1 Introduction: Bacterial Genomes and Gene Expression

Bacteria are the most dominant form of free-living life on Earth, and represent a major part of its genetic diversity. Some bacteria are closely adapted to a single environment in which they reside, whereas many others are capable of thriving across multiple environments. They range from the most benign inhabitants of the Earth to being deadly, multi-drug-resistant human pathogens. The success of bacterial life is exemplified by the wide variation in their genetic content, not just across phyla, but even among members of the same species. In addition is the plethora of gene regulatory mechanisms that—to state a most cliched phrase—ensure that appropriate genes are expressed when required. These two features of bacterial biology form the crux of this book.

It will not be wrong to state that the advent of genomics has considerably advanced our knowledge of both evolutionary and gene regulatory mechanisms in bacteria. The ambitious—and now historical—genome sequencing projects, driven by whole genome shotgun sequencing using automated Sanger sequencers, taught us many things about bacterial genetic diversity and the mechanisms underlying its generation. For example, by sequencing the genomes of many members of a species, such as *Escherichia coli*, we learnt that some bacteria have what is called an *open pan genome*: every new genome of a member of that species will invariably identify many genes hitherto unknown in that species. Genome sequencing projects also stoked a controversy on the importance of horizontal gene transfer to genome growth and bacterial evolution itself: in the face of rampant horizontal transfer, does the concept of a bacterial species have any meaning? At the other end of the scale, genome reduction emerged as an important phenomenon underlying the evolution of obligate parasites, including those of major human pathogens such as the *Rickettsia* and *Mycobacterium leprae*.

As databases of genome sequences grew exponentially in size, high-density microarrays, which use hybridisation-based techniques to probe nucleic acid content on a genomic scale, came into vogue. Though these were best known for

their application to the measurement of gene expression on a genome-wide scale, these were also used extensively to probe the genetic content of large numbers of bacterial isolates for which the complete genome sequence of a close relative was available. Such studies, besides cataloguing gene signatures of specific isolates of a pathogenic species, also enabled studies of the dynamics of gene gain and loss. In the more classical application of studying gene expression, microarrays—either from a single study or through metaanalyses of data from multiple studies—enabled the reconstruction of gene regulatory networks of model bacteria, revealing a much under-appreciated complexity in the transcriptional response of even some so-called simple bacteria.

Arguably, the most dramatic advance in genomics came about recently with the advent of what are known as 'next-generation' or 'deep' sequencing technologies. These dramatically reduced the cost and the time required for sequencing complete genomes, thus bringing genome sequencing out of the confines of large sequencing centres to the benchtop of common laboratories. For example, the sequence of the *E. coli* strain, which was responsible for the break-out of food poisoning in Europe in 2011, was fully obtained on a benchtop sequencer within weeks after the outbreak, in stark contrast to the years of meticulous work that sequencing by the Sanger method required. Deep sequencing technologies have permitted high-resolution phylogenetic analysis of bacterial pathogens—including the tracking of epidemics—at times within clinically-relevant timescales. For the basic sciences researcher, these techniques have enabled rapid identification of genetic variants associated with a phenotype of interest. Genome sequencing, both classical and next-generation, have also spawned and advanced the field of metagenomics, which pertains to the genetic characterisation of entire bacterial communities in a culture-independent manner. A particular contribution of the next-generation sequencing approaches to this field has been in allowing us to obtain the complete or near-complete genome sequence of a single bacterial species from within a complex community of meta-genomes. One can even isolate a single bacterial cell from a consortium of bacteria and sequence its genome to a near-complete stage.

Finally, the great depth of coverage that many next-generation sequencers offer also makes them quantitative, allowing us to measure the extent to which a particular genetic variant has spread in a population, as well as pursue genome-scale assays of gene expression and protein–nucleic acid interactions, previously performed using microarrays. In contrast to microarray technologies, deep sequencing can provide base-level resolution in defining transcripts and protein-binding regions on DNA or RNA. This has considerably enhanced research into experimental annotation of transcripts and regulatory regions in genomes, an approach that has used both microarrays and deep sequencing technologies. Single-molecule real-time sequencing technologies also help identify base modifications in a standard sequencing experiment, thus significantly advancing our ability to describe epigenetic modifications.

