Next-generation sequencing technologies and their impact on microbial genomics

Brian M. Forde and Paul W. O’Toole

Advance Access publication date 11 January 2013

Abstract

Next-generation sequencing technologies have had a dramatic impact in the field of genomic research through the provision of a low cost, high-throughput alternative to traditional capillary sequencers. These new sequencing methods have surpassed their original scope and now provide a range of utility-based applications, which allow for a more comprehensive analysis of the structure and content of microbial genomes than was previously possible. With the commercialization of a third generation of sequencing technologies imminent, we discuss the applications of current next-generation sequencing methods and explore their impact on and contribution to microbial genome research.

Keywords: NGS; prokaryotes; genome sequencing; resequencing; RNA-seq; metagenomics

INTRODUCTION

In 1995, almost 20 years after Sanger developed the chain termination sequencing strategy, researchers at the Institute of genomic Research (TIGR)—now the J. Craig Venter Institute (http://www.jcvi.org)—sequenced the first genomes of cellular organisms; the bacterial species Haemophilus influenzae [1] and Mycoplasma genitalium [2]. The publication of these genomes not only provided a glimpse of the complete genomes of a ‘living organism’ but revolutionized the field of genomics by introducing key improvements to sequencing strategies such as the usage of paired-end sequencing [3, 4] and adoption of the whole genome shotgun approach [5]. The complete sequences of these first bacterial genomes were quickly followed by the larger genomes of Bacillus subtilis [6] and Escherichia coli [7] and the genomes of the eukaryotes Saccharomyces cerevisiae [8], Caenorhabditis elegans [9], Arabidopsis thaliana [10], Drosophila melanogaster [11] and ultimately the human genome [12, 13]. However, despite advances in sequencing methodologies, sequencing cost remained relatively high and prohibitively expensive for most research groups. The high cost per base and low throughput of the traditional slab gel or capillary electrophoresis (CE) sequencing platforms prompted the development of so-called next-generation sequencing (NGS) technologies that provided a much greater throughput at a substantially lower cost [14]. The technical details of NGS technologies have been extensively reviewed elsewhere [14–16] and are not discussed here. Instead, this review will summarize recent developments of NGS, and explore their contribution to the field of microbial genomics. Furthermore, this review focuses on the application of NGS technologies to the sequencing and analysis
of bacterial genomes, the genome sequencing and genome analysis of viruses and other nonprokaryotic microbes is not discussed.

MICROBIAL GENOME SEQUENCING BY NGS METHODS

By 2004, before the introduction of NGS technologies, 192 bacterial genome sequences had been fully completed and published. However, since 2005 an additional 1566 bacterial genome sequences have been completed, published and deposited in online databases (Figure 1). As of 12 October 2012, 3173 (complete and draft) Bacterial, Archaeal and Eukaryal genomes have been deposited online of which 2847 are bacterial (Figure 1). In addition, there are a further 5156 genome projects classified as ‘in progress’ of which 4226 are bacterial (Figure 1) (http://www.genomesonline.org).

Prior to the development of NGS technologies, sequencing methodologies based on the Sanger sequencing chemistry dominated the genome sequencing industry. Automated Sanger capillary-based sequencing technologies, which rely on clone libraries, were too expensive, time consuming and labor intensive for the routine sequencing of bacterial genomes [17]. Consequently, bacterial sequencing projects focused on model organisms or those with practical applications, i.e. medically or industrially important species. Furthermore, this biased focus on single species and strains ignored the extreme diversity of the microbial world [18, 19] where even the most closely related species/strains can vary greatly in the composition of their ‘dispensable genes’ [20].

