Next-generation sequencing (NGS) in the microbiological world: how to make the most of your money

Antony T. Vincent[1,2,3], Nicolas Derome[1,4], Brian Boyle[1], Alexander I. Culley[1,2,5] and Steve J. Charette[1,2,3]*

1. Institut de Biologie Intégrative et des Systèmes (IBIS), Université Laval, Quebec City, QC, Canada, G1V 0A6

2. Département de biochimie, de microbiologie et de bio-informatique, Faculté des sciences et de génie, Université Laval, Quebec City, QC, Canada, G1V 0A6

3. Centre de recherche de l'Institut universitaire de cardiologie et de pneumologie de Québec, Quebec City, QC, Canada, G1V 4G5

4. Département de biologie, Faculté des sciences et de génie, Université Laval, Quebec City, Canada, G1V 0A6

5. Groupe de Recherche en Écologie Buccale (GREB), Faculté de médecine dentaire, Université Laval, Quebec City, QC, Canada, G1V 0A6

* Corresponding author: Steve J. Charette; steve.charette@bcm.ulaval.ca; telephone: (+1) 418-656-2131, ext. 6914; fax: (+1) 418- 656-7176.

Keywords: Next-generation sequencing; Bioinformatics; Microbial genomics; Metagenomics; Microbial transcriptomics

Abstract

The Sanger sequencing method produces relatively long DNA sequences of unmatched quality and has been considered for long time as the gold standard for sequencing DNA. Many improvements of the Sanger method that culminated with fluorescent dyes coupled with automated capillary electrophoresis enabled the sequencing of the first genomes. Nevertheless, using this technology to sequence whole genomes was costly, laborious and time consuming even for genomes that are relatively small in size. A major technological advance was the introduction of next-generation sequencing (NGS) pioneered by 454 Life Sciences in the early part of the 21th century. NGS allowed scientists to sequence thousands to millions of DNA molecules in a single machine run. Since then, new NGS technologies have emerged and existing NGS platforms have been improved, enabling the production of genome sequences at an unprecedented rate as well as broadening the spectrum of NGS applications. The current affordability of generating genomic information, especially with microbial samples, has resulted in a false sense of simplicity that belies the fact that many researchers still consider these technologies a black box. In this review, our objective is to identify and discuss four steps that we consider crucial to the success of any NGS-related project. These steps are: (1) the definition of the research objectives beyond sequencing and appropriate experimental planning, (2) library preparation, (3) sequencing and (4) data analysis. The goal of this review is to give an overview of the process, from sample to analysis, and discuss how to optimize your resources to achieve the most from your NGS-based research. Regardless of the evolution and improvement of the sequencing technologies, these four steps will remain relevant.

1. From a few nucleotide sequences to sequencing on a massive scale

Nucleic acid sequencing is now an integral part of modern science. We routinely use DNA sequencing in many fields in microbiology, including tracking infectious diseases (Gire et al., 2014) and studying the diversity of the microbial communities like the human microbiota (Guttman et al., 2014). But when did the sequencing era begin?

The first free-living organism to have its genome fully sequenced was the Gram-negative bacterium *Haemophilus influenzae* in 1995 by *The Institute for Genomic Research* (TIGR) (Fleischmann et al., 1995). The following year, a worldwide effort produced the first complete eukaryotic genome, the yeast *Saccharomyces cerevisiae* (Goffeau et al., 1996). Nevertheless, it is clear that the publication of the human genome at the beginning of the 21st century was the principal event in the rise of genomics and consequently marks the beginning of the sequencing era (Lander et al., 2001, Venter et al., 2001). It was now possible for scientists to study the hereditary molecule directly. However, there were many drawbacks to massive DNA sequencing. Among them were the expensive costs of reagents and significant human resources required to operate the sequencing platforms. This is still the case although on a different scale (see below).

In 2005, a revolution took place with the release of pyrosequencing technology (Margulies et al., 2005) by 454 Life Sciences (now part of Roche). This high-throughput technology, considered “next-generation sequencing” (NGS), allowed the generation of thousands to millions of short sequencing reads in a single machine run. Since then, many other NGS technologies have emerged, including the sequencing by synthesis technology used by Solexa/Illumina sequencers since 2006 that currently occupies a vast part of the NGS market. The field of next-generation sequencing is very dynamic due to the constant improvements of

the instruments and the continued emergence of new technologies. It is therefore difficult to predict the future of the market in the coming years. The initial human genome project had a cost of around 3 billion dollars. Fourteen years later, using current NGS technologies, we have almost attained the landmark price of $1000 per human genome (Hayden, 2014), with smaller bacterial genomes costing even less.

Since the cost of sequencing has become less prohibitive, many laboratories around the world are now able to conduct their own sequencing projects and even maintain their own sequencing apparatus. However, this new accessibility has led many non-specialists to use NGS without prior knowledge and consequently use the technology in a non-optimal way. This is particularly the case in the field of microbiology where the relatively smaller genome sizes of microbes can lead to the impression that sequencing these genomes is simple. The reality, even for microbial genomics and other derived fields, is that even smaller genomes require an adequate sequencing strategy. Without one, researchers are likely to be disappointed and frustrated at not being able to generate quality data due to bad planning, a lack of resources or unrealistic expectations. The goal of this review is to demystify the NGS process and provide guidelines on how to perform NGS efficiently. To get help on an individual basis or to find more information about specific NGS applications and tools, we recommend exploring two active NGS related resources: SEQanswers (Li et al., 2012) and BioStar (Parnell et al., 2011).

2. The conceptual workflow

A complete NGS related project involves a limited number of steps, of which some are crucial to the successful outcome of the project (Figure 1). The first and most important step is formulating a valid hypothesis that goes beyond sequencing and to develop an appropriate

97 experimental approach. In other words, the question to resolve is what is expected from the
98 sequence data, as this will subsequently determine how the library is prepared, influence the
99 choice of the sequencer and drive data analysis, the step that takes the most time to complete.
100 The first step (planning) is strictly conceptual while the second (library preparation) and third
101 (sequencing) involve laboratory work and the fourth (data analysis) involves computing
102 resources and the field of bioinformatics. Each step is discussed individually below.
103
104 **3. Step 1: Asking the right questions**
105 This step in the NGS workflow is the most crucial one in the process, but is often neglected
106 because microbial whole genome sequencing has gone from impossible to economical in a
107 relatively short period of time. The affordability of the technology should not drive research
108 and the ultimate goal is certainly not filling public databases. To use funding wisely, we must
109 first determine what scientific problem we want to resolve and then determine what dataset
110 will be the most useful for answering that question. In Table 1, you will find applications of
111 next-generation sequencing and their associated dataset types.
112
113 Most NGS technologies currently available are based on the following principle: sequence a
114 large number of DNA fragments (thousands to millions) in parallel in a single machine run.
115 To achieve this, nucleic acids (total DNA, genomic DNA, RNA, etc.), after their extraction
116 and purification, must be converted to machine sequenceable fragments in a process called
117 library preparation (Figure 2). After sequencing, the considerable amount of sequence
118 produced (from Mb to Gb of data) must be analyzed with bioinformatics procedures designed
119 to pull out the desired information in various applications (discussed later in the text). The
120 way libraries are prepared and the choice of the sequencing instrument and associated
121 technology have a large impact on the possible downstream analyses (Table 1). In the

following sections, the most important elements of library preparation, sequencing and data analysis, are presented.

4. Step 2: Choosing the sequencer and preparing libraries

The beginning of the sequencing workflow requires the conversion of the nucleic acids into instrument compatible libraries. The choice of sequencing instrument should be made prior to generating libraries because specific, proprietary sequences must be added at the library preparation stage.

4.1. Sequencer features

Things to consider when choosing an instrument are: (1) how the reads are generated (fragment vs. paired-ends), (2) read length, (3) read number (sequencing depth) and (4) error rate. Table 2 shows features of sequencers from ThermoFisher and Illumina, the two dominant technologies currently available. We will not discuss 454 pyrosequencing technology (and the corresponding sequencers) because Roche will discontinue the 454-sequencing platform in mid-2016.

Fragment reads are produced by the sequencer when a single read is generated per library molecule while paired-end reads (or paired reads) are generated from opposing ends of the same library molecule (Figure 3). Some instruments enable the choice of generating either fragments or paired-end reads, while others produce only fragments. Therefore, the decision to generate fragments or paired-end reads will not only determine sequencer choice, but also how the libraries are produced.

Read length is an important feature to consider before choosing a next-generation sequencer because it is directly linked to the amount of information that is obtained from a single molecule. For example, much more powerful analyses are possible when the whole PCR insert is sequenced, in amplicon-based studies for example (see section 6.4). The average read length produced by a sequencing instrument in a given run will also directly impact the quality of *de novo* assembly (i.e. the assembly of a genome without a reference) generally through the use of longer K-mers for longer read lengths (See box 1). For example, when the genome contains repeated elements such as insertion sequences (ISs), duplicated genes and ribosomal RNA operons that are larger than the average read length, these regions will cause breaks in the assembly (Vincent et al., 2014, Vincent et al., 2015).

The number of reads is important in determining coverage because during sequencing, different reads are generated from different library molecules and thus coverage is defined by the number of times a region, at the single base pair level, is covered by a read. The total number of reads is the most important parameter for quantitative applications like RNA-Seq (see section 6.7). The combination of read length and number of reads defines the throughput of an instrument in number of bases per run. If there are time sensitive issues, for example, for diagnostic applications, the throughput in numbers of bases per day of a particular sequencing platform must be taken into consideration. Theoretical coverage values of a particular instrument can be calculated by dividing throughput by genome size. The desired coverage will depend on the application. For example, for a *de novo* assembly, a coverage between 25 and 100 X is considered optimal. This means that a researcher should have an approximation of genome(s) size(s) in order to estimate the amount of sequencing required for an appropriate coverage. Additionally, it is important to keep in mind that it is possible to sequence multiple

samples with NGS (multiplexing) to optimize the machine run and reagents. This is achieved by adding multiple identifiers (MIDs) or barcodes (BC) at the library preparation stage.

As discussed below, NGS technologies are prone to sequencing errors, but these largely randomly occurring errors can be compensated by sequencing different molecules from the same region at multiple times (i.e. increased coverage). Consequently, increasing the coverage will also help to increase confidence in the validity of existing variations in sequence, for example with single-nucleotide polymorphism (SNP). However, too high coverage is also problematic because the absolute number of sequencing errors will increase with coverage (Ekblom and Wolf, 2014) and the accumulation of these errors will impact the quality of the genome assembly.