When it comes to next-generation or deep sequencing, the sky is the limit. Virtually any molecular technique that uses sequence-level information about a few genes can be ported to a genomic scale using these techniques. For example, one can rapidly quantify locus-dependent rates of transposon insertion on a genomic scale using deep sequencing techniques called TraDIS and InSEQ, among others.

The objective of this text is to be a celebration of the application of genomics to the study of genome architecture and gene expression in bacteria. The book takes the approach of introducing concepts underlying the generation and analysis of multiple types of genome-scale data, followed by the presentation of a few papers which have either pioneered the application of a technology to a particular problem, and/or described interesting biology using a combination of genomic techniques. In the context of data analysis, we have deliberately stayed away from providing tables of software that can be applied to a particular type of data. While mentioning a few popular software in the main text, we have emphasised more on the fundamental principles that underly some of these software. Lists of relevant software are generally available online, as well as in the reviews that we cite in this text. Even in a relatively young field like genomics, it is impossible to be comprehensive in a book of this size. Therefore, the selection of concepts and case studies is merely a reflection of my own biases, interests and familiarity towards a subset of the fast-growing literature in the field. In fact, there are whole areas of genomics that this book does not touch upon—proteomics and metabolomics, for example.

Technology moves forward faster than many of us can cope with; therefore, the power of the Internet. However, I do believe that many ideas discussed in this text, published in the traditional format, whether a part of history or contemporary, will remain relevant for a long time to come.

2 Comparative Genomics in the Era of Sanger Sequencing

2.1 Introduction

The first genome to be completely sequenced,[1] in 1976, was that of an RNA bacteriophage MS2, which encoded only three genes. Though the concept of comparative genomics had yet to emerge, the authors mentioned the possibility of understanding viral evolution by performing sequence searches against other viral genomes that were forthcoming. This was rapidly followed by the sequencing–by Sanger and colleagues–of the 5.4 kb-long genomic DNA of phage ΦX174,[2] and later in 1982, considerably larger sequence (>48 kb) of the genome of the molecular biology workhorse, bacteriophage lambda (λ).[3] As pointed out by Koonin and Galperin,[4] neither work presented the concept of sequence comparison and homology, which is now routine in any genome analysis pipeline. This may be surprising as the PIR protein sequence database[5] was already available, and the first sequence substitution matrix[6] constructed. However, sequence databases

[1] The sequence of the third and final gene was reported in Fiers W., Contreras R., Duerinck F., Haegeman G., Iserentant D., Merregaert J., Min Jou W., Molemans F., Raeymaekers A., Van den Berghe A., Volckaert G. and Ysebaert M. 1976. 'Complete nucleotide sequence of bacteriophage MS2 RNA: Primary and secondary structure of the replicase gene.' *Nature* 260: 500–07.

[2] Sanger F., Coulson A. R., Friedmann T., Air G. M., Barrell B. G., Brown N. L., Fiddes J. C., Hutchison C. A. 3rd, Slocombe P. M. and Smith M. 1977. 'Nucleotide sequence of bacteriophage ΦX174 DNA.' *Nature* 265: 687–95.

[3] Sanger F, Coulson A. R., Hong G. F., Hill D. F. and Petersen G. B. 'Nucleotide sequence of bacteriophage lambda DNA.' *Journal of Molecular Biology* 162: 729–73.

[4] Koonin and Galperin. 2003. *Sequence-Evolution-Function: Computational Approaches in Comparative Genomics.* Kluwer Academic Press, Boston.