Resequencing

Using NGS technologies to sequence bacterial genomes was not without attendant problems. Early in the development of NGS technologies read lengths were short, ranging between 35 bp (Illumina) and 100 bp (Roche 454); significantly shorter than the 900 bp obtainable with automated capillary sequencers. De novo genome assembly with short read technologies results in highly fragmented assemblies, because of the reduction in assembly quality with decreasing read lengths [21]. In assemblies derived from NGS reads, all gaps are typically as a direct result of unresolved repeats [22]. With short reads, repetitive segments longer than the read length

Figure I: Published genomes. (A) Published genome sequences for the three domains of life as of April 2012. (B) Distribution of completed and on-going genome projects amongst the three domains. (C) Phylogenetic distribution of bacterial genome sequencing projects. Source: http://www.genomesonline.org.
become more common, which increases the complexity of the assembly problem resulting in more fragmented assemblies. Consequently, it was believed that NGS-derived short read sequencing data would be unsuitable for de novo genome assembly. However, the low cost and high-throughput of these platforms was ideally suited to the resequencing of whole genomes.

Forty years of genome sequencing has resulted in an abundance of publically available genome sequences stored in online databases. This catalog of genomes—containing representatives from nearly all phyla—provides a bank of reference species to which reads are aligned to reconstruct the genome of the target organism [23, 24]. Accurately resequencing a genome requires that the reads must be long enough to allow for their correct mapping to the reference genome. Additionally, the number of reads which map to the reference increases with increasing read length, stabilizing at read lengths of approximately 40 nt [21, 25]. Although the Roche/454 platform has been used in resequencing projects [26, 27], the short read lengths, extremely high-throughput and lower per-base cost of the Illumina and Solid platforms has seen them become the most frequently used instruments for genome resequencing. For example: characterizing antibiotic resistance in Mycobacterium tuberculosis [28], investigating genome variation and diversity in Salmonella enterica enterica, serovar Typhi [29] or more recently estimating the mutation rate in M. tuberculosis during latent infection [30], have all benefited from usage of NGS platforms.

*M. tuberculosis* is a pathogenic bacterium and the causative agent of tuberculosis. The emergence of multidrug-resistant strains poses a particular global health risk. In active infections, *M. tuberculosis* is treated with multiple antibiotics to prevent the emergence of new drug resistant strains; in active infections, the presence of large numbers of replicating organisms is thought to increase the likelihood of the bacterium developing new drug-resistant mutations. However, in latent infection, it was believed that it was unlikely the bacterium would develop new mutations and treatment typically involved one antibiotic, isoniazid. In a recent study which sequenced and compared *M. tuberculosis* strains from active, latent and reactivated infections, Ford et al. [30] discovered that the mutation rate in strains isolated from latent and active infections is similar. The authors suggest, based on the pattern of polymorphisms they detected, that the *in vivo* mutation rate is due to DNA oxidation. Consequently, *M. tuberculosis* will continue to acquire mutations during latency. Moreover, treatment of latent infections with only isoniazid poses a significant risk and could result in the emergence of isoniazid-resistant strains. This study illustrates the power of microbial genome NGS for informing clinical practice.

**De novo sequencing and assembly**

Despite short read lengths, NGS technologies have been and continue to be successfully applied in de novo bacterial genome sequencing projects. With the release of the Roche/454 sequencing platform, Margulies et al. [31] demonstrated the practical applications of de novo genome sequencing using NGS technologies. The de novo NGS of the 580 kb genome of *M. genitalium* yielded 25 contigs covering 99.5% of the nonrepetitive portion of the genome; the original sequencing of *M. genitalium* yielded 28 contigs ranging in size from 606 to 73,351 bp [2]. Sixteen of the 25 gaps were as a direct result of unresolved repeats, highlighting the difficulties posed by these regions during the assembly process.

