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

Corresponding author. Paul W. O'Toole, Department of Microbiology, University College Cork, Cork, Ireland. Tel: +353 21 490 3000; Fax: +353 21 490 3997; E-mail: pwotoole@ucc.ie

Brian M. Forde is completing a PhD in Bioinformatics and Microbiology in the Laboratory of Paul W. O'Toole at University College Cork, Ireland. His primary interests are in genome assembly, and the comparative and evolutionary analysis of microbial genomes.

Paul W. O'Toole completed a PhD in microbial genetics in Trinity College, Dublin, and is currently a senior lecturer in Genetics at University College Cork, Ireland. He is also a Principal Investigator in the Genomics and Metagenomics core of the Alimentary Pharmobiotic Centre, with an interest in the functional genomics of commensal lactobacilli, and their interaction with the microbiota. He is an SFI Principle Investigator and leads the DAFF/HRB FHRI project ELDERMET, a metagenomic study of intestinal microbiota and health in the elderly.

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

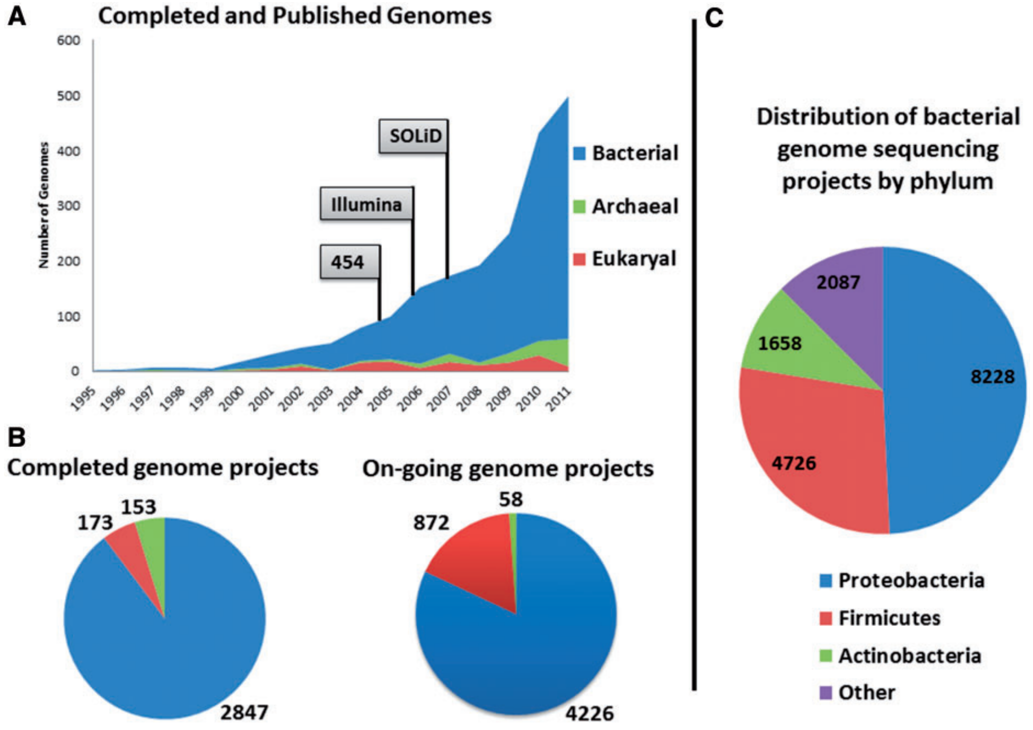

Figure 1: Published genomes. (**A**) Published genome sequences for the three domains of life as of April 2012. (**B**) Distribution of completed and on-going genome projects amongst the three domains. (**C**) Phylogenetic distribution of bacterial genome sequencing projects. *Source*: http://www.genomesonline.org.