The errors that occur during NGS can be classified as indels or base substitutions. Indels, or insertion/deletion errors, are defined as bases inserted (In) or absent (del) in the output sequence while base substitutions occur when one base is replaced by another base in the output sequence. Error rates can be estimated at the read level by comparing any given subset of reads to a reference sequence. Similarly, consensus error rates can be estimated by comparing the results of an assembly (consensus) to a reference sequence. Consensus error rates should be several magnitudes smaller than read error rates because coverage compensates for sequencing errors that occur randomly. However, some regions are more prone to sequencing errors, such as homopolymers and low complexity regions, and each sequencing technology has its own dominant error type. For an overview of sequencing errors and an error correction tool see (Marinier et al., 2015).

195 **4.2. Library features**

196 After the sequencer has been selected, the next step is to convert your sample into sequencer-
197 ready libraries by adding the sequencer brand proprietary sequences to library fragments
198 termini. We will not review this step extensively here (see (Head et al., 2014, van Dijk et al.,
199 2014). Nonetheless, Figure 4 summarizes the principle types of library preparations. The most
200 frequent library preparation method begins with the random fragmentation of genomic
201 segments into a target size range (thus the term "shotgun"), then the repair of the fragment
202 ends and the addition of a single dATP (or deoxyadenosine triphosphate) adenine to the 3'-
203 end of both strands, followed by the ligation of instrument specific adaptors to each molecule
204 to complete the process. Most protocols then recommend the PCR-amplification of adapter
205 containing molecules to enrich molecules with adaptors on both ends. The mate-pair library is
206 the procedure in which the ends of a large molecule (3 to 15 Kb) are brought together within a
207 single small fragment by a circularization step, which is then subjected to the shotgun
208 procedure described above. An enrichment of circularized adapter-containing molecules is
209 performed prior to the final amplification of the sequence library. The final molecules that are
210 sequenced contain the ends of large fragments (mate-pairs) interrupted by a circularization
211 adapter. Because mate-pair information comes from the same library molecule that is
212 sequenced, it is possible to use this information to link contigs during *de novo* assembly
213 where the relative order and orientation of each contig can be predicted. This process, named
214 "scaffolding", is implemented in a vast majority of modern *de novo* assemblers and can be
215 optimized for a specific data type (Vincent, et al., 2014). Another way to make instrument
216 compatible fragments is to add the proprietary sequences to the 5' –end of gene-specific
217 primers and perform a PCR. This way, the resulting amplicons will subsequently contain the
218 necessary adapters.
219

In most instances, library preparation is a relatively simple and robust process. Nevertheless, the following points should be taken into consideration. First, take care that the library insert size fits your instrument. Amplicon libraries will generally benefit from the sequencing of the entire molecules, but apart from this situation, there is little value in sequencing adapter sequences. Therefore matching insert size with instrument read length usually maximizes the instrument throughput. However, generating large insert libraries to be sequenced with short paired reads (Figure 3) can also improve scaffolding by jumping over small repeats or low complexity regions. On the other hand, since the quality of sequencing reads tends to decrease with length, overlapping forward and reverse reads is a way to increase overall sequence quality.

As stated earlier, most library preparation methods recommend the amplification of adapter containing molecules through a PCR step to enrich the reaction with molecules that contain adapters on both ends, particularly in those reactions where the adapters were added by ligation. Our experience with several commercial kits has shown that between 4 and 12 % of the molecules generated during the production of shotgun libraries contain adapters on both ends. Thus the quantification of library molecules by spectrophotometric or fluorescence methods prior to amplification would results in a highly biased number towards non-sequenceable molecules that contain no adapters or adapters on a single end. We therefore recommend the PCR amplification of libraries, however to avoid quantitative biases or the introduction of PCR errors, the number of cycles should be kept to a minimum (max 12 cycles). Alternatively, quantitative PCR can be used to quantify library molecules containing adapters on both ends, although it should be noted that quantitative PCR is not a linear scale technology and that some NGS instruments are more sensitive than others to small variations in the loading of template.

There are multiple ways to fragment DNA. Mechanical fragmentation is the most widely used method to shear DNA because it results in reproducible library synthesis and better control of the insert size from sample to sample, particularly, if DNA samples come from a wide variety of organisms, multiple users and/or several different DNA preparation methodologies. Mechanical fragmentation remains the most expensive method because it requires special instrumentation and associated consumables. Alternatives to mechanical shearing are either enzymatic or tagmentation (Marine et al., 2011). Although these methods are far less expensive and do not require special instrumentation, they have been shown to generate more variable results and are likely more prone to biases. The tagmentation procedure, which relies on a mutated Tn*5* transposase (cut-and-paste mechanism) (Picelli et al., 2014), is attractive because it is the most time and cost effective way to prepare libraries for resequencing applications in particular.

An important factor that is often ignored is the molecule diversity of a library. This is important because if the same molecule is sequenced repeatedly, it has no biological or statistical value in the analyses and thus can lead to an erroneous interpretation of the results. The only way to measure molecule diversity prior to sequencing is by qPCR. It is important that the diversity of molecules in a library is determined before the final PCR step, because afterwards, most of the molecules present will be the result of the amplification. Nonetheless, for most applications in microbiology, library diversity will generally not be an issue because of the relatively small size of microbial genomes. For example 1 ng of a 5 Mb bacterial genome represents 1.82×10^5 molecules, which translates to 1.82×10^9 fragments of 500 bp. Thus for normal shotgun applications in microbiology, molecule diversity will not be a limiting factor.

In contrast, the generation of large insert mate-pair libraries (15 Kb) is a very inefficient process and library diversity should be evaluated before money is spent on sequencing redundant molecules. The diversity in metagenomic surveys of rRNA genes (50K to 100K reads per sample) should not pose a problem because 10 ng of bacterial DNA contains approximately 300 000 bacterial genomes (avg. genomes size 3.5 Mb) where each genome has from 1 to 7 copies of the rRNA operon (Vetrovsky and Baldrian, 2013), resulting in a estimated total of 1 M distinct molecules. In contrast, the diversity of a library may become an issue in ultra deep sequencing projects (> 1M reads), particularly when the mass of the initial DNA template is low or when mixtures of DNA sources reduce the overall bacterial DNA content.

Deep sequencing provides a powerful means of investigating the low variant fraction (< 1%) of a microbial population (McElroy et al., 2014, Pulido-Tamayo et al., 2015). To detect this fraction of the population, the sample must be sequenced to a sufficient depth (generally an average coverage of hundreds to thousands X (McElroy, et al., 2014)). For these projects to be statistically meaningful, an initial estimate of molecule diversity would be highly beneficial.

5. Step 3: Sequencing

Sequencing is the most straightforward step in the NGS process because all brands of sequencers are relatively easy to operate and include comprehensive manufacturer support services. The most critical part of this step is loading the proper amount of library molecules onto the instrument. This can be accomplished by accurate library quantification and by following appropriate quality control procedures.

Another question related to the sequencing step is where to find a DNA sequencer? There are two answers: (1) purchase a sequencer or (2) outsource the sequencing to a core lab. Buying a sequencer requires not only sufficient funding to purchase the apparatus (prices range from several tens of thousands to just over a million dollars), but also to support the often overlooked costs of the reagents and instrument maintenance. In truth, only a small number of large-scale microbial genomic projects produce enough samples (Table 3) to justify the purchase and maintenance of an in-house sequencer. Moreover, sequencer technologies are evolving at such an accelerated pace that the instruments of today will likely be obsolete in just a few years from now. For the majority of NGS-based projects, the most cost effective and efficient approach is to employ a core-NGS facility. The advantages of an NGS center include the fact that they often possess different instruments and must maximize instrument usage to offer competitive pricing. Additionally, many core labs provide expertise ranging from experimental design to data analysis, offer a range of payment options and often guarantee sequence yield and quality. Most core labs are also able to accept raw nucleic acids for complete processing (library + sequencing) as well as prepared libraries ready for direct sequencing. Even if the number of private and university sequencing core labs around the world keeps increasing, it is important to consider the delays that can be encountered when employing a core facility; for example most facilities that maximize instrument usage are generally able to offer the lowest prices, however they also have longer sample queues that can result in delays in sample processing (Figure 1). Delays on the scale of months have convinced some researchers, to operate their own sequencer(s) and thus to pay a significant premium for faster processing.

6. Step 4: applications and data analysis

The analysis of NGS data is the last step before the final results and is considered second in importance after determining the objectives of the experiment. As indicated in Figure 1, this step is also the most time consuming. This is often the case because newcomers to NGS data analysis are unaware of the bioinformatic tools that are available and, more importantly, often lack the training to use them correctly. Indeed, many of these tools are only available on UNIX-like operating systems while having a command line interface.

Thankfully some free NGS data analysis platforms, such as GALAXY (Goecks et al., 2010) and Unipro UGENE (Golosova et al., 2014), integrate a suite of bioinformatics tools into an easy-to-use framework. These programs are an excellent starting place for neophytes and non-bioinformaticians, however, the users are limited to the tools included in the package, which are not necessarily optimized for the researchers analysis requirements

The goal of the following section is to provide an overview of NGS data analysis, including the pretreatment of sequencing reads, assembly with and without a reference genome, how to glean information from metagenomic data, and a brief overview of tools for subsequent downstream analyses.

6.1. The pretreatment of sequencing reads

NGS platforms are able to generate thousands to millions of sequencing reads in a single machine run. However, the quality of the sequences is not uniform among the dataset. Consequently, it is necessary to evaluate the quality of the sequence reads by different bioinformatics procedures. Quality control (QC) has led to significant improvements in *de*

novo assemblies (Salzberg et al., 2012), amplicon sequencing (Bokulich et al., 2013) and transcriptome assemblies (Macmanes and Eisen, 2013).

The reason that sequencing reads must be filtered (Zhou and Rokas, 2014) is to remove reads that will bias downstream analyses, such as sequence reads of low quality, adapter contaminants, as well as discordant and duplicate paired-end reads. The majority of NGS QC tools are only available for UNIX-based operating systems that lack a friendly user graphical interface. However, there are some exceptions such as the web-based tools GALAXY (Goecks, et al., 2010) and Prinseq (Schmieder and Edwards, 2011).

