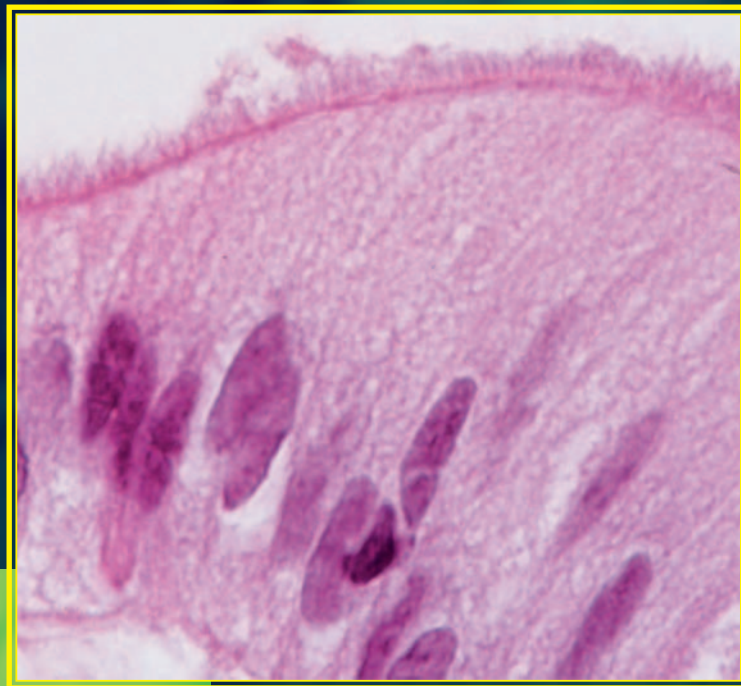


# FUNCTIONAL & INTEGRATIVE GENOMICS

THE journal for:

- Whole genome analysis/bioinformatics
- Comparative/functional genomics
- Expression profiling/integration with phenotype



Tim T. Binnewies · Yair Motro · Peter F. Hallin ·  
Ole Lund · David Dunn · Tom La · David J. Hampson ·  
Matthew Bellgard · Trudy M. Wassenaar ·  
David W. Ussery

## Ten years of bacterial genome sequencing: comparative-genomics-based discoveries

Received: 20 January 2006 / Revised: 24 February 2006 / Accepted: 7 March 2006  
© Springer-Verlag 2006

**Abstract** It has been more than 10 years since the first bacterial genome sequence was published. Hundreds of bacterial genome sequences are now available for comparative genomics, and searching a given protein against more than a thousand genomes will soon be possible. The subject of this review will address a relatively straightforward question: “What have we learned from this vast amount of new genomic data?” Perhaps one of the most important lessons has been that genetic diversity, at the level of large-scale variation amongst even genomes of the same species, is far greater than was thought. The classical textbook view of evolution relying on the relatively slow accumulation of mutational events at the level of individual bases scattered throughout the genome has changed. One of the most obvious conclusions from examining the sequences from several hundred bacterial genomes is the enormous amount of diversity—even in different genomes from the same bacterial species. This diversity is generated by a variety of mechanisms, including mobile genetic elements and bacteriophages. An examination of the 20 *Escherichia coli* genomes sequenced so far dramatically illustrates this, with the genome size ranging from 4.6 to 5.5 Mbp; much of the variation appears to be of phage origin. This review also addresses mobile genetic elements,

including pathogenicity islands and the structure of transposable elements. There are at least 20 different methods available to compare bacterial genomes. Metagenomics offers the chance to study genomic sequences found in ecosystems, including genomes of species that are difficult to culture. It has become clear that a genome sequence represents more than just a collection of gene sequences for an organism and that information concerning the environment and growth conditions for the organism are important for interpretation of the genomic data. The newly proposed Minimal Information about a Genome Sequence standard has been developed to obtain this information.

**Keywords** Bacterial genomics · Comparative genomics · Bioinformatics · Genomic diversity · Molecular evolution

### Introduction

The year 1995 marked the publication of two human pathogenic bacterial genome sequences: *Haemophilus influenzae* (Fleischmann et al. 1995, US patent number 6,528,289) and *Mycoplasma genitalium* (Fraser et al. 1995, US patent number 6,537,773). Since then, more than 300 bacterial genomes have been fully sequenced and become publicly available, including the sequence of a virulent form of *H. influenzae* (Harrison et al. 2005); the original *H. influenzae* strain sequenced in 1995 was from an isolate that does not cause disease. Although the majority of these several hundred genomes are from pathogenic organisms, some environmental bacterial genome sequences have also become available. This review article will provide a brief overview of sequenced bacterial genomes, their genomic diversity and some of the insights gained from analysis of this vast amount of data.

Bacteria are microscopic unicellular prokaryotes that inhabit a wide variety of environmental niches, broadly distributed in three ecosystems: the soil, marine environments and other living organisms. Although there are

T. T. Binnewies · P. F. Hallin · O. Lund · D. W. Ussery (✉)  
Center for Biological Sequence Analysis,  
Technical University of Denmark,  
2800 Lyngby, Denmark  
e-mail: dave@cbs.dtu.dk

Y. Motro · D. Dunn · M. Bellgard  
Center for Bioinformatics and Biological Computing,  
Murdoch University,  
Murdoch, Western Australia 6150, Australia

T. La · D. J. Hampson  
School of Veterinary and Biomedical Sciences,  
Murdoch University,  
Murdoch, Western Australia 6150, Australia

T. M. Wassenaar  
Molecular Microbiology and Genomics Consultants,  
Zotzenheim, Germany

literally millions of bacterial species, only a small proportion of these can be grown in the laboratory (Handelsman 2004). Bacteria (and Archaea) can be found almost anywhere in the environment: in the air, even in the International Space Station (Novikova et al. 2006), in thermal ducts found at great depths in the oceans (Alain et al. 2002; Vezzi et al. 2005), in the intestinal tracts of animals (Yan and Polk 2004; Backhed et al. 2005) and in soil and rocks, even thousands of meters deep (Torsvik et al. 1990). Bacteria live within unicellular eukaryotes, algae, plants or animals. This diversity is reflected in their physiology, morphology, metabolism and ecosystems. For example, from a physiological perspective, most intestinal bacteria such as *Escherichia coli* are motile by means of flagella, to overcome the peristalsis of the gut, whilst the soil bacterium *Clostridium perfringens* does not possess such motility machinery (Shimizu et al. 2002). From a metabolic perspective, the versatile *Burkholderia cepacia* (formerly *Pseudomonas cepacia*) can utilise approximately 100 different organic compounds as a sole energy source (Goldmann and Klinger 1986) compared to the strictly intracellular *Mycobacterium tuberculosis* which is dependent on only a few carbon sources produced by its involuntary host. From an inter-bacterial interaction perspective, sometimes bacteria cooperate. For example, *Enterobacter cloacae* and *Pseudomonas mendocina* positively interact to stimulate plant growth (Duponnois et al. 1999). On the other hand, there are also bacteria which not only “do not cooperate” but exhibit predatory behavior, such as *Bdellovibrio bacteriovorus* (Rendulic et al. 2004). As for bacteria–host interactions, for a given bacterial species both pathogenic and non-pathogenic strains can exist (Dobrindt and Hacker 2001; Penyalver and Lopez 1999), while other species may be exclusively parasitic (Goebel and Gross 2001), truly symbiotic (Gil et al. 2004) or commensal (Yan and Polk 2004) for their host. It is interesting to note that this diversity is somehow captured in the relatively small bacterial genomes.

The first complete viral genome ( $\phi$ X174) was published in 1977 (Sanger et al. 1977). To put this into perspective, to sequence the 4.6-Mbp *E. coli* K-12 genome at that time (about a thousand base pairs (bp) could be sequenced per year in 1977) would take more than a thousand years to finish, and to sequence the human genome would take more than a million years to complete. The automation of sequencing methods, the invention of polymerase chain reaction (PCR) (Mullis et al. 1986) and the shotgun cloning procedure reduced costs and time, and provided the capability for large-scale sequencing. These developments together have led to the sequencing of the first complete bacterial genome (Fleischmann et al. 1995) almost 20 years after the sequencing of  $\phi$ X174. The choice of the first bacterium to be completely sequenced (*H. influenzae* Rd KW20) was based on the following reasons: (1) the genome size was thought to be ‘typical’ among bacteria (1.8 Mbp), (2) the G + C base composition was close to that of the human genome (38%) and (3) the bacterium had important human health implications. In the absence of procedures to produce a genetic map for the species,

genome sequencing was proven to be a powerful alternative for genetic characterisation. This landmark work initiated the influx of genome sequence data which is now updated frequently and is publicly available. As of November 2005, there are more than 300 fully sequenced, publicly available bacterial genomes. Figure 1 shows this increase of sequence data over the past decade.<sup>1</sup>

