthermore, comparative analysis among different metagenomes will allow opportunities to address the relationship between different communities.

Contig assembly

If the purpose of the study is to recover the genome or full-length coding sequence (CDS) for genome level analysis from metagenome data, then short-read sequence data should be assembled into longer genomic contigs. High-throughput metagenome sequencing data include DNAs from numerous organisms of varied abundance. This unevenness of coverage makes it difficult to reconstruct contigs or genomes, moreover, chimeric assembly, caused by the similarity of closely related lineages further complicates the process. For these reasons, major *de novo* assemblers, which were designed to assemble single or clonal genomes, are not suited to the assembly of metagenomes with abundant heterogeneous sequences, and thus, their performance with metagenomic data sets varies significantly (Kunin et al. 2008). Therefore, many assemblers capable of assembling metagenome data have been developed, including MetaVelvet, IDBA-UD, MEGAHIT, and RayMeta (Table 4), although they are still at an early stage of development (Scholz et al. 2012, Thomas et al. 2012). Unlike traditional single genome assemblers, metagenome assemblers adopted the de Bruijn graph approach, which is reasonable for DNA assembly from mixed sequence of multiple species (Namiki et al. 2012).

Once the contigs are obtained, there are two possible approaches to analyze them: 1) genome construction and 2) contig annotation. If the sequences are sufficient to construct the whole genome, *de novo* assemblers are applied (Table 2). However, for many cases, genome construction is highly restricted due to the poor coverage of each taxon and the unevenness of community composition. If the target genome sequence is available, direct read mapping onto the reference sequence (reference-based assembly) is another approach (Table 2). Otherwise, the annotation process of individual contigs is suitable for community profiling. Based on the annotation data, the overall taxonomic composition and functional diversity of the given environment can be profiled. Annotation issues are discussed in the subsequent section.

Functional annotation

If the purpose of the study is to explore environmental community characterization, including taxonomic classification and functional diversity, direct annotation to the reads or contigs is a suitable approach; this is referred to as functional annotation. Essentially, functional annotation is focused on three questions: 1) who is living there, 2) what are they doing, and 3) how do they differ from each other (Mitra et al. 2011)? Addressing 'who is living there?' is based on investigation of the microbial community structure. It includes efforts to survey which taxa are included in the community, and how their composition is distributed. The question 'what are they doing?' addresses which functional genes are contained in the microorganisms of the environment, surveys relative abundance of each functional group, and ultimately focuses on understanding functional dynamics in the given environment by reconstructing the metabolic pathway. The third approach 'how do they differ?' relies upon comparing the different metagenome (community). The metagenome comparison has contributed to understanding the biological meaning by revealing the population level differences in multiple environments or population change process over time.

Basically, metagenomic functional annotation means classifying sequences into known functions or operational taxonomic units (OTUs) based on homology searches against existing reference data. Therefore, in general, annotation of metagenomic sequence data requires two kinds of bioinformatic tools: 1) a homology search program and 2) a reference database. Details of this process vary depending on the type of aligners, but the overall

