frontiers in Microbiology

REVIEW
published: 17 December 2021
doi: 10.3389/fmicb.2021.766364

Advances in Metagenomics and Its Application in Environmental Microorganisms

Lu Zhang[1†], FengXin Chen[1†], Zhan Zeng[2†], Mengjiao Xu[1], Fangfang Sun[1], Liu Yang[1], Xiaoyue Bi[1], Yanjie Lin[2], YuanJiao Gao[1], HongXiao Hao[1], Wei Yi[3], Minghui Li[1,2]* and Yao Xie[1,2]**

[1] Department of Hepatology Division 2, Beijing Ditan Hospital, Capital Medical University, Beijing, China, [2] Department of Hepatology Division 2, Peking University Ditan Teaching Hospital, Beijing, China, [3] Department of Gynecology and Obstetrics, Beijing Ditan Hospital, Capital Medical University, Beijing, China

OPEN ACCESS

Edited by:
Xuejun Ma,
Chinese Center for Disease Control and Prevention, China

Reviewed by:
Xiyang Wu,
Jinan University, China
Ahmed Abd El Wahed,
Leipzig University, Germany

***Correspondence:**
Minghui Li
wuhm2000@sina.com
orcid.org/0000-0003-3233-5473
Wei Yi
yiwei1215@163.com
orcid.org/0000-0003-4241-8205
Yao Xie
xieyao00120184@sina.com
orcid.org/0000-0003-4108-7037

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Food Microbiology, a section of the journal Frontiers in Microbiology

Received: 29 August 2021
Accepted: 18 November 2021
Published: 17 December 2021

Citation:
Zhang L, Chen F, Zeng Z, Xu M, Sun F, Yang L, Bi X, Lin Y, Gao Y, Hao H, Yi W, Li M and Xie Y (2021) Advances in Metagenomics and Its Application in Environmental Microorganisms. Front. Microbiol. 12:766364. doi: 10.3389/fmicb.2021.766364

Metagenomics is a new approach to study microorganisms obtained from a specific environment by functional gene screening or sequencing analysis. Metagenomics studies focus on microbial diversity, community constitute, genetic and evolutionary relationships, functional activities, and interactions and relationships with the environment. Sequencing technologies have evolved from shotgun sequencing to high-throughput, next-generation sequencing (NGS), and third-generation sequencing (TGS). NGS and TGS have shown the advantage of rapid detection of pathogenic microorganisms. With the help of new algorithms, we can better perform the taxonomic profiling and gene prediction of microbial species. Functional metagenomics is helpful to screen new bioactive substances and new functional genes from microorganisms and microbial metabolites. In this article, basic steps, classification, and applications of metagenomics are reviewed.

Keywords: metagenomics, shotgun sequencing, next-generation sequencing, third generation sequencing, antibiotics resistance genes, bioremediation

INTRODUCTION

Traditionally, research on microorganisms is based on cultured organisms, which has several disadvantages (Kellenberger, 2001). In 1998, Metagenome, also known as microbial environmental genome (Handelsman et al., 1998), is defined as "the genome of the total microbiota found in nature." Metagenome currently refers to the sum of the genomes of bacteria and fungi in environmental samples. Metagenomics is the study of a collection of genetic material (genomes) from a mixed community of organisms. In short, metagenomics is a new way to study microorganisms in a specific environment by using functional gene screening or sequencing analysis. The main areas of concern of metagenomics research are microbial diversity, population structure, genetic and evolutionary relationships, functional activity, and cooperative relationships, and relationship with the environment. Metagenomics research is developing rapidly in medicine, agriculture, environmental protection, and other fields.

This article reviews the development and application of metagenomics detection technology. The review is divided into three parts. The first part focuses on molecular biology technology from the extraction of all microbial genes, heterologous expression of gene fragments, to the screening of target genes or active products, and finally the sequencing of target genes. The second part

Tecan Ex. 2058, Page 1 of 15
IDT, Inc. v. Tecan Genomics - IPR2025-00016

introduces the main classification or branch of metagenomics from the research purpose or objects. The third part is about the application of metagenomics in various fields.

BASIC STEPS OF METAGENOMICS

Metagenomics is based on gene cloning. The basic procedure of metagenomics consists of the following steps. First, all genes in environmental microbial samples are extracted and enriched. Second, the genes are cloned into the vector, which is transformed into host bacteria to establish the metagenomic library. Finally, the metagenomic library is screened and analyzed. The extraction of metagenomic DNA and construction and screening of metagenomic library are essential.

Extraction and Enrichment of Metagenomic DNA

The traditional approach is to culture microorganisms and then extract the DNA, but 99% of the microorganisms in the environment cannot be cultured (Torsvik et al., 1990; Handelsman et al., 1998). Extracting high concentration and large fragment environmental microbial total DNA is the first and most critical step in building a metagenomic library. There are two key points to extracting metagenomic DNA. First, to extract all the genes of all microorganisms in the sample. Second, to keep the integrity and purity of the fragment. Sample collection must follow a strict procedure to extract as much DNA as possible from environmental microorganisms and maintain large DNA fragments.

According to whether or not the cells are isolated before DNA extracting, it can be divided into direct extraction and indirect extraction. Direct extraction uses physical (such as freeze-thaw) and chemical (such as the addition of protease or SDS) to lyse microorganisms directly to release microbial DNA (Gabor et al., 2003). The total DNA in all samples can be obtained by the direct extraction method, which is suitable for extracting 1–50 kb DNA. The direct extraction method is simple and efficient, but the purity is low (Schneegurt et al., 2003; Wang et al., 2011). The indirect extraction method requires isolating the microbial cells at first and then extracting DNA using a gentle method. The DNA obtained by indirect extraction is of high purity and suitable for the extraction of 20–500 kb DNA. The indirect extraction method is cumbersome and inefficient, which may cause the loss of some microbial DNA (Gabor et al., 2003).

The sample size depends on microbial concentration. The high density of microorganisms such as stool samples may require only an anal swab sample. To study the marine microbial community, which is a low-density microorganism, large numbers of water samples are captured and concentrated through filters. Samples need to be decontaminated. For example, humic acid in soil is often tightly bound to DNA, so humic acid needs to be removed during sample preparation (Daniel, 2005). When studying microorganisms in humans or animals, the DNA contamination in samples and host must be removed.

For DNA samples with low abundance, enrichment at the cell level and gene level can be carried out after DNA fragment extraction, but this may cause certain deviations (Probst et al., 2015). The common methods of gene enrichment are stable isotope probing (SIP) (Chen and Murrell, 2010), suppression subtractive hybridization (SSH) (Galbraith et al., 2008; Chew and Holmes, 2009), DNA chip technology, etc. (Avarre et al., 2007).

Construct Metagenome Libraries

In the construction of a metagenome library, microbial DNA should first be cut mechanically or digested into a certain size of DNA, and then randomly recombined with an appropriate clone vector, and finally transformed into suitable host cells. The goal is to isolate some target genes or preserve the entire DNA of an organism. The strategy of building microorganism DNA library changes with the research objective: low abundance genes or high abundance genes and small segments of a single gene or large segments of gene clusters in a metabolic pathway.

The vector is selected only if it is conducive to the amplification of the target DNA fragment and is convenient for screening after the transformation and expression. The common vectors include plasmid, cosmid, fosmid, or bacterial artificial chromosome (BAC), yeast artificial chromosome (YAC), and phage. Plasmids are suitable for cloning DNA fragments below 10 kb. Cosmid and fosmid are suitable for cloning large genes or multiple gene fragments (20–40 kb). BAC and YAC can be used to clone DNA fragments of 200–450 kb. The BAC shuttle vector, which can be duplicated and expressed in a variety of hosts, can be used to expand the host range (Kakirde et al., 2011).

The host cells are generally the ones with good stability of the recombinant vector, high efficiency of transformation, and effective expression of the target gene fragment. Usually, different host strains are selected according to different research purposes. *Escherichia coli* is easy to culture and is the most common host, but it is a prokaryotic host with limitations. Some new host cells are gradually applied, such as host *Streptomyces* (Rebets et al., 2017), *pseudomonas* (Craig et al., 2010), and *mycobacterium*.

There are many barriers to the study of foreign gene expression in metagenomics. For example, the target gene may be lost in the process of library construction, and the host cell cannot recognize the unique promoter, failing the expression of heterologous gene fragments. To solve this problem, multi-host expressions can be adopted. A research group expressed DNA fragments in 6 different bacteria (Craig et al., 2010). If the target gene is in cluster form, large segments of DNA need to be inserted to obtain complete metabolic enzymes and pathways. At present, the failure of DNA cloning in large fragments may be caused by the destructive effects of physical and chemical actions during the extraction process or the limited receptive capability of host bacteria.

Screening of Metagenome Library

Currently, there are three approaches to environmental metagenomic library screening: function-driven screening, sequence-driven screening, and substrate-induced gene expression screening (SIGEX). Other techniques include DNA stable-isotope probing (DNA-SIP) (Chen and Murrell, 2010) and fluorescence *in situ* hybridization.

Tecan Ex. 2058, Page 2 of 15
IDT, Inc. v. Tecan Genomics - IPR2025-00016

There are two kinds of functional screening. One is to screen the expression of active products of exogenous genes in host cells according to special characterization (such as color and plaque) on various selective media to obtain target clones (Apolinar-Hernández et al., 2016; Popovic et al., 2017). The other is to screen the host strain and its mutants based on their complementary growth characteristics under selective conditions (Cheng et al., 2017). The function-driven screening method is fast, simple, and does not rely on any known sequence information. Most biocatalysts, such as protease (Apolinar-Hernández et al., 2016), esterase (Popovic et al., 2017), are obtained by this method. The biggest disadvantage is that it depends on the expression of functional genes in foreign hosts, so the success rate of screening is very low. Thus, selecting the right host and cloning the full length of a gene or gene cluster is required. Functional analysis is suitable for gene screening of large fragment DNA libraries.

Sequence analysis usually uses oligonucleotide primers or probes to screen target genes by molecular biological methods such as PCR and gene hybridization. Only target clones with known sequences can be screened. Sequence analysis is suitable for screening highly conserved sequence enzyme genes, such as polyketide synthase gene (Ginolhac et al., 2004) and cellulase genes (Voget et al., 2003). Sequence analysis is also suitable for gene screening of small DNA fragments. The advantage of sequence analysis is that it need not rely on host cells to express the cloned genes, but the disadvantage is that it is unable to screen the unknown genes with completely different sequences from the existing genes.