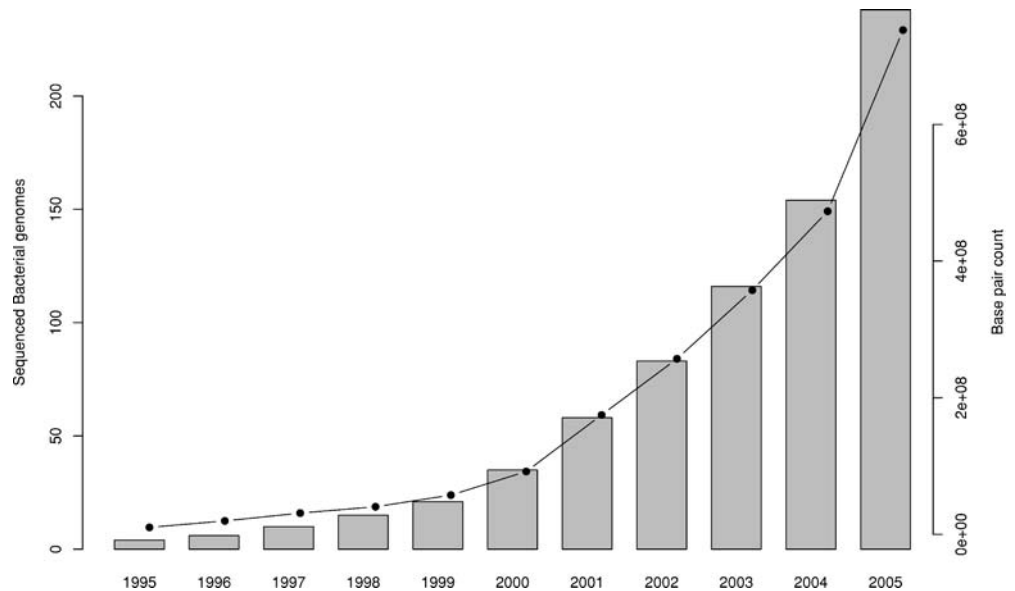
The total number of completed bacterial genome sequences has more than doubled over the past 2 years and, at the time of writing, there are 855 publicly listed bacterial and archaeal genome projects that are in various stages of progress.<sup>2</sup> In addition to new species, multiple strains of the same bacterial species are being sequenced. The amount of genomic data currently available has provided significant advances in our understanding of a number of important themes, including bacterial diversity, population characteristics, operon structure, mobile genetic elements (MGE) and horizontal gene transfer (HGT). It has also provided a number of challenges in understanding the ecology of, as yet, undiscovered bacterial worlds. The availability of whole genome sequences for pathogenic and commensal bacterial species has allowed a more detailed analysis of the complex interactions that occur with their plant or animal hosts. Figure 2a is a phylogenetic tree of 300 sequenced bacterial genomes (available at the time of writing). Many of these genomes are from pathogenic bacteria living in complex ecosystems, such as the spirochaete *Brachyspira pilosicoli* labelled in red in the phylogenetic tree shown in Fig. 2b. This bacterium attaches to enterocytes to form a “false brush border” in the colon.

Most genome sequencing projects are currently carried out using automated applications of the sequencing technique developed by Sanger et al. (1973), but newly developed methodologies may enable even more rapid sequencing in the future. Two papers have been published about two different methods for high-throughput sequencing of bacterial genomes (Pennisi 2005). One method is essentially a “do-it-yourself kit”, which uses a laser confocal microscope and other “off-the-shelf” components to build a sequencing machine capable of sequencing an *E. coli* genome in less than a day (Shendure et al. 2005). The second method is a commercial machine, based on pyrosequencing methodologies to generate many short pieces of DNA; this method was used to sequence a bacterial genome within a few hours (Margulies et al. 2005). Although there are still some technical problems with both of these methods, it is clear that, in the near future, it will be possible to quickly sequence a bacterial genome at a considerably low cost.

<sup>1</sup> Completed genome statistics obtained from the CBS atlas web pages <http://www.cbs.dtu.dk/services/GenomeAtlas>

<sup>2</sup> <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj>

**Fig. 1** Cumulative number of complete published sequenced bacterial genomes (*bars*) and total number of basepairs (*line*) over the past decade (1995–2005)



## Genomic information

DNA codes for more than just proteins

The quality of annotation of bacterial genomes varies, although a survey based on three different methods to predict the expected number of genes in a genome has found that it is likely that, for most bacterial genomes, around 20% of the genes annotated might not be “real” (Skovgaard et al. 2001). Furthermore, some “real” genes, based on proteomics experiments, which were not originally predicted have been detected, highlighting the dynamic nature of annotation and that genes are missed (Jaffe et al. 2004). Over-annotation of bacterial genomes is a problem but, unfortunately, this cannot be easily avoided. On the one hand, no one wants to miss a gene and, on the other hand, small genes can be quite difficult to predict, as a short open reading frame could easily occur by statistical chance (Skovgaard et al. 2001).

There are currently several automated annotation systems and the BaSys system (Van Domselaar et al. 2005) provides a comprehensive annotation of a DNA sequence file. To conduct comparative genomics with several hundred genomes, quality databases are essential and the “GenomeAtlas” database, which was originally developed to store DNA structural information about the various sequenced genomes, is one example (Hallin and Ussery 2004). Approximately a hundred different features for each genome (such as percent AT, coding skew bias, length of genome and number of genes) are currently made available through <http://www.cbs.dtu.dk/services/GenomeAtlas/>.

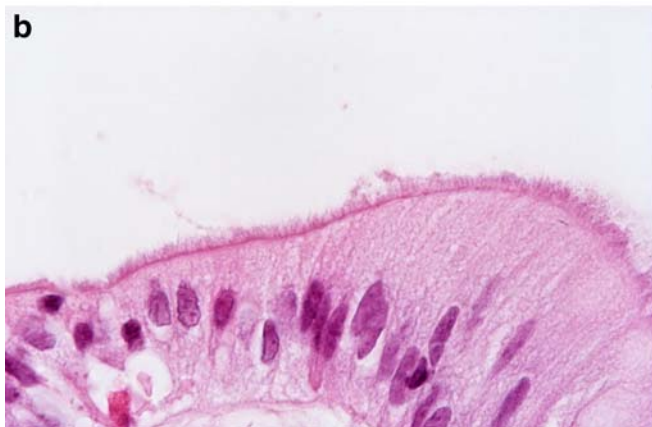
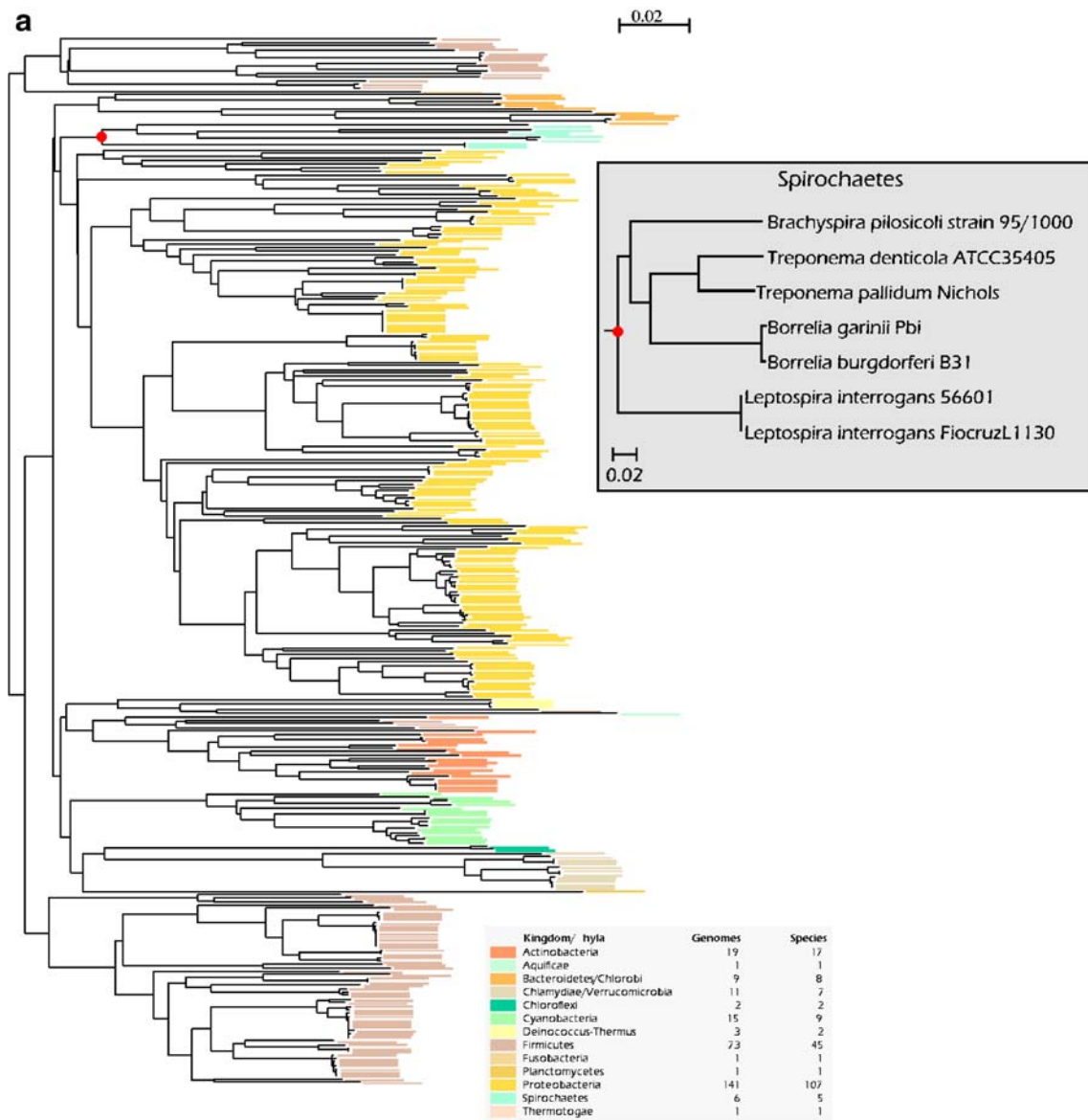
### Duplication of essentials

One of the features of genomic sequences that can be easily recognised is the presence of repeat sequences. The most obvious and extensive repeats present in many bacterial

genomes are the operons encoding the ribosomal RNA genes. These rRNA operons typically encode 16S and 23S rRNA separated by a short spacer, often followed by the 5S rRNA gene. All sequenced bacterial genomes possess at least one rRNA operon, and many (215 of 300) have two or more copies; the number of operons tends to correlate with bacterial division time. Thus, species that divide quickly (such as *Bacillus cereus*) have more copies of rRNA genes, so as to enable rapid production of ribosomes. In addition, species containing multiple rRNA operons appear to be more adaptable to changing environmental conditions (Acinas et al. 2004). The rRNA genes are a valuable tool for the estimation of taxonomic relationships (see Fig 2a). These genes evolve slowly, presumably because they play an essential role as the backbone of ribosomes while interacting with multiple proteins. Any changes in the shape (sequence) of rRNA would most likely be fatal.

