

Chapter 2

The Impact of Next-Generation Sequencing Technology on Bacterial Genomics

Avantika Lal and Aswin Sai Narain Seshasayee

Abstract For many decades, genomic studies were based on Sanger sequencing or the dideoxy chain termination method of sequencing DNA, along with microarray and hybridization-based techniques to understand genome function. Sanger sequencing was used to sequence the genomes of many organisms, from bacteria to humans. However, in recent years ‘Next-generation’ sequencing technologies have been developed that are cheaper and far more rapid. They produce great sequencing depth, making them applicable to quantitative studies such as gene expression measurements as well. As a result, these technologies have been used extensively to study the sequence, structure, function and evolution of both eukaryotic and bacterial genomes. Here we discuss next-generation sequencing and how it has been used to study a variety of areas from gene expression and protein-DNA interactions to bacterial community function and evolution, at the scale of whole bacterial genomes. We expect that further advances in DNA sequencing technology and methods for managing and analyzing the large volumes of data produced by these approaches will help to answer many more questions in this field.

Keywords Next generation sequencing · DNA · Microarray · Hybridization · Chain termination · Sanger sequencing · Sequence library · Comparative genome hybridization (CGH)

A. Lal · A. S. N. Seshasayee (✉)
National Centre for Biological Sciences, Tata Institute of Fundamental Research,
Bangalore, India
e-mail: aswin@ncbs.res.in

A. Lal
e-mail: avantika@ncbs.res.in

2.1 Introduction

DNA sequencing is the determination of the order of the four nucleotide bases, adenine, guanine, cytosine, and thymine, in a molecule of DNA. Some of the earliest efforts to sequence DNA used laborious methods based on two-dimensional gel chromatography. However, in 1977, Sanger and Coulson described a much easier and more reliable method of DNA sequencing based on chain termination [90], which soon became widespread. In time, their original method was improved by automation and advances in technology and for the next 30 years, ‘Sanger sequencing’ held a monopoly over DNA sequence determination.

The ability to sequence DNA proved to be a turning point in biological sciences. It enabled scientists to understand the genetic basis of many diseases and to trace evolution at the molecular level, among other applications. The genomes, or total cellular DNA contents, of several organisms were sequenced using this method [10, 27, 32]. Over 1990–2004, the International Human Genome Project used Sanger sequencing methods, coupled with the whole genome shotgun technique, to sequence the approximately three billion nucleotides of the human genome [43]. Today DNA sequencing has become a vital tool not only in basic biological research but also in applied fields such as diagnostics and forensics.

As genome sequencing projects matured, it became apparent that further large-scale experimental tools were required to understand the meaning of genome sequences. Several tools were developed to study genomes at the functional level: from gene expression, which is the first stage in the conversion of DNA sequence to a functional readout, to protein-nucleic acid interactions, which enable gene expression. The DNA microarray, developed in the early 1990s, allowed scientists to measure the expression levels of large numbers of genes simultaneously or to genotype multiple regions of a genome. Later developments in DNA microarrays, in the form of genome-tiling arrays, permitted experimental annotation of genomes leading to various descriptions of pervasive transcription in eukaryotic genomes [7, 20], and large numbers of intra-operonic transcriptional initiation events in the simple bacterial genome of *Escherichia coli* [16]. These techniques further allowed large-scale mapping of regulatory networks by interrogating regions of the genome bound by a protein of interest, leading to the emergence of large-scale network biology (See [36, 38, 55, 100] for experiments in yeast; see [3, 4] for reviews); these findings unraveled unanticipated complexity in the binding properties of transcriptional regulators even in model bacterial genomes [34].

However, both Sanger sequencing and DNA microarrays have several drawbacks. The major drawback of Sanger sequencing is that it is too slow and expensive for many applications. The human genome project took over thirteen years and more than two billion dollars to complete. Given these limitations, the ultimate goal of genome sequencing, which for many is the sequencing of personal genomes leading to personalized medicine, is unlikely to be met using Sanger sequencing.

As a result, there has been a large effort in science and industry to bring the cost of high-quality human genome sequencing down to a level that is affordable to

individuals. In 2006, the X PRIZE Foundation announced a \$10 million incentive for the group whose technology would enable a human genome to be sequenced for \$1,000 or less. This has spurred the rise of a number of ‘next-generation’ sequencing (NGS) technologies in the last decade [91]. These produce large quantities of sequence data in a short period of time, and at a reduced cost. This has encouraged ambitious projects such as the 1,000 human genome project [84] and the 10,000 vertebrate genome project [69]. It has also become routine for large sequencing centers to publish a single piece of work describing over 100 bacterial genomes, thus enabling fine-scale genomic studies of bacterial epidemiology and evolution [19, 72].

The second cornerstone of genomic studies in the early 2000s, DNA microarray technology, has the limitation that it can interrogate the properties of only those regions of the genome whose sequence is known. For example, the first generation gene expression microarrays probed only known gene sequences, and were unable to detect transcription from intergenic regions. Further, microarrays cannot be used effectively to investigate the functional properties of non-model organisms whose genomes are not known. Though sequences of related genomes have been used to design microarray probes for the study of non-model organisms, the limitations of such an approach become apparent in the light of the fact that non-conserved portions of a genome have a large effect on an organism’s biology. Moreover, even small variations between the two genomes in the regions probed by the microarrays could lead to unreliable findings. Further, saturation in the measurement of fluorescence limits the dynamic range of microarrays.

The great depth of sequencing afforded by NGS techniques allow us to quantify nucleic acids, thus making them applicable to various applications for which DNA microarrays had been used, while circumventing many of the problems associated with the latter.

In this chapter, we discuss the features and challenges of next-generation sequencing technologies, as well as various types of experimental studies that have been enabled by them. We focus particularly on how these sequencing methods have been used to study bacterial genomics.

2.2 Next-Generation Sequencing Technologies

The term ‘Next-generation sequencing’ applies to several commercially available platforms. The most commonly used NGS platforms are 454 pyrosequencing (Roche/454 Life Sciences), Illumina (erstwhile Solexa sequencing) and SOLiD (Applied Biosystems).

Although these platforms differ from each other in the procedure employed as well as the chemistry of sequencing, their basic strategy is similar. The DNA molecule to be sequenced is fragmented at random positions, and short adaptor sequences are ligated to the ends of each DNA fragment. The resultant set of molecules is called the sequencing library. Each molecule in this library is amplified to generate a cluster of amplicons. Each cluster of identical DNA molecules is spatially separated from the

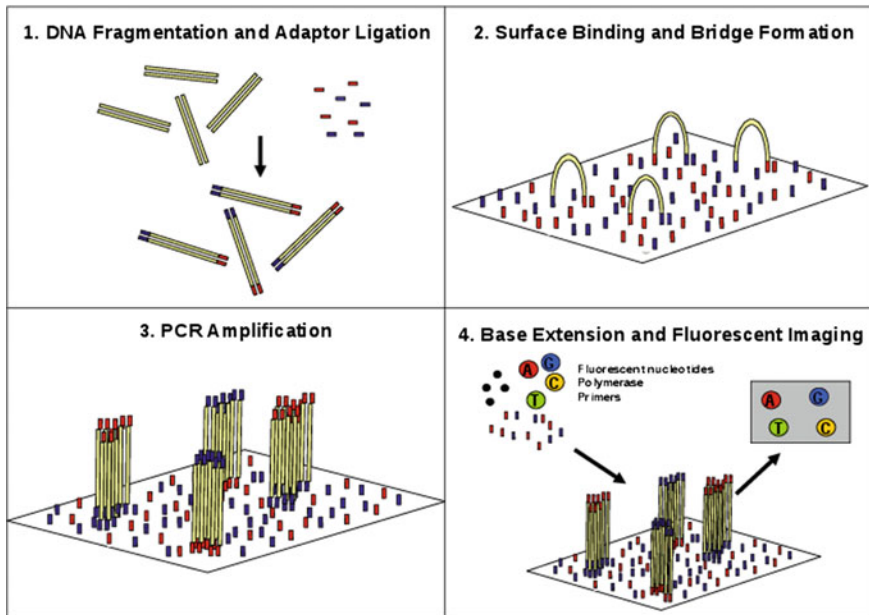


Fig. 2.1 The process of Illumina Sequencing. **1** The DNA to be sequenced is fragmented and short adaptor sequences are ligated to the ends of the fragments. **2** The DNA fragments are attached to a substrate coated with primers. **3** Bridge PCR amplifies each fragment to produce spatially separated clusters of fragments. **4** Fluorescently labeled nucleotides, polymerase and primers are added to synthesize complementary strands of DNA. After each step of nucleotide addition, the fluorescent signal from each cluster is read to generate the base sequence

others by tethering them to separate locations on a substrate. Primers complementary to the adaptor sequences at the end of each DNA fragment are added. The DNA molecules are then sequenced by extending these primers to produce DNA strands complementary to the template. Sequence data is acquired by imaging of the full array at the end of each cycle of nucleotide addition. These technologies have been described elsewhere [64, 107]. Figure 2.1 shows the process of DNA sequencing using Illumina technology.

The process of Sanger sequencing also begins with breaking the DNA molecule into smaller fragments. However, Sanger sequencing then requires each DNA fragment to be cloned into a vector and amplified in host cells (usually *E. coli*). This is time-consuming and expensive, and also, many cloned sequences are not stably maintained in the host [62]. In NGS, the sequencing library is constructed and amplified entirely in vitro, saving the trouble of cloning and colony picking.

Further, as the effective size of next-generation sequencing features can be in microns, millions of sequencing reads can be obtained in parallel by imaging of a small surface area. This makes the NGS strategy both faster and cheaper than Sanger sequencing. Also, because sequencing features are immobilized to a planar surface,

they can be enzymatically manipulated by a single reagent volume. This results in a much lower effective reagent volume per feature, and therefore lowers cost [93].