Table 4. Bioinformatic tools used in metagenome analysis

Category	Name	Web site	Short description	Reference
Assembler	IDBA-UD	http://i.cs.hku.hk/~alse/hkubrg/projects/idba_ud/	A *de novo* assembler for the short read data with highly uneven depth.	Peng et al. (2012)
	Meta-IDBA	http://i.cs.hku.hk/~alse/hkubrg/projects/metaidba/index.html	A short read assembler specially designed for *de novo* metagenomic assembly.	Peng et al. (2011)
	MEGAHIT	https://github.com/voutcn/megahit	A *de novo* assembler with ultra-fast speed for large and complex metagenomics assembly.	Li et al. (2015)
	MetaVelvet	http://metavelvet.dna.bio.keio.ac.jp/	An extension of Velvet assembler for *de novo* metagenome assembly.	Namiki et al. (2012)
	Omega	http://omega.omicsbio.org/	An overlap-graph metagenome assembler which was developed for assembling metagenome data of Illumina.	Haider et al. (2014)
	RayMeta	http://denovoassembler.sourceforge.net/	A metagenome assembler based on uniquely-colored k-mers which is coupled with Ray Communities.	Boisvert et al. (2012)
Aligner	BLAST	http://blast.ncbi.nlm.nih.gov/Blast.cgi	A most widely used sequence comparison tool. Both web-based tool and local standalone tool are available.	Altschul et al. (1990)
	PAUDA	http://ab.inf.uni-tuebingen.de/software/pauda/	A command line based aligner that with ultra-fast calculation speed.	Huson and Xie (2014)
	RapSearch2	http://omics.informatics.indiana.edu/mg/RAPSearch2/	A command line based aligner that with ultra-fast calculation speed.	Zhao et al. (2012)
	USEARCH	http://www.drive5.com/usearch/	A faster aligner which offers search and clustering algorithms.	Edgar (2010)
Database	nr (non-redundant)	ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/	A biggest sequence database which contains non-redundant sequences from GenBank, Refseq, PDB, SwissProt, PIR, and PRF.	Pruitt et al. (2007)
	COG	http://www.ncbi.nlm.nih.gov/COG/	A database for phylogenetic classification of proteins encoded in microbial genomes which is significantly smaller than the nr database.	Redfern et al. (2005)
	KEGG	http://www.kegg.jp/	A database tools for prediction of metabolic processes.	Kanehisa and Goto (2000)
	SILVA	http://www.arb-silva.de/	rRNA database of bacteria, archaea, eukaryotes.	Quast et al. (2013)
	SEED	http://www.theseed.org	A web-based database tool for high-throughput generation, optimization and analysis of genome-scale metabolic models	Disz et al. (2010)
	eggNOG	http://eggnogdb.embl.de/#/app/home	A database of orthologous groups and functional annotation for eukaryote, prokaryote, and virus.	Huerta-Cepas et al. (2015)
	uniprot	http://www.uniprot.org/	A database of protein sequence and functional information combining the Swiss-Prot, TrEMBL, and PIR-PSD databases.	UniProt Consortium (2015)
Annotation analysis	MG-RAST	http://metagenomics.anl.gov/	A web-based automatic functional annotation tool for metagenome data. Abundance profiling, phylogenetic classifications, metabolic reconstructions and metagenome comparison functions are provided.	Glass et al. (2010)
	IMG/M	https://img.jgi.doe.gov/cgi-bin/m/main.cgi	A web-based comparative analysis tool for metagenome data. Function-based metagenome comparison service is provided.	Markowitz et al. (2012)
	MEGAN	http://ab.inf.uni-tuebingen.de/software/megan5/	A standalone metagenome / metatranscriptome data comparison tool. Taxonomy and functional analysis is provided with text / graphical format.	Huson et al. (2007)

COG, Clusters of Orthologous Group; KEGG, Kyoto Encyclopedia of Genes and Genomes; MG-RAST, Metagenomics Rapid Annotation using Subsystem Technology; IMG/M, Integrated Microbial Genomes and Metagenomes.

steps for functional annotation are similar. First, the library needs to be constructed using a reference database. Then, individual reads are searched against the database using a homology search, and eventually, each is labeled with a taxonomic classification and functional group assignment. For the similarity search, considering the size of the sequence data and computational resources, an appropriate aligner must be used. Conceptually, the annotation is a simple process, so for the very small datasets (<10,000 sequences), manual curation can be used for better accuracy (Thomas et al. 2012). However, because metagenomic datasets are typically very large, automated annotation tools are recommended. Local BLAST is a highly accurate method as well (Scholz et al. 2012), but it requires significant calculation times. Therefore only if the sequence data is relatively small or computer resource is sufficient, local BLAST will be the best for similarity search. For more rapid work, PAUDA or RapSearch2 provide good alternatives. PAUDA is based on the bowtie aligner and shows extremely rapid calculation speed. RapSearch2 also provides a rapid speed of annotation (Table 4). For the calculation speed comparison, when annotating millions of reads using 40 cores of CPU, local BLAST takes several days, whereas PAUDA and RapSearch2 complete the work within a day. Along with the aligner, a suitable reference database is necessary. Many databanks are available, which provide reference sequence datasets for functional / taxonomic information assignment such as nr (non-redundant), Clusters of Orthologous Group (COG), Kyoto Encyclopedia of Genes and Genomes (KEGG), SEED, and so forth (Table 4). The nr database contains the greatest number of sequences (43 GB), however, because every reference sequence should be individually compared to each of the reads, considerable CPU time is required. COG, SEED, or uniprot contain smaller amounts of sequences than the nr database (only functionally identified sequences are included), thus provide rapid homology search with less computer power. KEGG database provides prediction of cellular metabolic processes, which is specialized for functional profiling. Because each reference databank contains different types of sequence sets, selecting proper database depending on the research aim is of critical for accurate population profiling. More detailed information of each databank is presented in the Table 4.