[5] http://pir.georgetown.edu/

[6] Dayhoff, Schwartz and Orcutt. 1978. 'A model for evolutionary change in proteins.' *Atlas of Protein Sequence and Structure* 5: 345–52.

were not sufficiently rich in information at that time, and sophisticated and rapid methods for searching sequence databases were not common, with the Smith–Waterman algorithm a then-recent innovation,[7] and the much faster BLAST not due to be reported for another eight years.[8]

The first tentative comparative genomics work, as identified by Koonin and Galperin, was published by Toh, Hayashida and Miyata in 1983.[9] This remarkable short paper identified homologues of retroviral reverse transcriptases encoded by the genome of two DNA viruses, whose life cycle was known to involve reverse transcription. In another study, Argos and coworkers[10] showed striking similarities in certain gene sequences between animal picornaviruses and plant cowpea mosaic virus, suggesting evolutionary relatedness either by independent abstraction of common host genes, or via vertical descent from a common ancestor. Finally, McGeoch and Davison[11] worked on considerably larger genomes of the varicella zoster virus (an α-herpesvirus), and the Epstein–Barr virus (a distantly related γ-herpesvirus), and found several homologues between them.

Despite several reports describing and comparing viral genomes, genomics of cellular organisms did not come to fruition till the publication of the genome of the human pathogenic bacterium *Haemophilus influenzae* in 1995,[12] almost 20 years after the completion of the genome of the bacteriophage MS2. Today, we have access to over 3,000 fully-sequenced bacterial genomes and many more draft genomes.

---

7 Smith and Waterman. 1981. 'Identification of common molecular subsequences.' *Journal of Molecular Biology* 147: 195–97.

8 Altschul S. F., Gish W., Miller W., Myers E. W. and Lipman D. J. 1990. 'Basic local alignment search tool.' *Journal of Molecular Biology* 215: 403–10.

9 Toh, Hayashida and Miyata. 1983. 'Sequence homology between retroviral reverse transcriptase and putative polymerases of hepatitis B virus and cauliflower mosaic virus.' *Nature* 305: 827–29.

10 Argos P., Kamer G., Nicklin M. J. and Wimmer E. 1984. 'Similarity in gene organization and homology between proteins of animal picornaviruses and a plant comovirus suggest common ancestry of these virus families.' *Nucleic Acids Research* 12: 7251–7267.

11 McGeoch and Davison. 1986. 'DNA sequence of the herpes simplex virus type 1 gene encoding glycoprotein gH, and identification of homologues in the genomes of varicella-zoster virus and Epstein-Barr virus.' *Nucleic Acids Research* 14: 4281–4292.

12 Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J., Dougherty, B. A., Merrick, J. M., McKenney, K., Sutton, G. G., FitzHugh, W., Fields, C. A., Gocayne, J. D., Scott, J. D., Shirley, R., Liu, L. I., Glodek, A., Kelley, J. M., Weidman, J. F., Phillips, C. A., Spriggs, T., Hedblom, E., Cotton, M. D., Utterback, T., Hanna, M. C., Nguyen, D. T., Saudek, D. M., Brandon, R. C., Fine, L. D., Fritchman, J. L., Fuhrmann, J. L., Geoghagen, N. S., Gnehm, C. L., McDonald, L. A., Small, K. V., Fraser, C. M., Smith, H. O. and Venter, J. C. 1995. 'Whoge-genome random sequencing and assembly of Haemophilus influenzae Rd.' *Science* 269: 496–98 + 507–12.

In this chapter, we provide a conceptual overview of the process of bacterial genome sequencing and annotation. This will be followed by a description of selected bacterial genomes and the lessons learnt from them. We will close the chapter by discussing some general trends and concepts that have emerged from large-scale comparative genomics of bacterial (and where appropriate, archaeal and eukaryotic) genomes.

2.2 The process of assembling and annotating bacterial genomes

DNA sequencing produces 'reads'–sequences of fragments of DNA of orders of magnitude shorter than the genome itself. A typical 500 nt read from a Sanger sequencing experiment is ~10^4-fold smaller than the >4 Mb genome of the model laboratory strain of *Escherichia coli*. Therefore, novel DNA fragmentation strategies had to be developed so that sequencing reads produced from such fragments could be *assembled* into a complete chromosome(s). The most straightforward procedure for generating fragments, without recourse to any genetic or physical mapping, is random shearing of the genomic DNA, called the whole genome shotgun method (Fig. 2.1; top). It follows from the theory of Lander and Waterman[13] that, given a sufficiently random fragment set, the probability that any nucleotide position in the genome is covered by k reads is approximated by a Poisson distribution, with deviations caused by biological factors such as clone lethality. On the basis of these calculations, for a sequencing coverage of m, the probability that a base is not sequenced is given by $P = e^{-m}$; this number equals 0.0067 for $m = 5$, and 0.000045 for $m = 10$.