Some new methods such as microarray, stable isotope labeling, and fluorescence *in situ* hybridization can also be applied to sequence screening. The DNA microarray technology, also known as gene chip technology, is based on nucleic acid hybridization (Avarre et al., 2007). First, DNA fragments are densely and orderly arranged on the solid support such as silicon wafer to form DNA microarray. Then the labeled samples are hybridized with the DNA microarray, and the genes are analyzed according to the hybridization sites and signal strength. DNA microarray technology can be used to study key genes in metabolic pathways. The DNA microarray technology has advantages of rapid detection and preliminary screening, but its specificity and sensitivity are only 1/100–1/10,000 of PCR method, so it is unsuitable for the detection of unknown functional genes (Call, 2005; Palka-Santini et al., 2009; Mauk et al., 2015).

The principle of substrate-induced gene expression screening is that metabolism-related genes or enzyme genes are usually expressed in the presence of substrate-induced conditions (Uchiyama et al., 2005; Yun and Ryu, 2005). SIGEX has several advantages, such as no modification of the substrate is required and functions of unknown enzymes and genes can be inferred from the substrate. SIGEX is a high-throughput screening method that is suitable for industrial use. But it is highly demanding on the structure and fitness of the target gene in the host and is only suitable for substrates that can enter the cytoplasm and target gene, which itself has a promoter.

What metagenomics need now are high-throughput screening methods. In 2015, a research team attempted to apply microfluidic chips integrating sample detection and automatic analysis (Colin et al., 2015), which is known as lab-on-chip (LOC). LOC only requires a sample of nanoliters or picoliter level, but the technical requirements are high. LOC requires advanced technologies such as chip sampling technology, microfluidic technology, and focusing electrophoresis technology. LOC implements various experimental steps in the laboratory, such as sample preparation, biochemical reaction, and sample detection, on a chip of several square centimeters. LOC meets the objective of overall miniaturization, automation, integration, and portability of sample testing.

Metagenomic Sequencing and Analysis Techniques

Rapid progress has been made in sequencing technologies, from shotgun sequencing to high-throughput, next-generation sequencing (NGS) (Mardis, 2008; van Dijk et al., 2014), and third-generation sequencing (TGS) (Branton et al., 2008).

Preparation of Metagenomic Sequencing Samples

Metagenomic DNA can be sequenced from DNA libraries obtained from target gene cloning or directly from samples. The purity and content of the DNA must be ensured before it is sequenced.

History of DNA Sequencing Technologies and Platforms

Second-generation DNA sequencing was established on PCR techniques. The first-generation sequencing is the shotgun sequencing based on the traditional Sanger sequencing technology (Olson, 1993; Bankier, 2001; Quince et al., 2017; **Figure 1**), and the sequencing using the dideoxy chain termination method is the gold standard of sequencing. The accuracy of shotgun sequencing is higher than those of second and TGS. It has a high accuracy of long reads from 700 to 1,000 bp and can deal with repeated sequences well. The disadvantage is that only one template can be detected at a time, which is slow and time-consuming.

Next-generation sequencing uses large-scale parallel sequencing technology, which can simultaneously synthesize millions of complementary chains of sequencing templates and acquire sequence data. High-throughput and low-cost NGS is used worldwide in large-scale genome sequencing and resequencing. The NGS mainly used sequencing by synthesis (SBS) technology and sequence by ligation (SBL). Roche has developed GS FLX and GS Junior sequencing platforms with 454 sequencing technology and SBS (Marsh, 2007; Harrington et al., 2013; **Figure 2**), which can read more than 400 bp. Illumina used SBS and a 3′ blocked reversible terminator for sequencing (Shendure and Ji, 2008; Ansorge, 2009; **Figure 3**). Based on this technology, the Genome Analyzer platform reads 125 bp. Illumina sequencing has become the mainstream product of NGS. Applied Biosystems, InC. (ABI) used SBL to develop the

Tecan Ex. 2058, Page 3 of 15
IDT, Inc. v. Tecan Genomics - IPR2025-00016

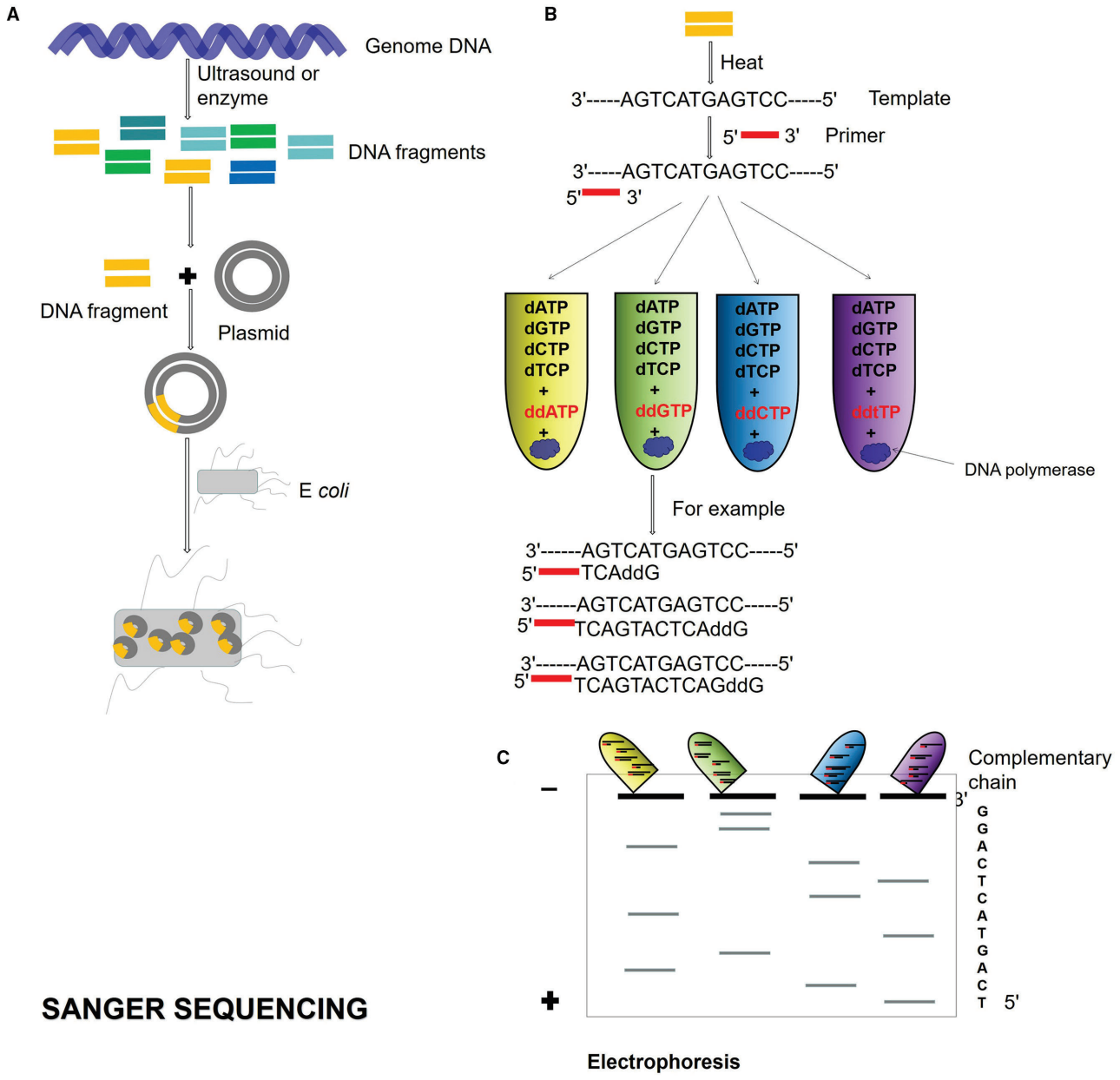


FIGURE 1 | Sanger sequencing process. **(A)** Genome DNA is cut into different DNA fragments by ultrasound or enzyme, connected into a plasmid, and transformed into *Escherichia coli*. **(B)** The Dideoxy chain termination method is used for sequencing. The reaction system is a mixture of template, primer, DNA polymerase, and dNTPs. Four kinds of ddNTP (ddATP, ddTTP, ddCTP, and ddGTP) with radioisotope marks are, respectively, added to four reaction systems. If ddNTP is incorporated into the oligonucleotide chain, the 2′ and 3′ of ddNTP do not contain hydroxyl groups and cannot form phosphodiester bonds, then the extension of the DNA chain is terminated. **(C)** Four PCR reaction system products are added to gel electrophoresis, and autoradiography was used to determine the DNA sequence according to the location of electrophoresis.

ABI Solid Sequencer (Mardis, 2008; Shendure and Ji, 2008; Ansorge, 2009). The advantage of sequencing by oligonucleotide ligation and detection (SOLiD) (**Figure 4**) is that each base can be detected two times by two-base encoding, and the detection accuracy is higher than the other NGS. The disadvantage is that only 40 bp can be read, which is only suitable for SNP mutation detection. In addition, ion torrent sequencing technology developed by Life Technologies/Thermo Fisher Scientific adopts semiconductor chip technology (**Figure 5**). Four kinds of dNTPs sequential flow through the chip in turn, if one dNTP is complementary to the DNA molecule in a specific microhole of the chip, the dNTP is synthesized into the DNA strand. Subsequently, the hydrogen ions are released as the DNA strand stretches, causing the change of the pH. Finally, voltage changes