Perhaps the greatest drawback of NGS technologies is their lower read length (35–400 bp, depending on the platform) compared to Sanger sequencing (650–800 bp). The raw accuracy of sequencing is also currently lower for NGS than Sanger sequencing. However, it is expected that these problems will reduce with further improvements in technology.

2.3 NGS in the Study of Bacterial Communities, Evolution and Epidemiology

Bacteria are the most predominant form of free-living life on earth. Many bacteria are disease causing and understanding their evolution and function might help in developing intervention procedures. From a basic science standpoint, they are excellent systems for studying adaptation, both at the level of genome content and at the level of controlling gene expression and protein activity in response to changing conditions. In the following sections, we explore how recent literature has explored these aspects of bacterial biology using NGS technologies.

2.3.1 Genome Sequencing and Re-sequencing

Despite the increased speed and lower cost of sequencing, *de novo* sequencing of a genome using NGS is challenging due to the low read length of these instruments. This makes it difficult to assemble the sequenced fragments into a complete genome. The low read-length of the sequencing data is offset by great depth; a single lane of sequencing on an Illumina HiSeq 1000 sequencer, with a 12x multiplexing of samples, will provide 3Gb of sequence data per sample, thus giving a 1,000-fold coverage of an ‘average’ bacterial genome. Our experience working with genomes of *Staphylococcus aureus* shows that extremely stringent filtering of sequencing reads could easily give >200-fold coverage [81]. Compare this with the ~10-fold coverage that was typically achieved with Sanger sequencing projects, after laborious experimental work!

These unique characteristics of NGS data, namely an extremely large number of short reads with unique error characteristics, have led to the development of dedicated software, such as Velvet [119], SOAPdenovo [57], and ALLPATHS [11], for genome assembly. These methods have developed to such an extent that with sufficient coverage, short-read sequencing data can be used to produce first-pass mammalian genome assemblies that are comparable to those obtained with traditional Sanger sequencing [31].

Smith et al. [97] first used pyrosequencing to sequence the ~4Mb genome of the human pathogen *Acinetobacter baumannii*, the causative agent of several infections including pneumonia and meningitis. To overcome the limitation of short DNA reads (~100 nt on average), they obtained more than 21-fold coverage of the genome. More recently, Bos et al. [8] used Illumina to sequence DNA samples from victims of the 'Black Death' pandemic that spread through Europe in the fourteenth century, and reconstructed the ancient genome of the bacterium *Yersinia pestis* that was responsible for the pandemic.

A more common application of NGS is genome resequencing, i.e. sequencing the genome of a member of a species for which a reference genome is already available (Figure 2.2a). This is often done to catalog variations such as single nucleotide polymorphisms (SNPs), insertions and deletions relative to a reference genome, for example to identify those that might lead to interesting phenotypes. Before the advent of NGS, such studies were pursued using DNA microarrays, designed on the basis of one or more fully-sequenced reference genomes, following a technique dubbed comparative genome hybridization (CGH).

CGH is a tool to detect variations in DNA copy number between a test and reference genome. In this method, samples of genomic DNA from the test and reference cells are hybridized to microarrays. If, for instance, a gene from the reference genome is absent in the test organism, then the fluorescence intensity from its corresponding probe will be considerably less in the test than in the reference. For example, Willenbrock et al. [115] designed a microarray with probes covering the total content of 32 *E. coli* genomes to characterize novel *E. coli* strains based on their genomic content. McCarthy et al. [67] developed a 62-strain *S. aureus* microarray and used it to compare the genomes of different isolates of *S. aureus* from pigs and humans. They found that while the core genomes of these isolates did not vary much, the distribution of mobile genetic elements was variable and several mobile elements were host-specific. McCarthy et al. [68] used the same method to compare mobile genetic elements between 40 isolates of methicillin-resistant *S. aureus* from a hospital and found a diverse range of MGEs, virulence and resistance genes in the population.

The CGH technique, however, suffers from the drawback of interrogating only genomic regions present in the reference genome(s). Further, it is hard to distinguish between regions that are absent in the test genome and those that are merely divergent in the probed loci. Though these can be partially overcome by the adoption of a larger number of reference genomes in the design of the microarray, the process becomes quickly complicated, in particular for genomes with what is called an open pan-genome [101]. However, the advent of NGS has circumvented these problems, leading to several studies on the genetic variation and evolution of pathogens.

He et al. [42] used 454 and Sanger sequencing to sequence the genomes of thirty isolates of *Clostridium difficile*, which causes diarrheal disease. A phylogenetic analysis of these genomes suggested that both horizontal gene transfer and large-scale recombination played a significant role in the evolution of this species, and that virulence evolved independently in multiple lineages. In another such study,

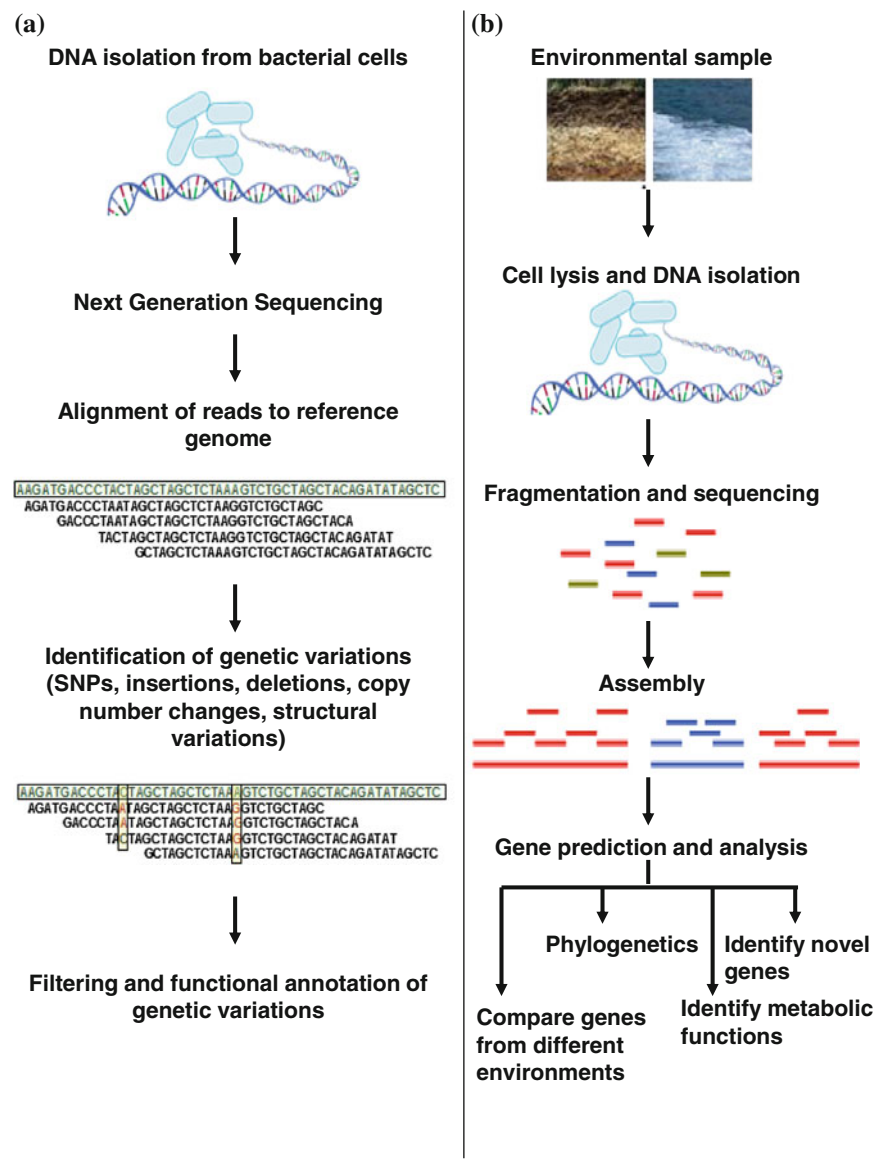


Fig. 2.2 **a** Identifying genome variations by resequencing. **b** Metagenomics

Mutreja et al. [72] studied the evolution of *Vibrio cholerae*, which causes millions of cases of cholera every year, and has caused seven recorded pandemics. They sequenced the genomes of 136 isolates of *V. cholerae* and identified SNPs to construct a phylogeny of this species. Their analysis suggested that the seventh cholera pandemic spread from the Bay of Bengal in at least three independent but overlapping

waves with a common ancestor in the 1950s, and identified several transcontinental transmission events. Similar genome-scale epidemiological studies have been carried out for *S. aureus* [39, 51, 66], and Casali et al. [12] sequenced 34 isolates of *Mycobacterium tuberculosis* to study the evolution of drug resistance in this pathogen. Such studies have given us valuable insights into the evolution and spread of human diseases, and may help us understand how to control epidemics in the future.

The ability to sequence large numbers of genomes also enabled studies in which the genomes of bacteria were sequenced over the course of long-term evolution experiments to identify regions that underwent evolutionary change. For example, Wielgoss et al. [114] sequenced 19 *E. coli* genomes from a 40,000-generation evolution experiment and estimated the point-mutation rate in the *E. coli* genome based on the accumulation of synonymous substitutions to be 8.9×10^{-11} per base pair per generation.

Other studies have examined the evolution of pathogens in their hosts. Yang et al. [118] studied the evolutionary dynamics of *Pseudomonas aeruginosa* as it adapted to its human hosts over 200,000 generations. These authors found that the population underwent limited genotypic diversification—most of which occurred early in the form of a few pleiotropic mutations followed by a landscape dominated by negative selection—despite the complex host environment. This was reported to be in contrast to *in vitro* studies, which documented continuous positive selection. Lieberman et al. [59] sequenced the genomes of 112 isolates of *Burkholderia dolosa* collected from human hosts over 16 years. They identified a set of genes that acquired nonsynonymous mutations in several individuals, suggesting that they experienced strong selection pressure during pathogenesis. These genes were involved in processes important for pathogenicity, such as antibiotic resistance and membrane biosynthesis, and might represent possible targets for therapy.