The read lengths and throughput of NGS technologies have steadily increased since 2005. Increasing read length improves assembly quality by reducing the number of gaps and increasing contig size [21, 22, 32]. Furthermore, due to the small size of bacterial genomes, increasing coverage can compensate for short read lengths and reduce the number of gaps which require closure, albeit at a greater cost [22, 33]. The current 454 instrument produces read lengths approaching those from capillary gel-based platforms. However, the most significant advancement in NGS technologies was the introduction of paired reads. Mate-pair information is critical in the identification and resolution of repeat induced assembly errors (Figure 2) [11, 34]. Collapsed and expanded repeats are readily identified by contraction or elongation of the distances between mate-pairs. Repeat induced excision errors typically result when a collapsed repeat forces contigs out of the assembly and as a consequence two contigs are created where there should be one. Assembly rearrangements typically arise when multiple copies of interspaced repeats are located close to each other. The incorrect assembly of these repetitive regions can result in errors in contig order. Repeat-induced rearrangements are typically identified by an elongation of the distance between mate-pairs. However, in order for mate-pairs to correctly resolve repeat-induced
Figure 2: The three main types of repeat-induced errors encountered in genome assembly projects. The errors can all be identified by abnormalities in the mate-pair information, e.g. elongated or truncated distance between mates or incorrect orientation. (A) Collapsed repeat; (B) excision; (C) repeat-induced rearrangement of contig order.
errors one read of the pair must be ‘anchored’ outside of the repetitive region [35].

One approach used to improve assembly quality is the adoption of a hybrid sequencing strategy. Initial hybrid strategies involved the combination of sequencing data from both CE and 454 platforms. Sequencing in this hybrid manner improves assemblies through increasing coverage, reducing the number of gaps and also improved existing Sanger assemblies by reducing the number of gaps resulting from cloning bias [36–38]. Hybrid sequencing strategies have been extended to use only NGS sequencing technologies; two second generation sequencing (SGS) platforms, such as Illumina and 454 [39], or combinations of SGS and third generation sequencing (TGS) platforms [40]. Hybrid assembly of bacterial genome sequences is most effective when using complementary sequencing technologies. For example, a hybrid approach using both 454 and Illumina platforms produces de novo assemblies whose quality is at least on a par with those produced using only Sanger sequencing [39]. Furthermore, the homopolymer errors inherent in 454 derived reads can be detected and corrected using the higher coverage Illumina platform; downstream annotation issues are now resolved during the assembly process [39, 41–43].

**METAGENOMICS**

NGS platforms have proven to be effective tools for the de novo sequencing and re-sequencing of bacterial genomes. However, culturable bacteria represent only a small fraction of the total microbial diversity which exists in the world [44]. To fully understand and investigate microbial diversity, researchers have turned to the field of metagenomics. Metagenomics refers to culture-independent methods used to explore the genetic diversity, population structures and interactions of microbial communities in their ecosystems. Initial metagenomics studies, exploiting traditional sequencing technologies, typically involved the examination of microbial diversity through targeted sequencing of 16S rRNA gene amplicons [45–47] or through whole community shotgun metagenomics [48, 49].

**16S rRNA gene-based community analysis**

The 16S rRNA gene is generally conserved in all bacteria but possess enough interspecies variability to allow for its use as a molecular tool for bacterial identification [50, 51]. With sufficiently long reads, obtainable through traditional slab gel or CE platforms, bacterial amplicons could be confidently assigned to genus- and in some cases species-level. However, Sanger sequencing is time consuming, labor intensive and the requirement of a cloning step can lead to a bias against cloned sequences that are not stably maintained in the heterologous host. Consequently, the shift toward metagenomics for microbial identification was slow. However, only small portions of the 16S rRNA gene are required for microbial identification. The 16S rRNA gene contains 9 hypervariable regions (V1–V9), ranging in length from 50 to 200 bases. High-throughput sequencing of a subset of these regions provides a rapid, cost-effective and less labor-intensive approach to microbial identification [52–58]. Furthermore, NGS platforms provide a depth of coverage which surpasses that affordably obtainable with Sanger sequencing allowing for the detection of rare organisms, which may otherwise be missed.