Multiple copies per genome of tRNA genes can also be found in some genomes, again tending to correlate with division time. However, for tRNAs, the duplication number is also dictated by the frequency with which particular codons are used (or vice versa, as cause and effect cannot be distinguished here). This enables a less obvious level of regulating gene activity: a gene using many codons for which only one tRNA gene is available will probably be translated at a rate-limiting step, whereas abundant proteins are more likely to use tRNAs for which multiple gene copies are available. This is the basis for the codon adaptation index, which is a measure of the adaptation of a gene’s codon usage towards the optimal tRNA pool (Sharp and Li 1987).

There are of course other duplications in bacterial genomes, some of which might appear at first glance to be less essential. For example, the ‘REP’ repetitive sequences frequently found in enterobacteriaceae can be used as unique identifiers of bacterial genomes (Tobes and Ramos 2005). It has been speculated that these repeats are meaningless, resulting from errors in replication, or that



◀ **Fig. 2 a** Phylogenetic tree of 287 sequenced bacterial genomes, based on alignments from the 16S rRNA gene sequence. The phyla are colour-coded; a more detailed view, with names of all the organisms can be found in the supplemental information: <http://www.cbs.dtu.dk/services/GenomeAtlas/suppl/FIG10yr/>. **b** Photomicrograph showing a dense fringe of anaerobic spirochaetes (*B. pilosicoli*) attached by one cell end to the luminal surface of human colonic enterocytes, forming a “false brush border”. Besides that of humans, *B. pilosicoli* colonises the large intestine of a variety of mammals and birds, causing diarrhoea and reduced growth rates. Genomic sequence from *B. pilosicoli* is being analysed to assist in understanding the genetic basis of this dense colonisation, including patterns of gene expression underlying the complex interactions that occur between individual bacterial cells and the colonised enterocytes. The photograph is courtesy of Dr. W. Bastiaan DeBoer, University of Western Australia, Perth, Western Australia

they may be a part of mobile elements that are able to translocate and duplicate themselves. These could alternatively be non-functional ‘molecular fossils’ of previous insertion events. Finally, it could well be that these repeats serve some as yet undiscovered useful purpose. It is possible, for example, that repetitive sequences and insertion sequence elements (ISs) contribute to genome plasticity through structural changes based on homologous recombination (Kennedy et al. 2001; Fraser-Liggett 2005).

#### A brief history of bacterial operons

Much of the early classical work in microbiology has been done with *E. coli*, as this bacterium is relatively easy to culture in the laboratory. As more and more genetic information was gathered, it was considered a ‘typical’ bacterium, although *E. coli* is not more typical for bacteria than a rabbit is for all eukaryotic organisms. More than 40 years ago, a model was proposed for gene regulation of the catabolism of lactose in *E. coli* (Jacob et al. 1960; Jacob and Monod 1961). The model described an operon as a cluster of genes with related functions (encoding, in this case, enzymes required for lactose degradation). This operon structure neatly allows regulation of gene expression by the concentration of lactose (Lewis et al. 1996; Reznikoff 1992). With the continuous expression of one small protein (a repressor), wasteful expression of several other catabolic enzymes in the absence of lactose is prevented.

Since the discovery of the *lac* operon, many more catabolic operons have been discovered, with positive and negative feedback strategies, and these illustrate the biological need to use resources as efficiently as possible. Many, if not all, bacterial genomes indeed display clusters of genes involved in a single process (be it co-jointly transcribed and regulated, as in classical operons, or with separate promoters and regulators), but the degree of operon gene organisation and gene clustering differs between species. In some bacteria, such as in *Helicobacter pylori*, operons are relatively unconserved, and genes involved in one cellular process can be dispersed

throughout the genome (Tomb et al. 1997; Alm and Trust 1999), although more recent work suggest that perhaps there are more operons in *H. pylori* than previously thought (Price et al. 2005). There are currently many resources for prediction of operons (Rogozin et al. 2004; Rosenfeld et al. 2004; Alm et al. 2005; Janga et al. 2005; Nishi et al. 2005; Price et al. 2005; Vallenet et al. 2006), including several databases, such as the Operon Database (Okuda et al. 2006), RegulonDB (Salgado et al. 2006a,b) and Gene-Chords (Zheng et al. 2005).

How did the first operon evolve? There have been historically three models proposed for the origins of gene clusters. The first model, which dates back to 1945, proposed the clustering of genes to be the direct result of gene duplication and evolution (Horowitz 1945, 1965). Gene duplication can occur during replication and, as a duplicated gene has more freedom to mutate, this is believed to be a classical mechanism for novel enzymes to evolve (Lazcano et al. 1995). However, although all genes within an operon may be involved in a single metabolic process, their function and structure can vary considerably, and a phylogenetic relationship between them is not always likely.

The second model proposed for the evolution of operons is that coregulation of genes under a common promoter could provide selective advantage (Jacob et al. 1960). However, we now know that, in fact, it is possible to have coregulation of genes that are not physically linked together. Furthermore, this model does not really provide a gradual step-by-step mechanism for the evolution of operons.

The third model for the evolution of an operon is that pre-existing genes moved together due to selective advantages of having genes involved in the same biochemical pathways or processes being physically close to each other. This hypothesis allows for structurally distinct genes to be part of one operon. This model requires both variation and frequent recombination and has been proposed as an explanation of clustering of genes in bacteriophage genomes (Stahl and Murray 1966; Juhala et al. 2000).

In addition to these three views, there are other alternatives. Gene clustering may be of selective advantage in the case of horizontal gene transfer (see section below) and, based on this idea, a fourth mechanism, ‘selfish operon’ model, was proposed (Lawrence and Roth 1996). This view has been recently called into question, based on the physical clustering of essential genes in the *E. coli* K-12 genome (Pal and Hurst 2004). Two other alternatives for operon evolution deal with chromatin structure and the physical location of genes in bacterial chromosomes, where transcription and translation are coupled (Pal and Hurst 2004). It is quite possible that, in fact, there is no one “correct” mechanism, but perhaps different mechanisms are involved at the same time. For example, the selective advantage of gene clustering during horizontal gene transfer is exemplified by the clustering of multiple antibiotic

resistance genes on mobile genetic elements (Carattoli 2001). In the era of antibiotic use, such genes are under strong selective pressure and are frequently passed on between bacteria by means of mobile elements. Whether these have directly contributed to the spread of catabolic and other operons between bacterial species is currently not known.

What separates genes in a genome?

In comparison to genes, the non-coding part of genomes receives far less attention. Some genomes are more densely packed than the others. The average coding density is about 90%, ranging from 95% for *Pelagibacter ubique* (Giovannoni et al. 2005) to 51% for *Sodalis glossinidius* (Toh et al. 2006). Bacterial genes are not spliced as they are in eukaryotes; that is, introns are absent from nearly all bacterial genes. The sequences separating genes (intergenic regions) can be thought of as spacers where information on regulation of transcription can be stored, although sometimes these intergenic regions can also be more than regulatory and spacer domains. Intergenic regions in the *E. coli* K-12 chromosome have been suggested to contain the sequences for several hundreds of small RNA genes which are transcribed but do

not code for proteins (Chen et al. 2002). Many of these small RNAs act as regulators (Gottesman 2005).

In general, the intergenic regions of bacterial genomes are more AT-rich, will melt more readily, are more curved and are more rigid than the chromosomal average (Pedersen et al. 2000; Hallin and Ussery 2004). This is true for nearly all of the several hundreds of bacterial genomes sequenced, regardless of AT content. These characteristics make sense in terms of mechanical properties needed for initiating transcription.

## Generation of genomic diversity in bacteria

Genomic diversity is far greater than expected

The view in many textbooks of biological diversity and evolution often envisions clonal bacteria which slowly evolve through the gradual accumulation of single-nucleotide changes. There might occasionally be a rare event where a new gene is duplicated but, in general, it has been commonly thought that if one were to sequence two different strains of a common bacterium like *E. coli*, the sequences would, for the most part, be similar and the two strains would share most (perhaps 90% or more) of their genes. At the time of writing, there are 20 different *E. coli*

**Table 1** Current *E. coli* genomes sequenced or in progress

<i>Escherichia coli</i> strain	Length (bp)	Number of genes	Number of tRNAs	Number of rRNAs	Number of contigs	Accessionnumber
O157_EDL93	5,528,445	5,349	100	7	1	AE005174
E22	5,516,16	4,788	NA	NA	109	AAJV00000000
O157_RIMD0509952	5,498,450	5,361	103	7	1	BA000007
E110019	5,384,084	4,746	NA	NA	119	AAJW00000000
B171	5,299,753	4,467	NA	NA	159	AAJX00000000
53638	5,289,471	4,783	NA	NA	119	AAKB00000000
042	5,241,977	4,899	93	7	2	Sanger Institute (unpublished)
CFT073	5,231,428	5,379	89	7	1	AE014075
H10407	~5,208,000	~5,000	NA	NA	225	Sanger Institute (unpublished)
F11	5,206,906	4,467	NA	NA	88	AAJU00000000
B7A	5,202,558	4,637	NA	NA	198	AAJT00000000
NMEC RS218	5,089,235	~4,900	NA	NA	1	Uni. Wisc. (unpublished)
E2348	5,072,200	4,594	71	7	4	Sanger Institute (unpublished)
E24377A	4,980,187	4,254	97	6	1	AAJZ00000000
UPEC 536	~4,900,000	~4800	NA	NA	1	Uni. Würzburg (unpublished)
101NA1	4,880,380	4,238	NA	NA	70	AAMK00000000
HS	4,643,538	3,689	89	6	1	AAJY00000000
K-12_W3110	4,641,433	4,390	88	7	1	AP009048
K-12_MG1655	4,639,675	4,254	88	7	1	U00096
B03	4,629,810	4,387	86	6	1	CNRS France (unpublished)

NA Currently not annotated

genomes which have been either completely sequenced or at least with an expected coverage of greater than 99% of the genome. Table 1 lists these genomes, and one of the surprising observations is the diversity just in size of the main chromosome, ranging from 5.5 to 4.6 Mbp—that is, close to a million base pairs present in some *E. coli* strains which are missing in others. Furthermore, if one were to pick any one of these 20 strains, there would be more than a hundred genes which are unique to that strain and are not found in the other 19 *E. coli* genomes. Studies have indicated that much of this diversity comes from bacteriophages (Ohnishi et al. 2001).