Once the annotation is complete, the result needs to be visualized for community profiling or community comparison. MEGAN is a great tool for visualization of annotation results. MEGAN analyzes the taxonomic content by placing the annotated reads onto the NCBI taxonomy, while functional distribution is analyzed by mapping the reads to the three different functional classifications (SEED, COG, and KEGG) (Huson and Weber 2013). This program supports various kinds of input file formats (BLAST, SAM, RDP, Silva, CSV, and BIOME) produced by alignment of the reads to a reference sequence database. Then the graphical and statistical output for each metagenome or the comparison of multiple metagenomes is created. However, due to the high requirement of computer resources, such a standalone analysis has limits for the researchers without an access to high performance computers.

For large-scale databases, web-based analysis tools such as Metagenomics Rapid Annotation using Subsystem Technology (MG-RAST) and Integrated Microbial Genomes and Metagenomes (IMG/M) provide powerful solutions, because these web portals offer large computational resources for data analysis. These servers have the automated analysis platforms, which are specialized for metagenome data. MG-RAST pipeline provides many analysis services including quality control, functional annotation, taxonomic assignment, metabolic pathway reconstruction and comparison of multiple metagenomes (Meyer et al. 2008). To use these services, sequencing data should be uploaded to the pipeline on the server. The raw sequence data formats such as FASTA or FASTQ are acceptable. The uploaded sequences then are normalized and annotated against the database that integrates information from several tools, including M5NR, GenBank, SEED, KEGG, SwissProt, M5RNA, Greengenes, and so forth. The analysis time alters from a few hours to a few weeks depending on the importance of the research theme and the size of the data. The results are produced in the form of organism / functional abundance profiles and are visualized in various formats (bar chart, tree, table, heat map, and so forth). Beyond the annotation, MG-RAST also provides comparative metagenomics tools. Users can use multiple data for metagenome comparison with lots of statistical analyses such as phylogenetic / metabolic reconstruction and abundance profiling. MG-RAST has more than 230,000 uploaded metagenomes (of which 32,000 are publicly accessible) and 97 Terabases of sequences at February 2016. IMG/M also provides similar analysis pipeline including automated genome annotation, individual metagenome abundance profiling, and comparative metagenomics (Markowitz et al. 2012).

Despite the development of diverse analysis tools, functional annotation is still restricted by several limitations. Short read length has the possibility of a higher error rate. Assembled contigs are better for the length, but

typically, large contigs are difficult to attain due to the technological limitations in DNA recovery and sequencing capacity (Scholz et al. 2012). Furthermore, due to the immense amount of sequence data, long computation times and sufficient hardware resources are required for the individual annotation of every read. In contrast, a limited amount of reference data makes it hard to confirm the accuracy of metagenomic data. Nevertheless, functional annotation has greatly contributed to understanding microbial community profiles such as the diversity of prokaryotes in surface ocean waters (Biers et al. 2009), the human gut microbiome (Qin et al. 2010), or metabolic dynamics in lacustrine ecosystems (Debroas et al. 2009). Furthermore, comparison of multiple communities (metagenomes) can reveal differences between environments or species composition at various time points. Metagenome comparison approach is of great importance for extending our understanding of the environment-driven effect on microbiota or the transition process of community structure over time. There are several examples of note. Sunagawa et al. (2015) investigated the change in oceanic microbial composition along with vertical stratification, which then revealed the impact of temperature on community variation; Warinner et al. (2015) studied the evolution of the microbial populations in the human body, and contributed to the evolutionary understanding of microbial population transition processes; the Human Microbiome Project Consortium (2012) revealed differences in microbiome community structure between different anatomical sites, individuals, or physical conditions, which can help describe healthy microbiome status in the human body; Jung et al. (2011) explored the changes in bacterial populations and functional dynamics during the fermentation of *kimchi*. All of these studies have provided valuable information, which has extended our understanding in the fields of ecology, evolution, and medical science.

CONCLUSIONS

We have described two classes of bioinformatic approaches that should prove helpful to beginners who wish to analyze high-throughput NGS data. The discussion of organelle genome reconstruction and analysis pipelines provides the necessary framework for researchers to greatly expand existing plastid and mitochondrial databases. In contrast, metagenome analysis is a useful approach for addressing whole community structure in natural settings. Both of these computational methods will be of great value to biologists interested in the application of high-throughput genome data to various fields of research from phylogenetics to ecosystem analysis.