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

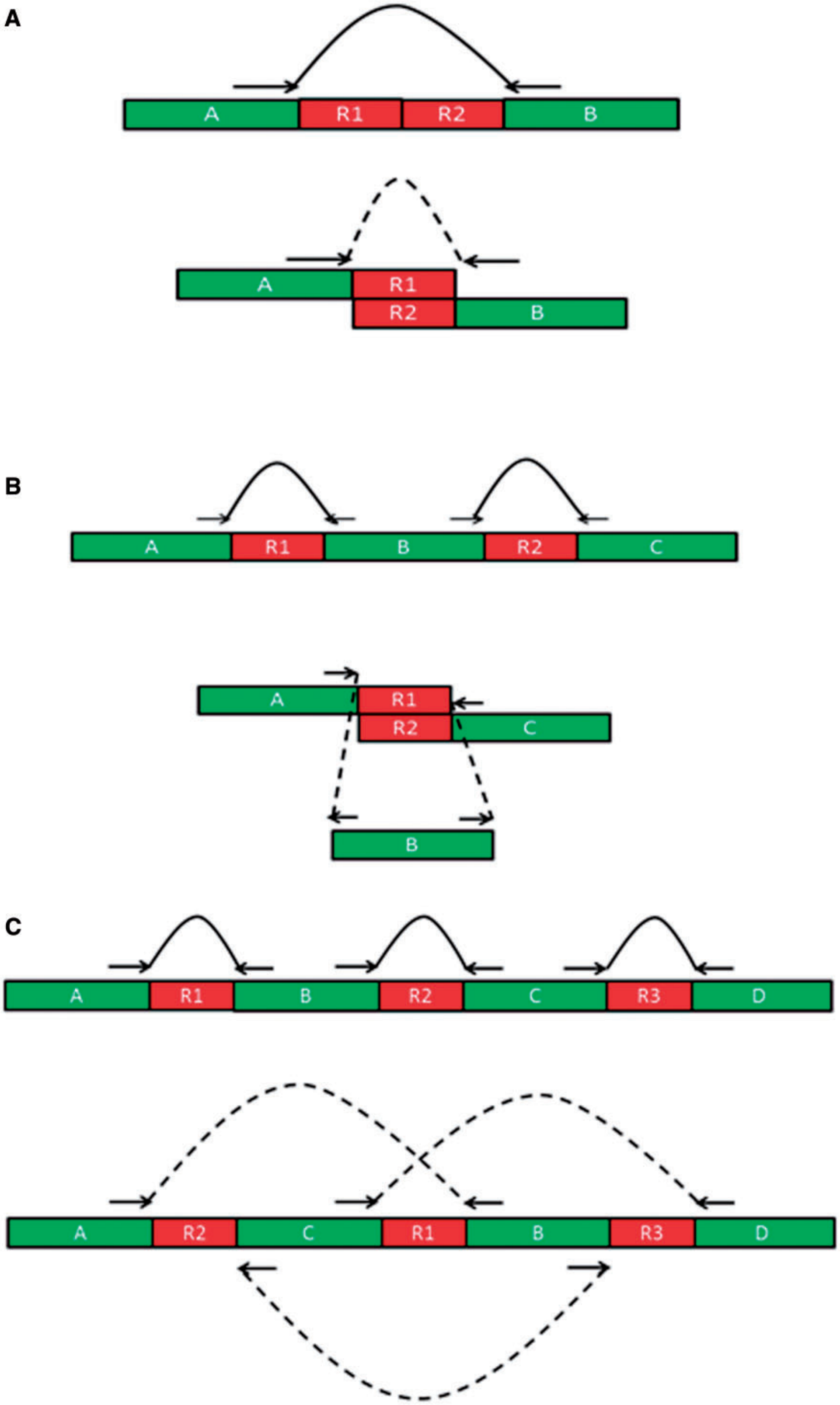

Figure 2: The three main types of repeat-induced errors encountered in genome assembly projects. The errors can all be identified by abnormalities in the mate-pair information, e.g. elongated or truncated distance between mates or incorrect orientation. (**A**) Collapsed repeat; (**B**) excision; (**C**) 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 *Rosburia* 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 *Coprobacillus* 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 prophages 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 structure 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 alternative to traditional CE sequencing platform, NGS instruments 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 expression 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 pangenome arrays—arrays based on multiple genomes for the comparison of different strains—can add to the background noise, complicating data analysis [86]. Furthermore, comparing results from different experiments is complicated, often requiring complex normalization procedures [87]. Finally, microarrays can only measure the relative abundance of transcripts; they cannot distinguish between *de novo* and modified mRNAs, nor can they be used to identify the promoter used in *de novo* transcription [88]. Through the use of NGS technologies, 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 identification of protein binding sites, gene expression profiling, discovery of small RNAs (sRNAs) and methylation analysis.