Tecan Ex. 2058, Page 4 of 15
IDT, Inc. v. Tecan Genomics - IPR2025-00016

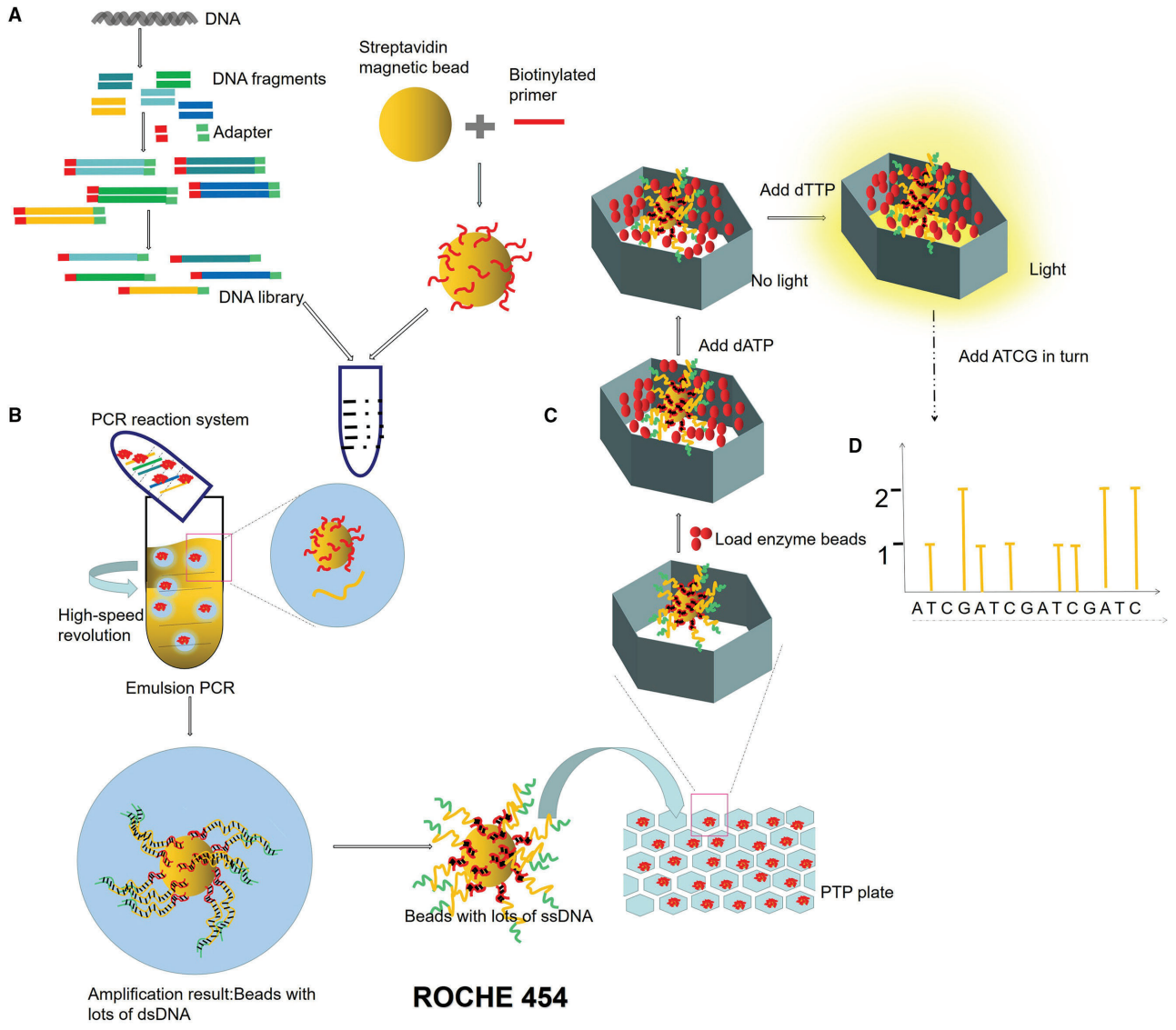


FIGURE 2 | ROCHE 454 sequencing process. **(A)** Construction of micro-reaction system. DNA library is constructed by adding an adapter at the ends of DNA fragments. The magnetic beads combined with primers are mixed in the PCR reaction system (DNA fragments, enzymes, dNTPs). **(B)** Emulsion PCR. PCR microreaction systems are formed by injecting water (PCR mixture) into the oil, and each system contained only one template and one bead. **(C)** Pyrophosphate sequencing. Pyrophosphate technology is an enzyme cascade chemiluminescence reaction in the same reaction system catalyzed by four enzymes. In each circle of the sequencing reaction, only one dNTP is added. If it just matches the next base of the DNA template, it will be added to the 3′ end of the sequencing primer under the action of DNA polymerase and release a molecular PPi at the same time. Under the action of ATP sulfurylase, luciferase, and apyrase, an enzyme cascade chemiluminescence reaction is triggered by the PPi. **(D)** Light signal reception.

are caused. Ion Torrent has the advantages of low cost, only an electrical signal detection system, and no light/fluorescence detection system (Pennisi, 2010).

Third-generation sequencing also called *de novo* sequencing technology, is a milestone in sequencing technology, which is based on single-molecule sequencing technology (SMS) (Xu et al., 2009) and large-scale parallel sequencing technology. There are two methods for the TGS. The first is single-molecule fluorescence sequencing. The representative technology is the American Helicos SMS (True single molecular sequencing, TSMS) and single molecular real-time sequencing (SMRT) technology from Pacific Bioscience (Xu et al., 2009; Rhoads and Au, 2015; Midha et al., 2019). SMRT Principle is SBS. DNA polymerase is anchored to the nanopore. The four fluorescent-labeled dNTPs emit different light in the base-pairing stage. The type of entering base can be determined according to the light wavelength and peak value. The phosphate group of dNTP is labeled with fluorescence in SMRT. The long reads are determined by DNA polymerase activity. SMRT requires a high-efficient optical detection system.

Tecan Ex. 2058, Page 5 of 15
IDT, Inc. v. Tecan Genomics - IPR2025-00016

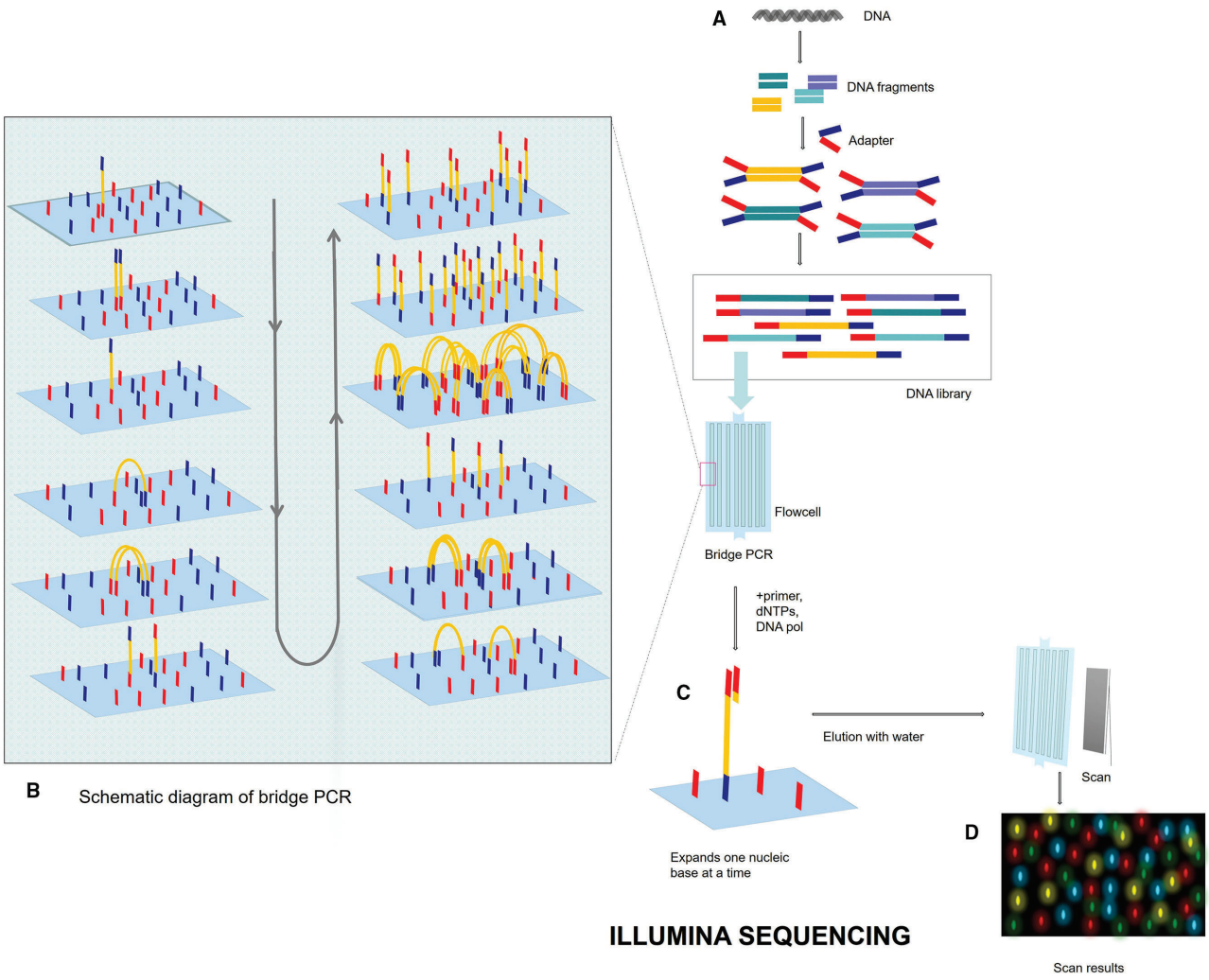


FIGURE 3 | Illumina sequencing process. **(A)** DNA library. Breaks the genome DNA to form DNA fragments, adds adapters at both ends, and constructs a single-stranded DNA (ssDNA) library. After adding DNA library into flowcell, DNA fragments will be attached to the surface of flowcell when passing through flowcell. **(B)** Bridge PCR. Each DNA fragment is clustered in its position. After bridge amplification, each cluster contains many copies of the ssDNA template. **(C)** SBS: DNA polymerase, primers, and four dNTPs with specific fluorescence are added to the reaction system at the same time. The 3′ end of these dNTPs is connected with an azide group, which can block the incorporation of the next base, so only one base can be extended at a time. After washing off the remaining dNTP and enzyme with water, photos are scanned. After scanning, add chemical reagent, cut off the azide group and quenching fluorescence, and next cycle. **(D)** Fluorescent signal reception.

The second is nanopore sequencing, which is based on electrical signals and single-molecule long-read detection. The representative company is Oxford Nanopore (Branton et al., 2008; Midha et al., 2019; **Figure 6**). The principle of nanopore sequencing is to use electrophoresis technology to drive individual molecules through nanopores one by one for sequencing. The electrification properties of each ATCG base are different during nanopore sequencing, and the difference of electrical signals detected can correspond to different bases. Compared with NGS, TGS is single-molecule sequencing. TGS has no PCR amplification preference or GC preference and can be sequenced directly without PCR amplification. The TGS has a very long read length, with an average read length of 10–15 kb and a maximum read length up to 40 kb. TGS can directly detect DNA methylation. The disadvantages of TGS are mainly shown in the following three aspects. First, sequencing errors occur randomly, with error rates of each base as high as 15%, which can be overcome and corrected by multiple sequences. Second, the cost of sequencing is high. Third, SMRT sequencing technology requires enzymes, and currently, better enzymes with much more activity and stability are needed.