For a more in depth discussion of the importance of QC in the analysis of NGS reads that includes topics such as performing quality assessment and describing workflows, see (Watson, 2014).

6.2. *De novo* assembly

De novo assembly is the process in which sequence reads are assembled without a reference sequence. *De novo* assembly is challenging and computationally demanding and thus the development of *de novo* assembly tools has been a top priority in the field of bioinformatics. This goal is exemplified in the Assemblathon program, a contest in which each assembler is evaluated based on its performance in the assembly of known datasets (Bradnam et al., 2013).

There are three main algorithm classes for *de novo* assemblers: Greedy, Overlap-layout-consensus (OLC) and De Bruijn graph (Nagarajan and Pop, 2013). Even with recent advances that have reduced memory requirements (Conway and Bromage, 2011, Simpson and Durbin, 2012) as well as the development of highly parallelizable algorithms (Boisvert et al., 2010,

Liu et al., 2013, Liu et al., 2011), *de novo* assembly is a non-deterministic polynomial-time hard (NP-hard) mathematical challenge, (Pop, 2009), meaning that the assembly cannot be solved in polynomial time (Medvedev et al., 2007).

The variety of *de novo* assemblers presently available raises the question of which one produces the best assembly. Unfortunately, the answer is that no tool will produce the best assembly for all datasets. In fact, given that all tools have their restrictions, strengths and weaknesses, *de novo* assembly should be considered an iterative process in which the assembly parameters are optimized with consecutive runs and different assemblers are employed to cross-validate the final assembly (Ekblom and Wolf, 2014). Towards this end, an integrative *de novo* assembly workflow tool, named RAMPART, has been recently developed that allows the user to test different parameters on various free tools (Mapleson et al., 2015).

To illustrate the variability between *de novo* assemblers, we assembled Illumina reads from *Aeromonas salmonicida* subsp. *salmonicida* 01-B526 with three assemblers, A5 (Coil et al., 2015), Ray (Boisvert, et al., 2010) and SPAdes (Bankevich et al., 2012) (Table 4). The *de novo* assembly of the reads with Ray produced the fewest number of contigs, while SPAdes produced 167% more contigs than Ray and A5 fell in between the two. The largest contig was almost identical for Ray and SPAdes while A5 assembled a largest contig that was smaller than the two other assemblers. However, the N50 value, which is the length for which all contigs of that length or longer covers at least half an assembly, is essentially the same for all assemblers. Finally, when we compared the three assemblies with the reference chromosome of *A. salmonicida* subsp. *salmonicida* A449 (Reith et al., 2008), we found that the Ray assembly had the lowest amount of coverage (genome fraction) relative to the other assemblers in our example. The above comparison underlines the importance of performing

multiple assemblies with several different *de novo* assemblers. It is clear that relying on only one assembler without testing others is risky and could lead to an incorrect interpretation of the data.

Many genomes contain large-repeated-elements that increase the complexity of the assembly process and results in assemblies with a high number of contigs. For example, the Gram-negative bacterium *A. salmonicida* subsp. *salmonicida* is known to contain many large insertion-sequences (ISs) that are responsible for a majority of the breaks in an assembly (Vincent, et al., 2014, Vincent, et al., 2015). At present there are two primary strategies to sequence genomes with a high IS content: (1) using mate-pair information to build genomic scaffolds and (2) using long-read sequencing technology.

As previously stated in the section "library features", a genomic scaffold is a series of contigs whose relative position and orientation are predicted. The information required for the scaffolding process is contained in mate-pair sequence reads where the most important parameter is the number of positive mate-pair reads that confirm a particular junction. This can be problematic because the gaps between contigs are sometimes rough estimates and therefore contain stretches of undetermined bases or "Ns".

The second approach is to use long-read (> 7-kb) sequencing technology. At present, these "third-generation" sequencing platforms are most commonly used to finish genome sequences and thus avoiding the time consuming process of amplifying gap regions followed by Sanger sequencing. A discussion of third-generation sequencing is beyond the scope of this review. See (Miyamoto et al., 2014) and (Koren and Phillippy, 2015) for more information.

6.3. Assembly using a reference

Another method to produce an assembly is by using a reference genome to "guide" the assembler. Relative to *de novo* assembly, producing an assembly using a reference genome is a simpler process. There are two principle methods of assembly with a reference genome: (1) produce *in silico* scaffolds by mapping the contigs from a *de novo* assembly onto a reference genome, and (2) guide the contig assembly process by mapping the individual reads onto the reference genome (Ekblom and Wolf, 2014). However, a major drawback of using a guided-assembly is that a reference sequence must be available. Although the number of genomes sequenced is growing rapidly, approximately 90% of the genome sequences deposited in GenBank remain incomplete (Land et al., 2015). Moreover, approximately half of the sequenced genomes in the database are related to the phylum *Proteobacteria* (Land, et al., 2015), which suggests that most microbial taxa are likely underrepresented and thus lack a reference sequence. Finally, it has been repeatedly demonstrated that published sequences can contains errors that will result in discrepancies during the mapping stage of assembly. Consequently, it is important to choose a reference sequence from a phylogenetically related organism that is well studied and curated when possible.

6.4. Amplicon based studies

The most common amplicon based studies of microbial communities focus on universal taxonomic markers such as 16S SSU rRNA (for bacteria), 18S SSU rRNA (for microeukaryotes and unicellular eukaryotes) or internal transcribed spacers (ITS - for fungal communities) to survey both microbial diversity and community structure (i.e., quantify the relative abundance of each taxon in a particular assemblage). Amplicon libraries are particularly useful in the context of comparative investigations and correlations between

444 community structure and metadata (i.e. for measuring community response to contrasted
445 biological, chemical or physical parameters) can provide insight into community dynamics
446 and adaptation (i.e. taxa replacement). Even though these studies are based on one or a few
447 molecular markers, it is possible to infer functional repertories from these data based on the
448 availability of reference genome databases (Langille et al., 2013). This is especially the case
449 for 16S libraries, as this bacterial taxonomic marker is the most extensively annotated.
450
451 Different genomic loci provide differential power to resolve taxa due to differences in their
452 genetic diversity distribution. Consequently, estimates of diversity will vary according to the
453 particular molecular marker selected. For any given genomic locus, resolution power is
454 proportional to the sequence length and level of polymorphism. Early studies based on 1.5 kb-
455 long 16S SSU rRNA sequences produced with Sanger-sequencing enabled the identification
456 of many individual genera and species. For most current NGS methods, amplicons lengths are
457 more restricted, where long-length and short-length reads vary between 450 and 700 and 50
458 and 200 bp, respectively. Even with these relatively short read lengths, current NGS
459 applications for microbial identification still focus on the 16S rRNA gene. This gene consists
460 of conserved sequences interspersed with nine variable sequences, called variable regions
461 (Ashelford et al., 2005). The lengths of these variable regions range from approximately 50 to
462 100 bases. Thus, depending on the read length, one or several variable regions can be
463 targeted. More importantly, the conservation of the flanking regions targeted by the primers
464 commonly used in these types of studies is critical to a comprehensive characterization of
465 bacterial diversity (Hartmann et al., 2010). Because of specific mismatches between primer
466 and target, some bacterial classes may be erroneously over-represented due to higher
467 sequence identity in the primer binding region. Therefore, the 16S variable regions that are
468 targeted should be selected based on factors such as the class of bacteria under investigation

and the required level of taxonomic resolution (order, family, genus, species, etc.) (Engelbrektson et al., 2010). For instance, the 16S V1 and V2 regions are highly variable, but their flanking regions are less conserved than those of the other variable regions. Thus while the use of the V1 and V2 region results in a higher level of taxonomic resolution, the estimate of both diversity and evenness are relatively more biased due to primer mismatches (Klindworth et al., 2013). In contrast, the V3 and V4 regions are less variable but their respective flanking regions are more conserved than the V1 and V2 regions. Thus in comparison, although the level of taxonomic resolution is less, the estimate of sample diversity and evenness are also less biased. Finally, it has been shown that the target molecular marker can be transferred between both closely and distantly related taxa. For example Acinas et al. (2004) demonstrated that 16S rRNA loci can be transferred between bacterial genotypes (Acinas et al., 2004), leading to individual 16S polymorphism and ultimately an overestimation of community diversity.

Strain typing has now reached the next-generation level (Boers et al., 2012) with the development of high throughput multi-locus-sequence-typing (MLST) based on next-generation sequencing. In this case, instead of targeting a conserved gene to survey the diversity of microbial communities, multiple amplicons for genes of interest are produced for individual strains, where each strain is uniquely barcoded. In MLST, the locus targets are endless and this approach is now being developed with third generation sequencing platforms (Chen et al., 2015).

ThermoFisher has pushed MLST to higher grounds by coupling its IonTorrent line of NGS products (PGM, Ion Proton and S5) with its Ampliseq Technology. Existing panels contain hundreds to hundreds of thousands of amplicons designed originally to target human genes.

494 Thermofisher is now offering Ampliseq panels for *Mycobacterium tuberculosis* and Ebola
495 virus typing and it can be expected that more and more typing panels will emerge in the next
496 few years.
497
498 **6.5. Metagenomic surveys**
499 The metagenome provides insight into the overall functional repertoire of a microbial
500 community, including information on the metabolic capabilities of the community and the
501 potential functional interactions among its members (Chistoserdovai, 2010). Metagenomics is
502 a non-targeted approach that results in the description and quantification of the copy number
503 and allelic variants of genes that could potentially be expressed by the microbial community
504 of interest. Various sequencing platforms can be employed for metagenomics: platforms
505 generating long read lengths facilitate the assembly and annotation processes, but fail to
506 accurately quantify copy number and allelic variants of genes because they produce a
507 relatively low read count. Conversely, platforms designed to produce high read counts of
508 shorter read length allow the accurate quantification of copy number and allelic variants of
509 various genes, but the process of assembly and annotation becomes particularly challenging
510 (Prakash and Taylor, 2012), especially if the ultimate goal is to assemble and recover single
511 genome. These data have proven to be best suited for comparative analysis of functional
512 repertoires in contrasting environmental conditions. Additionally, metagenomic data can also
513 provide invaluable reference sequences for assembling and mapping metatranscriptomic reads
514 (Ye and Tang, 2015). Metagenomics is a particularly effective approach to characterize
515 taxonomic profiles. Taxonomic annotation is based on hundreds of unique clade-specific
516 marker genes identified from reference genomes, and thus the broad sequence data generated
517 with metagenomics allows very accurate and unambiguous taxonomic assignments (ex.
518 MetaPhlAn (Segata et al., 2012)).