Chromatin immunoprecipitation sequencing

Chromatin immunoprecipitation (ChIP), the enrichment of protein-DNA complexes using antibodies specific for a particular protein, is a functional genomics 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 sequencing, 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 prokaryotic genome projects for the analysis of transcription 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 successful 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. tuberculosis.* These loci include genes encoding cell wall functions and virulence. Despite the limited application to date of this method in the analysis of prokaryotic genomes, ChIP-seq could become routinely used in microbial genomics; perhaps, in the genome-wide characterization of changes in transcription 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 espressed 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

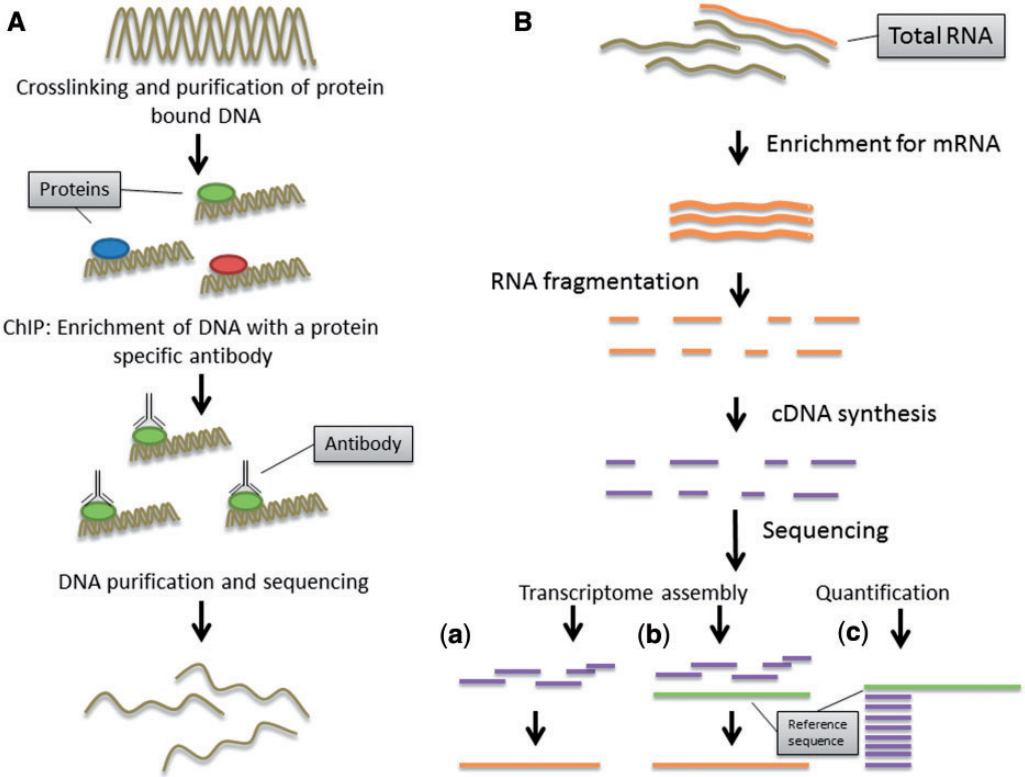

Figure 3: The two main assay-based applications of NGS technologies. (**A**) ChIP-seq. ChIP is combined with NGS to identify protein binding sites on DNA. First, crosslinks between the DNA and proteins are formed. Next, antibodies specific to a protein are used to selectively co-immunoprecipitate the protein and bound DNA. Finally, the DNA is purified and sequenced. (**B**) RNA-seq. Total RNA is extracted from the cell. In prokaryotes, mRNA constitutes as little as 1%–5% of total RNA. Consequently, mRNA requires enrichment prior to sequencing. Enrichment of mRNAs may include rRNA capture, processed RNA degradation and selective polyadenylation of mRNAs following enrichment, mRNAs are fragmented, converted to cDNA and sequenced. Sequenced cDNA is then used for (a) de novo transcriptome assembly, (b) transcriptome re-sequencing or (c) transcriptome quantification.

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 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