When selecting different sequencing methods, there are many factors to consider, including the length of the target gene, the requirement of sequencing accuracy, the purpose of sequencing (whole genome sequencing, resequencing, or SNP), and the gene abundance of the sample. The sequencing platform should

Tecan Ex. 2058, Page 6 of 15
IDT, Inc. v. Tecan Genomics - IPR2025-00016

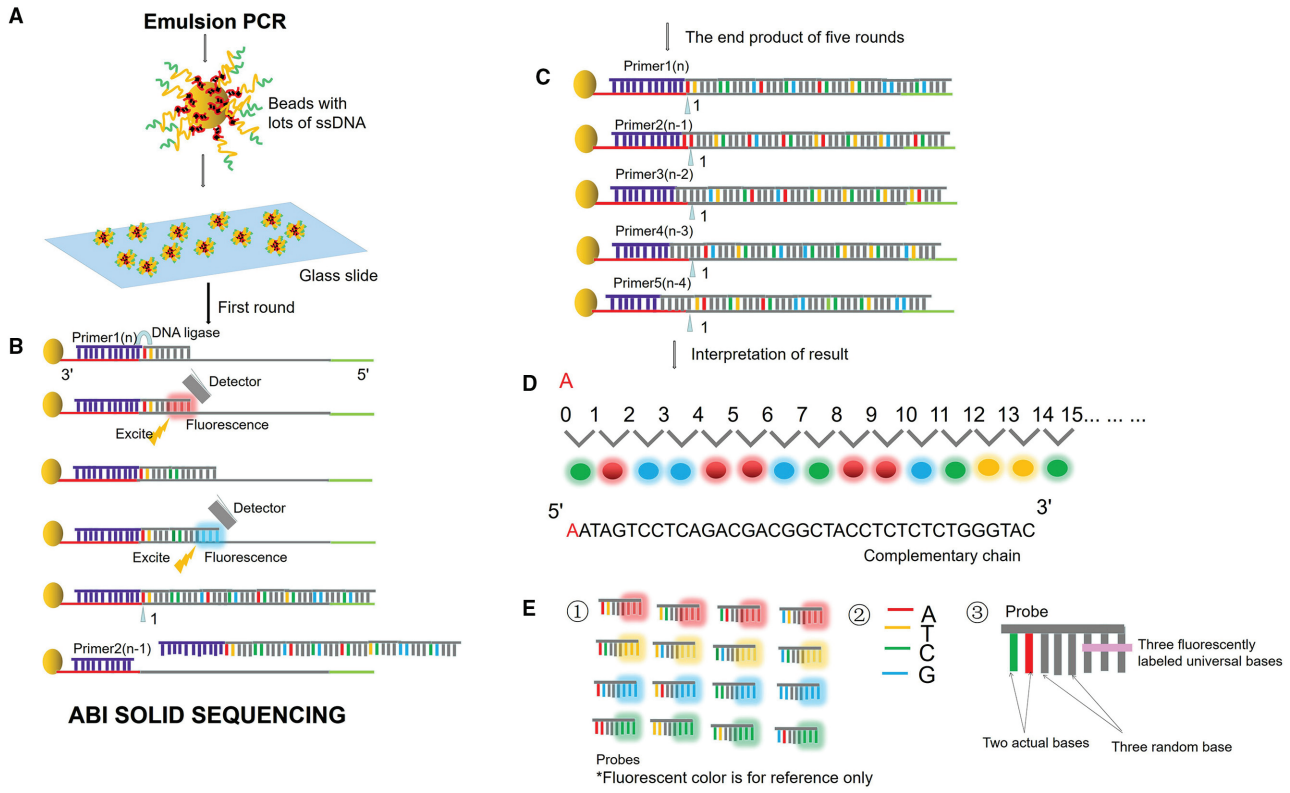


FIGURE 4 | ABI sequencing by oligonucleotide ligation and detection (SOLiD). **(A)** Emulsion PCR. After the emulsion is amplified, a large number of amplified products from the DNA template are attached to the magnetic beads. Then, the magnetic beads are deposited on the glass slide. **(B)** SLB. The first circle of sequencing reaction includes: the sequencing primer is combined with the template; the probe is combined with the template; the ligase connects the sequencing primer and the probe, excites the probe fluorescence, and removes the last three bases of the probe. Then plus the second probe, connects the two probes with the ligase, excites the probe fluorescence, and removes the last three bases of the probe. Repeat the above reaction to synthesize the whole chain. **(C)** Solid template shift. After cutting off the whole synthetic ligation, add sequencing primer n-1, repeat the above steps and start the second circle of reaction. Unidirectional solid sequencing requires five circles of sequencing reactions. The primers are n, n-1, n-2, n-3, and n-4, respectively, and the bases at each position are detected two times. **(D)** Fluorescent signal reception and interpretation of results. **(E)** Two-base encoding principle. ① "two-base encoding" specifies the corresponding relationship between different base pairs and four probe colors in the coding region (e.g., red fluorescence represents AT, TC, CA, GT). ② Four bases. ③ The probe is an 8-base single-chain fluorescent probe. The base pairs at positions first and second of the probe are determined, positions from third to fifth are random bases, which can be any of the four bases of "A, T, C, and G," and positions from sixth to eighth mark four fluorescent dyes of different colors.

consider factors such as sequencing speed, read length, cost, error rate, methylation, and others. At present, NGS is the mainstream sequencing method, but TGS is still in the initial stage of development due to its high cost and high error rate in a single base. In practical applications, the advantages in long reads of 10–15 kb with low assembly difficulty of TGS and the characteristics of high quality in short reads of NGS can be utilized (Kumar et al., 2019; **Table 1**).

Different sequencing methods have their advantages and disadvantages. Shotgun sequencing is still in use today and serves as the gold standard because of its high accuracy. NGS has parallel sequencing after PCR amplification of the template and shows high throughput. At present, the market share of NGS is the highest. Illumina is currently the most widely used platform in the NGS. The Roche 454 sequencing technology has been withdrawn from the market due to its high cost. The TGS sequence directly, and the read length is longer, but the error rate is higher. Therefore, below are some suggestions for sequencing method selection. (1) For SNP detection of genome or insertion DNA sequencing of positive clones, shotgun sequencing is recommended. (2) If whole genome sequencing is carried out, Illumina is a better choice among NGS. (3) If cost is not considered, the combination of NGS and TGS is the best choice, which has high-quality short reads of NGS to correct wrong bases generated by TGS. The future direction of sequencing is to expand the microbial reference gene database and increase the accuracy of long reads.

Processing of DNA Sequencing Information Data

After a large number of reads are obtained by preliminary metagenomic DNA sequencing, the reads should be assembled into a complete genome sequence by software, and taxonomic profiling should be obtained at the same time. Then gene prediction and function annotation are carried out.

Tecan Ex. 2058, Page 7 of 15
IDT, Inc. v. Tecan Genomics - IPR2025-00016

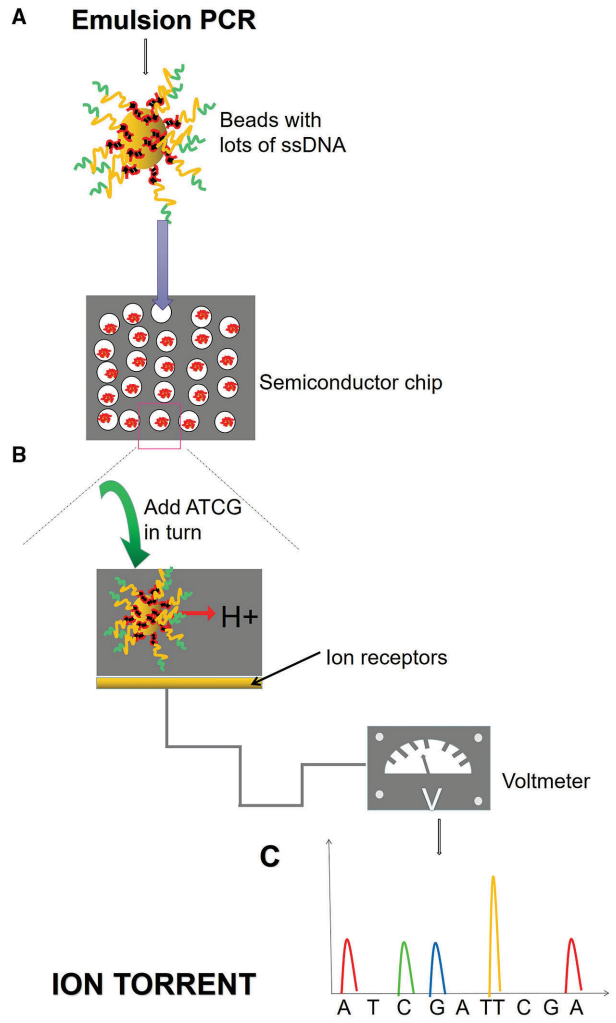


FIGURE 5 | Ion torrent process. **(A)** Emulsion PCR. After the emulsion is amplified, a large number of amplified products from the DNA template are attached to the magnetic beads. Magnetic beads are fixed in the micropores of the semiconductor chip, and each micropore can only contain one magnetic bead. **(B)** SBS. Fix the ssDNA in the micropore of the semiconductor chip and add different dNTPs in turn. When the DNA polymerase polymerizes the nucleotide onto the DNA strand in the extension, it will release an H^+, causing the change of pH value and voltage in the reaction system. **(C)** The electrical signal is accepted and converted into base information.