519 6.5.1. Read annotation and assembly procedure

520 Metagenomic sequences (i.e., reads) are classified into discrete clusters commonly referred to
521 as bins. Binning attempts to assign every metagenomic sequence to a taxonomic group (e.g.,
522 OTU, genus, family). As with amplicon analysis, binning accuracy improves with sequence
523 length (Charuvaka and Rangwala, 2011, McHardy et al., 2007). There are currently three
524 types of binning algorithms. These are either based on supervised learning procedures (i.e. the
525 similarity of metagenomic sequences to annotated sequences from a database) or based on
526 unsupervised learning procedures, which bin's reads in a given dataset based on their mutual
527 composition (Strous et al., 2012) or similarity (Huson et al., 2011, Kislyuk et al., 2009,
528 Krause et al., 2008, Mande et al., 2012). Similarity based binning tools provide higher
529 annotation accuracy and resolution compared to compositional binning tools. However,
530 similarity based binning tools require greater computational resources because they align
531 every single read to an immense number of annotated sequences. Conversely, compositional
532 and diversity binning tools require relatively fewer computational resources because they use
533 metagenome sequence characteristics (e.g., tetranucleotide patterns, codon usage, and GC
534 content) to cluster or classify sequences into taxonomic groups (Dick et al., 2009, Saeed et al.,
535 2012, Teeling et al., 2004). Also, this approach is useful for clustering contigs into groups that
536 can be subsequently assembled into nearly complete genomes of uncharacterized organisms.
537 Therefore, a straightforward strategy is to combine strengths of both supervised and
538 unsupervised learning procedures: using an unsupervised method to cluster the data, and then
539 assigning taxonomic groups to the bins by querying sequence databases. Such strategy speeds
540 up the analysis by annotating sequence clusters instead of single sequences.
541
542 The assembly of individual genomes from a metagenomic library can be accomplished
543 directly through *de novo* assembly or by using a reference genome (Hugerth et al., 2015, Luo

et al., 2012, Mehrshad et al., 2016). The assembly of whole genomes from a metagenome is only possible if the coverage of the genomes in the sample is sufficient. However, the efficiency of the assembly can be confounded by non-uniform coverage of the sample library, resulting either from gene abundance variation between taxa (evenness), and/or compositional bias of sequencing technologies. Therefore, whole genome assembly tends to be limited to the most abundant taxa in the community, and thus, very high coverage (above 20 terabases per metagenome) is required to assemble rare taxa (Luo, et al., 2012).

6.5.2. Normalization of metagenomic data

Characterizing the functional capacity of a microbial community necessitates building a list of gene functions and formulating an accurate estimate of the relative abundance of every gene, resulting in the identification of the proportion of genomes harboring a trait of interest (e.g., antibiotic or heavy metal resistance, nitrogen or carbon fixation). As contigs are treated as single sequences in most downstream analyses, the quantitative information for each taxon based on the number of unassembled reads assigned to particular taxa is lost. Therefore, assessing the relative abundance of contigs assigned to every single taxon in a metagenome is a crucial step in accurately characterizing the functional properties of a given microbial community. Basically, the number of reads mapped to each annotated gene is used as a proxy for its abundance in the sample (Luo et al., 2013). However, the resulting read counts are highly dependent on the sequencing approach (i.e. sequencing instrument), because the coverage biases across samples can vary significantly depending on the sequencing platform employed. Normalization of the sequence data is thus unavoidable in comparative metagenomics (Angly et al., 2009, Frank and Sorensen, 2011). Several approaches are commonly used. The compositional normalization approach is the most intuitive (Qin et al., 2010); it calculates relative abundance for every gene by dividing the abundance value

associated with each gene by the sum of abundance values for all genes identified in the metagenomic sample. The main issue of this "within sample" normalization method is that relative abundance for each gene is heavily dependent on the abundance of the total number of genes determined from the same metagenome, a factor that can lead to differential scaling across metagenomic samples (Manor and Borenstein, 2015).

Estimation of average genome sizes (AGS) is another normalization approach. The purpose of calculating AGS is to normalize the relative abundance of every gene in a given metagenome (Frank and Sorensen, 2011). AGS values can be biased because the probability of sampling a gene from a community varies with the size of the AGS for that community, i.e. the larger the AGS value, the higher the probability of sampling a given gene. Therefore differences in AGS between samples can lead to the spurious quantification of a given gene between metagenomes, i.e. genes present at an equal copy number per cell may appear variable across samples, while genes varying in copy number per cell may appear stable (Nayfach and Pollard, 2015).

To circumvent this normalization issue, another approach is based on read subsampling (Carcer et al., 2011). This normalization strategy aims to subsample *n* times (e.g. 10) an equal number of reads without replacement (e.g. 1 million of reads) from each metagenomic sample, in order to assess the data distribution uniformity of the iterated subsampling procedure, and thus control for subsampling bias that may occur between biological samples. Then, to investigate the biological meaning of differential abundances of genus/phylum across biological samples, the computed average of each diversity index will be compared between metagenome samples.

As the uniformity in terms of taxonomic diversity of subsampled reads was observed in several studies using simulated metagenomes (Garcia-Etxebarria et al., 2014, Mavromatis et al., 2007, Mende et al., 2012, Pignatelli and Moya, 2011), the normalization by read subsampling is definitely a promising approach.

6.6. Metatranscriptomic studies

The metatranscriptome, including both messenger and non-coding RNAs (rRNA, siRNA, etc.), provides information about the functional activity of a microbial community at a given time. As with other phenotypic traits, the characteristics of the metatranscriptome result from the interaction between the functional repertoire of the community (metagenotype) and biotic and abiotic environmental factors. Metatranscriptomic profiling is a powerful approach because it can provide insight into the regulatory networks and gene expression of a microbial community at the time of sampling.

One factor that must be addressed in the construction of a metatransciptome is the purification of mRNA from other RNA species present in the sample. Targeting bacterial mRNA is challenging because, unlike eukaryotic mRNA, bacterial transcripts are not polyadenylated and thus the classic oligo-dT-based method of mRNA capture cannot be employed. Furthermore, as the majority of RNA in a cell is composed of ribosomal and transfer RNAs (> 95%), metatranscriptomics typically requires a rRNA depletion step to enrich the mRNA fraction. Ribosomal RNA depletion techniques are based on rRNA specific probes (attached to biotin-streptavidin beads or columns) that capture rRNA molecules while mRNA and sRNA molecules are eluted. Until recently, the performance of these techniques was poor, particularly with complex bacterial communities such as the microbiota. Indeed, up to 60% of the resulting sequence data from some samples after depletion comprised rRNA reads. The

618 efficiency of subtractive hybridization can be improved for complex bacterial communities by
619 using customized, sample-specific rRNA probes (Stewart, 2013). After rRNA depletion,
620 enough mRNA must be recovered so that reamplification, which will restore the
621 overdominance of rRNAs in the sample, is avoided. Finally, another strategy is to skip the
622 rRNA removal step entirely and allocate more resources to a deeper sequencing effort. The
623 rRNA sequences can then be removed *in silico* (Urich et al., 2008).
624
625 Another challenging step is to prevent extensive RNA degradation during metatranscriptome
626 processing because mRNA stability can differ between microbial species and genes (Stewart,
627 2013). Therefore, it is crucial to snap-freeze samples in liquid nitrogen or, if liquid nitrogen is
628 unavailable, use a RNA preservation solution immediately after sampling. For example, when
629 harvesting microbial community RNA from aquatic environments, water samples must be
630 filtered immediately (10 min according to (Stewart, 2013, Tsementzi et al., 2014)) after
631 collection and frozen directly in liquid nitrogen in the field. In general, 1–3 L of
632 environmental sample will yield a minimum of 200 ng of total RNA (Stewart, 2013).
633 Importantly, it is recommended that additional samples are collected for DNA analysis in
634 order to perform downstream normalization of transcript abundance relative to gene or taxon
635 abundance (i.e. RNA:DNA expression ratios) (Stewart et al., 2012).
636
637 6.6.1. Transcripts abundance estimation
638 Transcript abundance of a given gene depends both on the number of gene copies (i.e.,
639 relative abundance of the taxon encoding the gene in the microbial community) and the
640 expression level of the individual gene. In other words, a given level of transcript abundance
641 may either result from a low expression of a gene belonging to several dominant taxa, or from
642 a high expression of a gene belonging to rare taxa. To accurately quantify the relative

abundance of a taxon specific transcript in a cDNA dataset, it is therefore crucial to map transcript sequences to the assembled genes of the corresponding metagenome.

A truly accurate quantification of the expression level of a given gene in order to detect real (i.e., biological) differential expression across samples must involve a normalization step. Indeed, using total read counts to estimate transcript abundance will result in a spurious estimate of expression level differences. Normalization consists of computing a relative expression ratio (Anders and Huber, 2010), defined as the transcript abundance divided by the abundance of its corresponding genomic sequence (i.e., cDNA/DNA).

6.6.2 Statistical methods to detect differentially expressed genes

The statistical power to detect differentially expressed genes depends essentially on the number of technical and more importantly biological replicates (true replicates) in an experiment. If a large number of replicates is available, issues related to data distribution can be avoided by using non-parametric methods such as rank-based or permutation tests. For experiments with a smaller numbers of replicates per condition, using distribution families, such as normal, Poisson and negative binomial distributions is a straightforward option (Oberg et al., 2012). Specifically, a Fisher's exact test, or a likelihood ratio test (Bullard et al., 2010, Marioni et al., 2008) are the most appropriate means of testing for genetic differential expression. However, the former should be interpreted with caution, as it is sensitive to the over-dispersion of data, and can underestimate the effect of biological variability for highly expressed genes (Anders and Huber, 2010). Therefore, the negative binomial distribution, by allowing larger variance, is better suited to cope with the strong variability for highly expressed genes (Oberg, et al., 2012, Tsementzi, et al., 2014). Subsequently, Bonferroni or FDR post-hoc corrections are necessary to resolve any false positive incidences.