ACKNOWLEDGEMENTS

This work was supported by the Polar Academic Program of the Korea Polar Research Institute (KOPRI), the Korean Rural Development Administration Next-generation BioGreen21 (PJ011121), the National Research Foundation of Korea (MEST: 2014R1A2A2A01003588), and Marine Biotechnology Program (PJT200620) funded by Ministry of Oceans and Fisheries, Korea to HSY and MR.

REFERENCES

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

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389-3402.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A. & Pevzner, P. A. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19:455-477.

Beck, N. & Lang, B. F. 2010. MFannot, organelle genome annotation websever. Available from: http://megasun.bch.umontreal.ca. Accessed May 2, 2016.

Bernt, M., Donath, A., Jühling, F., Externbrink, F., Florentz, C., Fritzsch, G., Pütz, J., Middendorf, M. & Stadler, P. F. 2013. MITOS: Improved *de novo* metazoan mitochondrial genome annotation. Mol. Phylogenet. Evol. 69:313-319.

Biers, E. J., Sun, S. & Howard, E. C. 2009. Prokaryotic genomes and diversity in surface ocean waters: interrogating the global ocean sampling metagenome. Appl. Environ. Microbiol. 75:2221-2229.

Biosoft, H. 2012. DNA Baser Sequence Assembler v3x. Heracle BioSoft SRL Romania, Pitesti.

Boisvert, S., Laviolette, F. & Corbeil, J. 2010. Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. J. Comput. Biol. 17:1519-1533.

Boisvert, S., Raymond, F., Godzaridis, É., Laviolette, F. & Corbeil, J. 2012. Ray Meta: scalable *de novo* metagenome assembly and profiling. Genome Biol. 13:R122.

Boom, R., Sol, C. J. A., Salimans, M. M. M., Jansen, C. L., Wertheim-van Dillen, P. M. E. & Van der Noordaa, J. 1990. Rapid and simple method for purification of nucleic acids. J. Clin. Microbiol. 28:495-503.

Chaisson, M. J. & Pevzner, P. A. 2008. Short read fragment assembly of bacterial genomes. Genome Res. 18:324-330.

Chevreux, B., Wetter, T. & Suhai, S. 1999. Genome sequence assembly using trace signals and additional sequence information. *In* Beyer, A. & Schroeder, M. (Eds.) German Conference on Bioinformatics, Dresden, Germany, Vol. 99, pp. 45-56.

Conant, G. C. & Wolfe, K. H. 2008. GenomeVx: simple web-based creation of editable circular chromosome maps. Bioinformatics 24:861-862.

Conklin, K. Y., Kurihara, A. & Sherwood, A. R. 2009. A molecular method for identification of the morphologically plastic invasive algal genera *Eucheuma* and *Kappaphycus* (Rhodophyta, Gigartinales) in Hawaii. J. Appl. Phycol. 21:691-699.

Csaikl, U. M., Bastian, H., Brettschneider, R., Gauch, S., Meir, A., Schauerte, M., Scholz, F., Sperisen, C., Vornam, B. & Ziegenhagen, B. 1998. Comparative analysis of different DNA extraction protocols: a fast, universal maxi-preparation of high quality plant DNA for genetic evaluation and phylogenetic studies. Plant Mol. Biol. Rep. 16:69-86.

Darling, A. C. E., Mau, B., Blattner, F. R. & Perna, N. T. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 14:1394-1403.

David, M., Dzamba, M., Lister, D., Ilie, L. & Brudno, M. 2011. SHRiMP2: sensitive yet practical short read mapping. Bioinformatics 27:1011-1012.

Debroas, D., Humbert, J. -F., Enault, F., Bronner, G., Faubladier, M. & Cornillot, E. 2009. Metagenomic approach studying the taxonomic and functional diversity of the bacterial community in a mesotrophic lake (Lac du Bourget–France). Environ. Microbiol. 11:2412-2424.

Delcher, A. L., Kasif, S., Fleischmann, R. D., Peterson, J., White, O. & Salzberg, S. L. 1999. Alignment of whole genomes. Nucleic Acids Res. 27:2369-2376.

Disz, T., Akhter, S., Cuevas, D., Olson, R., Overbeek, R., Vonstein, V., Stevens, R. & Edwards, R. A. 2010. Accessing the SEED genome databases via Web services API: tools for programmers. BMC Bioinformatics 11:319.