1. Fleischmann RD, Adams MD, White O, *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 1995;**269**:496–512.
2. Fraser CM, Gocayne JD, White O, *et al.* The minimal gene complement of *Mycoplasma genitalium*. *Science* 1995;**270**: 397–403.
3. Edwards A, Voss H, Rice P, *et al.* Automated DNA sequencing of the human HPRT locus. *Genomics* 1990;**6**:593–608.
4. Roach JC, Boysen C, Wang K, *et al.* Pairwise end sequencing: a unified approach to genomic mapping and sequencing. *Genomics* 1995;**26**:345–53.
5. Weber JL, Myers EW. Human whole-genome shotgun sequencing. *Genome Res* 1997;**7**:401–9.
6. Kunst F, Ogasawara N, Moszer I, *et al.* The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature* 1997;**390**:249–56.
7. Blattner FR, Plunkett G, Bloch CA, *et al.* The complete genome sequence of *Escherichia coli* K-12. *Science* 1997;**277**: 1453–74.
8. Goffeau A. [The yeast genome]. *Pathol Biol (Paris)* 1998;**46**: 96–7.
9. *C. elegans* sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* 1998;**282**:2012–8.
10. *Arabidopsis* Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 2000;**408**:796–815.
11. Myers EW, Sutton GG, Delcher AL, *et al.* A whole-genome assembly of *Drosophila*. *Science* 2000;**287**:2196–204.
12. Lander ES, Linton LM, Birren B, *et al.* Initial sequencing and analysis of the human genome. *Nature* 2001;**409**: 860–921.
13. Venter JC, Adams MD, Myers EW, *et al.* The sequence of the human genome. *Science* 2001;**291**:1304–51.
14. Metzker ML. Sequencing technologies—the next generation. *Nat Rev Genet* 2010;**11**:31–46.
15. Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. *Hum Mol Genet* 2010;**19**: R227–R240.
16. Mardis ER. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet* 2008;**9**:387–402.
17. Metzker ML. Emerging technologies in DNA sequencing. *Genome Res* 2005;**15**:1767–76.
18. Kyrpides NC. Fifteen years of microbial genomics: meeting the challenges and fulfilling the dream. *Nat Biotechnol* 2009; **27**:627–32.
19. Parkhill J. Time to remove the model organism blinkers. *Trends Microbiol* 2008;**16**:510–1.
20. Medini D, Donati C, Tettelin H, *et al.* The microbial pan-genome. *Curr Opin Genet Dev* 2005;**15**:589–94.
21. Whiteford N, Haslam N, Weber G, *et al.* An analysis of the feasibility of short read sequencing. *Nucleic Acids Res* 2005; **33**:e171.
22. Cahill MJ, Köser CU, Ross NE, *et al.* Read length and repeat resolution: exploring prokaryote genomes using next-generation sequencing technologies. *PLoS One* 2010; **5**:e11518.
23. Flicek P, Birney E. Sense from sequence reads: methods for alignment and assembly. *Nat Methods* 2009;**6**: S6–S12.
24. Pop M. Genome assembly reborn: recent computational challenges. *Brief Bioinformatics* 2009;**10**:354–66.
25. Chaisson MJ, Brinza D, Pevzner PA. De novo fragment assembly with short mate-paired reads: does the read length matter? *Genome Res* 2009;**19**:336–46.
26. Kennemann L, Didelot X, Aebischer T, *et al.* *Helicobacter pylori* genome evolution during human infection. *Proc Natl Acad Sci U S A* 2011;**108**:5033–8.
27. Chen PE, Willner KM, Butani A, *et al.* Rapid identification of genetic modifications in *Bacillus anthracis* using whole genome draft sequences generated by 454 pyrosequencing. *PLoS One* 2010;**5**:e12397.
28. Koenig R. Tuberculosis. Few mutations divide some drug-resistant TB strains. *Science* 2007;**318**:901–2.
29. Holt KE, Parkhill J, Mazzoni CJ, *et al.* High-throughput sequencing provides insights into genome variation and evolution in *Salmonella typhi*. *Nat Genet* 2008;**40**:987–93.
30. Ford CB, Lin PL, Chase MR, *et al.* Use of whole genome sequencing to estimate the mutation rate of *Mycobacterium tuberculosis* during latent infection. *Nat Genet* 2011;**43**:482–6.
31. Margulies M, Egholm M, Altman WE, *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 2005;**437**:376–80.
32. Chaisson M, Pevzner P, Tang H. Fragment assembly with short reads. *Bioinformatics* 2004;**20**:2067–74.
33. Pop M, Salzberg SL. Bioinformatics challenges of new sequencing technology. *Trends Genet* 2008;**24**:142–9.
34. Pevzner PA, Tang H. Fragment assembly with double-barreled data. *Bioinformatics* 2001;**17**(Suppl 1):S225–S233.

35. Pop M. Shotgun sequence assembly. *Adv Comput* 2004;**60**: 193–248.

36. Goldberg SMD, Johnson J, Busam D, *et al*. A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. *Proc Natl Acad Sci U S A* 2006;**103**:11240–5.

37. Schatz MC, Delcher AL, Salzberg SL. Assembly of large genomes using second-generation sequencing. *Genome Res* 2010;**20**:1165–73.