Despite the simplicity and force of this theoretical argument, there were justifiable fears that the scale of data produced might be too complex to handle during assembly. These fears were laid to rest by the publication of the genome of *H. influenzae*, a genome obtained *de novo* without any reference maps. Besides establishing the validity of the whole genome shotgun procedure, the work also laid down a data analysis pipeline, which includes read assembly, gap closure, gene finding and gene annotation, in that order. We will now briefly look at each of these steps (Fig. 2.1).

[13] Lander and Waterman. 1988. 'Genomic mapping by fingerprinting random clones: A mathematical analysis.' *Genomics* 2: 231–39. Described in simple terms in the *H. influenzae* genome paper.

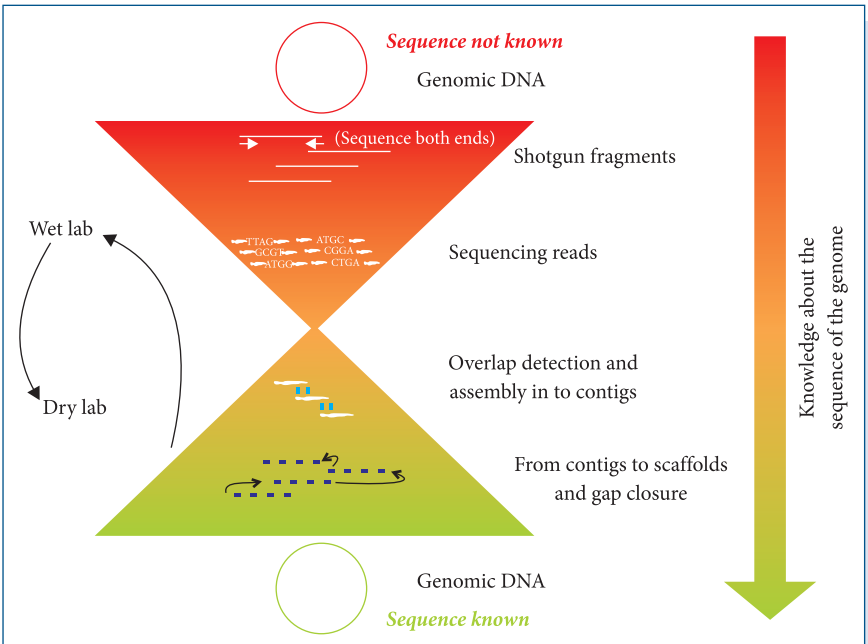

Figure 2.1 *A schematic overview of the process of sequencing a bacterial genome.* The top cone represents the fragmentation of genomic DNA molecules to produce *shotgun fragments*, the ends of which are sequenced. Typically, both ends of each fragment are sequenced, producing what are known as *paired-end* reads. The bottom cone shows the computation-intensive part of a genome sequencing experiment where sequence reads are assembled to produce larger *contigs* (and *scaffolds*, in which adjacent contigs are linked together); at this stage, there are un-resolved gaps between contigs/scaffolds, and the linear order of contigs is not necessarily known. A combination of computation and laboratory experiments is then used to stitch contigs together in a process called *gap closure*. The unique advantage of paired-end reads during the assembly of reads into a genome is mentioned in Chapter 4. The width of the cone is also indicative (not-to-scale) of the size of the average DNA molecule being analysed at any given stage of the experiment.