Doorduin, L., Gravendeel, B., Lammers, Y., Ariyurek, Y., Chin-A-Woeng, T. & Vrieling, K. 2011. The complete chloroplast genome of 17 individuals of pest species *Jacobaea vulgaris*: SNPs, microsatellites and barcoding markers for population and phylogenetic studies. DNA Res. 18:93-105.

Drummond, A., Ashton, B., Buxton, S., Cheung, M., Cooper, A., Duran, C., Field, M., Heled, J., Kearse, M. & Markowitz, S. 2011. Geneious v5. 4. Biomatters Limited., Auckland.

Edgar, R. C. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460-2461.

Ferrarini, M., Moretto, M., Ward, J. A., Šurbanovski, N., Stevanović, V., Giongo, L., Viola, R., Cavalieri, D., Velasco, R., Cestaro, A. & Sargent, D. J. 2013. An evaluation of the PacBio RS platform for sequencing and *de novo* assembly of a chloroplast genome. BMC Genomics 14:670.

Glass, E. M., Wilkening, J., Wilke, A., Antonopoulos, D. & Meyer, F. 2010. Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. Cold Spring Harb. Protoc. 2010:pdb.prot5368.

Gordon, D., Abajian, C. & Green, P. 1998. Consed: a graphical tool for sequence finishing. Genome Res. 8:195-202.

Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S. R. & Bateman, A. 2005. Rfam: annotating non-coding RNAs in complete genomes. Nucleic Acids Res. 33(Suppl. 1):D121-D124.

Haider, B., Ahn, T. -H., Bushnell, B., Chai, J., Copeland, A. & Pan, C. 2014. Omega: an Overlap-graph *de novo* Assembler for Metagenomics. Bioinformatics 30:2717-2722.

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J. & Goodman, R. M. 1998. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem. Biol. 5:R245-R249.

Hernandez, D., François, P., Farinelli, L., Østerås, M. & Schrenzel, J. 2008. *De novo* bacterial genome sequencing: millions of very short reads assembled on a desktop computer. Genome Res. 18:802-809.

Huang, Y., Gilna, P. & Li, W. 2009. Identification of ribosomal RNA genes in metagenomic fragments. Bioinformatics 25:1338-1340.

Huddleston, J., Ranade, S., Malig, M., Antonacci, F., Chaisson, M., Hon, L., Sudmant, P. H., Graves, T. A., Alkan, C., Dennis, M. Y., Wilson, R. K., Turner, S. W., Korlach, J. & Eichler, E. E. 2014. Reconstructing complex regions of genomes using long-read sequencing technology. Genome Res. 24:688-696.

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M. C., Rattei, T., Mende, D. R., Sunagawa, S., Kuhn, M., Jensen, L. J., von Mering, C. & Bork, P. 2015. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res. 44:D286-D293.

Human Microbiome Project Consortium. 2012. Structure, function and diversity of the healthy human microbi-