6.7 RNA-Seq

There are many annotation tools currently available such as Prokka (Seemann, 2014) or even RAST (available as a web-server) (Aziz et al., 2012) that will accurately predict protein-coding genes and RNAs from a genome sequence. However, it is necessary to go beyond simple presence or absence of these features to gain deeper insight into the function of a particular organism. A powerful tool to examine the relationship between a genome and an organism's biological function is transciptomics. It is the study of the transcriptome, which is defined as "the complete set of transcripts in a cell, and their quantity, for a specific developmental stage or physiological condition" (Wang et al., 2009).

The first high-throughput technology applied to study the transcriptome was the microarray. In this assay, RNAs are extracted, reverse-transcribed into cDNA, coupled with a fluorescent dye and hybridized onto a chip (Miller and Tang, 2009). However, as discussed elsewhere (Wang, et al., 2009), even though this technology is medium-throughput and affordable, it has majors limitations including: (1) it requires special instrumentation for hybridization and scanning, (2) the dynamic range of fluorescence scanners are unable to cover the full range of gene expression because some signals will saturate while others are too close to background to be detected, (3) it is technically challenging to perform and (4) the microarray requires that the sequence targets are already known and it is thus not suitable for *de novo* discovery.

A method using NGS technologies, RNA-seq, has allowed researchers to overcome the limitations of the microarrays. Preparation for RNA-seq requires that the RNA is extracted and purified from a sample, and then sheared and converted into cDNA. The pool of cDNA is subsequently directly sequenced by NGS. Gene transcription levels are determined by mapping the cDNA reads to a reference sequence. More information on RNA-seq, including

693 an interesting list of tools for each step, can be found in (Creecy and Conway, 2015, Oshlack
694 et al., 2010). For a discussion of the challenges associated with transcriptomics using NGS
695 technologies see (Capobianco, 2014).
696
697 **6.8. Single cell sequencing**
698 We have already discussed the process of sequencing a single organism and a community of
699 organisms (metagenomics). It should not be overlooked, however, that the usual process of
700 sequencing the genome of a single organism can also be considered a community sequencing
701 project. That is to say that the multiple genomes extracted from a culture of a particular
702 microbe are not identical, as is often assumed, but a community of subtypes of the same
703 strain. Therefore the final genome sequence is in fact a consensus of every sub-strain genome
704 sequenced from the extracted sample. The drawback of this approach is that the heterogeneity
705 that exists among substrain genomes is lost (Barrick and Lenski, 2009, Lang et al., 2011).
706
707 The recent emergence of single-cell sequencing methods, nominated for method of the year in
708 2013 by *Nature Methods* (2014), grant us the ability to characterize genomic heterogeneity on
709 a cell to cell basis. Additionally, single-cell sequencing has been used to examine bacterial
710 pathogens and host cells directly from clinical samples without cultivation. Single-cell
711 sequencing has also been used to explore “microbial dark matter”, the large fraction of
712 microbes in nature that cannot be cultured (Rinke et al., 2013). Since a single-cell does not
713 contain enough DNA to prepare a sequencing library, a whole-genome amplification step is
714 necessary before sequencing can take place. There are two main methods presently employed
715 to amplify DNA in preparation for sequencing: Multiple Displacement Amplification (MDA)
716 and Multiple Annealing and Looping-Based Amplification Cycles (MALBAC). A
717 comparison of these methods can be found in (Chen et al., 2014).

6.9. Other applications

A comprehensive overview of all the NGS applications currently being used in microbiology is not feasible in one review. Although not discussed here, NGS is being used in a wide range of other applications, including tRNA sequencing (Zheng et al., 2015), epigenomic profiling (Chen et al., 2014, Lee et al., 2014), ribosome profiling (Ingolia, 2014), as well as in the detection of structural variations (SVs) (Chen et al., 2009) (Fan et al., 2014). Finally, the chromatin immunoprecipitation sequencing (ChIP-seq) procedure is a method designed to generate information on the location of genomic protein-DNA interactions by using NGS. Please see (Landt et al., 2012) for a detailed review of this method, including the guidelines produced by the ENCODE project. An in depth review of ChIP-seq and related methods (histone modification ChIP-seq, DNase-seq and FAIRE-seq) can be found in (Furey, 2012).

7. Conclusion

It is certain that sequencing technologies will continue to evolve, resulting in platforms that are more powerful and cheaper to use. Nonetheless, researchers interested in sequencing-based studies must make informed choices on the sequencing platform that can best aid them to achieve their research objectives. The four steps (and corresponding discussion) identified in this review (planning, library preparation, sequencing and data analysis) provide a framework that is relevant now and will remain relevant in the future, even as sequencing technology continues to advance. For the growing number of newcomers to the sequencing field, it is important to clearly define the objectives of your project and seek information from NGS experts such as experienced colleagues and core facility application specialists. Ultimately, this is the best way to save time and money and the most efficient way to achieve the desired results.

Acknowledgements

ATV holds a scholarship from Fonds de recherche du Québec - Nature et technologies (FRQNT). SJC is a research scholar of the Fonds de Recherche du Québec - Santé (FRQS). This work was funded by a grant from the Natural Sciences and Engineering Research Council of Canada (NSERC) to SJC, ND and AIC.

References

2014. Method of the Year 2013. Nat. Methods. 11, 1-1.

Acinas, S.G., Marcelino, L.A., Klepac-Ceraj, V., Polz, M.F., 2004. Divergence and redundancy of 16S rRNA sequences in genomes with multiple rrn operons. J Bacteriol. 186, 2629-2635.

Anders, S., Huber, W., 2010. Differential expression analysis for sequence count data. Genome Biol. 11, R106.

Angly, F.E., et al., 2009. The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. PLoS Comput Biol. 5, e1000593.

Ashelford, K.E., Chuzhanova, N.A., Fry, J.C., Jones, A.J., Weightman, A.J., 2005. At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. Appl Environ Microbiol. 71, 7724-7736.

Aziz, R.K., et al., 2012. SEED servers: high-performance access to the SEED genomes, annotations, and metabolic models. PLoS One. 7, e48053.

Bankevich, A., et al., 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 19, 455-477.

Barrick, J.E., Lenski, R.E., 2009. Genome-wide mutational diversity in an evolving population of *Escherichia coli*. Cold Spring Harb Symp Quant Biol. 74, 119-129.

Boers, S.A., van der Reijden, W.A., Jansen, R., 2012. High-throughput multilocus sequence typing: bringing molecular typing to the next level. PLoS One. 7, e39630.

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.

Bokulich, N.A., Subramanian, S., Faith, J.J., Gevers, D., Gordon, J.I., Knight, R., Mills, D.A., Caporaso, J.G., 2013. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. Nat Methods. 10, 57-59.

Bradnam, K.R., et al., 2013. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. Gigascience. 2, 10.

Bullard, J.H., Purdom, E., Hansen, K.D., Dudoit, S., 2010. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics. 11, 94.

Capobianco, E., 2014. RNA-Seq Data: A Complexity Journey. Comput Struct Biotechnol J. 11, 123-130.

Carcer, D.A., Denman, S.E., McSweeney, C., Morrison, M., 2011. Evaluation of subsampling-based normalization strategies for tagged high-throughput sequencing data sets from gut microbiomes. Appl Environ Microbiol. 77, 8795-8798.

Charuvaka, A., Rangwala, H., 2011. Evaluation of short read metagenomic assembly. BMC Genomics. 12 Suppl 2, S8.

Chen, K., et al., 2009. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods. 6, 677-681.

Chen, M., Song, P., Zou, D., Hu, X., Zhao, S., Gao, S., Ling, F., 2014. Comparison of multiple displacement amplification (MDA) and multiple annealing and looping-based amplification cycles (MALBAC) in single-cell sequencing. PLoS One. 9, e114520.

Chen, P., Jeannotte, R., Weimer, B.C., 2014. Exploring bacterial epigenomics in the next-generation sequencing era: a new approach for an emerging frontier. Trends Microbiol. 22, 292-300.

Chen, Y., Frazzitta, A.E., Litvintseva, A.P., Fang, C., Mitchell, T.G., Springer, D.J., Ding, Y., Yuan, G., Perfect, J.R., 2015. Next generation multilocus sequence typing (NGMLST) and the analytical software program MLSTEZ enable efficient, cost-effective, high-throughput, multilocus sequencing typing. Fungal Genet Biol. 75, 64-71.

Chikhi, R., Medvedev, P., 2014. Informed and automated k-mer size selection for genome assembly. Bioinformatics. 30, 31-37.
Chistoserdovai, L., 2010. Functional metagenomics: recent advances and future challenges. Biotechnol Genet Eng Rev. 26, 335-352.
Coil, D., Jospin, G., Darling, A.E., 2015. A5-miseq: an updated pipeline to assemble microbial genomes from Illumina MiSeq data. Bioinformatics. 31, 587-589.
Conway, T.C., Bromage, A.J., 2011. Succinct data structures for assembling large genomes. Bioinformatics. 27, 479-486.
Creecy, J.P., Conway, T., 2015. Quantitative bacterial transcriptomics with RNA-seq. Curr Opin Microbiol. 23, 133-140.
Dick, G.J., Andersson, A.F., Baker, B.J., Simmons, S.L., Thomas, B.C., Yelton, A.P., Banfield, J.F., 2009. Community-wide analysis of microbial genome sequence signatures. Genome Biol. 10, R85.
Ekblom, R., Wolf, J.B., 2014. A field guide to whole-genome sequencing, assembly and annotation. Evol Appl. 7, 1026-1042.
Engelbrektson, A., Kunin, V., Wrighton, K.C., Zvenigorodsky, N., Chen, F., Ochman, H., Hugenholtz, P., 2010. Experimental factors affecting PCR-based estimates of microbial species richness and evenness. ISME J. 4, 642-647.
Fan, X., Abbott, T.E., Larson, D., Chen, K., 2014. BreakDancer - Identification of Genomic Structural Variation from Paired-End Read Mapping. Curr Protoc Bioinformatics, 45:15.46.41-15.46.11.
Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M., et al., 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science. 269, 496-512.
Frank, J.A., Sorensen, S.J., 2011. Quantitative metagenomic analyses based on average genome size normalization. Appl Environ Microbiol. 77, 2513-2521.
Furey, T.S., 2012. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. Nat Rev Genet. 13, 840-852.
Garcia-Etxebarria, K., Garcia-Garcera, M., Calafell, F., 2014. Consistency of metagenomic assignment programs in simulated and real data. BMC Bioinformatics. 15, 90.
Gire, S.K., et al., 2014. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science. 345, 1369-1372.
Goecks, J., Nekrutenko, A., Taylor, J., Galaxy, T., 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 11, R86.
Goffeau, A., et al., 1996. Life with 6000 genes. Science. 274, 546, 563-547.
Golosova, O., Henderson, R., Vaskin, Y., Gabrielian, A., Grekhov, G., Nagarajan, V., Oler, A.J., Quinones, M., Hurt, D., Fursov, M., Huyen, Y., 2014. Unipro UGENE NGS pipelines and components for variant calling, RNA-seq and ChIP-seq data analyses. PeerJ. 2, e644.
Guttman, D.S., McHardy, A.C., Schulze-Lefert, P., 2014. Microbial genome-enabled insights into plant-microorganism interactions. Nat Rev Genet. 15, 797-813.
Hartmann, M., Howes, C.G., Abarenkov, K., Mohn, W.W., Nilsson, R.H., 2010. V-Xtractor: an open-source, high-throughput software tool to identify and extract hypervariable regions of small subunit (16S/18S) ribosomal RNA gene sequences. J Microbiol Methods. 83, 250-253.
Hayden, E.C., 2014. Technology: The $1,000 genome. Nature. 507, 294-295.
Head, S.R., Komori, H.K., LaMere, S.A., Whisenant, T., Van Nieuwerburgh, F., Salomon, D.R., Ordoukhanian, P., 2014. Library construction for next-generation sequencing: overviews and challenges. Biotechniques. 56, 61-77.