### Gene order conservation

When comparing bacterial genomes, two features are frequently analysed: gene presence and gene order. The presence or absence of genes is particularly interesting when two closely related species or strains that have different phenotypes, such as a pathogenic and a commensal strain of the same species, are compared (Hayashi et al. 2001). As for the actual process leading to the difference, the direction of the insertion/deletion event is not always clear; the nature of the indel (INsertion/DEletion) is generally kept neutral.

There are various models of how the gene order within operons may have changed throughout evolution. It may be that the gene order in ancient ancestral operons has been maintained, such that all (or many) of the operons in studied genomes would be expected to have a similar gene structure. However, this view has been contradicted by data from whole genome studies. Examining the stability of operon structures over evolutionary distance shows that the majority of the gene orders within operons could be shuffled frequently during evolution, with the ribosomal protein operons as an exception (Itoh et al. 1999). Such observations support the alternative possibility that operons are multiple evolutionary inventions. A more recent study has examined the evolution of the histidine operon in Proteobacteria and found evidence for indeed a gradual merging of genes with similar function into operons, at least in this case (Fani et al. 2005).

Comparisons of gene order can also be informative of chromosomal translocations and inversions, which frequently happen in bacterial genomes (Kuwahara et al. 2004). Such events are mostly neutral in terms of evolution, as they do not change the total genetic content of the cell, but translocations and inversions frequently coincide with insertions or deletions. Any of these processes can result from inaccurate excision of mobile genetic elements and, as such elements are frequently

**Table 2** Types of mobile genetic elements found in bacterial genomes

MGE	Description	References
Plasmids	Circular, self-replicating DNA molecules that exist in cells as extra-chromosomal replicons. Some plasmids can insert into the chromosome.	(Dobrindt et al. 2004)
Transposons	DNA molecules that frequently change their chromosomal localisation, either within or between replicons. They usually code for a transposase and some other genes (such as antibiotic resistance genes), and are flanked by inverted repeat DNA sequences.	(Dobrindt et al. 2004)
Conjugative transposons	Transposons that also carry genes related to plasmid-encoded conjugation, thus, providing the ability to transfer between cells via conjugation	(Dobrindt et al. 2004)
Bacteriophages	Prokaryote-infecting viruses, which can modify the host genome by coding new functions or by modifying existing functions. They are also capable of inserting into the genome (prophages). These are also agents of HGT.	(Dobrindt et al. 2004)
Integrans	Genetic elements composed of a gene encoding an integrase (int gene; excises and integrates the gene cassettes from and into the integron), gene cassettes (become part of the integron upon integration; consist of a promoterless gene and a recombination site termed attC) and an integration site for the gene cassettes (attI gene)	(Fluit and Schmitz 2004; Holmes et al. 2003; Peters et al. 2001)
Insertion sequence elements	Small, genetically compact DNA sequences, generally less than 2.5 kbp in length, encoding functions involved in their translocation, and transpose both within and between genomes. IS elements are a subset of a general group of elements named transposable elements. These transposable elements are defined as elements of DNA segments that carry the genes required for this process (and, in some cases, other genes), and consequently move about chromosomes and, more generally, genomes.	(Mahillon et al. 1999; Ou et al. 2006)
Genomic islands	Large chromosomal regions that contain a cluster of functionally related genes, an operon or a number of operons, flanked by direct repeat sequences, and located near an integrase or transposase gene and a tRNA gene.	(Dobrindt et al. 2004)

involved in generating diversity in bacteria, they deserve to be treated in a separate section.

### Mobile genetic elements

MGEs are genomic elements that are capable of translocating themselves within or between genomes. When moving to a new genome, they may confer a new characteristic on the recipient. Their size ranges from hundreds of base pairs to more than 100 kbp. Plasmids, transposons, conjugative transposons, bacteriophages, integrons, insertion sequence elements and genomic islands (GEIs) are all considered MGEs (Table 2). Bacteriophages are the most sophisticated, as they produce their own protein coat to protect the genetic material (which can be DNA or RNA). Conjugative transposons induce conjugation between cells, a process in which cellular membranes merge to produce a bridge through which the transposon can move. Some plasmids can also induce conjugation (a transposon always encodes transposase whereas a conjugative plasmid replicates without integration in the chromosome). Some of the definitions for the various MGEs partly overlap, as indeed these terms are flexible. For instance, transposons can integrate in plasmids, and bacteriophages may contain insertion sequence elements (Burrus and Waldor 2004).

MGEs constitute potentially foreign DNA located in a conceptual ‘flexible’ gene pool, from where ‘donated’ DNA is made available for recipient cells. Once the MGE is transferred into the recipient cell, the DNA will either insert into a region on the chromosome or it will start to evoke its own replication machinery. If the MGE is integrated into the genome, for example, like a pathogenicity island (PAI), the genes (or operon) will start to be expressed, thus adding a new characteristic to the cell. The MGE may later initiate ‘donation’ of DNA either to a next receptor (for which the trigger is as yet unknown) or to the flexible gene pool, perhaps taking with it a ‘new’ or additional gene or function. The integrated MGE may also become immobile as a result of chromosomal re-arrangements, duplications or sequence insertions/deletions. In the case of such rendered immobility, the integrated MGE becomes a permanent genomic element or genomic island. At a later stage, the genomic island may be modified and rendered mobile again, making it available for transfer to the flexible gene pool once again.

As the subject of all MGEs listed in Table 2 would suffice a review paper on its own, this review focuses on two, namely, insertion sequence elements and GEIs. These two MGEs are of particular interest because our knowledge of them has improved dramatically as a direct result of genome sequence availability and due also to their impact on the diversity of bacteria.

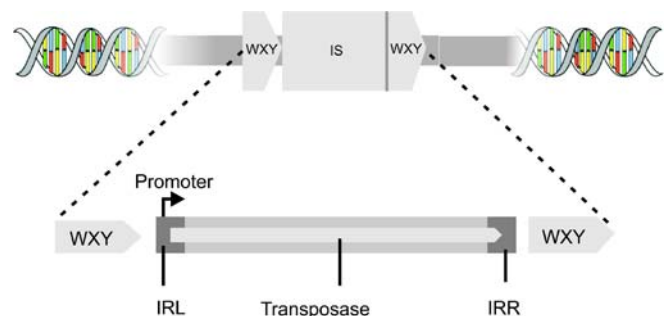
### Insertion sequence elements

IS elements are small DNA sequences, generally less than 2.5 kb in length, encoding functions involved in their own

translocation and can transpose both within and between genomes (Mahillon et al. 1999). IS elements were originally described as a subset of transposable elements (Prescott et al. 1999). IS elements are the simplest form of MGE and a key component of a majority of the more complex transposable elements, found both in bacterial and eukaryotic genomes. A number of reviews deal with IS elements in greater depth (van Belkum et al. 1998; Mahillon et al. 1999; Galun 2003).

An IS contains a transposase gene, flanked by terminal inverted repeats (the sequence of one flank is encoded on the opposite strand of the other flank). One of these repeats classically contains the promoter for the transposase gene (Fig. 3; Galun 2003). The IS elements are also flanked by short, directly repeated sequences, which are generated in the recipient DNA as a result of insertion.

The activity of transposable elements in genomes was first noted by McClintock (1950) in maize, although at that time the mechanism behind the observed genetic changes was not understood. Starlinger and Saedler (1976) provided the first review of IS elements in bacterial genomes. As noted by Lupski and Weinstock (1992), the first ISs were classified before their function, origin and dispersion mechanisms were understood. The present genomic era has resulted in advances in their classification, understanding of mechanisms of dispersion and identification of their role in evolution (van Belkum et al. 1998; Mahillon et al. 1999). Although the classical ISs are considered to be evolutionary neutral, as each can only translocate their own transposase, they are the means by which genomic islands (for example PAIs and metabolic islands) are transferred, and they also play a role in plasmid integration (Rocha et al. 1999). Variation in the excision of ISs promotes genome rearrangements (including deletions, inversions and replicon fusions; Mahillon et al. 1999). Antibiotic resistance genes are frequently spread within bacterial populations with the aid of ISs, which gives these simple elements clinical relevance. Finally, in special cases, IS elements can indirectly cause antigenic variation, a process in which a gene is switched off and on in a reversible manner within a bacterial population (Talarico et al. 2005). IS sequences that



**Fig. 3** Organisation of a typical insertion sequence. The IS is represented as an *open box* in which the terminal inverted repeats are shown as *blue boxes* labelled *IRL* (left IR) and *IRR* (right IR). An open reading frame encoding the transposase (*grey box*) is located in the IS. *WXY* boxes flanking the IS represent short directly repeated sequences generated in the target DNA as a consequence of insertion. The transposase promoter is localised in IRL.

are present in the first part of a gene can cause slippage during replication, as DNA polymerase has difficulties with correct replication of short multiple repeats. The result can be a frame shift with consequential inactivation, but the next frame shift can restore gene function. Such slippage can also vary the distance and, thus, activity of a promoter and its gene. Examples involving genes with a role in pathogenicity, with antigenic variation of surface exposed proteins, and environmental adaptation have been described (van Belkum et al. 1998; Rocha et al. 1999).