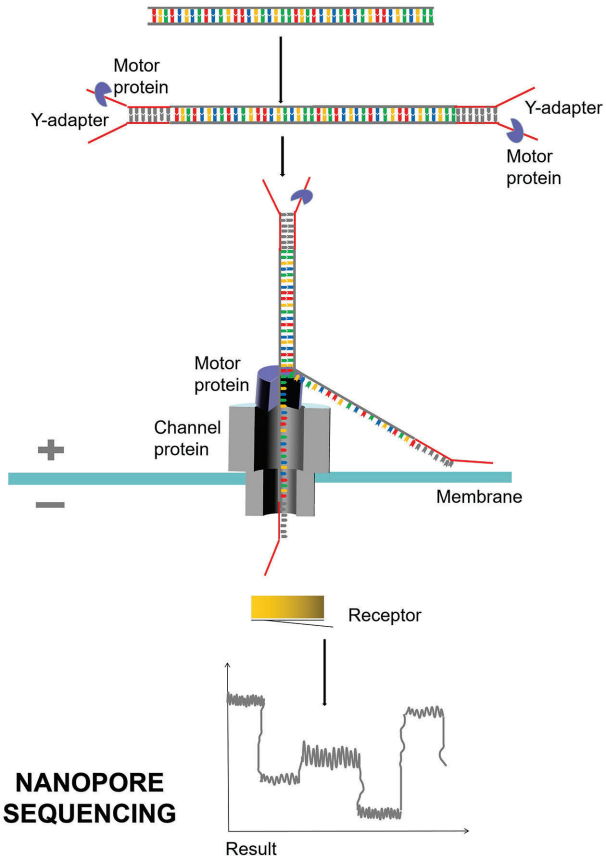


FIGURE 6 | Nanopore sequencing process. Nanopore sequencing mainly uses nanopore protein, also known as reader protein, which is a transmembrane channel protein. Different voltages are applied on both sides of the membrane. DNA double strand is decoiled under the traction of motor protein, and a single molecule is driven by electrophoresis to pass through reader protein one by one. Different electrical signals are produced and detected when the bases of ATCG pass through.

Genome sequencing mainly includes quality control (QC), sequence assembly, sequence binning, taxonomic profiling, gene prediction, and function annotation.

Quality control mainly includes two parts: de-noising and de-host sequence. It is necessary to establish a database of host sequences to be used for comparing and removing by sequence alignment (Schmieder and Edwards, 2011; Hong et al., 2014). Data analysis includes assembly-based and assembly-free metagenomic profiling. If the microbial genetic information database is complete, the post-QC data can be directly analyzed without assembly. There are currently two approaches to sequence assembly, relying on reference sequence assembly and *de novo* assembly (Iqbal et al., 2012; Quince et al., 2017; Breitwieser et al., 2019; Olson et al., 2019; Anyansi et al., 2020). Binning refers to the sorting and clustering of metagenome sequenced segments by species. Binning operations may result in bins of unknown microbial genome sequences that cannot be cultured in the laboratory. Binning can be divided according to different types of aggregation sequence: reads binning before assembly, contig binning after assembly, and gene binning based on annotation (Alneberg et al., 2014; Cleary et al., 2015; Quince et al., 2017). Contig binning usually classifies contigs with similar composition and consistent abundance patterns into the same species. In addition, contigs binning has a better effect on contigs with higher abundance and is often applied by many software. It is possible to obtain low abundance species with reads binning, and even cluster species with abundance as low as 0.00001% (Cleary et al., 2015). Gene binning method is used in many published metagenome-wide association studies, and gene abundance is calculated (Wang and

Tecan Ex. 2058, Page 8 of 15
IDT, Inc. v. Tecan Genomics - IPR2025-00016

TABLE 1 | Performance COMPARISON of seven sequencing instruments.

Platform	Sequencing rules	Advantages	Disadvantages	Signal received
Shotgun sequencing	Dideoxy chain termination method, capillary electrophoresis	Accuracy 99.999%, Reads 1,000 bp, High accuracy of processing repeated base, *De novo*	Low-throughput, high cost	Fluorescence
Roche 454	Pyrophosphate sequencing, SBS, Emulsion PCR	Accuracy 99%, Reads 400–600 bp, Parallel sequencing	Base insertion and deletion errors	Light
Illumina	SBS, Bridge PCR, Reversible terminator	Accuracy 99%, Reads 75–300 bp, Parallel sequencing	Base substitution error	Fluorescence
ABI solid	SBL, Emulsion PCR, Solid phase template shift, Two base encoding	Accuracy 99.94%, Reads 40 bp, Parallel sequencing, each base for sequencing twice	Prone to continuous base interpretation errors	Fluorescence
ABI/Ion torrent	Semiconductor sequencing, SBS, Emulsion PCR	Accuracy 99%, reads 200 bp, Parallel sequencing No light/fluorescence detection system	Difficult to identify homopolymers >8 bases, Base insertion and deletion errors	Electrical signal
Pacific bioscience	SMRT, SBS	Sequencing without PCR Detecting methylated bases average reads 10–15 kb	Random errors (5–15%) for a base, repeated sequencing can correct random errors	Fluorescence
Oxford nanopore	Nanopore sequencing Single-molecule sequencing, Direct detection of bases	Sequencing without PCR, No light/fluorescence detection system, Detecting methylated bases average reads 10–15 kb, longest to 40 kb	Random errors (5–15%) for a base, Repeated sequencing can correct random errors	Electrical signal

SBS, sequencing by synthesis; SLB, sequence by ligation; ssDNA, single-stranded DNA; SMRT, single molecular real time sequencing. Solid (sequencing by oligonucleotide Ligation and Detection).

Jia, 2016). Currently, gene binning and contig binning are most widely used.

If the sequence assembly and sequence sorting are not correct, it is often impossible to identify new microorganisms. In the metagenomic sequencing report, species annotation and abundance statistics are required. The former refers to the analysis of species in the phylum, while the latter refers to the relative abundance of various microorganisms.

Metagenomic gene prediction is to use prediction tools to identify potential open reading frames (0RF) in genomic libraries and to predict the structure and function of unknown sequences of genes. Metagenomic gene prediction includes homology prediction and *de novo* prediction. Homology prediction can be obtained by alignment of homologous sequences with gene database, but the disadvantages are dependent on the existing gene database, and new genes cannot be annotated. *De novo* prediction is based on the sequence characteristics obtained by sequencing, which can predict both known and unknown genes based on different characteristics in the coding region and the non-coding region.

Function annotation is to compare a gene or protein sequence in a specific functional database, associate a gene or protein with a specific function, and help understand the relevant metabolic pathway. Through gene prediction and functional annotation analysis, the information of metabolic pathways can be obtained. To improve the quality of sequencing analysis, Bowers RM formulated Genome Sketch Standards of Metagenome-Assembled Genome (MAG) in 2017. High-quality drafts require gene integrity >90%, contamination <5%, and inclusion of 16S, 23S, and other highly conserved genes in assembly quantity (Bowers et al., 2017).

Commonly Used Software for Data Analysis of Sequencing Information

Data analysis relies on continuous improvement of existing databases and analysis software. Each process such as QC, binning, assembly, gene prediction, and function annotation has its own software with its own advantages. Megan is selected for the analysis and comparison of multiple groups of metagenome data, while the software GeneMark and metagenemark are often selected for gene prediction. In 2017, researchers compared and evaluated several software regarding assembly, binning and taxonomic profiling, to help researchers to select appropriate software (Sczyrba et al., 2017). The assembly software (MEGAHIT, Minia, and Meraga) performs well, assembling significant portions of the genome over a wide range and reducing misassembly. The MaxBin2.0 genome binner has the highest good performance (70–80% integrity, >92%

Tecan Ex. 2058, Page 9 of 15
IDT, Inc. v. Tecan Genomics - IPR2025-00016

purity), followed by other programs (MetaBAT, CONCOCT, and MetaWatt). The taxonomic software of PhyloPythiaS and Kraken have good results before the family level.

CLASSIFICATION OF METAGENOMICS

Metagenomics technology does not rely on isolation and culture but directly extracts the genetic material of microorganisms from the natural environment. Then, the characteristics of microbial communities can be studied through genetic analysis. Metagenomics is classified according to research purpose and research object.

Metagenomics Can Be Divided Into Functional Metagenomics and Sequencing Metagenomics for Research Purposes

Functional Metagenomics Is Mainly Used for the Discovery of New Bioactive Substances and the Screening of New Microbial Functional Genes

Metagenomics technology has found many new genes, including the biocatalyst gene (Voget et al., 2003; Uchiyama et al., 2005), polyketide synthase coding gene (Ginolhac et al., 2004), antibiotic resistance genes (ARGs) (Riesenfeld et al., 2004), and others. In 2004, researchers used functional metagenomics to find that microorganisms in the soil had aminoglycoside and tetracycline antibiotic resistance genes that were significantly different from the original gene sequences in the database (Riesenfeld et al., 2004).

Active products produced by heterologous expression of metagenomic library genes, such as antibiotics, antitumor substances, and enzymes, are screened, sequenced, and analyzed. Metagenomics has great potential for application in the discovery of new enzymes without culture. Various enzymes found in metagenomes may be derived from microbial metabolism or microbial secondary metabolites. In the last two decades, metagenomics has shortened the discovery cycle of new enzymes, such as esterase, glycosidase, and oxido-reductase. Zhang et al. (2007) screened a specific low-temperature lipase at the South Pole. A strain showing extracellular lipolytic activity toward tributyrin was isolated from the deep-sea sediment of Prydz Bay and identified as a *Psychrobacter* species. The genomic DNA library was constructed from the strain with the highest lipolytic activity and the recombined plasmids were transformed into *E. coli*. The clones with lipase activity from 28,000 clones were identified. The insert DNA sequence of the selected clone was sequenced and an open reading frame of 954 bp coding was identified for a lipase gene by BLAST analysis. The effects of temperature on enzyme activity indicated that it was a typical cold-adapted enzyme.

There are many advantages to functional metagenomics. (1) Functional metagenomics can obtain target genes or active products through screening without culture conditions. (2) The development of new drugs was accelerated from marine and extreme environmental microbial resources. (3) Functional metagenomes help us to discover new members of existing enzyme families or enzymes that function only under specific physicochemical conditions, which can be better used in industry. (4) Functional metagenomics combined with metabolomics helped study the C, N, and S cycle metabolism of microorganisms in the environment.

There are many challenges in functional metagenomics. (1) The living conditions of the microbial community in the environment are complex, and the genes of the microorganisms with low abundance may not be extracted. (2) There is a loss of DNA fragments in the gene cloning process. (3) The frequency of heterologous expression of foreign genes is low. (4) The existing screening methods are limited and cannot meet all the enzyme screening requirements. (5) Screening is low efficiency, and only a few positive clones can be selected from tens of thousands of clones. Therefore, we need to develop higher throughput, high sensitivity, rapid, and simple screening methods in the future. (6) Due to temperature, pH, and other limitations, only a small part of the newly discovered enzymes can be used for industrial needs. (7) Current functional metagenomic studies are geographically uneven. It is mainly concentrated in developed countries. The vast microbial resources of developing countries are underexploited.