2.2.1 Genome assembly and gap closure

In a typical genome sequencing effort, either ends (*paired-end sequencing*) of DNA fragments of length ~2 kb are sequenced to produce *reads*, which in Sanger sequencing are typically ~500 nt long; for example, sequencing the genome of *H. influenzae* produced ~24,000 fragments of an average length of ~450 nt. The first step, after the production of these DNA sequence reads, is to assemble the reads together into longer fragments called *contigs* and *scaffolds*, and eventually

into a complete genome.[14] This is made possible by the fact that the shearing of the genomic DNA is random and that DNA sequencing produces multiple reads covering each nucleotide. These mean that pairwise sequence alignments could be made between reads, and read pairs with a significant sequence overlap at the ends be stitched together. In a greedy algorithm, this process could be iterated until all sequences are merged into a single long sequence, or at least into as few sequences are possible. However, false positive and false negative overlaps are possible and should be eliminated. In a false positive identification, an overlap could be small enough to be caused by mere random chance, or due to errors inherent to the sequencing technology. False negatives emerge when a true overlap is too short not to be called a false positive, or when the sequencing error is so high in a real overlap region that a merge cannot be made. Therefore, it is essential to exercise caution in identifying criteria for calling overlaps and merging reads.

The assembled fragments produced thus are called *contigs*. The ~24,000 reads produced by the *H. influenzae* genome sequencing project were assembled into 140 contigs. However, further sequencing reactions using longer DNA fragments[15] could be used to make relationships between contigs and arrange them into *scaffolds*. While assembling the *H. influenzae* genome, the authors identified DNA fragments whose forward and reverse sequences (as determined by paired-end sequencing) were deemed to be present in different post-assembly contigs; such contig pairs were linked together and eventually, the 140 contigs were organised into 42 groups separated by gaps. These gaps could be finally closed using PCR reactions primed by sequences from the edges of each pair of scaffolds, followed by the sequencing of appropriate amplicons. These were helped by analyses such as searches against protein databases, wherein two contigs that find a match against two different portions of a single protein sequence could be tentatively deemed to lie adjacent to each other. Finally, gaps that could not be closed because of lethality resulting from the cloning of certain DNA fragments in the host *Escherichia coli* could be sequenced using a phage library; in fact, 23 of the 42 physical gaps in the *H. influenzae* genome were closed using this approach.

2.2.2 Genome-scale computational identification of features

Once the complete genome sequence of an organism is obtained, the next step is to identify various functional features. First, the genomic G+C content is described. This simple measure is useful as it could be an overall indicator of the various mutational pressures operating on a genome.[16] It also allows a first-pass identification of genes with a different evolutionary history from the rest

[14] Sutton and Dew. 2006. 'Shotgun fragment assembly.' *Systems Biology Volume 1: Genomics.* Oxford University Press, USA.

[15] See reference to mate-pair library sequencing in Chapter 4.

[16] See discussion on genome reduction later in this Chapter.

of the genome, as many of these lie in windows whose G+C content is different from the genomic average (Fig. 2.2 a). Similarly, ribosomal operons, which are highly conserved, also tend to lie within regions of high G+C content. Patterns of G/C usage are also useful in identifying the origin of replication. Since the leading and the lagging strand are replicated differently, and may therefore come under different mutational regimes, they have different G/C skews as calculated by G−C/G+C. This is called strand asymmetry. Any strand of DNA will be lagging on one side of the origin and leading on the other; therefore, a position around which there is a strong switch in the G/C-skew value is a candidate for being the origin of replication (Fig. 2.2 b). This measure is usually supplemented