ome. Nature 486:207-214.
Huson, D. H., Auch, A. F., Qi, J. & Schuster, S. C. 2007. MEGAN analysis of metagenomic data. Genome Res. 17:377-386.
Huson, D. H. & Weber, N. 2013. Microbial community analysis using MEGAN. Methods Enzymol. 531:465-485.
Huson, D. H. & Xie, C. 2014. A poor man's BLASTX: high-throughput metagenomic protein database search using PAUDA. Bioinformatics 30:38-39.
Ingman, M., Kaessmann, H., Pääbo, S. & Gyllensten, U. 2000. Mitochondrial genome variation and the origin of modern humans. Nature 408:708-713.
Janouškovec, J., Liu, S. L., Martone, P. T., Carré, W., Leblanc, C., Collén, J. & Keeling, P. J. 2013. Evolution of red algal plastid genomes: ancient architectures, introns, horizontal gene transfer, and taxonomic utility of plastid markers. PLoS One 8:e59001.
Jeck, W. R., Reinhardt, J. A., Baltrus, D. A., Hickenbotham, M. T., Magrini, V., Mardis, E. R., Dangl, J. L. & Jones, C. D. 2007. Extending assembly of short DNA sequences to handle error. Bioinformatics 23:2942-2944.
Jung, J. Y., Lee, S. H., Kim, J. M., Park, M. S., Bae, J. -W., Hahn, Y., Madsen, E. L. & Jeon, C. O. 2011. Metagenomic analysis of kimchi, a traditional Korean fermented food. Appl. Environ. Microbiol. 77:2264-2274.
Kanehisa, M. & Goto, S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28:27-30.
Kim, K. M., Park, J. -H., Bhattacharya, D. & Yoon, H. S. 2014*a*. Applications of next-generation sequencing to unravelling the evolutionary history of algae. Int. J. Syst. Evol. Microbiol. 64(Pt. 2):333-345.
Kim, S. Y., Yang, E. C., Boo, S. M. & Yoon, H. S. 2014*b*. Complete mitochondrial genome of the marine red alga *Grateloupia angusta* (Halymeniales). Mitochondrial DNA 25:269-270.
Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K. & Hugenholtz, P. 2008. A bioinformatician's guide to metagenomics. Microbiol. Mol. Biol. Rev. 72:557-578.
Lagesen, K., Hallin, P., Rødland, E. A., Stærfeldt, H. -H., Rognes, T. & Ussery, D. W. 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 35:3100-3108.
Lang, B. F., Laforest, M. -J. & Burger, G. 2007. Mitochondrial introns: a critical view. Trends Genet. 23:119-125.
Langmead, B. & Salzberg, S. L. 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9:357-359.
Laslett, D. & Canbäck, B. 2004. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res. 32:11-16.
Laslett, D. & Canbäck, B. 2008. ARWEN: a program to detect tRNA genes in metazoan mitochondrial nucleotide sequences. Bioinformatics 24:172-175.
Le Corguillé, G., Pearson, G., Valente, M., Viegas, C., Gschloessl, B., Corre, E., Bailly, X., Peters, A. F., Jubin, C., Vacherie, B., Cock, J. M. & Leblanc, C. 2009. Plastid genomes of two brown algae, *Ectocarpus siliculosus* and *Fucus vesiculosus*: further insights on the evolution of red-algal derived plastids. BMC Evol. Biol. 9:253.
Lee, J., Kim, K. M., Yang, E. C., Miller, K. A., Boo, S. M., Bhattacharya, D. & Yoon, H. S. 2016. Reconstructing the complex evolutionary history of mobile plasmids in red algal genomes. Sci. Rep. 6:23744.
Lee, J. -M., Boo, S. M., Mansilla, A. & Yoon, H. S. 2015. Unique repeat and plasmid sequences in the mitochondrial genome of *Gracilaria chilensis* (Gracilariales, Rhodophyta). Phycologia 54:20-23.
Lee, J. -Y., Yoo, C., Jun, S. -Y., Ahn, C. -Y. & Oh, H. -M. 2010. Comparison of several methods for effective lipid extraction from microalgae. Bioresour. Technol. 101(Suppl.):S75-S77.
Li, D., Liu, C. -M., Luo, R., Sadakane, K. & Lam, T. -W. 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 31:1674-1676.
Li, H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:1303.3997.
Li, H. & Durbin, R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25:1754-1760.
Li, R. Q. 2009. Short Oligonucleotide Analysis Package: SOAPdenovo 1.03. Beijing Genomics Institute, Beijing.
Liu, C., Shi, L., Zhu, Y., Chen, H., Zhang, J., Lin, X. & Guan, X. 2012. CpGAVAS, an integrated web server for the annotation, visualization, analysis, and GenBank submission of completely sequenced chloroplast genome sequences. BMC Genomics 13:715.
Liu, Y., Schmidt, B. & Maskell, D. L. 2011. Parallelized short read assembly of large genomes using de Bruijn graphs. BMC Bioinformatics 12:354.
Lohse, M., Drechsel, O. & Bock, R. 2007. OrganellarGenomeDRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. Curr. Genet. 52:267-274.
Loomis, E. W., Eid, J. S., Peluso, P., Yin, J., Hickey, L., Rank, D., McCalmon, S., Hagerman, R. J., Tassone, F. & Hagerman, P. J. 2013. Sequencing the unsequenceable: expanded CGG-repeat alleles of the fragile X gene. Genome Res. 23:121-128.
Lowe, T. M. & Eddy, S. R. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic

sequence. Nucleic Acids Res. 25:955-964.

Ludwig, M. & Bryant, D. A. 2011. Transcription profiling of the model cyanobacterium *Synechococcus* sp. strain PCC 7002 by Next-Gen (SOLiD™) sequencing of cDNA. Front Microbiol. 2:41.

Lunter, G. & Goodson, M. 2011. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res. 21:936-939.

Maréchal, A. & Brisson, N. 2010. Recombination and the maintenance of plant organelle genome stability. New Phytol. 186:299-317.