The acquisition of antibiotic resistance by bacterial pathogens is a growing problem, and NGS has been used to investigate this phenomenon. Zhang et al. [120] grew *E. coli* in the presence of the antibiotic ciprofloxacin and sequenced the genomes of the bacteria that survived, to identify mutations that gave rise to resistance. Whole-genome sequencing revealed that four single-nucleotide polymorphisms, including one in the *gyrA* gene encoding gyrase, were fixed in the resistant population. Similarly, Toprak et al. [102] grew *E. coli* with several different drugs and studied the evolution of resistance over 20 days. Sequencing the genomes of the resistant populations revealed mutations that conferred resistance to specific drugs and to multiple drugs.

2.3.2 Metagenomics

Metagenomics, or community genomics, is an approach to analyze the total genomic content of a microbial community. The total DNA from a population is isolated, sequenced and compared with previously known sequences. Metagenomic studies allow researchers to discover new species, and also to identify the types of biological

processes that occur in a specific environment. (Figure 2.2b) These methods permit bird's eye-level genetic characterization of unculturable bacteria, which represent most of the bacterial populations on the planet.

Early metagenomic explorations were based on painstaking Sanger sequencing experiments [30, 106, 111]. However, the main difficulty was presented by the sheer number of microbes in an environmental sample—it is estimated that there are $\sim 10^6$ bacterial species per gram of soil. The DNA of bacteria in the human gut, which has been a subject of keen interest among biologists and medics alike, can be expected to harbor much more diversity than the genome of its host. This represents a vast amount of DNA to sequence. Further, sampling rarer constituents of the microbial population requires great depth of sequencing, which is difficult to achieve with traditional sequencers. The high-throughput capability, relatively low cost and depth of next-generation sequencing makes such an approach much easier. Next-generation sequencers have been used to sequence the metagenome of diverse environments such as soil, oceanic communities, and the human gut.

Edwards et al. [26] used pyrosequencing to sequence genomes from two adjacent but chemically and geologically different sites in an iron mine in Minnesota. The microbial communities at the two sites were found to be functionally and metabolically different from each other, in pathways such as carbon utilization, iron acquisition, nitrogen assimilation, and respiration. Dinsdale et al. [22] used pyrosequencing to compare microbial and viral DNA sequences from nine biomes including marine, freshwater, subterranean, and host-associated, and found strongly discriminatory metabolic profiles across different environments.

It is estimated that up to 100 trillion microbial cells reside in an average human body [5]. Most of these microbes are present in the gut, where they are thought to influence human physiology, nutrition, and health [17, 87]. To understand and exploit the functioning of the gut microbial community it is necessary to understand its content and diversity. Qin et al. [83] used Illumina to sequence the total genomic DNA from faecal samples of 124 European individuals. They assembled and characterized 3.3 million microbial genes from 576.7 Gb of sequence, and found that each individual harbored at least 160 bacterial species, which were largely common across individuals. Further, they compared samples from healthy individuals and inflammatory bowel disease patients, and showed that the gut microbiomes of the two groups differed in terms of overall bacterial diversity as well as the relative abundances of various species.

'Metatranscriptomic' studies aim to sequence the total RNA expressed by microbes in an environmental sample, instead of DNA. This is particularly interesting as bacteria have several untranslated small RNAs that regulate environmentally important processes, including amino acid biosynthesis, starvation responses, and quorum sensing [99]. Because studies on sRNAs have focused on a few model microorganisms, the diversity and ecological function of sRNAs in natural communities is little understood. Compared to protein-coding sequences, these are also difficult to identify from DNA sequences. Shi et al. [94] analyzed metatranscriptomic data sets from ocean water and found that a large fraction of cDNA sequences detected comprised small RNAs. They also identified several new classes of putative sRNAs.

Metagenomics has given scientists access to unculturable microbial communities and their activities in a wide variety of systems. Though this approach can tell us about the species and functions present in a microbial community, it is very difficult to assemble individual genomes in the community from such data. However, Iverson et al. [45] were able to reconstruct a nearly complete genome of the uncultured marine group II *Eurarchaeota* entirely *de novo* from the metagenome of seawater samples.

2.3.3 Single-Cell Genomics

The various limitations of metagenomics, particularly difficulties in accessing rare components of the microbiota and ability to map genes to individual species or organisms can be overcome by single-cell genome sequencing. Methods have now been developed to isolate single cells and amplify DNA for sequencing. The higher throughput of NGS makes it possible to finish several single-cell genomes in a reasonable time, and single-cell sequencing has now been applied to many environmental microbes.

Marcy et al. [63] developed a microfluidic device for isolating single cells and amplifying their genomes, and used it to isolate bacteria of a little-understood phylum from the human oral microbiota. They were able to assemble the sequences of over 1,000 genes by pyrosequencing, providing insight into the physiology of members of this phylum. Woyke et al. [116] isolated DNA from individual cells of two marine flavobacteria from the Gulf of Maine that were phylogenetically distant from existing cultured strains. With a combination of pyrosequencing and Sanger sequencing they recovered 91 % and 78 % of the two genome sequences, and analyzed the genome content, metabolic adaptations, and biogeography of these taxa.

2.4 Studying Genome Function

2.4.1 Gene Expression Control in Bacteria

Bacteria do not express all the genes present on their genome all of the time. Instead, they produce those gene products that are important for them to survive in the environment they face. Bacteria may accomplish this by modulating any step of a gene's expression and protein activity, from transcription to translation to post-translational modification of a protein, though they most commonly control gene expression by regulating the level of transcription.

Transcription is regulated at multiple levels, including: (a) variations in the sequence of the promoter to which RNA polymerase, the enzyme responsible for transcription, binds, thus ensuring that different genes have different inherent ability to be transcribed; (b) three-dimensional topology of the DNA where DNA supercoiling controls many DNA transactions including transcription, by controlling the extent to which the DNA is unwound; (c) sigma factors, which tightly associate with the

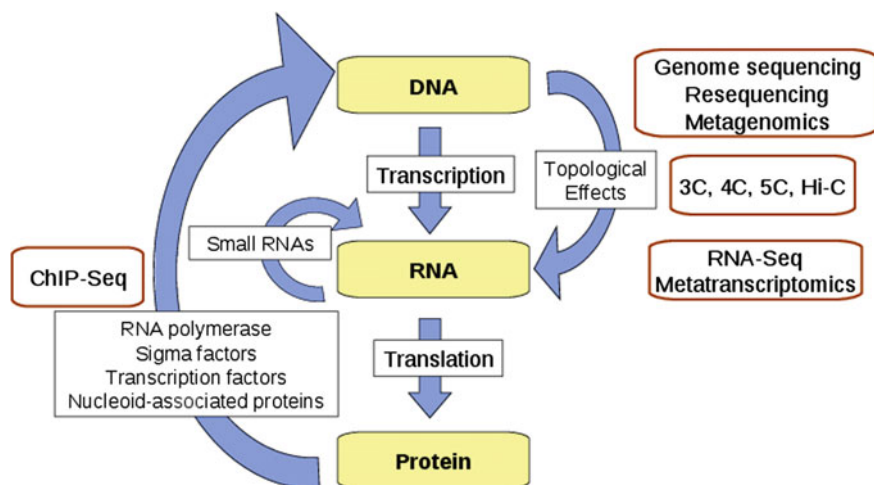


Fig. 2.3 NGS—based techniques are used to understand bacterial genomics at multiple levels

RNA polymerase and recognize promoter sequences; (d) transcription factors, which are proteins that alter the affinity of a promoter to the RNA polymerase by binding close to the promoter and either enabling or blocking access of the promoter to the RNA polymerase; (e) various small molecule and RNA-based ligands that bind to components of the RNA polymerase thus altering its availability to transcribe genes; (f) DNA modifications such as methylation.

Details of these regulatory mechanisms have been reviewed elsewhere [9, 44, 73]. Here we review research that has investigated the above aspects of gene regulation in bacteria using techniques based on NGS (See Figure 2.3). Where applicable, we point to relevant research based on microarrays indicating that similar work can be pursued using sequencing as well.

2.4.2 Describing Transcriptomes

The transcriptome of a cell is the total content of RNA transcripts expressed in the cell at a given time. Studying the transcriptome of a bacterium helps us to understand how it responds to different environmental conditions. For example, Nicolas et al. [74] investigated the transcriptomes of *Bacillus subtilis* grown under 104 conditions (nutrients, aerobic and anaerobic growth, stresses, etc.) that the bacterium might encounter in nature.

Before the development of NGS, most methods to study gene expression required hybridization of specific oligonucleotides to the loci of interest—either primers binding to complementary cDNA in quantitative reverse transcription polymerase chain reactions (qRT-PCR), or labeled probes binding to RNA in Northern blotting, or hybridization of cDNA to probes on microarray chips. Of these, only DNA

microarrays had a high throughput. However, sequencing of RNA by next-generation sequencing (RNA-seq) is different, in that fragments of DNA are matched to genes by sequence alignment instead of physical hybridization. (See [29] for a review of bacterial transcriptomics). These avoid problems typical of microarrays, such as (a) a limitation to interrogate only regions of the chromosome that have been used as probes; (b) saturation of fluorescence signals that limit the dynamic range; (c) background noise in fluorescence; (d) artifacts due to probe characteristics such as base composition.

Global transcriptional analyses using RNA-Seq have been carried out for many bacteria. Güell et al. [35] analyzed the transcriptome of one of the smallest self-replicating organisms, *Mycoplasma pneumoniae*, using a combination of microarrays and sequencing. They found that even this simple bacterium has several antisense transcripts, alternative transcripts, and multiple regulators per gene. This suggests a dynamic and regulated transcriptome, more similar to that of eukaryotes than was previously thought.