Sequencing Analysis of Metagenomics

To study the diversity and abundance of microbial community composition, species identification by microbial taxonomy is required. The traditional identification of microbial species composition is performed using the 16S/18S ribosomal DNA identification (16S/18S rDNA) in prokaryotic and eukaryotic cells, respectively; 16S rDNA sequencing is a commonly used molecular clock for bacterial taxonomic studies, but 16S rDNA provides limited information and may not be able to successfully identify species. Metagenomic sequencing is capable of identifying the species or the strain of a microbial community. For example, a research team used a combination of metagenome sequencing and single-cell sequencing to identify an extremely thermophilic bacterium *Candidatus kryptonia*, a new bacterium that could not be detected by 16S rDNA (Eloe-Fadrosh et al., 2016).

Metagenomic sequencing can be applied to bacteria, fungi, virus, and other microorganisms. In the absence of any available sequence information, metagenomic sequencing mainly uses *de novo* sequencing for microbial identification. Bioinformatics analysis methods were used to obtain the genomic sequence of microorganisms, such as COVID-19 (Wahba et al., 2020; Wu et al., 2020). In the presence of reference sequences, metagenome sequencing can focus on genetic variation information such as SNP, insertion/deletion of microbial significance, for example, the detection of YMDD (tyrosine—methionine—aspartic acid—aspartic acid) caused by point mutation of HBV polymerase gene region. At present, one of the hotspots of metagenomics research is disease-related human whole-exome sequencing (Wei et al., 2011).

There are many advantages to studying sequencing metagenomics. (1) Microorganisms in extreme environments are difficult to culture. Metagenomics can find such unculturable microorganisms. (2) Species, genetic, and evolutionary information of microorganisms can be obtained by studying

Tecan Ex. 2058, Page 10 of 15
IDT, Inc. v. Tecan Genomics - IPR2025-00016

the diversity of the microbial community. (3) New pathogenic microbial evidence can improve the current diagnosis of medical infections. (4) In terms of emerging infectious diseases, the target of pathogenic microorganisms can be quickly identified.

The sequencing of metagenomics is facing many challenges. (1) The DNA of environmental microorganisms cannot be extracted completely. (2) Sequencing process may miss low-abundance microorganisms. (3) There is no "gold standard" for sequencing processing software. (4) Because of the imperfection of the microbial database, many sequencing data might not be analyzed, including species annotation and functional analysis. (5) A sequencing platform with long reads and high accuracy is urgently needed. (6) The determination of the abundance of microorganisms in a microbial community is influenced by many factors.

According to the Microorganism Studied, Metagenomics Can Be Divided Into Bacterial Metagenomics, Fungal Metagenomics, and Viral Metagenomics

At present, bacterial metagenomics is the most studied one, while the development of sequence analysis metagenomics is the fastest-growing one. Fungal and viral metagenomics are mainly used in sequence analysis metagenomics. Virus metagenomics can be used for systematic analysis and identification of overly scattered and low abundance viruses from the environment, a virus can be obtained directly from the environment without isolation and culture. With the advance in metagenome sequencing, most of the new information on virus databases comes from metagenomics rather than cultures. The Bundibugyo Ebola virus was discovered in 2007 in patients having Ugandan hemorrhagic fever using viral metagenomics (Towner et al., 2008).

APPLICATION FIELDS OF MICROORGANISM METAGENOMICS

Applications of environmental metagenomics include the study of microorganisms in oceans, soils, deep oceans, glaciers, craters, and other environments. Metagenomics also has many applications in the medical field, such as the identification of pathogenic microorganisms in the intestinal tract, bloodstream infections, lung infections, central nervous system infections, and other infections. At present, metagenomic research focuses on antibiotic-resistant bacteria (ARB) and antibiotic-resistant genes (ARGs) (Gillespie et al., 2002; Riesenfeld et al., 2004; Arango-Argoty et al., 2018; Karkman et al., 2018; Nnadozie and Odume, 2019), biocatalysts, drugs, and others.

Environmental Object of Microorganism Metagenomics Continues to Expand

Metagenomics is used in research related to agriculture, biology, pollution control, energy, environment, atmosphere, and other fields. Microbial subjects range from land to sea and extreme environments. There are huge microorganisms in the soil, and metagenomics studies have shown that soil is a potential cornucopia of antibiotics and antifungal agents. In 2002, the antibiotics Turbomycin A and Turbomycin B with broad-spectrum antibacterial activity were screened and named from the soil metagenomic library (Gillespie et al., 2002).

Compared with soil microorganisms, marine microorganisms have a broad prospect. Oceans, rivers, and lakes account for more than 70% of the earth's area, presenting great differences in the species, genes, and biological characteristics of microorganisms in the water. Marine microorganisms have special metabolites and metabolic regulation mechanisms. Metagenomic techniques are used to study the synthetic pathways of various marine secondary metabolites. A good example is marine actinomycetes-related metabolites (polyketone biosynthesis gene cluster), which can be used as a source of new drugs. In terms of ARGs, ARB levels in most river and lake systems are high, in particular, vancomycin-resistant *enterococcus* (VRE) and methicillin-resistant *Staphylococcus aureus* (MRSA) (Nnadozie and Odume, 2019).

The majority of extremophiles that survive in various types of extreme environments cannot be cultured artificially. In such extreme environments, it is more likely to obtain unique enzymes and abundant microbial resources. The active compounds can be screened by direct sequencing, sequence alignment with database, and functional analysis, some of which may be selected out for commercial development (Zhang et al., 2007).

Metagenomics for Prevention and Control of Environmental Pollution

Bioremediation is a new bioengineering technology that uses the natural purification ability of environmental microorganisms. Bioremediation is expected to reduce pollutants or make them harmless by using the metabolic activities of microorganisms in the treatment system. Bioremediation needs several conditions: (1) Pollutants are organic and can be degraded by microorganisms. (2) The microorganisms possess certain metabolic activities. (3) The environment is suitable for degradation, meanwhile, for the maintenance of microbial activity. Metagenomics may help explore new functional microorganisms and genes, and construct and screen out new microbial strains with high degradation efficiency, broad applicability, and stable expression.

In the sewage treatment project, not only the organic matter in the sewage needs to be degraded but also the nitrogen and phosphorus elements in the sewage need to be removed. The wastewater treatment process deals with diverse microbial populations, such as anaerobic ammonium oxidation bacteria, phosphorus-accumulating bacteria, and electrochemically active bacteria. Metagenomic studies of water microorganisms contribute to biological nitrogen removal processes and enhanced biological phosphorus removal. Functional genes have been found in microorganisms that can degrade pesticides, plastics, polycyclic aromatic hydrocarbons, petroleum hydrocarbons, and other organic pollutants, which can play an important role in environmental bioremediation. Researchers screened a variety of new proteolytic enzymes from yard compost to degrade pesticides (Lämmle et al., 2007). In

Tecan Ex. 2058, Page 11 of 15
IDT, Inc. v. Tecan Genomics - IPR2025-00016

2005, a researcher screened benzoate anaerobic degradation genes from Black Sea sediment, which contributes to the anaerobic degradation of aromatic compounds (Krüger et al., 2003). Microbial electrochemical systems (MES) use electroactive bacteria enriched in anodes to recover electrical energy directly from organic waste (Wang and Ren, 2013; Venkata Mohan et al., 2014). As an emerging sewage treatment technology, MES has been developed rapidly.

Metagenomics Are Used in the Medical Diagnosis of Pathogenic Microorganisms

Metagenomics findings suggest that the intestinal microbiome is associated with a variety of diseases, such as non-alcoholic fatty liver disease (Aron-Wisnewsky et al., 2020), autoimmune diseases (Svoboda, 2021), and tumors (Andrews et al., 2021). Metagenomics can be used to detect drug-resistant genes for pathogens and monitor outbreaks of infectious diseases in hospitals and communities. In 2021, the Laboratory Medicine Branch of the Chinese Medical Association published an expert consensus, recommending the use of NGS on the samples of suspected infection sites in case that no evidence of etiology is obtained by routine biochemical examination or microbial culture, and empirical anti-infection therapy is ineffective. It is recommended to select sequencing metagenomics when the cause and site of infection are unknown and the pathogenic microorganisms cannot be determined by the existing detection technology. NGS can deal with a wide range of pathogenic microorganisms, including bacteria, viruses, fungi, and parasites. Samples may be collected in case of suspected central nervous system infection, respiratory tract infection, bloodstream infection, etc. Now metagenomic sequencing has been widely used in the detection of medical pathogens. Respiratory RNA viruses such as Picornaviridae, Coronaviridae, Paramyxoviridae, and Orthomyxoviridae can be diagnosed by NGS (Thorburn et al., 2015). Five organisms (*Neisseria meningitidis*, *Streptococcus agalactiae*, *Candida albicans*, *Mycobacterium fortuitum*, and *Mycobacterium abscessus*) were detected by mNGS testing in the cerebrospinal fluid (Miller et al., 2019).

Limitations of metagenomics in the medical diagnosis of pathogenic microorganisms are as follows: (1) Pathogenic microorganisms detected by metagenomics might be dead or dormant microorganisms, and their clinical significance should be carefully judged by combining with other clinical data. (2) Real-time quantitative PCR has been applied in clinical practice, and its stability and accuracy have been confirmed. While the accuracy of metagenomics on microbial abundance needs to be verified by large sample data. (3) In the case of intracellular bacterium infection, the false-negative result may occur due to an incorrect cell wall rupture method. (4) It is recommended that metagenomics should not be the first choice when immunoassay or PCR methods are available for pathogen diagnosis. (5) When interpreting whether the SNP in microbial genes is related to function, large-scale clinical trials are needed to verify the conclusion. (6) It is important to remove DNA contamination from environmental microorganisms or hosts before data analysis.

Applications of Metagenomics in Other Aspects

Metagenomics studies the diversity, species composition, and genetic evolution of microorganisms in different environments. People have applied metagenomics to winemaking, farming, and authentication of Chinese medicine. It is found that the diversity of microbial composition and population characteristics in a wine cellar is very important to improve the flavor and quality of the wine (Stefanini and Cavalieri, 2018; Liu et al., 2021). The pharmaceutical value of Chinese medicine from different producing areas is different, which may be related to the composition of microorganisms under their unique climate and soil. The increase in crop yield is inseparable from the correlation between plant roots and soil microbial community (Sharma et al., 2018).

OUTLOOK

At present, metagenomics is advancing rapidly in many fields. The bottlenecks of functional metagenomics lie in the failure of large fragment DNA extraction and expression of genes from metagenome libraries in heterologous hosts. High-throughput screening in metagenomic libraries is also a problem to be solved in the future. The development of NGS and TGS technology has promoted the progress of metagenomics. Although the cost of NGS sequencing is reducing and the speed of sequencing is increasing, the microbiological database is incomplete, so it is difficult to correctly assemble and bin without sequencing data for reference. For example, most viral sequences obtained by viral metagenomics have no similarity to any known sequences in the existing databases and, therefore, cannot be annotated for gene and function. In the future, with the progress of TGS sequencing technology, the accuracy rate will be greatly improved. The emergence of new computer algorithms can better process post-sequencing data.