Monitoring of these elements has provided insights into bacterial genome molecular processes and the nature of IS elements. For example, understanding the regulatory mechanisms of IS elements has provided insights into the importance of the compromises adopted by IS elements (and MGEs, in general) between a stable host genome and in endangering the survival of the host, through too much transposition activity (Nagy and Chandler 2004). It has also been suggested that IS expansion occurs during an evolutionary bottleneck, which reduces effective population size and the degree of intraspecies competition (Parkhill et al. 2003).

### Genomic islands

GEIs, also referred to as integrative and conjugative elements or ICElands (van der Meer and Sentchilo 2003), are large chromosomal regions that cluster functionally related genes, are flanked by direct repeat sequences and are located near an integrase or transposase gene and often also near a tRNA. Furthermore, GEIs must have a GC composition different from the rest of the genome. GEIs include pathogenicity islands, symbiosis islands (SYIs), metabolic islands (MEIs), antibiotic resistance islands (REIs) and secretion system islands (SEIs) (Zhang and Zhang 2004). This remarkable variety of GEIs demonstrates the power of horizontal gene transfer, as they are believed to be the result of interspecies DNA transfer. With multiple genes neatly clustered in functional groups including all necessary regulatory and secretory genes, the power of transferring such 'adaptive genetic bombs' can be easily imagined.

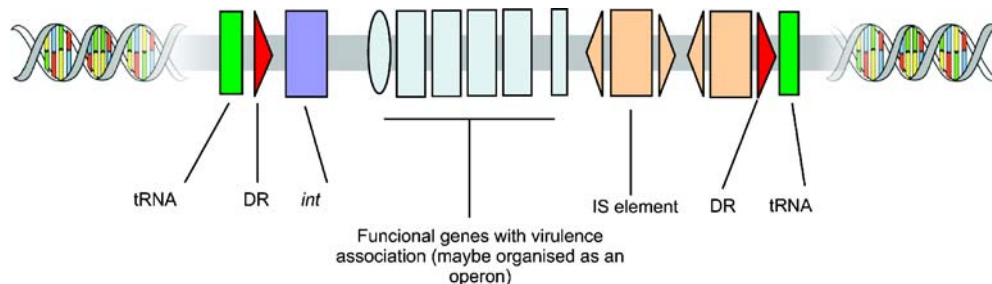
Genome sequences have revealed that GEIs are common in bacteria as a result of successful horizontal transfers of

DNA from a donor genome to a recipient genome. In most cases, the nature of the donor is unfortunately unknown. Even when an identified GEI bears a high resemblance to a section of another sequenced organism, one should not assume (though frequently this mistake has been made) that the GEI was directly received from that other organism. The transfer could well have involved a third unidentified species, serving either as an intermediate between the first two or as the donor for the others. These possibilities are frequently not recognised, as people can be misled by the available genome sequences and are not sufficiently aware of all those bacterial genomes for which we are currently lacking sequence information.

The discovery of abundant genomic islands is strengthening the concept of a bacterial genome being quite dynamic and consisting of a backbone genome supplemented with adaptive genome modules, which may or may not be present in a given strain of the species (Fraser-Liggett 2005). All modules available to the species (but never all present in one strain) would comprise the gene pool of that organism. This concept clearly does not apply to strictly clonal species, in which case all isolates or strains closely resemble each other (as is the case, for instance, with *Bacillus anthracis*), but it better describes the situation for frequently observed highly diverse species, such as *E. coli* or *Streptomyces*. Nevertheless, the timescale at which these events take place should not be ignored. Genomes are the sum of thousands of years of evolution. Observations of evolutionary events taking place in 'real time' are still relatively seldom.

### Pathogenicity islands

PAIs are now considered a subtype of genomic islands but were among the earliest islands to be described. PAIs harbour pathogenicity-related genes, thus potentially conferring a pathogenic phenotype on a recipient genome. Figure 4 illustrates a generalised model of a PAI. As with other GEIs, PAIs are commonly inserted into tRNA genes, which may be preferred sites of insertion due to their relative conservation and redundancy (Dobrindt et al. 2004). PAIs are flanked by direct repeat sequences and contain an integrase gene that enables the integration into the



**Fig. 4** Generalised diagrammatic representation of a pathogenicity island. Commonly inserted into a tRNA gene sequence, flanked by direct repeat sequences, containing an integrase (*int*) gene, commonly containing insertion sequence elements, and harbouring

functional genes (with virulence associated properties), which may be organised into an operon structure. Sometimes, a type III secretion system is also present

recipient DNA. A feature observed for many PAIs (and originally included in their definition although not always present) is the presence of a type III secretion system, a set of genes building an apparatus to specifically inject virulence factors into the host cell (Jores et al. 2004). Numerous investigations have identified and analysed PAIs (McGillivray et al. 2005; Middendorf et al. 2004; Paulsen et al. 2003; Schneider et al. 2004; Zubrzycki 2004; Schmidt and Hensel 2004).

#### Horizontal gene transfer and restriction modification systems

Evidence of HGT (also referred to as lateral gene transfer LGT) dates back more than 30 years (Falkow 1975), with the finding of transposable elements. Although such events were considered only exceptional cases at that time, it is now evident that HGT events can make a substantial contribution to the generation of genetic diversity. As with all other features, the degree of horizontal transfer varies amongst species. Ochman et al. (2000) assessed 19 completely sequenced bacterial genomes and reported that the proportion of foreign proteins vary from 0% (*Mycoplasma genitalium*) to about 17% (*Synechocystis* spp). These findings were supported by others including Dufraigne et al. (2005). Ortutay et al. (2003) undertook a genomic-scale phylogenetic analysis of protein-encoding genes from five closely related *Chlamydia* spp and identified a set of sequences that have arisen via HGT as the divergence of the *Chlamydia* lineage. These data illustrate the significant role of HGT in the evolution of particular bacterial species. It is not surprising that obligate intracellular pathogens show less evidence of recent HGT: they will not easily encounter other bacterial species with which to share DNA.

Doolittle (1999a) listed three observations that can only be explained by HGT. The first observation is that phylogenetic trees based on individual protein-coding genes frequently differ substantially from the rRNA tree or from each other. The second observation comes from analysis, within a genome, of variation in G + C content, codon usage and gene order. The third observation is a result of between-genome comparisons, which show that all genomes contain particular genes that are more similar to homologues in distant genomes than to homologues in closer relatives or indeed that are absent from all known genomes of closer relatives. Combining this evidences, Doolittle (1999b) proposed an alternative to the tree of life to describe the evolutionary history of living organisms. His model of a web-like structure takes into account the influence of HGT, where interactions occur between ancestral organisms and descendants (branches) as well as between branches. A similar concept of a biological network has been further explored by Kunin et al. (2005). Such a concept is difficult to work with, and currently many microbiologists still accept a tree-like phylogenetic relationship, at least for an artificial 'backbone' of the species. Independent of the source (strain or species) of the

genes, phylogenetic trees can indeed be correctly produced for many genes and gene families and may describe evolutionary relationships that do not date back very far. Going back further in time, the vertical lineages become weaker and the phylogenetic trees are less meaningful. The paradoxical conclusion is that, by elucidating more of the evolutionary history of bacteria, their history has become less clear.

If it is really true that horizontal gene transfer is so general, how is it still possible to recognise bacterial species? First, HGT is not so frequent that it can be easily observed as DNA exchange in 'real time' (other than the uptake of plasmids, spread of antibiotic resistance genes or transfection of phages). Evidence for past HGT events can be seen in many bacterial genomes and exemplifies its importance in evolution but, without a time scale, the frequency of such events cannot be estimated. Second, there are barriers that restrict HGT. It is obvious that not all bacteria share the same gene pool and only bacteria that share an ecological niche are likely to encounter and share each other's DNA. Even under circumstances that favour DNA exchange, internal factors restrict the success of HGT, notably bacteriophage specificity, plasmid incompatibility, and the activity of restriction modification (RM) systems. Finally, not all putatively HGT genes from *E. coli* are actually translated into proteins, perhaps because of incompatibility of translational machinery (Taoka et al. 2004).

The discovery of restriction enzymes which could cleave specific DNA sequences provided the basis for driving the "biotechnology revolution" in the 1970s. RM systems are popular in molecular genetics and are routinely used by most molecular biology laboratories throughout the world. The RM systems encode a modification enzyme that chemically modifies a specific short DNA sequence and a restriction endonuclease that will digest the DNA at that same specific recognition sequence unless the sequence has been modified (usually by methylation). Bacterial species (and frequently strains within a species) all have their own combination of RM systems (Roberts et al. 2005). Incoming DNA with a different modification pattern will be recognised by the endonuclease of the recipient strain, and the fate of such DNA is to be degraded. This is seen as a serious restriction for the spread of DNA through populations unless their RM systems are compatible.

The analysis of RM systems at a comparative genomics level (particularly the type restriction II endonucleases) has shown the dynamic state of the respective genes (Lin et al. 2001) and posed a number of questions to the view that RM genes restrict gene flow. For example, *H. pylori* and *Campylobacter jejuni* are competent to take up DNA and have a large set of genes to maintain this property. The dynamic nature of the *H. pylori* genome and its natural competence is consistent with the weakly clonal population structure of *H. pylori*. Nevertheless, studies on *H. pylori* identified at least eight type II RM systems across two strains with an active restriction endonuclease and methylase (Kong et al. 2000; Lin et al. 2001). In addition, there were several active methylase genes without an active

endonuclease. The occurrence of RM systems that are not shared between the strains suggests that new RM systems are readily acquired and subsequently lost as a result of mutation or recombination (Lin et al. 2001). But that these would pose restriction barriers in gene flow is difficult to envisage with the dynamic population structure. RM genes possibly have other advantages to the cell. For methylation genes missing their matching restriction gene, it has been suggested that they may be used for regulating gene expression (as for DAM methylation in *E. coli*; Lobner-Olesen et al. 2005; Robbins-Manke et al. 2005) and for keeping track of which parts of the chromosome have been recently replicated (Maas 2004).