However, it is not possible to directly locate the transcription start sites of genes using RNA-Seq, as it is not possible to distinguish between primary transcripts and processed transcripts. Sharma et al. [92] used a differential or dRNA-seq approach to discriminate these two. Primary transcripts have a 5' triphosphate group, whereas processed transcripts such as ribosomal and transfer RNAs have a 5' monophosphate. They carried out 454 pyrosequencing of two cDNA libraries—one prepared from untreated total bacterial RNA, and the other enriched for primary transcripts by treatment with an exonuclease that degrades 5' monophosphate but not 5' triphosphate RNA. They were able to map ~217 million bases of cDNA to the *H. pylori* genome, and construct a genome-wide map of *H. pylori* transcriptional start sites and operons. They discovered hundreds of transcriptional start sites within operons, and opposite to annotated genes, indicating that antisense transcription takes place throughout the genome. They also discovered ~60 small RNAs of different classes.

RNA-Seq involves reverse transcribing single-stranded RNA into double stranded cDNA, which is then sequenced. The result is the sequences of both DNA strands of the gene that encodes the RNA. Hence RNA-Seq does not tell us which strand of the DNA is transcribed into RNA, which is important to resolve overlapping genetic features and detect antisense transcription. However, there are methods to identify the directionality of transcription. These generally involve modifying the RNA molecules before reverse transcription, or modifying the first cDNA strand before the synthesis of its complementary strand. Croucher et al. [18] developed strand-specific cDNA sequencing, in which they reverse transcribed the RNA into only one strand of cDNA and directly sequenced the library of single-stranded cDNA molecules. Perkins et al. [79] used this to analyze the transcriptome of *Salmonella enterica* serovar Typhi in a strand-specific manner. This allowed them to identify many transcribed regions within prophages, pseudogenes, and UTRs of other genes.

As no pre-existing knowledge of the RNA sequence to be detected is necessary, RNA-Seq has been especially useful for discovering new species of RNA. Sit-tka et al. [96] carried out a transcriptome analysis in *Salmonella enterica* serovar Typhimurium, and also sequenced the small RNAs associated with the regulatory

protein Hfq. They discovered several novel small RNAs, and found that Hfq regulates the expression of nearly one-fifth of all *Salmonella* genes, including several pathogenicity islands.

Mraheil et al. [71] sequenced cDNAs from RNA less than 500 nucleotides long, in the intracellular pathogen *Listeria monocytogenes*, during intracellular and extracellular growth. They discovered 150 putative regulatory RNAs, including 29 that were expressed only during intracellular growth. Some of these were found to be required for efficient intracellular growth and infection by this pathogen. Lasa et al. [54] sequenced long and short (<50 nt) RNAs from *S. aureus*. They found short RNAs that were produced by RNase III digestion of double stranded RNAs formed by overlapping sense and antisense transcripts throughout the genome. This suggested that antisense transcription is used to suppress expression of some genes by producing double-stranded RNA that is degraded.

2.4.3 Promoter Sequences and Their Affinity to RNA Polymerase

Cho et al. [16] performed a ChIP-chip experiment—in which fragments of DNA bound to a protein of interest are isolated using an antibody and the resulting DNA fragment hybridized to a microarray—for RNA polymerase in the presence of the antibiotic rifampicin which blocks transcription elongation. This was shown to provide a static picture of RNA polymerase occupancy at promoters. Though this was used primarily to define promoters and transcription start sites, one can envisage these data being used as a measure of the inherent affinity of a promoter to the RNA polymerase.

2.4.4 Structure of the DNA and Its Effect on Transcription

The three-dimensional structure of the bacterial genome both reflects and regulates the functional state of the cell. However, until recently it had not been possible to study the three-dimensional conformation of the chromosome on a genomic-scale with high resolution. Recent techniques under the general category of “chromosome conformation capture” (3C [21], 4C [95], 5C [25] and Hi-C [60]), build interaction maps in which spatially proximal regions of the chromosome are linked together. This network is subsequently used to build a three-dimensional model of the chromosome.

Umbarger et al. [105] used 5C to construct a 3D model of the *Caulobacter crescentus* genome in wild and genetically modified strains. They found the chromosome to be ellipsoidal with periodically arranged arms, and identified a short region of the genome that affected the orientation of the entire chromosome.

An important topological property of the chromosome that affects transcription is supercoiling. Though it has not been possible to define local supercoiling at high resolution on a genomic scale yet, Peter et al. [80] used DNA microarrays to identify genes that respond to perturbations to the global supercoiling levels. They found that

negative supercoiling activates expression of ~200 genes in the *E. coli* chromosome, while repressing that of ~100. Genes that were activated by negative supercoiling tended to have higher G/C content than average, whereas the opposite was true of those that were repressed. This has potential implications for transcription during stationary phase.

2.4.5 Sigma Factors and Transcription

Bacterial RNA polymerase is a multisubunit enzyme. The core RNA polymerase, composed of five subunits, is capable of transcribing DNA. However, this core polymerase is not capable of binding tightly and specifically to promoter sequences. This ability is conferred by the sigma subunit [9].

The number of sigma subunits varies between bacterial species. *E. coli* has seven, of which RpoD is the 'housekeeping' sigma factor that transcribes most of the cellular genes in growing cells. The other sigma subunits are activated under different environmental conditions and direct the transcription of genes needed to survive in those conditions. For example, RpoH is activated under heat stress and transcribes chaperones and other genes involved in the heat stress response. Nicolas et al. [74] examined the transcriptome of *B. subtilis*, and concluded that approximately 66 % of the variance in gene expression between different environmental conditions can be explained by variation in the expression of different sigma factors.

Wade et al. [109] used ChIP-chip to identify the genomic binding sites of RpoH in *E. coli*. Interestingly, a quarter of the RpoH targets were found within coding regions. Also, most of the targets overlapped with those of RpoD, suggesting extensive overlap between the functions of different sigma factors. However this result remains controversial [110]. Patten et al. [78] and Weber et al. [113] used microarrays to compare gene expression of wild-type *E. coli* with a strain lacking a functional RpoS sigma factor. This sigma factor controls the expression of many stationary-phase genes. They identified hundreds of genes that are regulated by RpoS during the transition into stationary phase.

NGS can be a powerful tool to study the functions of sigma factors in detail. Recently, Dong and Mekalanos [24] used ChIP-Seq and RNA-Seq to define the regulon of the alternative sigma factor RpoN in *Vibrio cholerae*. They identified a consensus sequence for RpoN binding and showed that RpoN regulates the expression of flagellar genes and secreted proteins.

2.4.6 Global Transcription Factors

Many crucial processes in the cell, including chromosome organization, replication, and regulation of gene expression are orchestrated through the interaction of proteins with their binding sites on the bacterial genome. Understanding these processes on a global scale requires mapping of protein-DNA interactions across the entire genome.

Vora et al. [108] profiled the binding sites of all proteins on the entire *E. coli* genome using a technique they named *in vivo* protein occupancy display (IPOD). They isolated protein-DNA complexes and used DNA microarrays to identify protein-bound domains on the *E. coli* chromosome. They found extensive (longer than 1 kb) protein occupancy domains (EPODs), many of which were located in highly curved and transcriptionally silent regions of the genome. They suggested that these EPODs bind nucleoid-associated proteins and act as organizing centers that isolate the domains of the chromosome. However, this technique does not allow identification of the protein bound at each site, which has been addressed with ChIP (described earlier).

As with transcriptomes, ChIP studies were first carried out on a genomic scale using DNA microarrays. However, in a manner similar to RNA-Seq, ChIP followed by NGS (ChIP-Seq) can produce significantly better data than microarrays, including single base-pair resolution when appropriately modified (ChIP-exo), lower noise, a larger dynamic range, and of course it is not limited by fixed probe sequences. Although the short reads (~35 bp) of NGS platforms are disadvantageous for applications like de novo genome assembly, they are acceptable for ChIP-Seq.

ChIP-Seq has been used to study the DNA binding of many proteins that are involved in global gene regulation, including RNA polymerase, nucleosomes and transcription factors, in eukaryotes [6, 13] and to a much lesser extent in prokaryotes.

Bacterial transcription factors can be broadly classified into local and global transcription factors. Local transcription factors regulate the expression of a small number of genes, generally directed toward a single cellular function. On the other hand, global transcription factors regulate a large number of genes which belong to multiple functional categories, act under several different environmental conditions, and bind extensively to chromosomal DNA. For example, there are 187 known transcription factors in *E. coli*, but nine of these regulate 63 % of the target genes and are responsible for 52 % of the regulatory interactions [89]. Seven of these are considered to be global transcription factors: CRP, FNR, ArcA, LRP, FIS, IHF and H-NS [65].

Genome-wide approaches are necessary to understand the function of global transcription factors, as these proteins influence transcription across the entire genome. Grainger et al. [34] used ChIP-chip and microarrays to study the binding pattern of the global transcription factor CRP across the *E. coli* genome and identify its regulated genes. They found that while CRP does not have many strong binding sites (around 70), it binds to several weaker sites throughout the genome. Cho et al. [14] carried out ChIP-chip to identify binding sites of the *E. coli* transcription factor Lrp and RNA polymerase along with a comparison of gene expression in wild-type and an Lrp knockout. They showed that Lrp regulates a large number of genes involved in diverse functions and has three modes of regulation at different promoters.

More recently, NGS techniques have been used to study transcription factor function. Kahramanoglou et al. [50] used ChIP-Seq to map the binding sites of the nucleoid-associated proteins H-NS and Fis throughout the *E. coli* chromosome. These proteins were previously studied using ChIP-chip [15, 33, 76]. Fis affects the expression of over 20 % of all genes, mostly by increasing transcription. However, there was little correlation between Fis binding regions and regions where the

protein affected transcription. But, regions where such a correlation existed were characterized by multiple Fis binding sites in operon-upstream regions. This in turn was correlated with high A/T content of the binding region and possible DNA bending. H-NS binds to longer stretches of DNA, and shows mild or strong repression of its target genes depending on the length of the binding site. Prieto et al. [82] identified the binding sites of HU and IHF in *E. coli* on a genome-wide scale using ChIP-seq, along with microarray analysis of gene expression in single- and double-deletion mutants of each protein. They found that the binding of IHF was sequence specific and included ~30% of all operons in the genome, demonstrating its role as a global regulator. HU was seen to bind non-specifically to the chromosome, though with a preference for A/T-rich DNA.