Hugerth, L.W., Larsson, J., Alneberg, J., Lindh, M.V., Legrand, C., Pinhassi, J., Andersson, A.F., 2015. Metagenome-assembled genomes uncover a global brackish microbiome. Genome Biol. 16, 279.
Huson, D.H., Mitra, S., Ruscheweyh, H.J., Weber, N., Schuster, S.C., 2011. Integrative analysis of environmental sequences using MEGAN4. Genome Res. 21, 1552-1560.
Ingolia, N.T., 2014. Ribosome profiling: new views of translation, from single codons to genome scale. Nat Rev Genet. 15, 205-213.
Kislyuk, A., Bhatnagar, S., Dushoff, J., Weitz, J.S., 2009. Unsupervised statistical clustering of environmental shotgun sequences. BMC Bioinformatics. 10, 316.
Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., Glockner, F.O., 2013. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Res. 41, e1.
Koren, S., Phillippy, A.M., 2015. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. Curr Opin Microbiol. 23, 110-120.
Krause, L., Diaz, N.N., Goesmann, A., Kelley, S., Nattkemper, T.W., Rohwer, F., Edwards, R.A., Stoye, J., 2008. Phylogenetic classification of short environmental DNA fragments. Nucleic Acids Res. 36, 2230-2239.
Land, M., et al., 2015. Insights from 20 years of bacterial genome sequencing. Funct Integr Genomics. 15, 141-161.
Lander, E.S., et al., 2001. Initial sequencing and analysis of the human genome. Nature. 409, 860-921.
Landt, S.G., et al., 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. 22, 1813-1831.
Lang, G.I., Botstein, D., Desai, M.M., 2011. Genetic variation and the fate of beneficial mutations in asexual populations. Genetics. 188, 647-661.
Langille, M.G., et al., 2013. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotechnol. 31, 814-821.
Lee, Y.K., Jin, S., Duan, S., Lim, Y.C., Ng, D.P., Lin, X.M., Yeo, G., Ding, C., 2014. Improved reduced representation bisulfite sequencing for epigenomic profiling of clinical samples. Biol Proced Online. 16, 1.
Li, J.W., Schmieder, R., Ward, R.M., Delenick, J., Olivares, E.C., Mittelman, D., 2012. SEQanswers: an open access community for collaboratively decoding genomes. Bioinformatics. 28, 1272-1273.
Liu, X., Pande, P.R., Meyerhenke, H., Bader, D.A., 2013. PASQUAL: Parallel Techniques for Next Generation Genome Sequence Assembly. IEEE Trans. Parallel Distrib. Syst. 24, 977-986.
Liu, Y., Schmidt, B., Maskell, D.L., 2011. Parallelized short read assembly of large genomes using de Bruijn graphs. BMC Bioinformatics. 12, 354.
Luo, C., Rodriguez, R.L., Konstantinidis, K.T., 2013. A user's guide to quantitative and comparative analysis of metagenomic datasets. Methods Enzymol. 531, 525-547.
Luo, C., Tsementzi, D., Kyrpides, N.C., Konstantinidis, K.T., 2012. Individual genome assembly from complex community short-read metagenomic datasets. ISME J. 6, 898-901.
Macmanes, M.D., Eisen, M.B., 2013. Improving transcriptome assembly through error correction of high-throughput sequence reads. PeerJ. 1, e113.
Mande, S.S., Mohammed, M.H., Ghosh, T.S., 2012. Classification of metagenomic sequences: methods and challenges. Brief Bioinform. 13, 669-681.
Manor, O., Borenstein, E., 2015. MUSiCC: a marker genes based framework for metagenomic normalization and accurate profiling of gene abundances in the microbiome. Genome Biol. 16, 53.

897 Mapleson, D., Drou, N., Swarbreck, D., 2015. RAMPART: a workflow management system
898 for de novo genome assembly. Bioinformatics. 31, 1824-1826.
899 Margulies, M., et al., 2005. Genome sequencing in microfabricated high-density picolitre
900 reactors. Nature. 437, 376-380.
901 Marine, R., Polson, S.W., Ravel, J., Hatfull, G., Russell, D., Sullivan, M., Syed, F., Dumas,
902 M., Wommack, K.E., 2011. Evaluation of a transposase protocol for rapid generation of
903 shotgun high-throughput sequencing libraries from nanogram quantities of DNA. Appl
904 Environ Microbiol. 77, 8071-8079.
905 Marinier, E., Brown, D.G., McConkey, B.J., 2015. Pollux: platform independent error
906 correction of single and mixed genomes. BMC Bioinformatics. 16, 10.
907 Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M., Gilad, Y., 2008. RNA-seq: an
908 assessment of technical reproducibility and comparison with gene expression arrays.
909 Genome Res. 18, 1509-1517.
910 Mavromatis, K., et al., 2007. Use of simulated data sets to evaluate the fidelity of
911 metagenomic processing methods. Nat Methods. 4, 495-500.
912 McElroy, K., Thomas, T., Luciani, F., 2014. Deep sequencing of evolving pathogen
913 populations: applications, errors, and bioinformatic solutions. Microb Inform Exp. 4, 1.
914 McHardy, A.C., Martin, H.G., Tsirigos, A., Hugenholtz, P., Rigoutsos, I., 2007. Accurate
915 phylogenetic classification of variable-length DNA fragments. Nat Methods. 4, 63-72.
916 Medvedev, P., Georgiou, K., Myers, G., Brudno, M., 2007. Computability of models for
917 sequence assembly. In: R. Giancarlo, S. Hannenhalli (Eds.), Algorithms in Bioinformatics,
918 Proceedings, Vol. 4645, Springer-Verlag Berlin, Berlin, pp. 289-301.
919 Mehrshad, M., Amoozegar, M.A., Ghai, R., Shahzadeh Fazeli, S.A., Rodriguez-Valera, F.,
920 2016. Genome reconstruction from metagenomic datasets reveals novel microbes in the
921 brackish waters of the Caspian Sea. Appl Environ Microbiol.
922 Mende, D.R., Waller, A.S., Sunagawa, S., Jarvelin, A.I., Chan, M.M., Arumugam, M., Raes,
923 J., Bork, P., 2012. Assessment of metagenomic assembly using simulated next generation
924 sequencing data. PLoS One. 7, e31386.
925 Miller, M.B., Tang, Y.W., 2009. Basic concepts of microarrays and potential applications in
926 clinical microbiology. Clin Microbiol Rev. 22, 611-633.
927 Miyamoto, M., et al., 2014. Performance comparison of second- and third-generation
928 sequencers using a bacterial genome with two chromosomes. BMC Genomics. 15, 699.
929 Nagarajan, N., Pop, M., 2013. Sequence assembly demystified. Nat Rev Genet. 14, 157-167.
930 Nayfach, S., Pollard, K.S., 2015. Average genome size estimation improves comparative
931 metagenomics and sheds light on the functional ecology of the human microbiome.
932 Genome Biol. 16, 51.
933 Oberg, A.L., Bot, B.M., Grill, D.E., Poland, G.A., Therneau, T.M., 2012. Technical and
934 biological variance structure in mRNA-Seq data: life in the real world. BMC Genomics.
935 13, 304.
936 Oshlack, A., Robinson, M.D., Young, M.D., 2010. From RNA-seq reads to differential
937 expression results. Genome Biol. 11, 220.
938 Parnell, L.D., Lindenbaum, P., Shameer, K., Dall'Olio, G.M., Swan, D.C., Jensen, L.J.,
939 Cockell, S.J., Pedersen, B.S., Mangan, M.E., Miller, C.A., Albert, I., 2011. BioStar: an
940 online question & answer resource for the bioinformatics community. PLoS Comput Biol.
941 7, e1002216.
942 Picelli, S., Bjorklund, A.K., Reinius, B., Sagasser, S., Winberg, G., Sandberg, R., 2014. Tn5
943 transposase and tagmentation procedures for massively scaled sequencing projects.
944 Genome Res. 24, 2033-2040.
945 Pignatelli, M., Moya, A., 2011. Evaluating the fidelity of de novo short read metagenomic
946 assembly using simulated data. PLoS One. 6, e19984.