## Methods for comparing bacterial genomes

There are at least 20 methods to compare bacterial genomes, as shown in Table 3. Some methods are more commonly used than the others, and it is beyond the scope of this review to provide a detailed analysis of each method. A few of these methods are discussed in this section.

## Chromosome alignment and size comparison

Perhaps one of the easiest ways to compare genomes is by their sizes, as shown in Fig. 5. Although different phyla have different average sizes, it must be kept in mind that many of the phyla have currently few representatives and that there is a strong economic bias towards sequencing the smallest genome, so the size distributions shown here for the sequenced genomes could well be shorter than what

exist in natural ecosystems. Another way of comparing chromosomes is to do a simple alignment of the DNA sequences. There are two versions of the alignment programmes. One involves downloading some scripts and running them on a local computer such as the Sanger Centre's (Cambridge, UK) Artemis Comparison Tool (ACT, Carver et al. 2005) and the other is web-based such as "WebACT", a web-based version of ACT with pre-computed comparisons between several hundred bacterial genomes. The latter might be easier to use for those biologists who are less computationally inclined (Abbott et al. 2005).

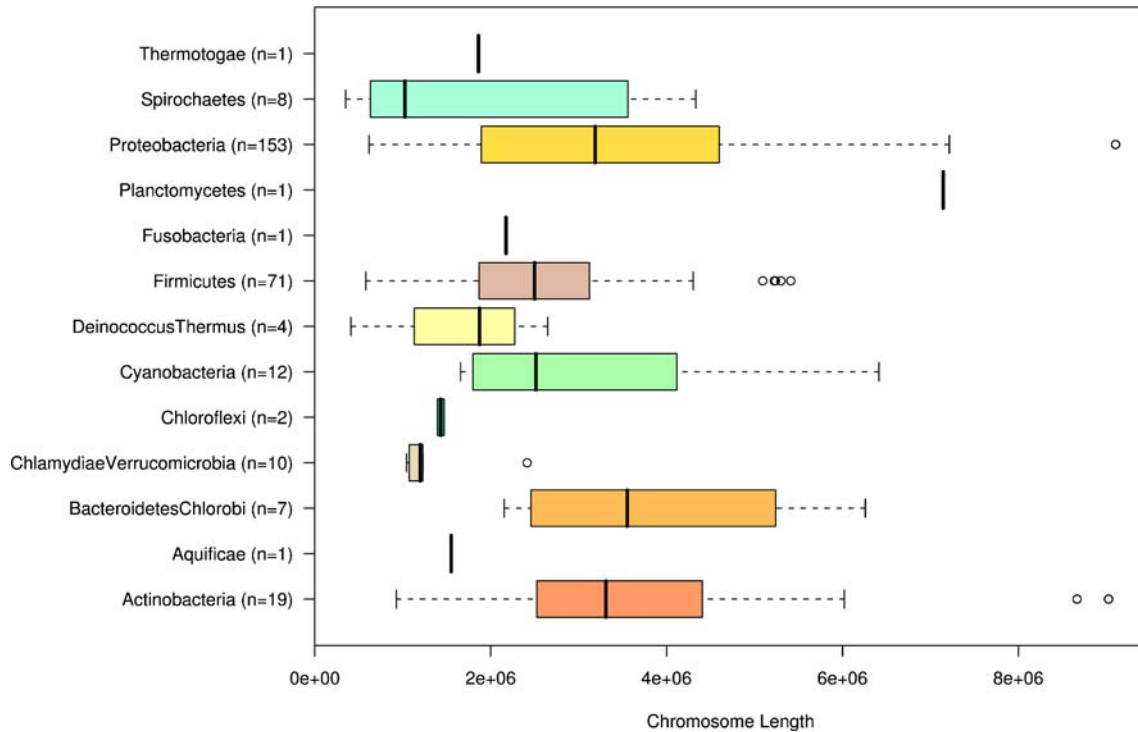
## AT content in genomes and promoter analysis

Another relatively easy method to compare genomes is by their AT content, which ranges from 78% (*Wigglesworthia glossinidia*) to 27% (*Clavibacter michiganensis*) for the 300 genomes sequenced at the time of writing. In addition to the average AT content for a whole genome, if the variation of the AT content within a given genome is examined, two general trends can be seen for nearly all of the bacterial genomes. First, on a more global chromosomal level, there is a tendency for the region around the origin of DNA replication to be more GC rich (i.e. less AT rich) and the region around the replication terminus to be more AT rich (Hallin et al. 2004b). Second, the average AT content for DNA about 400 bp upstream of the translation start site for all the genes in a genome is higher than 400 bp downstream (Hallin et al. 2004b). This makes sense in that the DNA will need to melt more easily in order for transcription to start.

**Table 3** Approaches to comparing bacterial genomes

Level	Method	Reference
Genome	Chromosome alignment	Carver et al. 2005
	AT content in the genome and upstream of genes	Ussery and Hallin 2004a
	Oligomer bias on leading or lagging strands	Worning et al. 2006
	Repeats (local and global)	Ussery et al. 2004a
	Periodicity of DNA structural properties	Worning et al. 2000
	Length comparison	Ussery and Hallin 2004b
	Promoter analysis	Ussery et al. 2004d
	Transcriptome	Organisation of rRNA operons
tRNAs and codon usage		Ussery et al. 2004c
Third nucleotide position bias in codon usage		Ussery et al. 2004c
Annotation quality		Skovgaard et al. 2001
Proteome	Amino acid usage	Ussery et al. 2004c
	BLAST atlases	Hallin et al. 2004a
	BLAST matrices	Binnewies et al. 2004
	Sigma factors	Kiil et al. 2005a
	Transcription factors	Kummerfeld 2006
	Secreted proteins	Bendtsen et al. 2005a
	Membrane proteins	Bendtsen et al. 2005b
	2-D correlation of properties	Willenbrock et al. 2005
	Two component signal transduction systems	Kiil et al. 2005b

Length Distribution of Bacterial genomes (n=290)



**Fig. 5** Genome length distribution for 287 bacterial chromosomes, shown as *box* and *whiskers* plot for each phyla. The number of chromosomes in each phylum is shown on the *axis*. Most of the bacterial genomes shown are either Proteobacteria (156 genomes) or

Firmicutes (70). At the time of writing, the largest complete bacterial genome sequenced is that of *Burkholderia xenovorans*, which is consists of 9,703,676 bp within two chromosomes, and the smallest is that of *M. genitalium* genome of 580,074 bp

### tRNAs, codon usage and amino acid

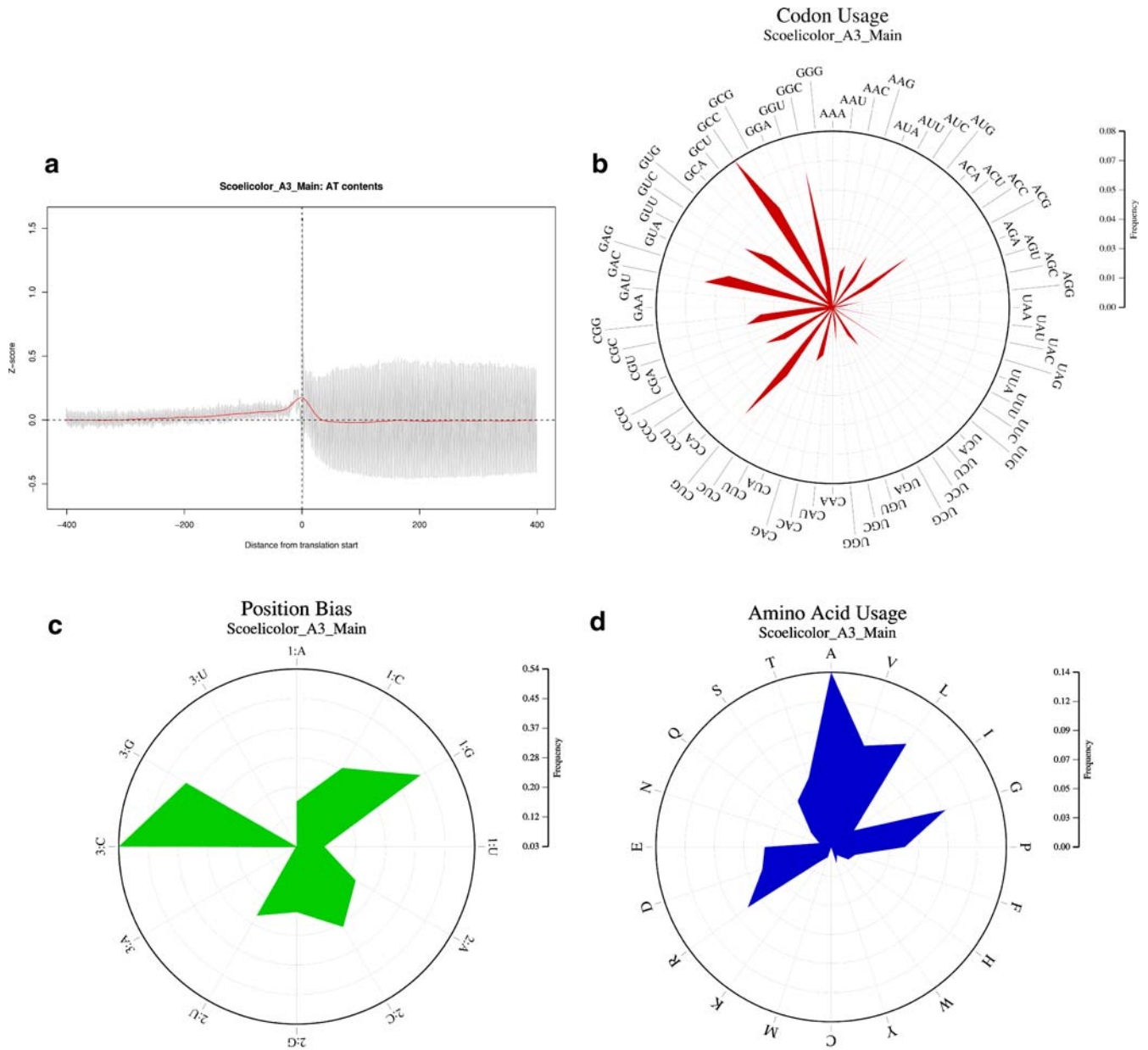
As mentioned above, the 200 bp upstream of translation start sites is more AT rich, on average, than the 200 bp downstream. However, if the unsmoothed data is examined (the grey lines in Fig. 6, panel a), there is much “noise” in the coding sequence, compared to the upstream, noncoding DNA. This is due to bias in codon usage, as shown in Fig. 6, panel b. The genome for a given organism will tend to show a preference towards certain codons and can be seen as a bias in the third codon position (Fig. 6, panel c). Finally, these codon biases also are in part affected by which amino acids an organism uses, as shown in panel d of Fig. 6. The amino acid usage for different *E.coli* proteomes differ: for example, *E. coli* K-12 shows the same amino acid usage as *Salmonella enterica* LT2, while the usage in *E.coli* O157 resembles that of *Shigella flexeneri*. Thus, two different *E. coli* genomes can have quite different amino acid usage (which might not be that surprising in view of the differences between strains of this species, see Table 1).