2.4.7 DNA Modifications and Transcription

Methylation of the nucleotides of DNA is another important means of regulation in a cell. In bacteria, the Dam methylase adds a methyl group to the adenine residue in the sequence 5'-GATC-3'. Dam methylation regulates DNA replication, mismatch repair and transcription in bacteria, by modulating protein-DNA binding. A microarray study by Robbins-Manke et al. [85] found an upregulation of over 200 genes in the absence of Dam in *E. coli*. They suggested that this might be due to Dam changing the binding sites of transcription factors and RNA polymerase and hence modulating the binding of these proteins.

A second DNA methylase in *E. coli* is DNA Cytosine Methylase or Dcm, which methylates the internal cytosine in CCWGG sequences. Kahramanoglou et al. [49] carried out bisulfite sequencing of *E. coli* genomic DNA using Illumina to identify sites of cytosine methylation. They also compared gene expression of wild-type *E. coli* with a Dcm knockout strain and identified over 500 differentially expressed genes. Methylation by Dcm progressively increases from exponential to stationary phase, and Dcm may also regulate the stationary phase sigma subunit RpoS. Another study [70] had previously shown that Dcm regulates the expression of ribosomal proteins in stationary phase.

2.5 Computational Challenges of Next-Generation Sequencing

Because of the higher parallelism and lower cost of sequencing, the widespread use of NGS in biology has made massive amounts of sequence data available. Also, as single-cell genomics becomes more widely used, it is likely that even greater amounts of genomic data will become available. The large datasets produced by NGS experiments require large amounts of storage space. Apart from the sequence data itself, an NGS experiment initially produces terabytes of raw image files. Once base calling is done to convert these images of fluorescent light into DNA sequences,

these images are discarded due to lack of storage space. It would be an interesting challenge to store and use these images, possibly to improve base calling algorithms. Several tools and algorithms have been designed to analyze data generated by Next-Generation Sequencing experiments (See Table 2.1).

2.5.1 Reference Mapping

For genome resequencing as well as techniques like ChIP-Seq and RNA-Seq, the short sequence reads obtained from NGS have to be mapped back to their position on a reference genome. This remains a challenge for large and complex genomes like that of humans. Commonly used alignment algorithms like BLAST have drawbacks for this application. Primarily, NGS data may contain many millions of short reads, which BLAST would be very slow to align [62]. Therefore there was a need for methods that were designed to work with short sequences and save time by operating on compressed data.

The Burrows-Wheeler Transform is an algorithm used to permute the order of characters in a sequence, which allows sequence data to be greatly compressed. This technique also allows searching for subsequences in the original sequence while operating on the compressed file, making this technique suitable for mapping NGS data to reference genomes. BWA [58] and BOWTIE [53] are successful short-read aligners which are based on this technique.

2.5.2 Genome Assembly

De novo genome assembly is done by piecing together sequence fragments to join them into contigs or contiguous sequences. Many algorithms designed for whole-genome sequence assembly from Sanger sequencing data use an approach of representing each read as a node and each overlap between sequences as an arc between the two nodes. However, NGS data contains a much larger number of very short reads. The number of reads would make such an overlap graph very large and difficult to compute [119].

A different approach is based on de Bruijn graphs, which are not based on reads, but on k -mers (words that are k nucleotides long). Reads are mapped as paths through the graph, going from one word to the next word in a determined order. Velvet is a short read assembler based on de Bruijn graphs that has been used to assemble bacterial genomes [119].

2.5.3 Analysis of ChIP-Seq Data

When ChIP-Seq is carried out for global transcription factors, sequences over the entire genome are sampled. These sequences are mapped back to their position on a

Table 2.1 Computational tools to analyze NGS data

Category	Program	Type	Reference
de novo Genome assembly	1. Velvet	1–3. De Bruijn assemblers	1. Zerbino and Birney (2008) [119]
	2. ALLPATHS	4–6. Greedy extension assemblers	2. Butler et al. (2008) [11]
	3. SOAPdenovo		3. Li et al. (2010) [61]
	4. SSAKE	7. Overlap-layout-consensus assembler	4. Warren et al. (2007) [112]
	5. SHARCGS		5. Dohm et al. (2007) [23]
	6. VCAKE		6. Jeck et al. (2007) [46]
	7. Edena		7. Hernandez et al. (2008) [41]
Reference mapping	1. MAQ	1–4. Burrows-Wheeler transform based	1. Li et al. (2008) [56]
	2. BWA		2. Li and Durbin (2009) [58]
	3. Bowtie	5. Hash table based	3. Langmead et al. (2009) [53]
	4. SOAP		4. Li et al. (2008) [57]
	5. SSAHA		5. Ning et al. (2001) [75]
RNA-Seq	1. Scripture	1. Transcriptome reconstruction.	1. Guttman et al. (2010) [37]
	2. Cufflinks	2. Transcript assembly, estimation, differential expression testing	2. Trapnell et al. (2010) [104]
	3. TopHat		3. Trapnell et al. (2009) [103]
ChIP-Seq	4. SpliceMap	3, 4. Splice junction discovery	4. Au et al. (2010) [2]
	1. SISSRS	1–4. Peak identification	1. Jothi et al. (2008) [48]
	2. MACS		2. Zhang et al. (2008) [121]
	3. BayesPeak	5, 6. Visualization, peak detection, gene-peak association	3. Spyrou et al. (2009) [98]
	4. PeakSeq		4. Rozowsky et al. (2009) [88]
	5. CisGenome		5. Ji et al. (2008) [47]
	6. CASSys		6. Alawi et al. (2011) [1]
Conformation capture	1. my5C	A web tool for design, visualization and analysis of 5C studies	1. Lajoie et al. (2009) [52]

reference genome. After reference mapping, the next step in ChIP-Seq data analysis is peak calling, i.e. identifying the genomic regions that produced a relatively high number of sequence reads, indicating that they are the binding sites of the protein of interest. However, ChIP-Seq data often contains a high amount of noise, and a major challenge in analysis is how to distinguish peaks from background noise.

One way to reduce this problem is to perform a ‘mock IP’ or control experiment in which the procedure is carried out without using specific antibodies [77]. If a peak in the experimental data co-localizes with a peak in the control data, it would lower the likelihood of having detected a binding site. Several algorithms have been developed to carry out peak calling for Chip-Seq data, including BayesPeak [98] and MACS [121].

2.5.4 Analysis of RNA-Seq Data

In an RNA-Seq experiment, the abundance of a gene in the sample is measured by the number of reads that map to that gene. The number of reads for a gene in the raw sequence data is generally normalized by the length of a gene and by the total number of reads. However, the number of reads of a given gene depends not only on its abundance, but also on the abundance of other genes. For example, if an RNA sample includes a very highly expressed gene, most of the reads may be taken up by its transcripts, leaving very little sequencing space for less abundant transcripts. The number of reads for a transcript in such conditions may not reflect its actual abundance, and this can be a difficulty in comparing transcriptomes of different samples. Robinson and Oshlack [86] proposed a normalization method using a weighted trimmed mean of the log expression ratio to estimate the ratio of RNA production.

2.5.5 Chromosome Conformation

Yaffe et al. [117] studied several sources of experimental bias in Hi-C experiments. These were: (1) Ligation occurs between nonspecific cleavage sites as well as restriction fragment ends. (2) The efficiency of ligation of restriction fragments may depend on their length. (3) The GC content near the ends of the ligated fragments may influence the processing and sequencing of the DNA. (4) The mappability or genomic uniqueness of the fragment ends affects the estimated probability of contact between sequences.

There is a need to minimize these experimental biases and incorporate them into algorithms to analyze chromosome conformation capture data.

2.6 The Future of Sequencing Technologies

Just as NGS technologies did in the last decade, a number of emerging DNA sequencing technologies promise to influence the study of genomes in the coming years. A major advance in this field is single-molecule sequencing, which allows the sequence of a single molecule of DNA to be read without the need for any amplification step at all.

One example of a single-molecule sequencing technology is Single molecule real time sequencing (SMRT). In this, a single DNA polymerase enzyme is presented with a template DNA molecule and substrate nucleotides. Each of the four DNA bases is attached to a different fluorescent dye. The fluorescent signal of the incorporated nucleotides is read to generate the sequence. SMRT data was first published by Harris et al. [40], who used this method to resequence the M13 phage genome. This method can also be used to detect DNA methylation, and, in theory, other DNA modifications [28].

Another promising recent method is nanopore sequencing. A nanopore is a hole with diameter of the order of 1 nm. When a nanopore is immersed in a conducting fluid and a potential is applied across it, an electric current passes through it. The DNA to be sequenced is forced through the pore one base at a time. As it does so, each nucleotide of the DNA obstructs the nanopore. The degree of obstruction, and hence the amount of current through the nanopore, varies depending on whether the nucleotide blocking the nanopore is an A, C, G or T. The change in current through the nanopore as the DNA passes through it can therefore be used to directly read the DNA sequence.

At present, Oxford Nanopore Technologies is developing a commercial nanopore sequencing system. They have recently announced that they have used this technology to sequence the 5.4 kb genome of the Phi X phage in one continuous read.

2.7 Lessons Learnt

1. Next-Generation technologies to sequence DNA are significantly faster and cheaper than Sanger sequencing. However, they have lower read length and accuracy of sequencing.
2. NGS technologies have made a great difference to the field of bacterial genomics by allowing whole genomes and large sets of DNA to be sequenced at a reasonable speed and cost. This has allowed sequencing of many bacterial genomes and metagenomic sequencing of DNA from whole bacterial communities. Ecological and evolutionary studies based on NGS have given us valuable insights.
3. Bacteria respond to changing environmental conditions by regulating the expression of their genes at many levels. Methods of gene regulation used by bacteria include DNA topology, promoter sequences, sigma subunits, transcription factors, RNA polymerase ligands and small RNAs.