In the future, metagenomics will be combined with metatranscriptomics, proteomics, and metabolomics. Metagenomics provides genetic information of all the environmental microorganisms. Metatranscriptomics is to study the species, functions, and expression activities of transcripts in a microbial community at the RNA level. Metaproteomics can also provide real-time information on the functional expression of environmental genes, while metabolomics studies the metabolic pathways of metabolites. The linking of these omics will lead us to study from genes to proteins and from structure to function.

AUTHOR CONTRIBUTIONS

YX and ML contributed to the study concept and design. LZ, FC, ZZ, MX, FS, LY, XB, YL, YG, and HH collected and sorted out the literature. LY, YX, and ML drew pictures. LZ, FC, and ZZ wrote the first draft. WY, ML, and YX edited and approved the English version of the article. All authors contributed to the article and approved the submitted version.

Tecan Ex. 2058, Page 12 of 15
IDT, Inc. v. Tecan Genomics - IPR2025-00016

FUNDING

This research was supported by the Beijing Hospitals Authority Clinical Medicine Development of Special Funding Support (Nos. XMLX 201706 and XMLX 202127), the Special Public Health Project for Health Development in Capital (2021-1G-4061), the Digestive Medical Coordinated Development Center of Beijing Hospitals Authority (Nos. XXZ0302 and XXT28), the National Science and Technology Major Project of China (Nos. 2017ZX10201201-001-006 and 2017ZX10201201-002-006, and 2018ZX10715-005-003-005), the Beijing Municipal Science & Technology Commission (No. Z151100004015122), the Beijing Science and Technology Commission (No. D161100002716002), and the National Key R & D Projects No. 2020YFC0846200.

REFERENCES

Alneberg, J., Bjarnason, B. S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., et al. (2014). Binning metagenomic contigs by coverage and composition. *Nat. Methods* 11, 1144–1146. doi: 10.1038/nmeth.3103

Andrews, M. C., Duong, C., Gopalakrishnan, V., Iebba, V., Chen, W. S., Derosa, L., et al. (2021). Gut microbiota signatures are associated with toxicity to combined CTLA-4 and PD-1 blockade. *Nat. Med.* 27, 1432–1441. doi: 10.1038/s41591-021-01406-6

Ansorge, W. J. (2009). Next-generation DNA sequencing techniques. *New Biotechnol.* 25, 195–203. doi: 10.1016/j.nbt.2008.12.009

Anyansi, C., Straub, T. J., Manson, A. L., Earl, A. M., and Abeel, T. (2020). Computational methods for strain-level microbial detection in colony and metagenome sequencing data. *Front. Microbiol.* 11:1925. doi: 10.3389/fmicb.2020.01925

Apolinar-Hernández, M. M., Peña-Ramírez, Y. J., Pérez-Rueda, E., Canto-Canché, B. B., De Los Santos-Briones, C., and O'Connor-Sánchez, A. (2016). Identification and *in silico* characterization of two novel genes encoding peptidases S8 found by functional screening in a metagenomic library of Yucatán underground water. *Gene* 593, 154–161. doi: 10.1016/j.gene.2016.08.009

Arango-Argoty, G., Garner, E., Pruden, A., Heath, L. S., Vikesland, P., and Zhang, L. (2018). DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome* 6:23. doi: 10.1186/s40168-018-0401-z

Aron-Wisnewsky, J., Vigliotti, C., Witjes, J., Le, P., Holleboom, A. G., Verheij, J., et al. (2020). Gut microbiota and human NAFLD: disentangling microbial signatures from metabolic disorders. *Nat. Rev. Gastroenterol. Hepatol.* 17, 279–297. doi: 10.1038/s41575-020-0269-9

Avarre, J. C., de Lajudie, P., and Béna, G. (2007). Hybridization of genomic DNA to microarrays: a challenge for the analysis of environmental samples. *J. Microbiol. Methods* 69, 242–248. doi: 10.1016/j.mimet.2006.11.007

Bankier, A. T. (2001). Shotgun DNA sequencing. *Methods Mol. Biol.* 167, 89–100. doi: 10.1385/1-59259-113-2:089

Bowers, R. M., Kyrpides, N. C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T., et al. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* 35, 725–731. doi: 10.1038/nbt.3893

Branton, D., Deamer, D. W., Marziali, A., Bayley, H., Benner, S. A., Butler, T., et al. (2008). The potential and challenges of nanopore sequencing. *Nat. Biotechnol.* 26, 1146–1153. doi: 10.1038/nbt.1495

Breitwieser, F. P., Lu, J., and Salzberg, S. L. (2019). A review of methods and databases for metagenomic classification and assembly. *Brief. Bioinform.* 20, 1125–1136. doi: 10.1093/bib/bbx120

Call, D. R. (2005). Challenges and opportunities for pathogen detection using DNA microarrays. *Crit. Rev. Microbiol.* 31, 91–99. doi: 10.1080/10408410590921736

Chen, Y., and Murrell, J. C. (2010). When metagenomics meets stable-isotope probing: progress and perspectives. *Trends Microbiol.* 18, 157–163. doi: 10.1016/j.tim.2010.02.002

Cheng, J., Romantsov, T., Engel, K., Doxey, A. C., Rose, D. R., Neufeld, J. D., et al. (2017). Functional metagenomics reveals novel β-galactosidases not predictable from gene sequences. *PLoS ONE* 12:e0172545. doi: 10.1371/journal.pone.0172545

Chew, Y. V., and Holmes, A. J. (2009). Suppression subtractive hybridisation allows selective sampling of metagenomic subsets of interest. *J. Microbiol. Methods* 78, 136–143. doi: 10.1016/j.mimet.2009.05.003

Cleary, B., Brito, I. L., Huang, K., Gevers, D., Shea, T., Young, S., et al. (2015). Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning. *Nat. Biotechnol.* 33, 1053–1060. doi: 10.1038/nbt.3329

Colin, P. Y., Kintses, B., Gielen, F., Miton, C. M., Fischer, G., Mohamed, M. F., et al. (2015). Ultrahigh-throughput discovery of promiscuous enzymes by picodroplet functional metagenomics. *Nat. Commun.* 6:10008. doi: 10.1038/ncomms10008

Craig, J. W., Chang, F. Y., Kim, J. H., Obiajulu, S. C., and Brady, S. F. (2010). Expanding small-molecule functional metagenomics through parallel screening of broad-host-range cosmid environmental DNA libraries in diverse proteobacteria. *Appl. Environ. Microbiol.* 76, 1633–1641. doi: 10.1128/AEM.02169-09

Daniel, R. (2005). The metagenomics of soil. *Nat. Rev. Microbiol.* 3, 470–478. doi: 10.1038/nrmicro1160

Eloe-Fadrosh, E. A., Paez-Espino, D., Jarett, J., Dunfield, P. F., Hedlund, B. P., Dekas, A. E., et al. (2016). Global metagenomic survey reveals a new bacterial candidate phylum in geothermal springs. *Nat. Commun.* 7:10476. doi: 10.1038/ncomms10476

Gabor, E. M., de Vries, E. J., and Janssen, D. B. (2003). Efficient recovery of environmental DNA for expression cloning by indirect extraction methods. *FEMS Microbiol. Ecol.* 44, 153–163. doi: 10.1016/S0168-6496(02)00462-2

Galbraith, E. A., Antonopoulos, D. A., and White, B. A. (2008). Application of suppressive subtractive hybridization to uncover the metagenomic diversity of environmental samples. *Methods Mol. Biol.* 410, 295–333. doi: 10.1007/978-1-59745-548-0_16

Gillespie, D. E., Brady, S. F., Bettermann, A. D., Cianciotto, N. P., Liles, M. R., Rondon, M. R., et al. (2002). Isolation of antibiotics turbomycin a and B from a metagenomic library of soil microbial DNA. *Appl. Environ. Microbiol.* 68, 4301–4306. doi: 10.1128/AEM.68.9.4301-4306.2002

Ginolhac, A., Jarrin, C., Gillet, B., Robe, P., Pujic, P., Tuphile, K., et al. (2004). Phylogenetic analysis of polyketide synthase I domains from soil metagenomic libraries allows selection of promising clones. *Appl. Environ. Microbiol.* 70, 5522–5527. doi: 10.1128/AEM.70.9.5522-5527.2004

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., and 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. doi: 10.1016/s1074-5521(98)90108-9

Harrington, C. T., Lin, E. I., Olson, M. T., and Eshleman, J. R. (2013). Fundamentals of pyrosequencing. *Archiv. Pathol. Lab. Med.* 137, 1296–1303. doi: 10.5858/arpa.2012-0463-RA

Hong, C., Manimaran, S., and Johnson, W. E. (2014). PathoQC: computationally efficient read preprocessing and quality control for high-throughput sequencing data sets. *Cancer Inform.* 13, 167–176. doi: 10.4137/CIN.S13890

Iqbal, Z., Caccamo, M., Turner, I., Flicek, P., and McVean, G. (2012). De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nat. Genet.* 44, 226–232. doi: 10.1038/ng.1028

Kakirde, K. S., Wild, J., Godiska, R., Mead, D. A., Wiggins, A. G., Goodman, R. M., et al. (2011). Gram negative shuttle BAC vector for heterologous expression of metagenomic libraries. *Gene* 475, 57–62. doi: 10.1016/j.gene.2010.11.004

Karkman, A., Do, T. T., Walsh, F., and Virta, M. (2018). Antibiotic-resistance genes in waste water. *Trends Microbiol.* 26, 220–228. doi: 10.1016/j.tim.2017.09.005

Kellenberger, E. (2001). Exploring the unknown. The silent revolution of microbiology. *EMBO Rep.* 2, 5–7. doi: 10.1093/embo-reports/kve014

Tecan Ex. 2058, Page 13 of 15
IDT, Inc. v. Tecan Genomics - IPR2025-00016

Krüger, M., Meyerdierks, A., Glöckner, F. O., Amann, R., Widdel, F., Kube, M., et al. (2003). A conspicuous nickel protein in microbial mats that oxidize methane anaerobically. *Nature* 426, 878–881. doi: 10.1038/nature02207