### BLAST atlases

The GenomeAtlas is a method to visualise structural features of an entire bacterial genome sequence as one plot. The plots are created using the “GeneWiz” programme,

developed at CBS (Pedersen et al. 2000). A more recent extension of this method is the development of the “genome BLAST atlas”, in which genes from different genomes are blasted against a reference genome and visualised using an atlas plot. BLAST atlases can provide additional contextual information about regions which contain few conserved genes. For example, a new genome might have a few small islands of unique proteins, and these regions might be more AT rich or might be expected to be potentially highly expressed, based on chromosomal structural information also provided in the plots. As mentioned above, when the 20 *E. coli* sequenced genomes in Table 1 are compared, an enormous amount of diversity is found. A BLAST atlas for *E.coli* O157 is shown in Fig 7a. Several regions of the chromosome have “holes” representing large segments of missing genes in some organisms, compared to the reference genome. In a sense, this information is somewhat similar to that obtained by the ACT plots mentioned above, although now the comparisons are being made at the level of presence/absence of clusters of proteins. In Fig. 7b, some of the regions containing gaps are more AT rich, some contain repeats and a few (marked) contain genes that might be highly expressed, based on chromatin properties. Thus, this tool can give a quick overview of the comparison of many genomes.

In Fig. 7a, the gaps correspond to regions of missing genes in the *E. coli* O157 genome. Similar patterns can be

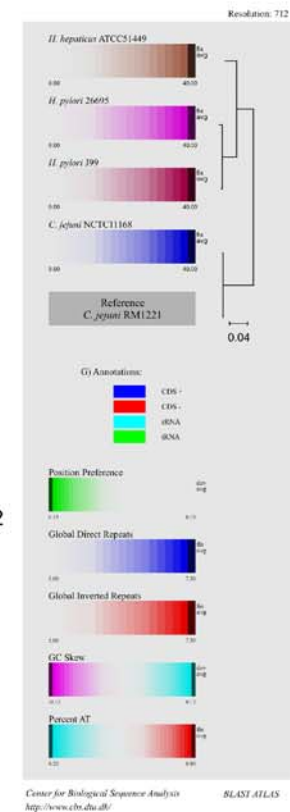
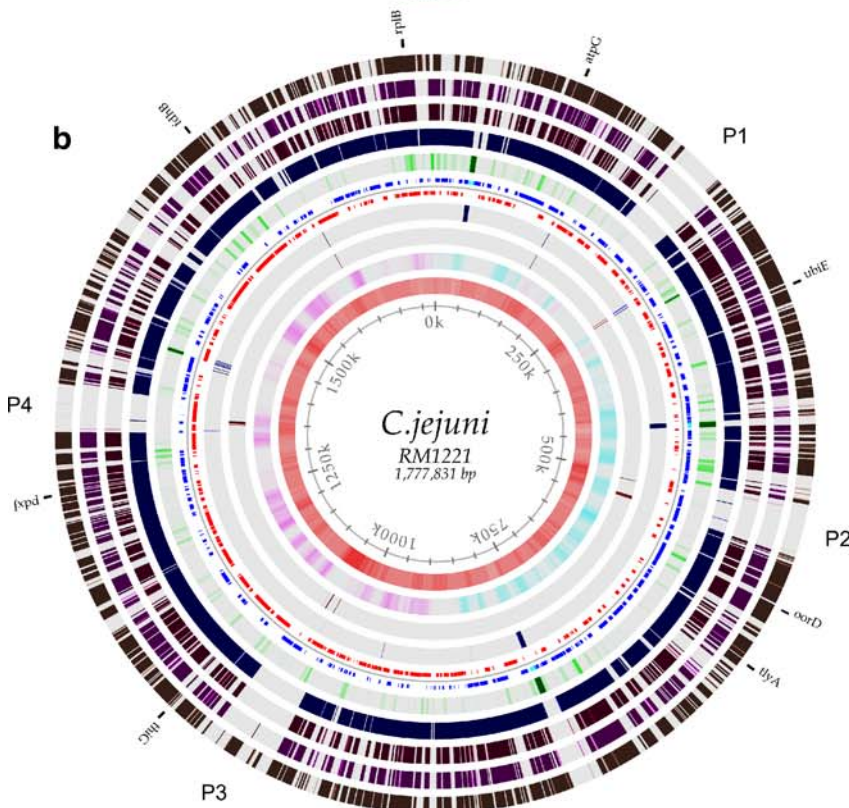
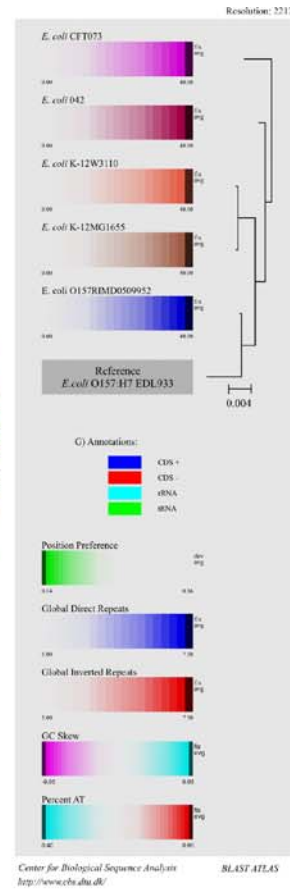
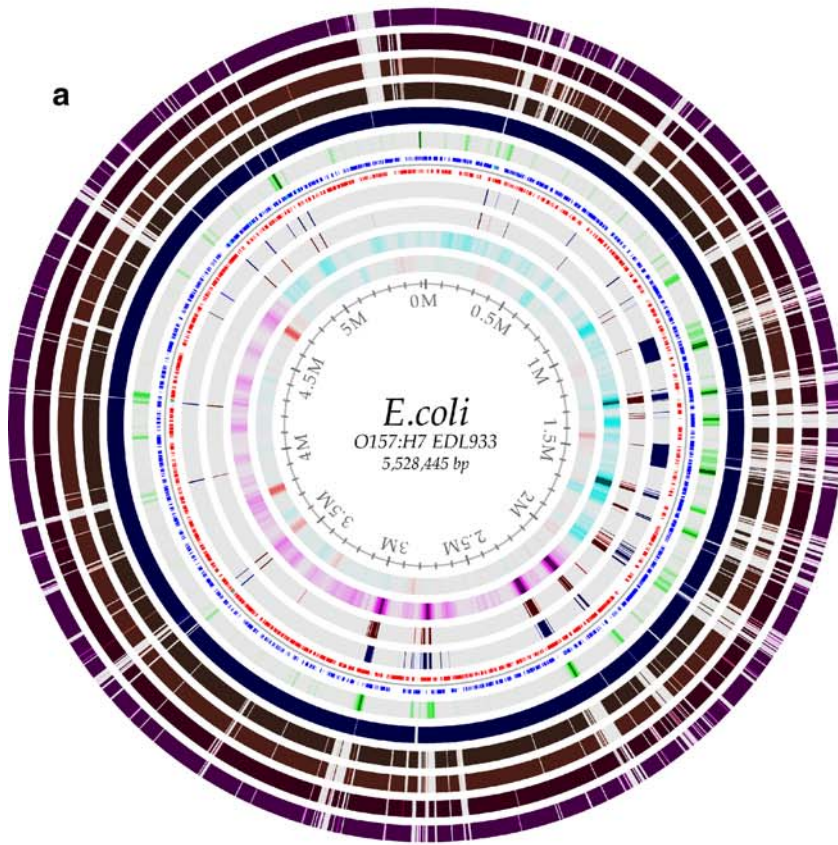


**Fig. 6** Genomic properties of *Streptomyces coelicolor* A3. **a** Comparison of AT content upstream and downstream of all 7,825 genes; the genes are all oriented in the same direction and aligned such that the translation start site is in the middle. Z-scores of standard deviations from the chromosomal average are plotted, as described previously (Ussery and Hallin 2004a). **b** Codon usage of the same set of 7825 genes. The frequency of occurrence of each of the 64 codons is plotted in a *star plot*; note that most codons have a relatively low frequency of usage. **c** Bias in the codon position are plotted as frequencies; note that

there is a strong tendency for Cs and Gs in third position. **d** Amino acid usage of each of the 20 amino acids for the entire *S. coelicolor* proteome is plotted as frequency of the total; the amino acids in this plot are grouped according to their properties; for example, all the aliphatic amino acids (A, V, L, I and G) are together and, in general, there is a general trend for this proteome to favour aliphatic amino acids, with the exception of isoleucine. The *three star plots* are as described previously (Ussery et al. 2004c)

seen for many other bacterial genomes. For example, in Fig. 7b, there are four large gaps in the *C. jejuni* RM1221 genome compared to other epsilon Proteobacteria. These correspond to phage insertion sites in *C. jejuni* RM1221, as described in the original genome sequence publication (Fouts et al. 2005). Similar results have been observed for

*Streptococcus* (Hallin et al. 2004a). In all three of these cases, there are large regions which contain many genes which are missing in other genomes of the same species. These clusters of genes often contain evidence that they came from phages, which appears to be an efficient method of bringing new DNA into a genome.