4. NGS has enabled studies of gene regulation on a global scale in bacteria, for example by mapping the topology of the bacterial chromosome, studying the transcriptome of bacteria under different conditions, and by mapping sites at which transcription factors bind to the DNA.
5. NGS leads to the generation of vast amounts of sequence data. The storage and analysis of this data is difficult. Several algorithms have been developed specifically for analyzing NGS data.
6. Several new and more advanced techniques for DNA sequencing are being developed, which may lead to further and more interesting developments in the field of genomics.

Acknowledgments We thank Dr. Dasaradhi P. for proofreading the manuscript.

References

1. Alawi M, Kurtz S, Beckstette M (2011) CASSys: an integrated software-system for the interactive analysis of ChIP-seq data. *J Integr Bioinf* 8(2):155
2. Au KF, Jiang H, Lin L, Xing Y, Wong WH (2010) Detection of splice junctions from paired-end RNA-seq data by splicemap. *Nucleic Acids Res* 38(14):4570–4578
3. Babu MM, Teichmann SA, Aravind L (2006) Evolutionary dynamics of prokaryotic transcriptional regulatory networks. *J Mol Biol* 358(2):614–633
4. Babu MM, Luscombe NM, Aravind L, Gerstein M, Teichmann SA (2004) Structure and evolution of transcriptional regulatory networks. *Curr Opin Struct Biol* 14(3):283–291
5. Bäckhed F, Ley RE, Sonnenburg JL, Peterson DA, Gordon JI (2005) Host-bacterial mutualism in the human intestine. *Science* 307(5717):1915–1920
6. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K (2007) High-resolution profiling of histone methylations in the human genome. *Cell* 129(4):823–837
7. Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu X, Rinn JL, Tongprasit W, Samanta M, Weissman S, Gerstein M, Snyder M (2004) Global identification of human transcribed sequences with genome tiling arrays. *Science (New York, N.Y.)* 306(5705):2242–2246
8. Bos KI, Schuenemann VJ, Golding GB, Burbano HA, Waglechner N, Coombes BK, McPhee JB, DeWitte SN, Meyer M, Schmedes S, Wood J, Earn DJD, Herring DA, Bauer P, Poinar HN, Krause J (2011) A draft genome of *Yersinia pestis* from victims of the black death. *Nature* 478(7370):506–510
9. Browning DF, Busby SJ (2004) The regulation of bacterial transcription initiation. *Nat Rev Microbiol* 2(1):57–65
10. Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG, Blake JA, FitzGerald LM, Clayton RA, Gocayne JD, Kerlavage AR, Dougherty BA, Tomb J-F, Adams MD, Reich CI, Overbeek R, Kirkness EF, Weinstock KG, Merrick JM, Glodek A, Scott JL, Geoghagen NSM, Weidman JF, Fuhrmann JL, Nguyen D, Utterback TR, Kelley JM, Peterson JD, Sadow PW, Hanna MC, Cotton MD, Roberts KM, Hurst MA, Kaine BP, Borodovsky M, Klenk H-P, Fraser CM, Smith HO, Woese CR, Venter JC (1996) Complete genome sequence of the methanogenic archaeon, *methanococcus jannaschii*. *Science* 273(5278):1058–1073
11. Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, Lander ES, Nusbaum C, Jaffe DB (2008) ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome Res* 18(5):810–820

12. Casali N, Nikolayevskyy V, Balabanova Y, Ignatyeva O, Kontsevaya I, Harris SR, Bentley SD, Parkhill J, Nejentsev S, Hoffner SE, Horstmann RD, Brown T, Drobniewski F (2012) Microevolution of extensively drug-resistant tuberculosis in Russia. *Genome Res* 22(4):735–745
13. Chen X, Han X, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang W, Jiang J, Loh Y-HH, Yeo HCC, Yeo ZXX, Narang V, Govindarajan KRR, Leong B, Shahab A, Ruan Y, Bourque G, Sung W-KK, Clarke ND, Wei C-LL, Huck-Hui HNg (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* 133(6):1106–1117
14. Cho B-K, Barrett CL, Knight EM, Park YS, Bernhard (2008) Genome-scale reconstruction of the Lrp regulatory network in *Escherichia coli*. *Proc Natl Acad Sci* 105(49):19462–19467
15. Cho B-K, Knight EM, Barrett CL, Bernhard (2008) Genome-wide analysis of Fis binding in *Escherichia coli* indicates a causative role for A-/AT-tracts. *Genome Res* 18(6):900–910
16. Cho B-KK, Zengler K, Qiu Y, Park YSS, Knight EM, Barrett CL, Gao Y, Palsson BØ (2009) The transcription unit architecture of the *Escherichia coli* genome. *Nat Biotechnol* 27(11):1043–1049
17. Clemente JC, Ursell LK, Parfrey LW, Knight R (2012) The impact of the gut microbiota on human health: an integrative view. *Cell* 148(6):1258–1270
18. Croucher NJ, Fookes MC, Perkins TT, Turner DJ, Marguerat SB, Keane T, Quail MA, He M, Assefa S, Bähler J, Kingsley RA, Parkhill J, Bentley SD, Dougan G, Thomson NR (2009) A simple method for directional transcriptome sequencing using Illumina technology. *Nucleic Acids Res* 37(22):e148
19. Croucher NJ, Harris SR, Fraser C, Quail MA, Burton J, van der Linden M, McGee L, von Gottberg A, Song JHH, Ko KSS, Pichon B, Baker S, Parry CM, Lambertsen LM, Shahinas D, Pillai DR, Mitchell TJ, Dougan G, Tomasz A, Klugman KP, Parkhill J, Hanage WP, Bentley SD (2011) Rapid pneumococcal evolution in response to clinical interventions. *Science (New York, N.Y.)* 331(6016):430–434
20. David L, Huber W, Granovskaia M, Toedling J, Curtis PJ, Bofkin L, Jones T, Ronald DW, Lars SM (2006) A high-resolution map of transcription in the yeast genome. *Proc Natl Acad Sci* 103(14):5320–5325
21. Dekker J, Rippe K, Dekker M, Kleckner N (2002) Capturing chromosome conformation. *Science* 295(5558):1306–1311
22. Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM, Furlan M, Desnues C, Haynes M, Li L, McDaniel L, Moran MA, Nelson KE, Nilsson C, Olson R, Paul J, Brito BR, Ruan Y, Swan BK, Stevens R, Valentine DL, Thurber RV, Wegley L, White BA, Rohwer F (2008) Functional metagenomic profiling of nine biomes. *Nature* 452(7187):629–632
23. Dohm JC, Lottaz C, Borodina T, Himmelbauer H (2007) SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing. *Genome Res* 17(11):1697–1706
24. Dong TG, Mekalanos JJ (2012) Characterization of the RpoN regulon reveals differential regulation of T6SS and new flagellar operons in *Vibrio cholerae* O37 strain V52. *Nucleic Acids Res* 40(16):7766–7775
25. Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL, Honan TA, Rubio ED, Krumm A, Lamb J, Nusbaum C, Green RD, Dekker J (2006) Chromosome conformation capture carbon copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res* 16(10):1299–1309
26. Edwards R, Brito BR, Wegley L, Haynes M, Breitbart M, Peterson D, Saar M, Alexander S, Alexander EC, Rohwer F (2006) Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics* 7(1):57+
27. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM et al (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269(5223):496–512
28. Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, Turner SW (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat Methods* 7(6):461–465

29. GÅell M, Yus E, Lluch-Senar M, Serrano L (2011) Bacterial transcriptomics: what is beyond the RNA horizo-me? *Nat Rev Micro* 9(9):658–669
30. Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM, Nelson KE (2006) Metagenomic analysis of the human distal gut microbiome. *Science (New York, N.Y.)* 312(5778):1355–1359
31. Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, Berlin AM, Aird D, Costello M, Daza R, Williams L, Nicol R, Gnirke A, Nusbaum C, Lander ES, Jaffe DB (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci* 108(4):1513–1518
32. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG (1996) Life with 6000 genes. *Science* 274(5287):546–567
33. Grainger DC, Hurd D, Goldberg MD, Busby SJW (2006) Association of nucleoid proteins with coding and non-coding segments of the *Escherichia coli* genome. *Nucleic Acids Res* 34(16):4642–4652
34. Grainger DC, Hurd D, Harrison M, Holdstock J, Busby SJW (2005) Studies of the distribution of *Escherichia coli* cAMP-receptor protein and RNA polymerase along the *E. coli* chromosome. *Proc Natl Acad Sci USA* 102(49):17693–17698
35. Güell M, van Noort V, Yus E, Chen W-H, Leigh-Bell J, Michalodimitrakis K, Yamada T, Arumugam M, Doerks T, Kühner S, Rode M, Suyama M, Schmidt S, Gavin A-C, Bork P, Serrano L (2009) Transcriptome complexity in a genome-reduced bacterium. *Science* 326(5957):1268–1271
36. Guelzim N, Bottani S, Bourguin P, Kepes F (2002) Topological and causal structure of the yeast transcriptional regulatory network. *Nat Genetics* 31(1):60–63
37. Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C, Rinn JL, Lander ES, Regev A (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol* 28(5):503–510
38. Harbison CT, Gordon B, Lee TII, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne J-BB, Reynolds DB, Yoo J, Jennings EG, Zeitlinger J, Pokholok DK, Kellis M, Rolfe PA, Takusagawa KT, Lander ES, Gifford DK, Fraenkel E, Young RA (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature* 431(7004):99–104
39. Harris SR, Feil EJ, Holden MT, Quail MA, Nickerson EK, Chantratita N, Gardete S, Tavares A, Day N, Lindsay JA, Edgeworth JD, de Lencastre H, Parkhill J, Peacock SJ, Bentley SD (2010) Evolution of MRSA during hospital transmission and intercontinental spread. *Science (New York, N.Y.)* 327(5964):469–474
40. Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, Braslavsky I, Causey M, Colonell J, DiMeo J, Efcavitch JW, Giladi E, Gill J, Healy J, Jarosz M, Lapen D, Moulton K, Quake SR, Steinmann K, Thayer E, Tyurina A, Ward R, Weiss H, Xie Z (2008) Single-molecule DNA sequencing of a viral genome. *Science* 320(5872):106–109
41. Hernandez D, François P, Farinelli L, Østerås M, Schrenzel J (2008) De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome Res* 18(5):802–809
42. He M, Sebahia M, Lawley TD, Stabler RA, Dawson LF, Martin MJ, Holt KE, Seth-Smith HMB, Quail MA, Rance R, Brooks K, Churcher C, Harris D, Bentley SD, Burrows C, Clark L, Corton C, Murray V, Rose G, Thurston S, van Tonder A, Walker D, Wren BW, Dougan G, Parkhill J (2010) Evolutionary dynamics of *clostridium difficile* over short and long time scales. *Proc Natl Acad Sci* 107(16):7527–7532
43. International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature* 431(7011):931–945
44. Ishihama A (2000) Functional modulation of *Escherichia coli* RNA polymerase. *Ann Rev Microbiol* 54:499–518
45. Iverson V, Morris RM, Frazar CD, Berthiaume CT, Morales RL, Armbrust EV (2012) Untangling genomes from metagenomes: revealing an uncultured class of marine euryarchaeota. *Science* 335(6068):587–590