Compositional 16S rRNA gene sequencing has allowed for comprehensive quantitative and qualitative analysis of microbial diversity in a variety of ecosystems [59, 60] including living organisms, where it has been extensively used to characterize the composition of microbial communities present in a number of niches on the human body. These niches include the gut [61–64] oral cavity [65], skin [66] and vagina [67]. 16S rRNA gene sequencing of the bacterial habitats on the human body has shown that species composition is dependent on the site sampled and varies from individual to individual. For example, the species composition of the human digestive tract contains representatives of a small proportion of the known phyla, typically dominated by the *Firmicutes* and *Bacteroidetes*. However, there is much greater interindividual variation at lower taxonomic levels [68–70]. Additionally, 16S rRNA gene sequencing of the human microbiota has furthered our understanding of the impact stable microbial communities have on an individual’s health and how changes in this composition can result in a number of diseases and metabolic conditions [61, 63, 71, 72]. For example, in a recent study which profiled the composition of the intestinal microbiota of 174 elderly individuals, Claesson *et al.* [61] identified a clear correlation between intestinal microbiota, diet and health. They showed that the intestinal microbiota of elderly in the community...
was dominated by *Firmicutes* and unclassified bacteria with the genera *Coprococcus* and *Roshuria* being most proportionally abundant. However, for individuals in long stay residential care, intestinal microbiota was dominated by the *Bacteroidetes*. Additionally, for individuals in long stay residential care the genera *Parabacteroides*, *Eubacterium*, *Anaerotruncus*, *Lactonifactor* and *Coprococcus* were also present in high numbers. The authors suggest that difference in diet between elderly individuals residing in the community and those in long-term residential care, can alter the composition of the intestinal microbiota and result in an accelerated deterioration in health in these aging populations.

**Whole community shotgun metagenomics**

16S rRNA gene sequencing is effective at identifying bacterial taxa within communities but it does have limitations. Although 16S rRNA gene sequencing can provide an abundance of information on microbial diversity in a particular niche and the impact community composition has on health and disease, it can only provide minimal information regarding the contribution each species makes to the ecosystem. To fully discover the genetic potential of a particular microbiome, whole community shotgun (WS) metagenomics is required. In addition to characterizing the microbes in a community, WS metagenomics has allowed for the annotation of a diverse range of microbial genes and due to the massive volumes of data generated, numerous novel genes encoding new functions have also been identified [73]. Large-scale metagenomes projects, such as MetaHIT (http://www.metahit.eu), the HMP (http://www.hmpdacc.org) and the Global Ocean Survey (http://www.jcvi.org/cms/research/projects/gos/) have allowed for the analysis of microbial communities at a scale that was technically and financially unachievable using traditional sequencing technologies and have dramatically increased our knowledge of microbial gene diversity. For example, the MetaHIT project, which aims to establish an association between human health states and the genes of the intestinal microbiome, identified approximately 3.3 million different microbial genes present in over 1000 species; as expected, microbial species were dominated by members of the *Firmicutes* and *Bacteroidetes* [74]. On average, each individual was estimated to harbor 540,000 genes from 160 microbial species [74] and 40%–50% of the microbial genes in each individual were shared with at least half the other individuals in the study. However, only 10% of the 3.3 million genes were common to all individuals, suggesting large interindividual gene, and thus species, diversity. As part of the MetaHIT project the 3.3 million catalog genes were classified into 19,000 functional clusters. Although a large number (14,000) of these clusters has previously been defined, 5000 were novel and contained at least 20 genes. Furthermore, approximately 6000 clusters were common to all individuals and thus represent the core or minimal metagenome. Many of the genes which comprise the core metagenome are likely to be general housekeeping genes present in all bacteria. However, some of the genes in these clusters may be essential for a healthy and functioning intestinal ecosystem.