947 Pop, M., 2009. Genome assembly reborn: recent computational challenges. Brief Bioinform.
948 10, 354-366.
949 Prakash, T., Taylor, T.D., 2012. Functional assignment of metagenomic data: challenges and
950 applications. Brief Bioinform. 13, 711-727.
951 Pulido-Tamayo, S., Sanchez-Rodriguez, A., Swings, T., Van den Bergh, B., Dubey, A.,
952 Steenackers, H., Michiels, J., Fostier, J., Marchal, K., 2015. Frequency-based haplotype
953 reconstruction from deep sequencing data of bacterial populations. Nucleic Acids Res. 43,
954 e105.
955 Qin, J., et al., 2010. A human gut microbial gene catalogue established by metagenomic
956 sequencing. Nature. 464, 59-65.
957 Reith, M.E., et al., 2008. The genome of *Aeromonas salmonicida* subsp. *salmonicida* A449:
958 insights into the evolution of a fish pathogen. BMC Genomics. 9, 427.
959 Rinke, C., et al., 2013. Insights into the phylogeny and coding potential of microbial dark
960 matter. Nature. 499, 431-437.
961 Saeed, I., Tang, S.L., Halgamuge, S.K., 2012. Unsupervised discovery of microbial
962 population structure within metagenomes using nucleotide base composition. Nucleic
963 Acids Res. 40, e34.
964 Salzberg, S.L., et al., 2012. GAGE: A critical evaluation of genome assemblies and assembly
965 algorithms. Genome Res. 22, 557-567.
966 Schmieder, R., Edwards, R., 2011. Quality control and preprocessing of metagenomic
967 datasets. Bioinformatics. 27, 863-864.
968 Seemann, T., 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 30, 2068-
969 2069.
970 Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., Huttenhower, C., 2012.
971 Metagenomic microbial community profiling using unique clade-specific marker genes.
972 Nat Methods. 9, 811-814.
973 Simpson, J.T., Durbin, R., 2012. Efficient de novo assembly of large genomes using
974 compressed data structures. Genome Res. 22, 549-556.
975 Stewart, F.J., 2013. Preparation of microbial community cDNA for metatranscriptomic
976 analysis in marine plankton. Methods Enzymol. 531, 187-218.
977 Stewart, F.J., Ulloa, O., DeLong, E.F., 2012. Microbial metatranscriptomics in a permanent
978 marine oxygen minimum zone. Environ Microbiol. 14, 23-40.
979 Strous, M., Kraft, B., Bisdorf, R., Tegetmeyer, H.E., 2012. The binning of metagenomic
980 contigs for microbial physiology of mixed cultures. Front Microbiol. 3, 410.
981 Teeling, H., Waldmann, J., Lombardot, T., Bauer, M., Glockner, F.O., 2004. TETRA: a web-
982 service and a stand-alone program for the analysis and comparison of tetranucleotide usage
983 patterns in DNA sequences. BMC Bioinformatics. 5, 163.
984 Tsementzi, D., Poretsky, R., Rodriguez, R.L., Luo, C., Konstantinidis, K.T., 2014. Evaluation
985 of metatranscriptomic protocols and application to the study of freshwater microbial
986 communities. Environ Microbiol Rep. 6, 640-655.
987 Urich, T., Lanzen, A., Qi, J., Huson, D.H., Schleper, C., Schuster, S.C., 2008. Simultaneous
988 assessment of soil microbial community structure and function through analysis of the
989 meta-transcriptome. PLoS One. 3, e2527.
990 van Dijk, E.L., Jaszczyszyn, Y., Thermes, C., 2014. Library preparation methods for next-
991 generation sequencing: tone down the bias. Exp Cell Res. 322, 12-20.
992 Venter, J.C., et al., 2001. The sequence of the human genome. Science. 291, 1304-1351.
993 Vetrovsky, T., Baldrian, P., 2013. The variability of the 16S rRNA gene in bacterial genomes
994 and its consequences for bacterial community analyses. PLoS One. 8, e57923.

995 Vincent, A.T., Boyle, B., Derome, N., Charette, S.J., 2014. Improvement in the DNA
996 sequencing of genomes bearing long repeated elements. J Microbiol Methods. 107, 186-
997 188.
998 Vincent, A.T., Tanaka, K.H., Trudel, M.V., Frenette, M., Derome, N., Charette, S.J., 2015.
999 Draft genome sequences of two *Aeromonas salmonicida* subsp. *salmonicida* isolates
1000 harboring plasmids conferring antibiotic resistance. FEMS Microbiol Lett. 362, fnv002.
1001 Wang, Z., Gerstein, M., Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics.
1002 Nat Rev Genet. 10, 57-63.
1003 Watson, M., 2014. Quality assessment and control of high-throughput sequencing data. Front
1004 Genet. 5, 235.
1005 Ye, Y., Tang, H., 2015. Utilizing de Bruijn graph of metagenome assembly for
1006 metatranscriptome analysis. Bioinformatics.
1007 Zheng, G., Qin, Y., Clark, W.C., Dai, Q., Yi, C., He, C., Lambowitz, A.M., Pan, T., 2015.
1008 Efficient and quantitative high-throughput tRNA sequencing. Nat Methods. 12, 835-837.
1009 Zhou, X., Rokas, A., 2014. Prevention, diagnosis and treatment of high-throughput
1010 sequencing data pathologies. Mol Ecol. 23, 1679-1700.
1011

1012

1013

1014 **Table 1. Most common applications of next-generation sequencing**

Application	Library type	Relative importance of the sequencer features[a]	Recommended instrument
Genomic diversity and phylogeny	Shotgun	Consensus accuracy *** Throughput ** Read length **	All
Structural analysis of genome	Shotgun + mate pairs	Consensus accuracy **** Read length ***	MiSeq
Gene expression	Reverse transcription + shotgun	Throughput ****** Read accuracy *	HiSeq, Ion Proton
Population diversity studies - Species composition	Amplicons	Read accuracy **** Read length ***	MiSeq, Ion PGM
Population diversity studies - Gene function composition	Shotgun	Read length *** Read accuracy ** Throughput **	MiSeq for assembly HiSeq, Ion Proton for quantification
Multi-locus sequence typing	Amplicons	Consensus accuracy **** Read length ***	All

1015 a: Indicated by the number of asterisk on a total of seven.

1016

Table 2. The most common sequencer of next-generation sequencing (September 2015)

Apparatus	Throughput range (Gb) [a]	Read length range (bp)	Strength	Weakness
Sanger Sequencing ABI3730 96 capillary system	0.0003	Up to 1 kb	Sequence quality and length	Cost and throughput
ThermoFisher				
Ion PGM	0.08-2	Up to 400	Read length and speed	Long homopolymers
Ion Proton	10-15	Up to 200	Throughput and speed	Long homopolymers
Ion S5 or S5XL	0.6 – 15	Up to 400	Read length, throughput and speed	Long homopolymers
Illumina				
MiSeq	0.3-15	1x50 to 2x300	Read length	Run length
NextSeq	10-120	1x75 to 2x150	Throughput	Run length
HiSeq (2500)	10-800	1x50 to 2x125	Read accuracy and throughput	High initial investment, run length
HiSeq X Ten	900-1800	1x50 to 2x150	Read accuracy and throughput	Enormous initial investment, run length

a: the throughput ranges are determined by available kits and run modes on a per run basis.

1021 **Table 3. Examples of optimal number of samples per instrument[a]**

Application	Instrument	Throughput	How	Number of samples in a year
Bacterial genome sequencing 50X coverage	Illumina MiSeq	Paired-end 2 x 300 nt, 20 M reads per run, output 12 Gb	48 samples per run, 2 runs per week, 50 weeks a year	4800 samples
	Illumina HiSeq 2000	Paired-end 2 x 125 nt, 150 M reads per lane, 16 lanes per run, output 600 Gb	150 samples per lane, 16 lanes per instrument, 25 runs per year	60 000 samples
	Ion PGM, 318 chip	Single read > 300 nt avg, 4 M reads per run, output 1,2 Gb	5 samples per run, 2 runs per day, 4 days a week, 50 weeks a year	2000 samples
Bacterial RNA Sequencing 10M reads per sample	Illumina HiSeq 2000	Single read 100 nt, 150 M reads per lane, 16 lanes per run, output 300 Gb	15 samples per lane, 16 lanes per instrument, 35 runs per year	8400 samples
	Ion Proton, PI chip or Ion S5 540 chip	Single read, >100 nt avg, 60M reads per run, output	6 samples per chip, 2 runs per day, 4 days a week, 50 weeks a year	2400 samples
Amplicon analysis > 25K reads per sample	Illumina MiSeq	Paired-end 2 x 300 nt, 15 M reads per run, output 9 Gb	384 samples per run, 2 runs per week, 50 weeks a year	38 400 samples
	Ion PGM, 318 chip	Single read > 300 nt avg, 4 M reads per run, output 1,2 Gb	96 samples per run, 2 runs per day, 4 days a week, 50 weeks a year	38 400 samples

1022 a: Based on instrument available on August 2015.

Table 4. Features of *de novo* assemblies produced by different assemblers for the *A. salmonicida* subsp. *salmonicida* strain 01-B526.

	Assemblers		
Features	**A5**	**Ray[a]**	**SPAdes[b]**
# contigs (≥500 bp)	140	95	159
Largest contig (bp)	274 318	376 027	375 980
N50 (bp)	115 661	108 909	108 386
Genome fraction (%)	97.622	88.100	97.190

a: The kmer length (117) used for Ray was found with KmerGenie version 1.6663 (Chikhi and Medvedev, 2014).

b: The kmer lengths used with SPAdes were 21, 33, 55, 77, 99 and 127 as recommended in the manual for sequencing reads produced by a MiSeq apparatus. The coverage cutoff was turned ON and the threshold was auto-detected.

Box

BOX1. K-mers and read length

Most data analysis packages use K-mers, which are defined as all the possible substrings of K length found in a string. For example, the sequence GGATCTGATAC contains 4 K-mers of 8 nucleotides

```
                Sequence: GGATCTGATAC
   K-mers of K-length=8: GGATCTGA
                          GATCTGAT
                           ATCTGATA
                            TCTGATAC
```

The number of K-mers shared between two sequences defines how similar they are to each other.

Longer read length enables both the use of longer K-mers and a higher number of K-mers between related sequences to increase precision.