38. Chaisson MJ, Pevzner PA. Short read fragment assembly of bacterial genomes. *Genome Res* 2008;**18**:324–30.

39. Aury J-M, Cruaud C, Barbe V, *et al*. High quality draft sequences for prokaryotic genomes using a mix of new sequencing technologies. *BMC Genomics* 2008;**9**:603.

40. Bashir A, Klammer AA, Robins WP, *et al*. A hybrid approach for the automated finishing of bacterial genomes. *Nat Biotechnol* 2012;**30**:701–7.

41. Cerdeira LT, Carneiro AR, Ramos RTJ, *et al*. Rapid hybrid de novo assembly of a microbial genome using only short reads: *Corynebacterium pseudotuberculosis* I19 as a case study. *J Microbiol Methods* 2011;**86**:218–23.

42. Forde BM, Neville B, O'Donnell MM, *et al*. Genome sequences and comparative genomics of two *Lactobacillus ruminis* strains from the bovine and human intestinal tracts. *Microbial Cell Factories* 2011;**10**:S13.

43. Reinhardt JA, Baltrus DA, Nishimura MT, *et al*. De novo assembly using low-coverage short read sequence data from the rice pathogen *Pseudomonas syringae pv. oryzae*. *Genome Res* 2009;**19**:294–305.

44. Pace NR. Mapping the tree of life: progress and prospects. *Microbiol Mol Biol Rev* 2009;**73**:565–76.

45. Grice EA, Kong HH, Renaud G, *et al*. A diversity profile of the human skin microbiota. *Genome Res* 2008;**18**: 1043–50.

46. Delgado S, Suárez A, Mayo B. Identification of dominant bacteria in feces and colonic mucosa from healthy Spanish adults by culturing and by 16S rDNA sequence analysis. *Dig Dis Sci* 2006;**51**:744–51.

47. Hold GL, Pryde SE, Russell VJ, *et al*. Assessment of microbial diversity in human colonic samples by 16S rDNA sequence analysis. *FEMS Microbiol Ecol* 2002;**39**: 33–9.

48. Venter JC, Remington K, Heidelberg JF, *et al*. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 2004;**304**:66–74.

49. Tyson GW, Chapman J, Hugenholtz P, *et al*. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 2004; **428**:37–43.

50. Kolbert CP, Persing DH. Ribosomal DNA sequencing as a tool for identification of bacterial pathogens. *Curr Opin Microbiol* 1999;**2**:299–305.

51. Janda JM, Abbott SL. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *J Clin Microbiol* 2007;**45**:2761–4.

52. Luna RA, Fasciano LR, Jones SC, *et al*. DNA pyrosequencing-based bacterial pathogen identification in a pediatric hospital setting. *J Clin Microbiol* 2007;**45**:2985–92.

53. Petrosino JF, Highlander S, Luna RA, *et al*. Metagenomic pyrosequencing and microbial identification. *Clin Chem* 2009;**55**:856–66.

54. Ventura M, O'Flaherty S, Claesson MJ, *et al*. Genome-scale analyses of health-promoting bacteria: probiogenomics. *Nat Rev Microbiol* 2009;**7**:61–71.

55. Claesson MJ, O'Sullivan O, Wang Q, *et al*. Comparative analysis of pyrosequencing and a phylogenetic microarray for exploring microbial community structures in the human distal intestine. *PLoS One* 2009;**4**:e6669.

56. Claesson MJ, Wang Q, O'Sullivan O, *et al*. Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. *Nucleic Acids Res* 2010;**38**: e200.

57. Caporaso JG, Lauber CL, Walters WA, *et al*. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc Natl Acad Sci U S A* 2011; **108**(Suppl):4516–22.

58. Degnan PH, Ochman H. Illumina-based analysis of microbial community diversity. *ISME J* 2012;**6**:183–94.

59. Galand PE, Casamayor EO, Kirchman DL, *et al*. Ecology of the rare microbial biosphere of the Arctic Ocean. *Proc Natl Acad Sci U S A* 2009;**106**:22427–32.

60. Bowen JL, Morrison HG, Hobbie JE, *et al*. Salt marsh sediment diversity: a test of the variability of the rare biosphere among environmental replicates. *ISME J* 2012;**6**:2014–23.

61. Claesson MJ, Jeffery IB, Conde S, *et al*. Gut microbiota composition correlates with diet and health in the elderly. *Nature* 2012;**488**:178–84.

62. Alcaraz LD, Belda-Ferre P, Cabrera-Rubio R, *et al*. Identifying a healthy oral microbiome through metagenomics. *Clin Microbiol Infect* 2012;**18**(Suppl 4):54–7.