Included in this core metagenome are a number of functional clusters involved in amino acid and vitamin biosynthesis and the production of short chain fatty acids. Finally, 1200 clusters were present with sufficient frequency to be considered to represent the ‘minimal genome’; the ‘minimal genome’ is expected to contain genes required by all bacteria to survive and thrive in the intestinal environment. However, the ‘minimal genome’ contains a large proportion of genes whose functions have not yet been or are poorly characterized. Of those genes in the ‘minimal genome’, which have been characterized, 5% were homologous to genes from phages which may indicate an important role for bacteriophages in the maintenance of gut homeostasis [74]. The MetaHIT gene catalog provides both a population-scale view of the composition of the human gut microbiome and knowledge on the contribution each species makes to the gut ecosystem. Additionally, the catalog provides a comprehensive reference structure, which allows for correlations between gut microbial gene composition and human phenotypes. Knowledge of these associations may allow for the development of a new range of diagnostic techniques and therapeutics to modulate, enhance and maintain intestinal homeostasis and thus promote intestinal [75–77] and general health [78–81]. Similarly, the NIH-funded Human Microbiome project consortium (HMP) (http://www.hmpdacc.org) has produced population-scale 16S rRNA gene amplicon and WS metagenome data sets, which detail the composition of microbial communities populating a number of sites on the human body—the human microbiota [82]. This catalog of taxa extensively characterizes the normal microbiota.
of a healthy western human adult which can be data
mined to identify novel taxa and organism [83].
Furthermore, the catalog provides a reference struc-
ture to which the microbiota of a diseased individual
can be compared, allowing for correlations between
microbial composition and health and disease to be
identified [70, 84].

UTILITY APPLICATIONS OF NGS
TECHNOLOGIES
Although primarily developed as a low cost alterna-
tive to traditional CE sequencing platform, NGS in-
struments have been adapted to perform a number of
sequence-based assays and are rapidly replacing
microarrays as the technology of choice for a range
of genomic assays. Microarrays have long been the
standard for genome-wide transcriptome and expres-
sion analyses but they have technical limitations.
Hybridization-based techniques, such as microarrays,
are reliant on existing genome sequences and can
only provide information based on probes for
known genes in sequenced genomes. Additionally,
due to issues relating to high levels of background
noise, saturation, spot density and spot quality,
microarrays possess a limited dynamic range for the
detection of transcript levels [85]. Moreover, cross
hybridization in pan-genome arrays—arrays based
on multiple genomes for the comparison of different
strains—can add to the background noise, complicat-
ing data analysis [86]. Furthermore, comparing re-
results from different experiments is complicated,
often requiring complex normalization procedures
[87]. Finally, microarrays can only measure the rela-
tive abundance of transcripts; they cannot distin-
guish between de novo and modified miRNAs, nor can they
be used to identify the promoter used in de novo tran-
scription [88]. Through the use of NGS technolo-
gies, the limitations of microarrays can be bypassed
and a number of diverse genome-wide questions can
be answered through direct sequencing. Several
sequencing methods for functional genomics have
already been developed including those for the iden-
tification of protein binding sites, gene expression
profiling, discovery of small RNAs (siRNAs) and
methylation analysis.

Chromatin immunoprecipitation
sequencing
Chromatin immunoprecipitation (ChIP), the enrich-
ment of protein-DNA complexes using antibodies
specific for a particular protein, is a functional ge-
omics technique used to identify protein-binding
sites on DNA [89, 90]. Hybridization of ChIP-
derived DNA fragments to an array (ChIP-chip)
allowed for a genome-wide analysis of these binding
sites [91, 92]. ChIP-seq, ChIP followed by sequen-
cing, is the earliest assay-based application of NGS
technologies [93–95]. In ChIP-seq, ChIP-derived
fragments are sequenced rather than hybridized to
an array (Figure 3). The direct sequencing of these
fragments provides a number of advantages over
ChIP-chip, including higher resolution (single
base-pair), deeper coverage, and a large dynamic
range [96]. ChIP-seq was rapidly implemented for
use in the analyses of eukaryotic transcription factors
[94, 96] but has not been so readily adopted for the
analysis of prokaryotic genomes. Nevertheless,
ChIP-seq has been employed in a number of pro-
karyotic genome projects for the analysis of tran-
scription factors and their associated binding sites
[97–101]. In a recent CHiP-seq analysis of the
M. tuberculosis virulence regulator EspR, a key regulator
of the ESX-1 secretion system and required for suc-
cessful infection, Blasco et al. [101] discovered that
the EspR regulator is in fact a nucleoid-associated
protein. The authors identified that EspR has both
regulatory and architectural roles and binds to at least
165 different loci throughout the genome of M. tu-
berculosis. These loci include genes encoding cell
wall functions and virulence. Despite the limited ap-
lication to date of this method in the analysis of
prokaryotic genomes, ChIP-seq could become rou-
tinely used in microbial genomics; perhaps, in the
genome-wide characterization of changes in tran-
scription factor binding in response to environmental
stimuli or during pathogenesis.