46. Jeck WR, Reinhardt JA, Baltrus DA, Hickenbotham MT, Magrini V, Mardis ER, Dangel JL, Jones CD (2007) Extending assembly of short DNA sequences to handle error. *Bioinformatics* 23(21):2942–2944
47. Ji Hongkai, Jiang Hui, Ma Wenxiu, Johnson David S, Myers Richard M, Wong Wing H (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat Biotechnol* 26(11):1293–1300
48. Jothi R, Cuddapah S, Barski A, Cui K, Zhao K (2008) Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res* 36(16):5221–5231
49. Kahramanoglou C, Prieto AI, Khedkar S, Haase B, Gupta A, Benes V, Fraser GM, Luscombe NM, Seshasayee ASN (2012) Genomics of DNA cytosine methylation in *Escherichia coli* reveals its role in stationary phase transcription. *Nat Commun* 3:886+
50. Kahramanoglou C, Seshasayee ASN, Prieto AI, Ibberson D, Schmidt S, Zimmermann J, Benes V, Fraser GM, Luscombe NM (2011). Direct and indirect effects of H-NS and Fis on global gene expression control in *Escherichia coli*. *Nucleic Acids Res* 39(6):2073–2091
51. Köser CU, Holden MTG, Ellington MJ, Cartwright EJP, Brown NM, Ogilvy-Stuart AL, Hsu LY, Chewapreecha C, Croucher NJ, Harris SR, Sanders M, Enright MC, Dougan G, Bentley SD, Parkhill J, Fraser LJ, Betley JR, Schulz-Trieglaff OB, Smith GP, Peacock SJ (2012) Rapid Whole-genome sequencing for investigation of a neonatal MRSA outbreak. *N Engl J Med* 366(24):2267–2275
52. Lajoie BR, van Berkum NL, Sanyal A, Dekker J (2009) My5C: web tools for chromosome conformation capture studies. *Nat Methods* 6(10):690–691
53. Langmead B, Trapnell C, Pop M, Salzberg S (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10(3):R25–10
54. Lasa I, Toledo-Arana A, Dobin A, Villanueva M, de los Mozos IR, Vergara-Irigaray M, Segura V, Fagegaltier D, Penadés JR, Valle J, Solano C, Gingeras TR (2011) Genome-wide antisense transcription drives mRNA processing in bacteria. *Proc Nat Acad Sci* 108(50):20172–20177
55. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne J-B, Volkert TL, Fraenkel E, Gifford DK, Young RA (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* 298(5594):799–804
56. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* 18(11):1851–1858
57. Li R, Li Y, Kristiansen K, Wang J (2008) SOAP: short oligonucleotide alignment program. *Bioinformatics* 24(5):713–714
58. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14):1754–1760
59. Lieberman TD, Michel J-B, Aingaran M, Potter-Bynoe G, Roux D, Davis MR, Skurnik D, Leiby N, LiPuma JJ, Goldberg JB, McAdam AJ, Priebe GP, Kishony R (2011) Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. *Nat Genet* 43(12):1275–1280
60. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragozcy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326(5950):289–293
61. Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, Li S, Yang H, Wang J, Wang J (2010) De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res* 20(2):265–272
62. MacLean D, Jones JDG, Studholme DJ (2009) Application of 'next-generation' sequencing technologies to microbial genetics. *Nat Rev Microbiol* 7(4):287–296
63. Marcy Y, Ouverney C, Bik EM, Lösekann T, Ivanova N, Martin HG, Szeto E, Platt D, Hugenholtz P, Relman DA, Quake SR (2007) Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proc Nat Acad Sci* 104(29):11889–11894

64. Mardis ER (2008) The impact of next-generation sequencing technology on genetics. *Trends Genet (TIG)* 24(3):133–141
65. Martínez-Antonio A, Collado-Vides J (2003) Identifying global regulators in transcriptional regulatory networks in bacteria. *Curr Opin Microbiol* 6(5):482–489
66. McAdam PR, Templeton KE, Edwards GF, Holden MTG, Feil EJ, Aanensen DM, Bargawi HJA, Spratt BG, Bentley SD, Parkhill J, Enright MC, Holmes A, Girvan EK, Godfrey PA, Feldgarden M, Kearns AM, Rambaut A, Robinson DA, Fitzgerald JR (2012) Molecular tracing of the emergence, adaptation, and transmission of hospital-associated methicillin-resistant *Staphylococcus aureus*. *Proc Nat Acad Sci* 109(23):9107–9112
67. McCarthy AJ, Witney AA, Gould KA, Moodley A, Guardabassi L, Voss A, Denis O, Broens EM, Hinds J, Lindsay JA (2011) The distribution of mobile genetic elements (MGEs) in MRSA CC398 is associated with both host and country. *Genome Biol Evol* 3:1164–1174
68. McCarthy AJ, Breathnach AS, Lindsay JA (2012) Detection of mobile-genetic-element variation between colonizing and infecting hospital-associated methicillin-resistant *Staphylococcus aureus* isolates. *J Clin Microbiol* 50(3):1073–1075
69. McGuire J et al (2009) Genome 10K community of scientists. Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. *J Hered* 100(6):659–674
70. Militelio KT, Simon RD, Qureshi M, Maines R, Horne ML, Hennick SM, Jayakar SK, Pounder S (2012) Conservation of Dcm-mediated cytosine DNA methylation in *Escherichia coli*. *FEMS Microbiol Lett* 328(1):78–85
71. Mraheil MA, Billion A, Mohamed W, Mukherjee K, Kuenne C, Pischmarov J, Krawitz C, Retey J, Hartsch T, Chakraborty T, Hain T (2011) The intracellular sRNA transcriptome of *Listeria monocytogenes* during growth in macrophages. *Nucleic Acids Res* 39(10):4235–4248
72. Mutreja A, Kim DWW, Thomson NR, Connor TR, Lee JHH, Kariuki S, Croucher NJ, Choi SYY, Harris SR, Lebens M, Niyogi SKK, Kim EJJ, Ramamurthy T, Chun J, Wood JL, Clemens JD, Czerkinsky C, Nair GB, Holmgren J, Parkhill J, Dougan G (2011) Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature* 477(7365):462–465
73. Narain AS, Seshasayee S (2011) An overview of prokaryotic transcription factors: a summary of function and occurrence in bacterial genomes. *Subcell Biochem* 52:7–23
74. Nicolas P, Mäder U, Dervyn E, Rochat T, Leduc A, Pigeonneau N, Bidnenko E, Marchadier E, Hoebeke M, Aymerich S, Becher D, Bisicchia P, Botella E, Delumeau O, Doherty G, Denham EL, Fogg MJ, Fromion V, Goelzer A, Hansen A, Härtig E, Harwood CR, Homuth G, Jarmer H, Jules M, Klipp E, Le Chat L, Lecointe F, Lewis P, Liebermeister W, March A, Mars RAT, Nannapaneni P, Noone D, Pohl S, Rinn B, Rügheimer F, Sappa PK, Samson F, Schaffer M, Schwikowski B, Steil L, Stülke J, Wiegert T, Devine KM, Wilkinson AJ, van Dijl JM, Hecker M, Völker U, Bessières P, Noirot P (2012) Condition-dependent transcriptome reveals high-level regulatory architecture in *Bacillus subtilis*. *Science* 335(6072):1103–1106
75. Ning Z, Cox AJ, Mullikin JC (2001) SSAHA: a fast search method for large DNA databases. *Genome Res* 11(10):1725–1729
76. Oshima T, Ishikawa S, Kurokawa K, Aiba H, Ogasawara N (2006) *Escherichia coli* histone-like protein H-NS preferentially binds to horizontally acquired DNA in association with RNA polymerase. *DNA Res Int J Rapid Publ Rep Genes Genomes* 13(4):141–153
77. Park PJ (2009) ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genetics* 10(10):669–680
78. Patten CL, Kirchhof MG, Schertzberg MR, Morton RA, Schellhorn HE (2004) Microarray analysis of RpoS-mediated gene expression in *Escherichia coli* K-12. *Mol Genet Genomics* 272(5):580–591
79. Perkins TT, Kingsley RA, Fookes MC, Gardner PP, James KD, Yu L, Assefa SA, He M, Croucher NJ, Pickard DJ, Maskell DJ, Parkhill J, Choudhary J, Thomson NR, Dougan G (2009) A strand-specific RNA-seq analysis of the transcriptome of the typhoid bacillus salmonella typhi. *PLoS Genet*, 5(7):e1000569+
80. Peter BJ, Arsuaga J, Breier AM, Khodursky AB, Brown PO, Cozzarelli NR (2004) Genomic transcriptional response to loss of chromosomal supercoiling in *Escherichia coli*. *Genome Biol* 5(11):R87+