Kumar, K. R., Cowley, M. J., and Davis, R. L. (2019). Next-generation sequencing and emerging technologies. *Semi. Thromb. Hemost.* 45, 661–673. doi: 10.1055/s-0039-1688446

Lämmle, K., Zipper, H., Breuer, M., Hauer, B., Buta, C., Brunner, H., et al. (2007). Identification of novel enzymes with different hydrolytic activities by metagenome expression cloning. *J. Biotechnol.* 127, 575–592. doi: 10.1016/j.jbiotec.2006.07.036

Liu, D., Legras, J. L., Zhang, P., Chen, D., and Howell, K. (2021). Diversity and dynamics of fungi during spontaneous fermentations and association with unique aroma profiles in wine. *Int. J. Food Microbiol.* 338:108983. doi: 10.1016/j.ijfoodmicro.2020.108983

Mardis, E. R. (2008). Next-generation DNA sequencing methods. *Ann. Rev. Genomics Hum. Genet.* 9, 387–402. doi: 10.1146/annurev.genom.9.081307.164359

Marsh, S. (2007). Pyrosequencing applications. *Methods Mol. Biol.* 373, 15–24. doi: 10.1385/1-59745-377-3:15

Mauk, M. G., Liu, C., Sadik, M., and Bau, H. H. (2015). Microfluidic devices for nucleic acid (NA) isolation, isothermal NA amplification, and real-time detection. *Methods Mol. Biol.* 1256, 15–40. doi: 10.1007/978-1-4939-2172-0_2

Midha, M. K., Wu, M., and Chiu, K. P. (2019). Long-read sequencing in deciphering human genetics to a greater depth. *Hum. Genet.* 138, 1201–1215. doi: 10.1007/s00439-019-02064-y

Miller, S., Naccache, S. N., Samayoa, E., Messacar, K., Arevalo, S., Federman, S., et al. (2019). Laboratory validation of a clinical metagenomic sequencing assay for pathogen detection in cerebrospinal fluid. *Genome Res.* 29, 831–842. doi: 10.1101/gr.238170.118

Nnadozie, C. F., and Odume, O. N. (2019). Freshwater environments as reservoirs of antibiotic resistant bacteria and their role in the dissemination of antibiotic resistance genes. *Environ. Pollut.* 254:113067. doi: 10.1016/j.envpol.2019.113067

Olson, M. V. (1993). The human genome project. *Proc. Natl. Acad. Sci. U.S.A.* 90, 4338–4344. doi: 10.1073/pnas.90.10.4338

Olson, N. D., Treangen, T. J., Hill, C. M., Cepeda-Espinoza, V., Ghurye, J., Koren, S., et al. (2019). Metagenomic assembly through the lens of validation: recent advances in assessing and improving the quality of genomes assembled from metagenomes. *Brief. Bioinform.* 20, 1140–1150. doi: 10.1093/bib/bbx098

Palka-Santini, M., Cleven, B. E., Eichinger, L., Krönke, M., and Krut, O. (2009). Large scale multiplex PCR improves pathogen detection by DNA microarrays. *BMC Microbiol.* 9:1. doi: 10.1186/1471-2180-9-1

Pennisi, E. (2010). Genomics. Semiconductors inspire new sequencing technologies. *Science* 327:1190. doi: 10.1126/science.327.5970.1190

Popovic, A., Hai, T., Tchigvintsev, A., Hajighasemi, M., Nocek, B., Khusnutdinova, A. N., et al. (2017). Activity screening of environmental metagenomic libraries reveals novel carboxylesterase families. *Sci. Rep.* 7:44103. doi: 10.1038/srep44103

Probst, A. J., Weinmaier, T., DeSantis, T. Z., Santo Domingo, J. W., and Ashbolt, N. (2015). New perspectives on microbial community distortion after whole-genome amplification. *PLoS ONE* 10:e0124158. doi: 10.1371/journal.pone.0124158

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., and Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* 35, 833–844. doi: 10.1038/nbt.3935

Rebets, Y., Kormanec, J., Luzhetskyy, A., Bernaerts, K., and Ann,é, J. (2017). Cloning and expression of metagenomic DNA in *Streptomyces lividans* and subsequent fermentation for optimized production. *Methods Mol. Biol.* 1539, 99–144. doi: 10.1007/978-1-4939-6691-2_8

Rhoads, A., and Au, K. F. (2015). PacBio sequencing and its applications. *Genomics Proteomics Bioinform.* 13, 278–289. doi: 10.1016/j.gpb.2015.08.002

Riesenfeld, C. S., Goodman, R. M., and Handelsman, J. (2004). Uncultured soil bacteria are a reservoir of new antibiotic resistance genes. *Environ. Microbiol.* 6, 981–989. doi: 10.1111/j.1462-2920.2004.00664.x

Schmieder, R., and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27, 863–864. doi: 10.1093/bioinformatics/btr026

Schneegurt, M. A., Dore, S. Y., and Kulpa, C. F. Jr. (2003). Direct extraction of DNA from soils for studies in microbial ecology. *Curr. Issues Mol Biol.* 5, 1–8.

Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Dröge, J., et al. (2017). Critical Assessment of metagenome interpretation-a benchmark of metagenomics software. *Nat. Methods* 14, 1063–1071. doi: 10.1038/nmeth.4458

Sharma, T. R., Devanna, B. N., Kiran, K., Singh, P. K., Arora, K., Jain, P., et al. (2018). Status and prospects of next generation sequencing technologies in crop plants. *Curr. Issues Mol. Biol.* 27, 1–36. doi: 10.21775/cimb.027.001

Shendure, J., and Ji, H. (2008). Next-generation DNA sequencing. *Nat. Biotechnol.* 26, 1135–1145. doi: 10.1038/nbt1486

Stefanini, I., and Cavalieri, D. (2018). Metagenomic approaches to investigate the contribution of the vineyard environment to the quality of wine fermentation: potentials and difficulties. *Front. Microbiol.* 9:991. doi: 10.3389/fmicb.2018.00991

Svoboda, E. (2021). Gut feeling yields evidence of microbial involvement in autoimmunity. *Nature* 595, S54–S55. doi: 10.1038/d41586-021-01837-8

Thorburn, F., Bennett, S., Modha, S., Murdoch, D., Gunson, R., and Murcia, P. R. (2015). The use of next generation sequencing in the diagnosis and typing of respiratory infections. *J. Clin. Virol.* 69, 96–100. doi: 10.1016/j.jcv.2015.06.082

Torsvik, V., Goksøyr, J., and Daae, F. L. (1990). High diversity in DNA of soil bacteria. *Appl. Environ. Microbiol.* 56, 782–787. doi: 10.1128/aem.56.3.782-787.1990

Towner, J. S., Sealy, T. K., Khristova, M. L., Albariño, C. G., Conlan, S., Reeder, S. A., et al. (2008). Newly discovered ebola virus associated with hemorrhagic fever outbreak in Uganda. *PLoS Pathog.* 4:e1000212. doi: 10.1371/journal.ppat.1000212

Uchiyama, T., Abe, T., Ikemura, T., and Watanabe, K. (2005). Substrate-induced gene-expression screening of environmental metagenome libraries for isolation of catabolic genes. *Nat. Biotechnol.* 23, 88–93. doi: 10.1038/nbt1048

van Dijk, E. L., Auger, H., Jaszczyszyn, Y., and Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends Genet.* 30, 418–426. doi: 10.1016/j.tig.2014.07.001

Venkata Mohan, S., Velvizhi, G., Vamshi Krishna, K., and Lenin Babu, M. (2014). Microbial catalyzed electrochemical systems: a bio-factory with multi-facet applications. *Bioresour. Technol.* 165, 355–364. doi: 10.1016/j.biortech.2014.03.048

Voget, S., Leggewie, C., Uesbeck, A., Raasch, C., Jaeger, K. E., and Streit, W. R. (2003). Prospecting for novel biocatalysts in a soil metagenome. *Appl. Environ. Microbiol.* 69, 6235–6242. doi: 10.1128/AEM.69.10.6235-6242.2003

Wahba, L., Jain, N., Fire, A. Z., Shoura, M. J., Artiles, K. L., McCoy, M. J., et al. (2020). An extensive meta-metagenomic search identifies SARS-CoV-2-homologous sequences in pangolin lung viromes. *mSphere* 5:e00160-20. doi: 10.1128/mSphere.00160-20

Wang, H., and Ren, Z. J. (2013). A comprehensive review of microbial electrochemical systems as a platform technology. *Biotechnol. Adv.* 31, 1796–1807. doi: 10.1016/j.biotechadv.2013.10.001

Wang, J., and Jia, H. (2016). Metagenome-wide association studies: fine-mining the microbiome. *Nat. Rev. Microbiol.* 14, 508–522. doi: 10.1038/nrmicro.2016.83

Wang, T. Y., Wang, L., Zhang, J. H., and Dong, W. H. (2011). A simplified universal genomic DNA extraction protocol suitable for PCR. *Genet. Mol. Res.* 10, 519–525. doi: 10.4238/vol10-1gmr1055

Wei, X., Walia, V., Lin, J. C., Teer, J. K., Prickett, T. D., Gartner, J., et al. (2011). Exome sequencing identifies GRIN2A as frequently mutated in melanoma. *Nat. Genet.* 43, 442–446. doi: 10.1038/ng.810

Wu, F., Zhao, S., Yu, B., Chen, Y. M., Wang, W., Song, Z. G., et al. (2020). A new coronavirus associated with human respiratory disease in China. *Nature* 579, 265–269. doi: 10.1038/s41586-020-2008-3

Xu, M., Fujita, D., and Hanagata, N. (2009). Perspectives and challenges of emerging single-molecule DNA sequencing technologies. *Small* 5, 2638–2649. doi: 10.1002/smll.200900976

Yun, J., and Ryu, S. (2005). Screening for novel enzymes from metagenome and SIGEX, as a way to improve it. *Microb. Cell Factor.* 4:8. doi: 10.1186/1475-2859-4-8

Tecan Ex. 2058, Page 14 of 15
IDT, Inc. v. Tecan Genomics - IPR2025-00016

Zhang, J., Lin, S., and Zeng, R. (2007). Cloning, expression, and characterization of a cold-adapted lipase gene from an antarctic deep-sea psychrotrophic bacterium, *Psychrobacter* sp. 7195. *J. Microbiol. Biotechnol.* 17, 604–610.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Zhang, Chen, Zeng, Xu, Sun, Yang, Bi, Lin, Gao, Hao, Yi, Li and Xie. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Tecan Ex. 2058, Page 15 of 15
IDT, Inc. v. Tecan Genomics - IPR2025-00016