63. O'Toole PW. Changes in the intestinal microbiota from adulthood through to old age. *Clin Microbiol Infect* 2012; **18**(Suppl 4):44–6.

64. Turnbaugh PJ, Hamady M, Yatsunenko T, *et al*. A core gut microbiome in obese and lean twins. *Nature* 2009;**457**: 480–4.

65. Crielaard W, Zaura E, Schuller AA, *et al*. Exploring the oral microbiota of children at various developmental stages of their dentition in the relation to their oral health. *BMC Med Genomics* 2011;**4**:22.

66. Kong HH. Skin microbiome: genomics-based insights into the diversity and role of skin microbes. *Trends Mol Med* 2011; **17**:320–8.

67. Aagaard K, Riehle K, Ma J, *et al*. A metagenomic approach to characterization of the vaginal microbiome signature in pregnancy. *PLoS One* 2012;**7**:e36466.

68. Eckburg PB, Bik EM, Bernstein CN, *et al*. Diversity of the human intestinal microbial flora. *Science* 2005;**308**:1635–8.

69. Segata N, Haake SK, Mannon P, *et al*. Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. *Genome Biol* 2012;**13**:R42.

70. The Human Microbiome Project consortium. Structure, function and diversity of the healthy human microbiome. *Nature* 2012;**486**:207–14.

71. Ley RE, Turnbaugh PJ, Klein S, *et al*. Microbial ecology: human gut microbes associated with obesity. *Nature* 2006; **444**:1022–3.

72. Jeffery IB, O'Toole PW, Öhman L, *et al*. An irritable bowel syndrome subtype defined by species-specific alterations in faecal microbiota. *Gut* 2012;**61**:997–1006.

73. Yooseph S, Sutton G, Rusch DB, *et al*. The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol* 2007;**5**:e16.

74. Qin J, Li R, Raes J, *et al*. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 2010;**464**:59–65.

75. Kostic AD, Gevers D, Pedamallu CS, *et al*. Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. *Genome Res* 2012;**22**:292–8.

76. Tana C, Umesaki Y, Imaoka A, *et al*. Altered profiles of intestinal microbiota and organic acids may be the origin of symptoms in irritable bowel syndrome. *Neurogastroenterol Motil* 2010;**22**:512–9, e114–5.

77. Turnbaugh PJ, Ley RE, Mahowald MA, *et al*. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* 2006;**444**:1027–31.

78. Islami F, Kamangar F. *Helicobacter pylori* and esophageal cancer risk: a meta-analysis. *Cancer Prevent Res* 2008;**1**: 329–38.

79. Chen Y, Blaser MJ. Inverse associations of *Helicobacter pylori* with asthma and allergy. *Arch Int Med* 2007;**167**:821–7.

80. Gao Z, Tseng C, Strober BE, *et al*. Substantial alterations of the cutaneous bacterial biota in psoriatic lesions. *PLoS One* 2008;**3**:e2719.

81. Wang Z, Klipfell E, Bennett BJ, *et al*. Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. *Nature* 2011;**472**:57–63.

82. The Human Microbiome Project consortium. A framework for human microbiome research. *Nature* 2012;**486**:215–21.

83. Wylie KM, Truty RM, Sharpton TJ, *et al*. Novel bacterial taxa in the human microbiome. *PLoS One* 2012;**7**:e35294.

84. Li E, Hamm CM, Gulati AS, *et al*. Inflammatory bowel diseases phenotype, *C. difficile* and NOD2 genotype are associated with shifts in human ileum associated microbial composition. *PLoS One* 2012;**7**:e26284.

85. Hinton JCD, Hautefort I, Eriksson S, *et al*. Benefits and pitfalls of using microarrays to monitor bacterial gene expression during infection. *Curr Opin Microbiol* 2004;**7**: 277–82.

86. Okoniewski MJ, Miller CJ. Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. *BMC Bioinformatics* 2006;**7**:276.

87. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009;**10**: 57–63.

88. van Vliet AHM. Next generation sequencing of microbial transcriptomes: challenges and opportunities. *FEMS Microbiol Lett* 2010;**302**:1–7.

89. Solomon MJ, Larsen PL, Varshavsky A. Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. *Cell* 1988;**53**:937–47.

90. Collas P, Dahl JA. Chop it, ChIP it, check it: the current status of chromatin immunoprecipitation. *Frontiers Biosci* 2008;**13**:929–43.

91. Pillai S, Chellappan SP. ChIP on chip assays: genome-wide analysis of transcription factor binding and histone modifications. *Methods Mol Biol* 2009;**523**:341–66.