81. Prabhakara S, Khedkar S, Loganathan RM, Chandana S, Gowda M, Arakere G, Seshasayee ASN (2012) Draft genome sequence of *Staphylococcus aureus* 118 (ST772), a major disease clone from India. *J Bacteriol* 194(14):3727–3728
82. Prieto AI, Kahramanoglou C, Ali RM, Fraser GM, Seshasayee ASN, Luscombe NM (2011) Genomic analysis of DNA binding and gene regulation by homologous nucleoid-associated proteins IHF and HU in *Escherichia coli* K12. *Nucleic Acids Res*
83. Qin J, Li R, Raes J, Arumugam M, Burgdorf KSS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR, Li J, Xu J, Li S, Li D, Cao J, Wang B, Liang H, Zheng H, Xie Y, Tap J, Lepage P, Bertalan M, Batto J-MM, Hansen T, Le Paslier D, Linneberg A, Nielsen HB, Pelletier E, Renault P, Sicheritz-Ponten T, Turner K, Zhu H, Yu C, Li S, Jian M, Zhou Y, Li Y, Zhang X, Li S, Qin N, Yang H, Wang J, Brunak S, Doré J, Guarner F, Kristiansen K, Pedersen O, Parkhill J, Weissenbach J, Consortium M, Bork P, Ehrlich SD, Wang J (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464(7285):59–65
84. Reid J (2010) A map of human genome variation from population-scale sequencing. *Nature* 467(7319):1061–1073
85. Robbins-Manke JL, Zdraveski ZZ, Marinus M, Essigmann JM (2005) Analysis of global gene expression and double-strand-break formation in DNA adenine methyltransferase- and mismatch repair-deficient *Escherichia coli*. *J Bacteriol* 187(20):7027–7037
86. Robinson MD, Oshlack A (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* 11(3):R25+
87. Round JL, Mazmanian SK (2009) The gut microbiota shapes intestinal immune responses during health and disease. *Nat Rev Immunol* 9(5):313–323
88. Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, Carriero N, Snyder M, Gerstein MB (2009) PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat Biotechnol* 27(1):66–75
89. Salgado H, Gama-Castro S, Peralta-Gil M, Díaz-Peredo E, Sánchez-Solano F, Santos-Zavaleta A, Martínez-Flores I, Jiménez-Jacinto V, Bonavides-Martínez C, Segura-Salazar J, Martínez-Antonio A, Collado-Vides J (2006) RegulonDB (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res* 34(suppl 1):D394–D397
90. Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proc Nat Acad Sci USA* 74(12):5463–5467
91. Schuster SC (2008) Next-generation sequencing transforms today's biology. *Nat Methods* 5(1):16–18
92. Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeisz S, Sittka A, Chabas S, Reiche K, Hackermuller J, Reinhardt R, Stadler PF, Vogel J (2010) The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature* 464(7286):250–255
93. Shendure J, Ji H (2008) Next-generation DNA sequencing. *Nat Biotechnol* 26(10):1135–1145
94. Shi Y, Tyson GW, DeLong EF (2009) Metatranscriptomics reveals unique microbial small RNAs in the ocean's water column. *Nature* 459(7244):266–269
95. Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, de Wit E, van Steensel B, de Laat W (2006) Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat Genet* 38(11):1348–1354
96. Sittka A, Lucchini S, Papenfort K, Sharma CM, Rolle K, Binnewies TT, Hinton JCD, Vogel J (2008) Deep sequencing analysis of small noncoding RNA and mRNA targets of the global post-transcriptional regulator, Hfq. *PLoS Genet* 4(8):e1000163+
97. Smith MG, Gianoulis TA, Pukatzki S, Mekalanos JJ, Ornston LN, Gerstein M, Snyder M (2007) New insights into *Acinetobacter baumannii* pathogenesis revealed by high-density pyrosequencing and transposon mutagenesis. *Genes Dev* 21(5):601–614
98. Spyrou C, Stark R, Lynch AG, Tavaré S (2009) BayesPeak: bayesian analysis of ChIP-seq data. *BMC Bioinf* 10(1):299+
99. Storz G, Vogel J, Wassarman KM (2011) Regulation by small RNAs in bacteria: expanding frontiers. *Mol Cell* 43(6):880–891

100. Tanay A, Sharan R, Kupiec M, Shamir R (2004) Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc Nat Acad Sci USA* 101(9):2981–2986
101. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, Deboy RT, Davidsen TM, Mora M, Scarselli M, y Ros IM, Peterson JD, Hauser CR, Sundaram JP, Nelson WC, Madupu R, Brinkac LM, Dodson RJ, Rosovitz MJ, Sullivan SA, Daugherty SC, Haft DH, Selengut J, Gwinn ML, Zhou L, Zafar N, Khouri H, Radune D, Dimitrov G, Watkins K, O'Connor KJ, Smith S, Utterback TR, White O, Rubens CE, Grandi G, Madoff LC, Kasper DL, Telford JL, Wessels MR, Rappuoli R, Fraser CM (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial “pan-genome”. *Proc Nat Acad Sci USA* 102(39):13950–13955
102. Toprak E, Veres A, Michel J-B, Chait R, Hartl DL, Kishony R (2012) Evolutionary paths to antibiotic resistance under dynamically sustained drug selection. *Nat Genet* 44(1):101–105
103. Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-seq. *Bioinformatics* 25(9):1105–1111
104. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L (2010) Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat biotechnol* 28(5):511–515
105. Umbarger MA, Toro E, Wright MA, Porreca GJ, Baù D, Hong S-HH, Fero MJ, Zhu LJ, Marti-Renom MA, McAdams HH, Shapiro L, Dekker J, Church GM (2011) The three-dimensional architecture of a bacterial genome and its alteration by genetic perturbation. *Mol Cell* 44(2):252–264
106. Venter CJ, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers Y-H, Smith HO (2004) Environmental genome shotgun sequencing of the Sargasso sea. *Science* 304(5667):66–74
107. Voelkerding KV, Dames SA, Durtschi JD (2009) Next-generation sequencing: from basic research to diagnostics. *Clin Chem* 55(4):641–658
108. Vora T, Hottes AK, Tavazoie S (2009) Protein occupancy landscape of a bacterial genome. *Mol Cell* 35(2):247–253
109. Wade JT, Roa DC, Grainger DC, Hurd D, Busby SJW, Struhl K, Nudler E (2006) Extensive functional overlap between σ factors in *Escherichia coli*. *Nat Struct Mol Biol* 13(9):806–814
110. Waldminghaus T, Skarstad K (2010) ChIP on chip: surprising results are often artifacts. *BMC Genomics* 11(1):414+
111. Warnecke F, Luginbuhl P, Ivanova N, Ghassemian M, Richardson TH, Stege JT, Cayouette M, McHardy AC, Djordjevic G, Aboushadi N, Sorek R, Tringe SG, Podar M, Martin HG, Kunin V, Dalevi D, Madejska J, Kirton E, Platt D, Szeto E, Salamov A, Barry K, Mikhailova N, Kyrpides NC, Matson EG, Ottesen EA, Zhang X, Hernandez M, Murillo C, Acosta LG, Rigoutsos I, Tamayo G, Green BD, Chang C, Rubin EM, Mathur EJ, Robertson DE, Hugenholtz P, Leadbetter JR (2007) Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature* 450(7169):560–565
112. Warren RL, Sutton GG, Jones SJ, Holt RA (2007) Assembling millions of short DNA sequences using SSAKE. *Bioinformatics* (Oxford, England) 23(4):500–501
113. Weber H, Polen T, Heuveling J, Wendisch VF, Hengge R (2005) Genome-wide analysis of the general stress response network in *Escherichia coli*: sigma S-dependent genes, promoters, and sigma factor selectivity. *J Bacteriol* 187(5):1591–1603
114. Wielgoss S, Barrick JE, Tenaillon O, Cruveiller S, Chane-Woon-Ming B, Médigue C, Lenski RE, Schneider D (2011) Mutation rate inferred from synonymous substitutions in a long-term evolution experiment with *Escherichia coli*. *G3: Genes Genomes Genet* 1(3):183–186
115. Willenbrock H, Hallin P, Wassenaar T, Ussery D (2007) Characterization of probiotic *Escherichia coli* isolates with a novel pan-genome microarray. *Genome Biol* 8(12):R267+
116. Woyke T, Xie G, Copeland A, González JM, Han C, Kiss H, Saw JH, Senin P, Yang C, Chatterji S, Cheng J-F, Eisen JA, Sieracki ME, Stepanauskas R (2009) Assembling the marine metagenome, one cell at a time. *PLoS ONE* 4(4):e5299+

117. Yaffe E, Tanay A (2011) Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat Genet* 43(11):1059–1065
118. Yang L, Jelsbak L, Marvig RL, Damkiær S, Workman CT, Rau MH, Hansen SK, Folkesson A, Johansen HK, Ciofu O, Høiby N, Sommer MOA, Molin S (2011) Evolutionary dynamics of bacteria in a human host environment. *Proc Nat Acad Sci* 108(18):7481–7486
119. Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 18(5):821–829
120. Zhang Q, Lambert G, Liao D, Kim H, Robin K, Tung C, Pourmand N, Austin RH (2011) Acceleration of emergence of bacterial antibiotic resistance in connected microenvironments. *Science* 333(6050):1764–1767
121. Zhang Y, Liu T, Meyer C, Eeckhoutte J, Johnson D, Bernstein B, Nusbaum C, Myers R, Brown M, Li W, Liu XS (2008) Model-based Analysis of ChIP-Seq (MACS). *Genome Biol* 9(9):R137+

A Systems Theoretic Approach to Systems and
Synthetic Biology II: Analysis and Design of Cellular
Systems

Kulkarni, V.; Stan, G.-B.; Raman, K. (Eds.)

2014, XVIII, 298 p. 107 illus., 77 illus. in color.,

Hardcover

ISBN: 978-94-017-9046-8