Transcriptome sequencing (RNA-seq)
Even prior to the development of NGS technologies,
sequence-based methods had been developed to
analyze transcriptomes. Initially, Sanger sequencing
was used to directly sequence cDNA or expressed
sequence tag (EST) libraries [102, 103]. However,
these sequence-based methods were subject to the
same previously described limitations imposed on
all Sanger sequencing-based experiments (slow, low
throughput and expensive). To overcome some of
these limitations, a number of tag-based sequencing
methods were developed [104, 105]. In contrast
to sequencing-methods, these tag-based methods
were high-throughput and could provide digital
quantification of transcript levels. However, tag-based methods are expensive and have been found to be of little use for transcriptome annotation [87].

RNA-seq is a NGS assay which, through the direct sequencing of cDNA, provides a rapid, potentially lower cost alternative to microarrays for the genome-wide analysis of the complete transcriptome of a living organism (Figure 3). Unlike microarrays which quantify transcript levels based on an abundance spectrum, RNA-seq measures expression simply by counting the number of reads for each transcript; thereby more accurately quantifying transcript levels over a larger dynamic range [106–108]. In a recent study on gene expression during infection, Mandlik et al. [109] used RNA-seq to quantify the expression of Vibrio cholerae genes during infection in animal models. V. cholerae is a Gram-negative pathogenic bacteria and the causative agent of cholera; of the 3–5 million cholera cases each year 100,000 to 120,000 results in death (http://www.who.int). In addition to identifying significant up-regulation of all known V. cholerae virulence-associated genes, the study identified several up-regulated sRNAs and noncoding RNAs, which were not previously linked to infection. These include several sRNAs that regulate quorum sensing and intestinal colonization. Furthermore, virulence gene induction was detectable even in samples where V. cholerae cells accounted for only a small proportion of the infected tissue. This approach allows for the transcriptome profiling of bacteria within infected tissue rather than isolating bacteria uncontaminated by host cells. Additionally, the transcriptome profiles of commensal microbiota can also be monitored in response to infection, as can changes in the physiology of the hosts infected tissue. In addition to quantitatively profiling gene expression in bacteria, RNA-seq can contribute considerably to the annotation process through the high resolution (single
base pair) mapping of transcriptional start sites. This also facilitates revision of gene boundaries of existing gene annotations and identifying previously unrecognized transcribed regions [110–114]. For example, whole transcriptome profiling of *Histophilus somni*, a causative agent of Bovine Respiratory Disease, which costs the cattle industry in the United States $3 billion annually, identified 38 novel protein coding regions with an average length of 60 amino acids. Although the majority of these proteins were homologous to conserved hypothetical proteins, several were homologs of previously characterized proteins including toxic membrane protein TnAC, DnaK, the putative *E. coli* toxic peptide IbsB3. Additionally, incorrect annotations of the start sites of five genes were identified and corrected [113]. Through the whole genome transcriptome profiling of *H. somni* 83 novel sRNAs were identified. These novel sRNAs were predicted to be involved in a range of functions, including housekeeping and virulence, and tended to form clusters suggesting functional relatedness [113].