92. Ren B, Robert F, Wyrick JJ, *et al*. Genome-wide location and function of DNA binding proteins. *Science* 2000;**290**: 2306–9.

93. Mikkelsen TS, Ku M, Jaffe DB, *et al*. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 2007;**448**:553–60.

94. Johnson DS, Mortazavi A, Myers RM, *et al*. Genome-wide mapping of in vivo protein-DNA interactions. *Science* 2007;**316**:1497–502.

95. Robertson G, Hirst M, Bainbridge M, *et al*. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* 2007;**4**:651–7.

96. Park PJ. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet* 2009;**10**:669–80.

97. Markel E, Maciak C, Butcher BG, *et al*. An extracytoplasmic function sigma factor-mediated cell surface signaling system in *Pseudomonas syringae* pv. tomato DC3000 regulates gene expression in response to heterologous siderophores. *J Bacteriol* 2011;**193**:5775–83.

98. Butcher BG, Bronstein PA, Myers CR, *et al*. Characterization of the Fur regulon in *Pseudomonas syringae* pv. tomato DC3000. *J Bacteriol* 2011;**193**:4598–611.

99. Davies BW, Bogard RW, Mekalanos JJ. Mapping the regulon of *Vibrio cholerae* ferric uptake regulator expands its known network of gene regulation. *Proc Natl Acad Sci U S A* 2011;**108**:12467–72.

100. Wilbanks EG, Larsen DJ, Neches RY, *et al*. A workflow for genome-wide mapping of archaeal transcription factors with ChIP-seq. *Nucleic Acids Res* 2012;**40**:e74.

101. Blasco B, Chen JM, Hartkoorn R, *et al*. Virulence regulator EspR of *Mycobacterium tuberculosis* is a nucleoid-associated protein. *PLoS Pathogens* 2012;**8**:e1002621.

102. Boguski MS, Tolstoshev CM, Bassett DE. Gene discovery in dbEST. *Science* 1994;**265**:1993–4.

103. Gerhard DS, Wagner L, Feingold EA, *et al*. The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). *Genome Res* 2004;**14**:2121–7.

104. Velculescu VE, Zhang L, Vogelstein B, *et al*. Serial analysis of gene expression. *Science* 1995;**270**:484–7.

105. Kodzius R, Kojima M, Nishiyori H, *et al*. CAGE: cap analysis of gene expression. *Nat Methods* 2006;**3**:211–22.

106. Yuan M, Chen M, Zhang W, *et al*. Genome sequence and transcriptome analysis of the radioresistant bacterium *Deinococcus gobiensis*: insights into the extreme environmental adaptations. *PLoS One* 2012;**7**:e34458.

107. Deng X, Li Z, Zhang W. Transcriptome sequencing of *Salmonella enterica* serovar Enteritidis under desiccation and starvation stress in peanut oil. *Food Microbiol* 2012;**30**: 311–5.

108. Westermann AJ, Gorski SA, Vogel J. Dual RNA-seq of pathogen and host. *Nat Rev Microbiol* 2012;**10**:618–30.

109. Mandlik A, Livny J, Robins WP, *et al*. RNA-Seq-based monitoring of infection-linked changes in *Vibrio cholerae* gene expression. *Cell Host Microbe* 2011;**10**:165–74.

110. Güell M, van Noort V, Yus E, *et al*. Transcriptome complexity in a genome-reduced bacterium. *Science* 2009;**326**: 1268–71.

111. Wurtzel O, Sapra R, Chen F, *et al*. A single-base resolution map of an archaeal transcriptome. *Genome Res* 2010;**20**: 133–41.

112. Wang Y, Li X, Mao Y, *et al*. Single-nucleotide resolution analysis of the transcriptome structure of *Clostridium*

beijerinckii NCIMB 8052 using RNA-Seq. *BMC Genomics* 2011;**12**:479.

113. Kumar R, Lawrence ML, Watt J, *et al.* RNA-seq based transcriptional map of bovine respiratory disease pathogen "*Histophilus somni* 2336". *PLoS One* 2012;**7**:e29435.

114. Passalacqua KD, Varadarajan A, Ondov BD, *et al.* Structure and complexity of a bacterial transcriptome. *J Bacteriol* 2009;**191**:3203–11.

115. Perkins TT, Kingsley RA, Fookes MC, *et al.* A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus *Salmonella typhi*. *PLoS Genet* 2009;**5**: e1000569.

116. Sharma CM, Hoffmann S, Darfeuille F, *et al.* The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature* 2010;**464**:250–5.