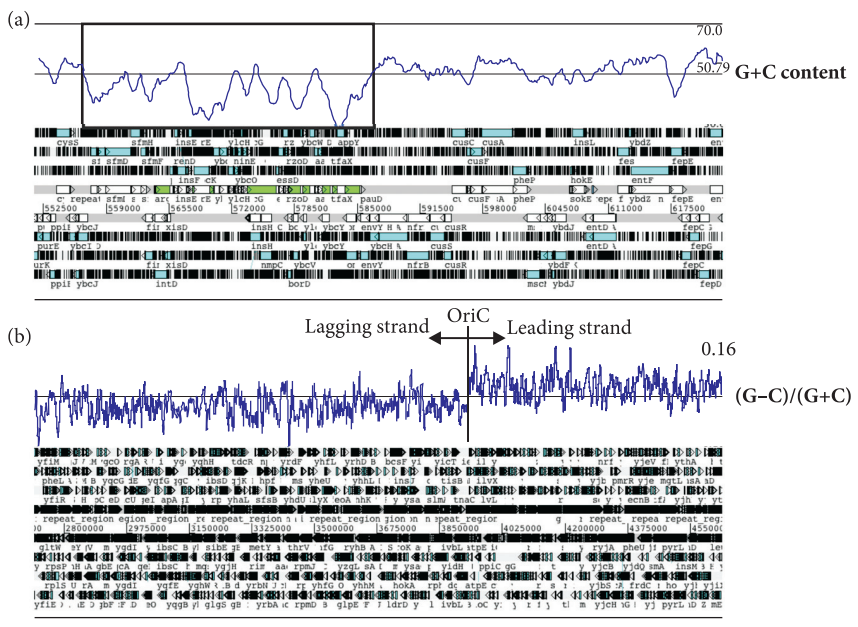

Figure 2.2 *Base composition patterns.* (a) The G+C content over short windows for a segment of the genome of *Escherichia coli* K12 is shown. The boxed region is one where the G+C content is considerably less than the average, which is marked by the thin black horizontal line as well as by the blue curve to the right of the box. This region of abnormally low A+T content is a horizontally-acquired piece of DNA, from a defective prophage called DLP-12. (b) The G/C-skew across a portion of the *E. coli* K12 genome is shown. The switch in the skew from negative (lagging strand) to positive (leading strand) at the origin of replication is marked.

by other predictors such as the presence of clusters of the binding site for the DNA replication initiator protein DnaA. The terminus is usually bipartite with

the two parts bisected by a position that is diametrically opposite to the origin of replication on the circular chromosome.

Next, highly conserved genes such as the ribosomal and transfer RNAs can typically be identified by homology to known examples.

Today, though a large number of protein coding sequences can be identified by homology to members of rich sequence databases, the possibility of missing novel proteins still remains large. Therefore, homology-independent, but pattern-dependent statistical methods are first used to identify regions that might potentially encode proteins, before these are annotated with putative functions, typically by homology. A simple pattern defining a protein-coding region may be any 'long' stretch of sequence between a START and a STOP codon, where 'long' may be defined by say 600 nt corresponding to 200 amino acids. This method can be expected to be highly specific, with nearly every identified candidate being a real protein-coding gene; however, it will suffer from low sensitivity as it will miss a large number of smaller protein-coding genes. Therefore, there is a need for more sophisticated probabilistic methods for gene identification, which are aptly described by Azad and Borodovsky[17] as follows: "The first step in developing an ab-initio gene-finding algorithm is to perform statistical analysis of DNA sequences of interest (protein-coding and non-coding) and to identify statistical determinants, such as in-frame frequencies of oligonucleotides, that help recognise sequences of these two types. The second step is to build statistical models, such as Markov models for all sequence categories, particularly gene models. The third step is to integrate the models into a pattern recognition algorithm". To elaborate, one could use high-confidence sequences identified by the previously described procedure of choosing long protein-coding regions, or using homology searches, to define properties of protein-coding regions. This property could, for example, be simply G+C content; in fact, in the genome of *E. coli*, protein-coding regions typically have a higher G+C content than intergenic regions. However, this property is not sufficiently discriminative. A statistical model that can lead to specific identification of codon-like sequences might be more suitable for discriminating protein-coding from other sequences. In this respect, a base composition profile that is calculated for each position within a codon is a sensible predictor; for example, in typical *E. coli* genes, the base 'G' is twice as common at the first base of a codon as it is in the second base. However, such position dependence is not seen in intergenic regions. These occurrence profiles could be extended to include immediate sequence context in a Markov model,[18]

[17] Azad and Borodovsky. 2004. 'Probabilistic methods of identifying genes in prokaryotic genomes: Connections to the HMM theory.' *Briefings in Bioinformatics* 5: 118–30.

[18] An m^{th} order Markov model in the context of a DNA sequence states that the probability of finding a particular base at position i is dependent only on m-preceding bases; a zeroth order Markov model is nothing beyond base composition, and a first order