Figures

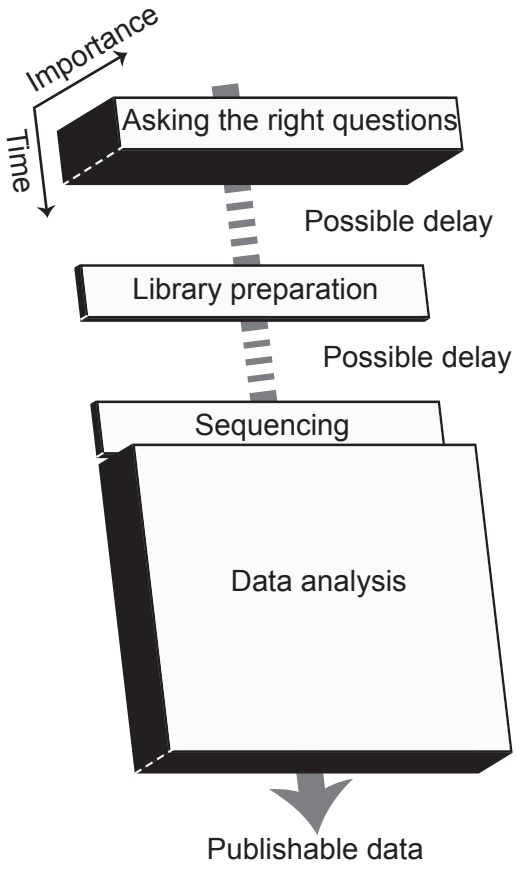
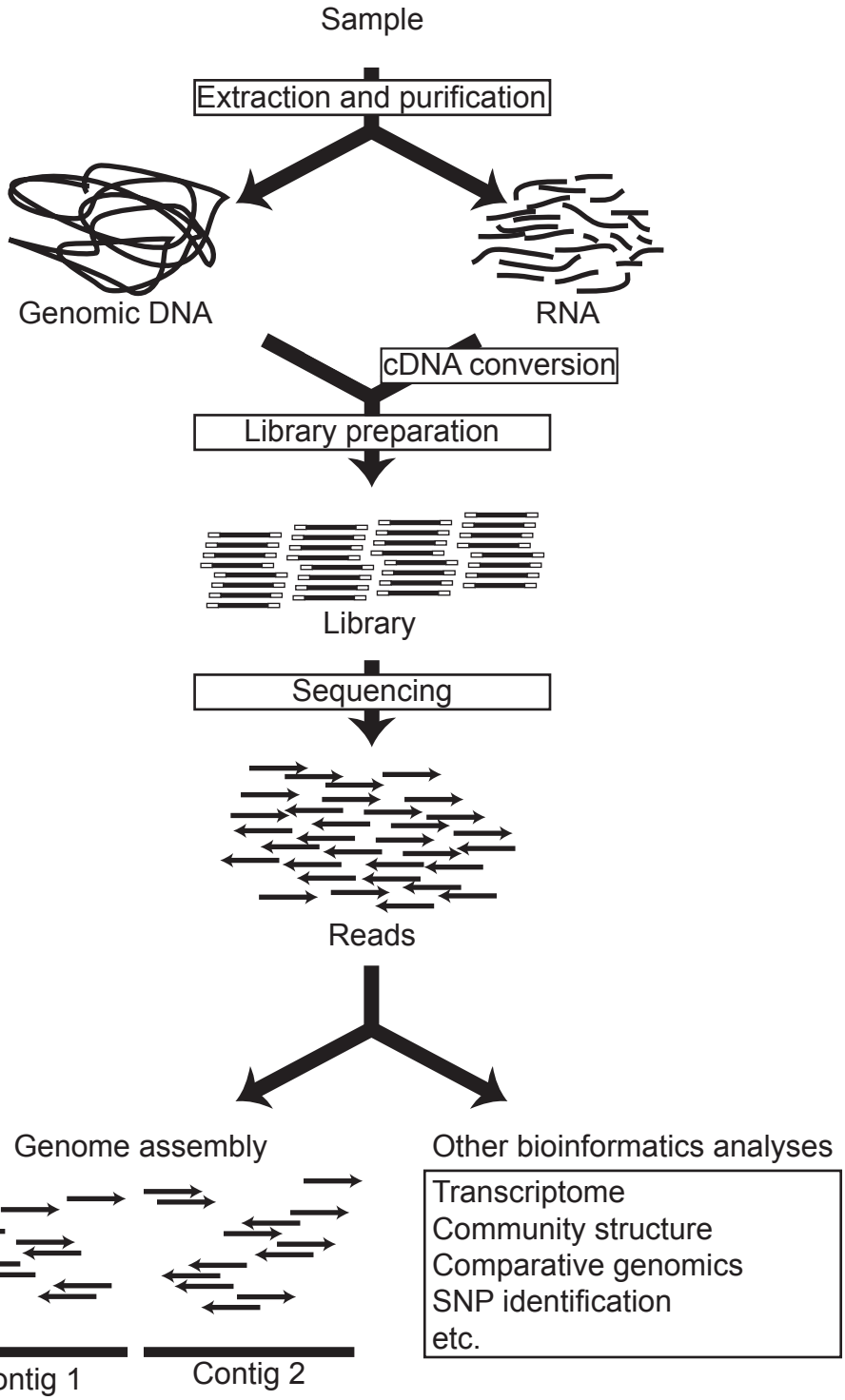


Figure 1. Conceptual workflow of a complete NGS based project with the relative importance and time spent for each step.



1056

1057 **Figure 2. General overview of the NGS procedure.**

1058

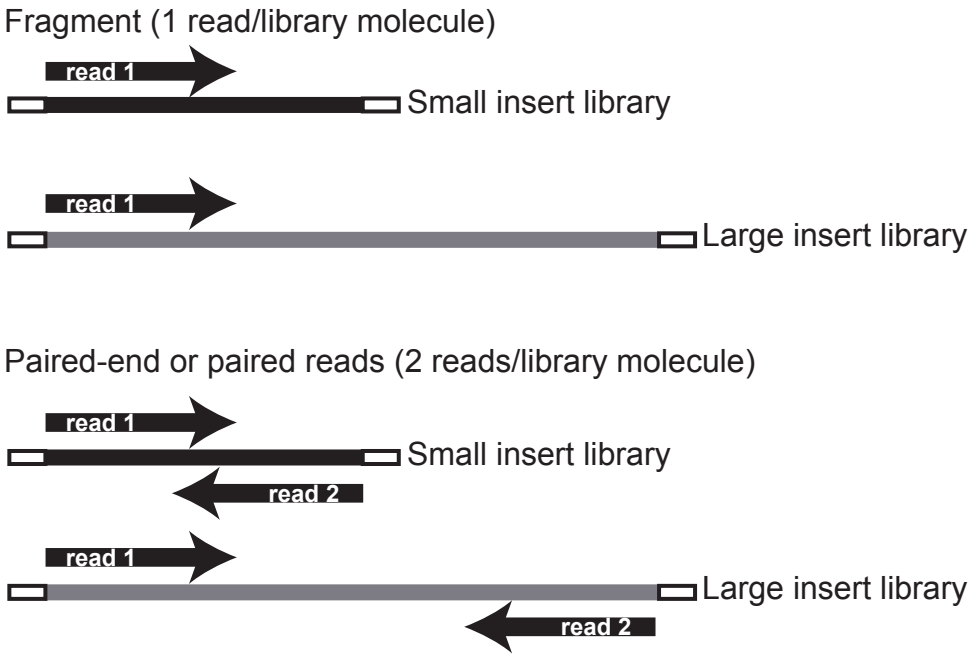


Figure 3. Fragment versus paired-end reads. Library molecules are illustrated with the technology specific adapters shown with white rectangles. Inserts are represented in grey (large libraries) or black (small libraries) while sequencing reads are represented by arrows.

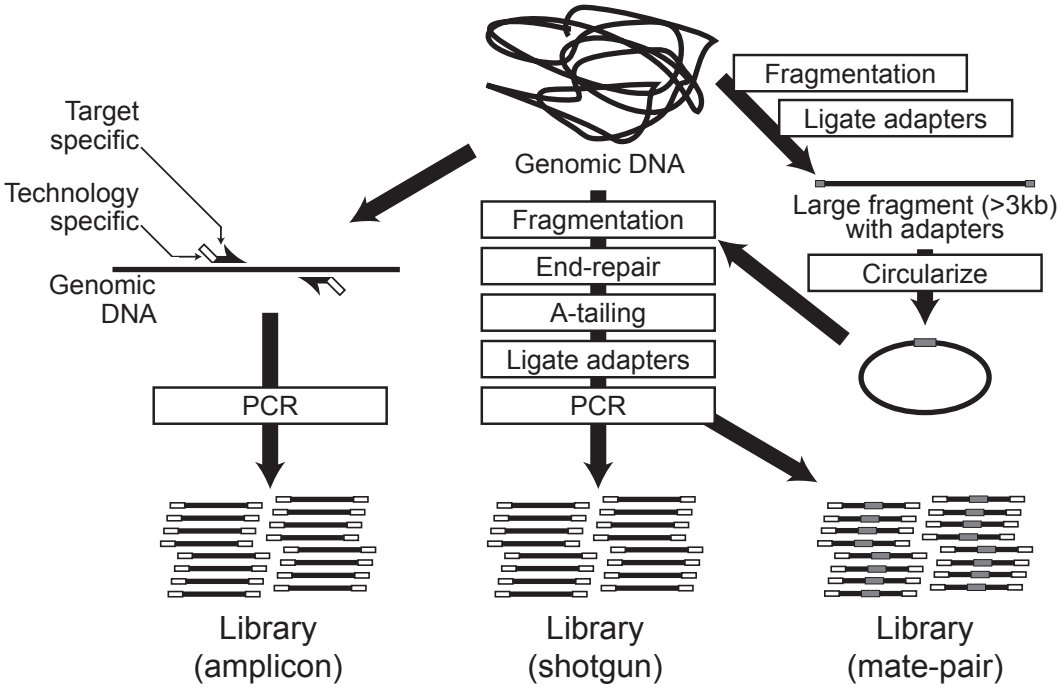


Figure 4. Overview of library preparation for NGS. Three kinds of libraries can be produced. The one on the left is produced by performing a PCR on purified genomic DNA. The primers used include in their 5' region additional sequences (white rectangles) required by the sequencing technology to perform NGS. These additional sequences vary from one technology to another. These sequences can also include barcodes to allow multiplex sequencing of many libraries in one machine run (barcodes are not shown on the figure). The shotgun library illustrated in the middle involves many steps to generate DNA fragments surrounded on each side by adapters required by the sequencing technology. Finally, the mate-pair libraries offer the possibility of including sequences that are physically linked together but at a certain distance in the same DNA fragment. This is possible by doing an initial fragmentation step followed by the addition of a set of circularization adapters (grey rectangles). These adapters allow circularization of the DNA fragments. These circular molecules are then re-fragmented as described previously.