The genome wide mapping of untranslated regulatory regions (UTRs) has identified a number of regulatory elements, including binding sites for sRNAs and riboswitches; more in depth research suggests that UTRs may have role in regulating virulence [115, 116]. Whole genome transcriptome profiling of *S. enterica enterica*, serovar Typhi identified a number of riboswitches and sRNAs in the 5′-UTRs of 127 genes. *S. enterica enterica*, serovar Typhi is a Gram-negative pathogenic bacterium transmitted through the ingestion of contaminated food and drink, and the causative agent of typhoid fever (typically only in developing countries) and gastroenteritis. *Salmonella* pathogenicity Islands (SPI) that encode type III secretion systems responsible for the injection of effector proteins into eukaryotic cells are major virulence determinants of this species. The localization of a number of these riboswitches and sRNAs to SPI-1 may indicate their role in the expression of virulence genes [115]. Additionally, through RNA-seq transcriptome profiling of *H. pylori*, a Gram negative human pathogen linked to peptic ulcers and gastric cancers [117], Sharma *et al.* [116] observed that the length of the 5′-UTR correlated to cellular function, with large 5′-UTR typically related to pathogenicity.

RNA-seq has greatly contributed to the discovery of small, antisense and noncoding RNAs [118, 119]. These sRNAs are very difficult to detect bioinformatically and often overlooked using normal annotation protocols. However, it is now known that sRNAs play an important regulatory role in bacterial genomes, particularly in bacterial physiology, where they regulate key process such as quorum sensing, virulence, niche switching, and the stress response [120–123]. Similarly, anti-sense RNA has been shown to perform a number of key regulatory functions [124], including repression of transposons [125] and toxic proteins [126], regulation of transcriptional regulator levels [127] and regulating the levels of virulence proteins [128]. sRNA or micro-RNAs (miRNA) were once believed to only play an important regulatory role in the genomes of complex multicellular organisms. However, with the discovery of large numbers of sRNAs, antisense RNA and miRNAs in microbial genomes, it is now believed that these elements provided a common form of regulation in prokaryotes [125, 129, 130] that may have originated in ancient unicellular organisms [131]. RNA-seq has also increased our understanding of the nature and structure of operons in bacterial genomes [121, 130]. Operon maps, based on polycistronic RNA are now available for a number of prokaryotes and suggest that up to 70% of bacterial mRNAs are polycistronic [110, 111, 125, 132]. Further to this, in a landmark study which analysed the transcriptome of *Mycoplasma pneumoniae* under 137 different growth conditions, operon structure was found to be context-dependent; the structure of the polycistron varied under different growth conditions [110]. The application of NGS technologies for transcriptome profiling has highlighted the dynamic nature of operons and further elucidates the complexity of prokaryotic transcriptomes through the provision of a regulatory function analogous to alternative splicing in eukaryotes [121].

In addition to circumventing the limitations of microarrays (discussed earlier), data produced in NGS assays are highly reproducible with little difference observed between replicates, provided the data are obtained from the same sequencing library [133]. However, NGS assays are not error free and associated biases can produce unwanted artifacts, which could affect downstream data analysis [134]. In general, sequencing errors still exist in NGS-derived data, particularly toward the ends of reads, although improvements in alignment algorithms have helped to mitigate this problem. Additionally, there is a biased selection of GC-rich fragments in library preparation and amplification, leading to false-positive
results during downstream analysis [119, 130]. Furthermore, the sample preparation steps in RNA-Seq experiments (mRNA fragmentation, enrichment, cDNA synthesis and size selection of fragments) have been shown to introduce a number of biases, particularly in read distribution, which can ultimately impact gene annotation and quantification of transcripts [87, 119, 130].

The diagnostic and clinical applications of bacterial WGS

Recently, whole genome sequencing (WGS) of bacterial genomes has been investigated as a diagnostic tool to assist in the management and control of infectious outbreaks. Outbreaks of infectious organisms, such as methicillin-resistant *Staphylococcus aureus* (MRSA) in hospitals can significantly increase recovery time with a corresponding increase in healthcare costs. Furthermore, outbreaks affecting critically ill patients or the vulnerable, such as the elderly or infants in neonatal care wards, can result in death; a recent *Pseudomonas* outbreak in neonatal wards in Northern Ireland resulted in the deaths of several infants (http://www.rqia.org.uk/cms_resources/RQIA%20Independent%20Review%20of%20Pseudomonas%20Interim%20Report.pdf) and in 2010, MRSA infections in US hospitals were associated with over 11,000 deaths (http://www.cdc.gov/abcs/reports-findings/survreports/mrsa10.html).