Markowitz, V. M., Chen, I. -M. A., Chu, K., Szeto, E., Palaniappan, K., Grechkin, Y., Ratner, A., Jacob, B., Pati, A., Huntemann, M., Liolios, K., Pagani, I., Anderson, I., Mavromatis, K., Ivanova, N. N. & Kyrpides, N. C. 2012. IMG/M: the integrated metagenome data management and comparative analysis system. Nucleic Acids Res. 40:D123-D129.

Martin, W. & Müller, M. 1998. The hydrogen hypothesis for the first eukaryote. Nature 392:37-41.

McKinnon, A. E., Vaillancourt, R. E., Tilyard, P. A. & Potts, B. M. 2001. Maternal inheritance of the chloroplast genome in *Eucalyptus globulus* and interspecific hybrids. Genome. 44:831-835.

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., Wilkening, J. & Edwards, R. A. 2008. The metagenomics RAST server: a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 9:386.

Mitra, S., Rupek, P., Richter, D. C., Urich, T., Gilbert, J. A., Meyer, F., Wilke, A. & Huson, D. H. 2011. Functional analysis of metagenomes and metatranscriptomes using SEED and KEGG. BMC Bioinformatics 12(Suppl. 1):S21.

Mocali, S. & Benedetti, A. 2010. Exploring research frontiers in microbiology: the challenge of metagenomics in soil microbiology. Res. Microbiol. 161:497-505.

Namiki, T., Hachiya, T., Tanaka, H. & Sakakibara, Y. 2012. MetaVelvet: an extension of Velvet assembler to *de novo* metagenome assembly from short sequence reads. Nucleic Acids Res. 40:e155.

Nguyen, L. -T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32:268-274.

Partek Inc. 1998. Partek software, version 2.0B7. Partek Inc., St. Peters, MO.

Paszkiewicz, K. & Studholme, D. J. 2010. *De novo* assembly of short sequence reads. Brief Bioinform. 11:457-472.

Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. 2010. IDBA: a practical iterative de Bruijn graph *de novo* assembler. *In* Berger, B. (Ed.) *Research in Computational Molecular Biology*. Springer Berlin, Heidelberg, pp. 426-440.

Peng, Y., Leung, H. C., Yiu, S. M. & Chin, F. Y. 2011. Meta-IDBA: a *de novo* assembler for metagenomic data. Bioinformatics 27:i94-i101.

Peng, Y., Leung, H. C., Yiu, S. M. & Chin, F. Y. 2012. IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28:1420-1428.

Pruitt, K. D., Tatusova, T. & Maglott, D. R. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 35:D61-D65.

Qiagen. 2015. CLC Genomics Workbench 8.0.3. Available from: https://www.qiagenbioinformatics.com/. Accessed May 2, 2016.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D. R., Li, J., Xu, J., Li, S., Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., Xie, Y., Tap, J., Lepage, P., Bertalan, M., Batto, J. -M., Hansen, T., Le Paslier, D., Linneberg, A., Nielsen, H. B., Pelletier, E., Renault, P., Sicheritz-Ponten, T., Turner, K., Zhu, H., Yu, C., Li, S., Jian, M., Zhou, Y., Li, Y., Zhang, X., Li, S., Qin, N., Yang, H., Wang, J., Brunak, S., Doré, J., Guarner, F., Kristiansen, K., Pedersen, O., Parkhill, J., Weissenbach, J., MetaHIT Consortium, Bork, P., Ehrlich, S. D. & Wang, J. 2010. A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464:59-65.

Qiu, H., Price, D. C., Weber, A. P. M., Reeb, V., Yang, E. C., Lee, J. M., Kim, S. Y., Yoon, H. S. & Bhattacharya, D. 2013. Adaptation through horizontal gene transfer in the cryptoendolithic red alga *Galdieria phlegrea*. Curr. Biol. 23:R865-R866.

Qiu, H., Price, D. C., Yang, E. C., Yoon, H. S. & Bhattacharya, D. 2015. Evidence of ancient genome reduction in red algae (Rhodophyta). J. Phycol. 51:624-636.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J. & Glöckner, F. O. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 41:D590-D596.

Redfern, O., Grant, A., Maibaum, M. & Orengo, C. 2005. Survey of current protein family databases and their application in comparative, structural and functional genomics. J. Chromatogr. B Analyt. Technol. Biomed. Life Sci. 815:97-107.

Rombel, I. T., Sykes, K. F., Rayner, S. & Johnston, S. A. 2002. ORF-FINDER: a vector for high-throughput gene identification. Gene 282:33-41.

Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M. & Nyrén, P. 1996. Real-time DNA sequencing using detection of pyrophosphate release. Anal. Biochem. 242:84-89.

Rothberg, J. M. & Leamon, J. H. 2008. The development and impact of 454 sequencing. Nat. Biotechnol. 26:1117-1124.

Samarasinghe, N., Fernando, S., Lacey, R. & Faulkner, W. B. 2012. Algal cell rupture using high pressure homogenization as a prelude to oil extraction. Renew. Energy 48:300-308.

Sanger, F., Nicklen, S. & Coulson, A. R. 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U. S. A. 74:5463-5467.

Schmidt, B., Sinha, R., Beresford-Smith, B. & Puglisi, S. J. 2009. A fast hybrid short read fragment assembly algorithm. Bioinformatics 25:2279-2280.

Scholz, M. B., Lo, C. -C. & Chain, P. S. G. 2012. Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. Curr. Opin. Biotechnol. 23:9-15.

Sharon, D., Tilgner, H., Grubert, F. & Snyder, M. 2013. A single-molecule long-read survey of the human transcriptome. Nat. Biotechnol. 31:1009-1014.

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M. & Birol, I. 2009. ABySS: a parallel assembler for short read sequence data. Genome Res. 19:1117-1123.

Stamatakis, A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688-2690.

Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D. R., Alberti, A., Cornejo-Castillo, F. M., Costea, P. I., Cruaud, C., d'Ovidio, F., Engelen, S., Ferrera, I., Gasol, J. M., Guidi, L., Hildebrand, F., Kokoszka, F., Lepoivre, C., Lima-Mendez, G., Poulain, J., Poulos, B. T., Royo-Llonch, M., Sarmento, H., Vieira-Silva, S., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Tara Oceans Coordinators, Bowler, C., de Vargas, C., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Jaillon, O., Not, F., Ogata, H., Pesant, S., Speich, S., Stemmann, L., Sullivan, M. B., Weissenbach, J., Wincker, P., Karsenti, E., Raes, J., Acinas, S. G. & Bork, P. 2015. Structure and function of the global ocean microbiome. Science 348:1261359.

Swindell, S. R. & Plasterer, T. N. 1997. SEQMAN. Contig assembly. *In* Swindell, S. R. (Ed.) *Sequence Data Analysis Guidebook*. Springer, New York, pp. 75-89.

Thomas, T., Gilbert, J. & Meyer, F. 2012. Metagenomics: a guide from sampling to data analysis. Microb. Inform. Exp. 2:3.

Tombácz, D., Sharon, D., Oláh, P., Csabai, Z., Snyder, M. & Boldogkői, Z. 2014. Strain kaplan of pseudorabies virus genome sequenced by PacBio single-molecule real-time sequencing technology. Genome Announc. 2:e00628-14.

Trapnell, C., Pachter, L. & Salzberg, S. L. 2009. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25:1105-1111.

UniProt Consortium. 2015. UniProt: a hub for protein information. Nucleic Acids Res. 43:D204-D212.

Vellai, T., Takács, K. & Vida, G. 1998. A new aspect to the origin and evolution of eukaryotes. J. Mol. Evol. 46:499-507.

Walsh, P. S., Metzger, D. A. & Higuchi, R. 1991. Chelex 100 as a medium for simple extraction of DNA for PCR-based typing from forensic material. Biotechniques 10:506-513.

Warinner, C., Speller, C., Collins, M. J. & Lewis, C. M. Jr. 2015. Ancient human microbiomes. J. Hum. Evol. 79:125-136.

Weinstock, G. M. 2012. Genomic approaches to studying the human microbiota. Nature 489:250-256.

Wyman, S. K., Jansen, R. K. & Boore, J. L. 2004. Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20:3252-3255.

Yang, E. C., Kim, K. M., Kim, S. Y., Lee, J., Boo, G. H., Lee, J.-H., Nelson, W. A., Yi, G., Schmidt, W. E., Fredericq, S., Boo, S. M., Bhattacharya, D. & Yoon, H. S. 2015. Highly conserved mitochondrial genomes among multicellular red algae of the Florideophyceae. Genome Biol. Evol. 7:2394-2406.

Zerbino, D. R. & Birney, E. 2008. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Res. 18:821-829.

Zeugin, J. A. & Hartley, J. L. 1985. Ethanol precipitation of DNA. Focus 7:1-2.

Zhao, Y., Tang, H. & Ye, Y. 2012. RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. Bioinformatics 28:125-126.