117. Kusters JG, van Vliet AHM, Kuipers EJ. Pathogenesis of *Helicobacter pylori* infection. *Clin Microbiol Rev* 2006;**19**: 449–90.

118. Gómez-Lozano M, Marvig RL, Molin S, *et al.* Genome-wide identification of novel small RNAs in *Pseudomonas aeruginosa*. *Environ Microbiol* 2012;**14**:2006–16.

119. Pinto AC, Melo-Barbosa HP, Miyoshi A, *et al.* Application of RNA-seq to reveal the transcript profile in bacteria. *Genet Mol Res* 2011;**10**:1707–18.

120. Yoder-Himes DR, Chain PSG, Zhu Y, *et al.* Mapping the *Burkholderia cenocepacia* niche response via high-throughput sequencing. *Proc Natl Acad Sci U S A* 2009;**106**:3976–81.

121. Sorek R, Cossart P. Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. *Nat Rev Genet* 2010;**11**:9–16.

122. Toledo-Arana A, Repoila F, Cossart P. Small noncoding RNAs controlling pathogenesis. *Curr Opin Microbiol* 2007; **10**:182–8.

123. Bejerano-Sagie M, Xavier KB. The role of small RNAs in quorum sensing. *Curr Opin Microbiol* 2007;**10**:189–98.

124. Thomason MK, Storz G. Bacterial antisense RNAs: how many are there, and what are they doing? *Annu Rev Genet* 2010;**44**:167–88.

125. Toledo-Arana A, Dussurget O, Nikitas G, *et al.* The Listeria transcriptional landscape from saprophytism to virulence. *Nature* 2009;**459**:950–6.

126. Fozo EM, Hemm MR, Storz G. Small toxic proteins and the antisense RNAs that repress them. *Microbiol Mol Biol Rev* 2008;**72**:579–89, Table of Contents.

127. Tramonti A, De Canio M, De Biase D. GadX/GadW-dependent regulation of the *Escherichia coli* acid fitness island: transcriptional control at the gadY-gadW divergent promoters and identification of four novel 42 bp GadX/GadW-specific binding sites. *Mol Microbiol* 2008;**70**: 965–82.

128. Lee E-J, Groisman EA. An antisense RNA that governs the expression kinetics of a multifunctional virulence gene. *Mol Microbiol* 2010;**76**:1020–33.

129. Bernick DL, Dennis PP, Lui LM, *et al.* Diversity of antisense and other non-coding RNAs in Archaea revealed by comparative small RNA sequencing in four *Pyrobaculum* species. *Front Microbiol* 2012;**3**:231.

130. Siezen RJ, Wilson G, Todt T. Prokaryotic whole-transcriptome analysis: deep sequencing and tiling arrays. *Microb Biotechnol* 2010;**3**:125–30.

131. Molnár A, Schwach F, Studholme DJ, *et al.* miRNAs control gene expression in the single-cell alga *Chlamydomonas reinhardtii*. *Nature* 2007;**447**:1126–9.

132. Koide T, Reiss DJ, Bare JC, *et al.* Prevalence of transcription promoters within archaeal operons and coding sequences. *Mol Syst Biol* 2009;**5**:285.

133. Marioni JC, Mason CE, Mane SM, *et al.* RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* 2008;**18**: 1509–17.

134. Dohm JC, Lottaz C, Borodina T, *et al.* Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res* 2008;**36**:e105.

135. Gault G, Weill FX, Mariani-Kurkdjian P, *et al.* Outbreak of haemolytic uraemic syndrome and bloody diarrhoea due to *Escherichia coli* O104:H4, south-west France, June 2011. *Euro Surveill* 2011;**16**:pii:19905.

136. Mellmann A, Harmsen D, Cummings CA, *et al.* Prospective genomic characterization of the german enterohemorrhagic *Escherichia coli* O104:H4 outbreak by rapid next generation sequencing technology. *PLoS One* 2011;**6**: e22751.

137. Grad YH, Lipsitch M, Feldgarden M, *et al.* Genomic epidemiology of the *Escherichia coli* O104:H4 outbreaks in Europe, 2011. *Proc Natl Acad Sci U S A* 2012;**109**: 3065–70.

138. Köser CU, Holden MTG, Ellington MJ, *et al.* Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. *N Engl J Med* 2012;**366**:2267–75.

139. Chan JZ-M, Pallen MJ, Oppenheim B, *et al.* Genome sequencing in clinical microbiology. *Nature Biotechnol* 2012;**3**:1068–71.