Traditional approaches to manage infectious outbreaks are slow, inefficient and costly. Accurate diagnosis can be difficult, particularly in neonatal cases, and outbreaks often result in unnecessary ward closure. In 2011, bacterial WGS using NGS technologies allowed for the rapid sequencing of four *E. coli* 0104:H4 strains from a deadly outbreak in Germany and France [135]. Genome sequences and optical maps of each of the strains were available within 62 h, demonstrating the power of WGS for the investigation of infectious outbreaks in real time [136]. Furthermore, NGS technologies facilitated an epidemiological analysis of the *E. coli* strains and identified difference which would be indistinguishable using standard molecular tools [137]. More recently, collaboration between the Wellcome Trust Sanger Institute and Illumina demonstrated the true diagnostic potential of whole genome sequencing. In this study, Köser *et al.* [138] used bacterial WGS for the rapid diagnosis of a MRSA outbreak in a neonatal ward. The study showed that WGS provides a number of benefits over traditional infection control methods. First, the genome scale data generated allowed for easy differentiation between different MRSA strains, currently unachievable with normal typing methods. Köser *et al.* [138] could thus distinguish between MRSA strains that were part of the outbreak and those that were not, preventing unnecessary treatment and ward closure. Additionally, it was demonstrated that the outbreak could have been identified earlier using WGS rather than clinical/microbiological testing. Furthermore, catalogs of antibiotic resistance genes (the resistome) and toxin genes (the toxome) were quickly established, which could allow for the tailored treatment of infected individuals. Before WGS can be implemented as a routine diagnostic tool, a number of issues would have to be resolved [139]. First, software must be developed which can convert sequencing data into clinically relevant information that is easily interpreted by healthcare professionals and the appropriate IT infrastructure needs to be in place. Additionally, a cost-benefit analysis would be required so support the use of more costly WGS over traditional clinical diagnostic techniques. However, despite these caveats, it is likely that WGS (and other NGS applications) will soon be routinely used as diagnostic tools in clinical laboratories, supporting or replacing traditional diagnostic techniques.

**FUTURE PERSPECTIVES**

Although they have been commercially available for less than 10 years, NGS technologies have already made a dramatic impact on the field of Microbiology. In addition to providing more cost-effective sequencing methods, the range of utility-based applications, which extended beyond the original scope of NGS technologies, will allow for a more accurate functional annotation of microbial genomes. The development of TGS technologies promises to further improve genome and utility-based sequencing applications. Single molecule sequencing will hopefully eliminate amplification biases, and longer read length will provide greater coverage and depth, enabling increased accuracy and profiling of more complex transcriptomes. Additionally, direct sequencing of RNA will remove/reduce many of the biases in RNA-seq data, produced during sample preparation [15, 87]. Current and future NGS technologies promise to provide new insights into individual microbial genomes, the structure of the communities they inhabit, and their impact on human health and
disease. This in turn will allow for the development of more accurate models of disease and infection and result in the development of a new range of diagnostic tools and therapeutics to combat infectious disease.

**Key points**
- Genome resequencing using NGS technologies has proven to be highly effective for characterizing genome variation and diversity and is potentially a powerful tool for informing clinical practice.
- Data generated by NGS metagenome analysis can correlate microbial composition with human phenotypes. Knowledge of the association of microbial community structure and human health and disease can be used to develop new therapeutics to enhance and maintain human health.
- NGS utility-based sequencing applications are rapidly replacing microarrays as the method of choice for genomic assays, such as transcriptome profiling. Techniques such as RNA-seq and CHIP-seq are providing novel insights into the structure and dynamics of bacterial genomes.

**FUNDING**

This work was supported by Science Foundation Ireland through a Principal Investigator award to Paul W. O’Toole.

**References**