◀ **Fig. 7** Genome BLAST atlases. The *outer circles* represent BLAST hits of a given genome (named in the *legend*) to the reference genome (named in the *center* of the atlas). The colours are scaled such that good BLAST hits ( $E=10-40$ ) are *darkly shaded*, whilst regions containing no hits are shown in *light grey*, as described previously (Hallin et al. 2004a). **a** Genome BLAST atlas of *E. coli* EO157 EDL933 vs four other sequenced *E. coli* strains (the *four outermost circles*; the genomes are, going from the outermost towards the center, *E. coli* K-12 MG1655, *E. coli* K-12 W3110, *E. coli* CFT1076 and *E. coli* O157 RIMD0509952). **b** Genome BLAST atlas of *C. jejuni* vs other epsilon Proteobacteria

## BLAST matrices

Figure 7a,b illustrates the use of BLAST atlases to compare genome sequences. However, with several hundred genomes available, there is a need for a faster way of getting an overview of genome similarity. One method is the use of reciprocal hits—that is, to BLAST all the proteins encoded in a genome of interest against those in another genome (Binnewies et al. 2004). First, the genomes of interest are selected (e.g. all genomes of Proteobacteria), then a BLAST matrix can be displayed from this selection. The results are pre-generated and the system keeps track of sequence updates by generating MD5 checksums of all sequences and the combinations in which they have been BLASTed. The MD5 (termed also a *message digest*) will

produce a 32-digit string that is unique to an input string, e.g. a genomic sequence. The system maintains an all-against-all BLAST database updating only the missing comparisons—that is, changing the sequence of a record or inserting a new record will cause a BLAST run of the sequence against all the existing sequences of the database. By having multiple genomes in a given selection, an all-against-all BLAST matrix can be presented showing the percentage of genes that are shared between sequences—both on a protein and on a nucleotide level. Each such percentage is supplied with a link to give a full listing from the BLAST report. Fig. 8 shows an example of such a BLAST matrix, with the diagonal (in red) reflecting the internal homologues of a given genome. The boxes are colour-coded such that the intensity represents the fraction of hits (Binnewies et al. 2004) (Fig. 8).

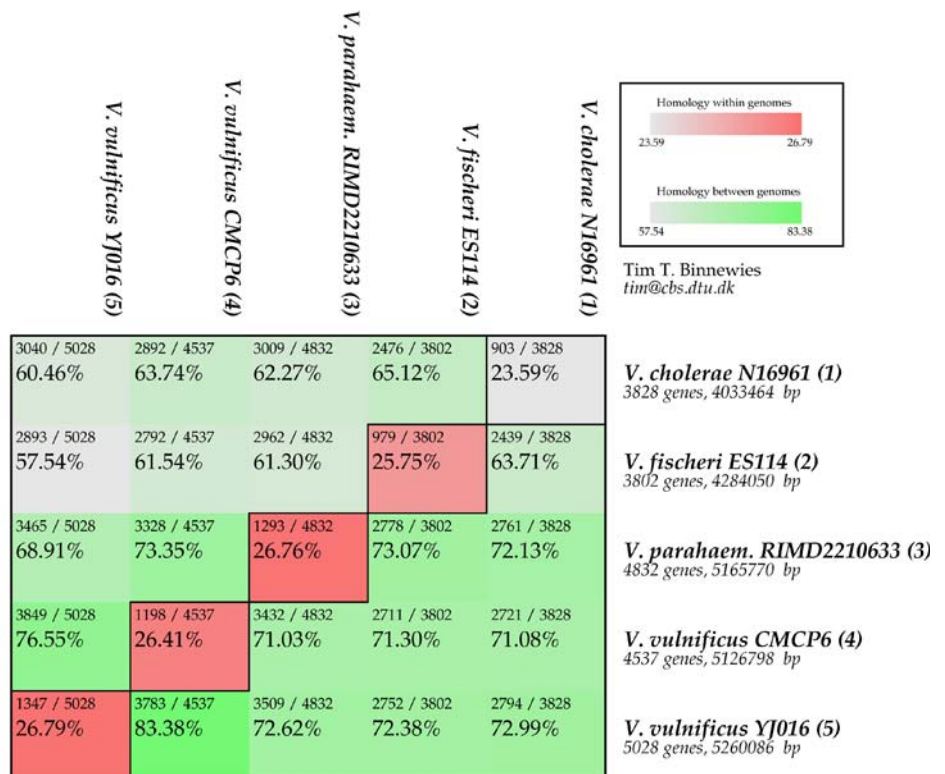
Meta-genomics: comparison of all the genomes in an ecosystem

The term “metagenomics” is used for genome sequencing projects in which many organisms are sequenced at once by shotgun cloning of all DNA present in a sample (Handelsman 2004). This enables microbial ecosystems containing microbes that are not (presently) culturable in pure form to be investigated (Handelsman 2004). The

**Fig. 8** The BLAST table shows the overall protein homology between all combinations of the five available *Vibrio* sequences. Only hits containing at least 80% of the length of the gene and with an  $E$ -value of  $1 \times 10^{-5}$  or better are counted. The *diagonal* (red/pink) indicates the fraction of proteins that have homologous hits within the proteome itself; the fraction is similar in all genomes, and the intensity is shown by the *red colour*, scaled from ~24% (grey) to ~27% (red). Note that the largest genome also has the highest fraction of internal homologs. The *green area* for the rest of the table, on each side of the diagonal, shows the number of proteins that have homologous hits between different *Vibrio* genomes. As before, the fraction is indicated by the intensity of the colour (*green*) scaled from ~57 (grey) to ~83% (green). In general, it is clear that these organisms share a high percentage of their genes with the other *Vibrio* species, which should be expected because they are from the same genus

## 5x5 Proteome Comparison (Vibrio)

ALR=80%, E-value <  $1 \times 10^{-5}$



reasons why organisms remain uncultured can be practical (e.g. thermophilic bacteria grow at a temperature above the melting point of agar), physiological (e.g. extremophiles that grow on pure culture can have very different properties from those observed in their true environment) or biological (symbiotic life forms cannot be cultured in microbiological pure form). The first genome sequence obtained from a non-culturable bacterium was indeed that of *Buchnera aphidicola*, a symbiont of aphids. This sequence was not obtained by meta-genomics at the total genome DNA level but rather at the rRNA level. Cell counts compared to plate counts showed that the latter can be orders of magnitude wrong: many viable bacteria refuse to grow on solid culture medium. The isolation of bulk RNA and the subsequent determination of rRNA sequences using specific primers allowed qualitative analysis to be performed for identifying novel bacterial species or ribotypes present in an ecosystem (Olsen et al. 1986). The application of PCR improved the sensitivity of such approaches but the limitation to rRNA sequences confined analyses to phylogenetic information only and little further knowledge was obtained about the new species. Metagenomics can be used to generate complete or fragmented genome sequences of organisms that might be abundant in nature but are not easily culturable.

The acid mine drainage sequencing project has shown the potential of meta-genomics (Tyson et al. 2004). The mine water of the Richmond mine is covered with a biofilm of bacteria despite its hostile environment: an extreme acid pH (between 0 and 1), high concentrations of metal ions, including copper, zinc and arsenic, and the absence of carbon or nitrogen sources (other than from air). The biofilm was composed of relatively few organisms, enabling the sequencing of shotgun-cloned DNA and the sorting of fragments according to their G + C content into nearly complete bacterial genomes. A dominant bacterial genus was identified, *Leptospirillum*, and a less abundant *Sulfobacillus* spp and some Archaea were also present. The findings greatly improved understanding of this ecosystem. The predominant bacteria were responsible for nitrogen and carbon fixation (*Leptospirillum* group III), whereas several species were able to generate energy from iron oxidation (*Ferroplasma* and *Leptospirillum* spp). As in this approach, each sequenced DNA fragment is obtained from a different individual (whereas in classical genome sequencing all DNA is obtained from one clone); information on polymorphisms also becomes available. As more complex ecosystems are studied, the puzzle of genome assembly becomes more difficult due to the presence of more species, genomic rearrangements and horizontal gene transfer events.

The largest attempt so far at metagenomics was initiated by C. Venter to sequence the microbial ecosystem in the Sargasso Sea (Venter et al. 2004). Seawater was sampled by filtering to specifically recover bacterial (and not viral or amoebal) DNA. Over 1 billion base pairs of sequence were generated, which was attributed to at least 1,800 species. As the abundance of individual species determines their coverage in shotgun cloning, this coverage (or rather the

mean of their Poisson distribution) was used to sort out DNA scaffolds (a scaffold is a reconstructed genomic region), and oligonucleotide frequencies were used to refine this sorting. Although the complexity of the investigated ecosystem did not allow complete assembly of individual genomes, the scaffolds belonging to the most abundant species could be attributed to *Burkholderia* and *Shewanella*-like species. As with the acid mine drainage project, polymorphisms were detected with varying frequencies. In fact, the dataset ranged from organisms belonging to a single species and clonal (few polymorphisms) to a population continuum in which some clonal complexes could be recognised. These observations illustrate the 'unnatural' approach of studying only pure bacterial cultures that have a strict clonal structure in contrast to natural environments where the population structure is much more fluid and the concept of clones or species is more elusive. The most impressive output of the Sargasso Sea study is the numbers of individual genes that were identified (69,901). Among the surprising findings was that rhodopsin (the bacterial protein required for carbon fixation) was abundant outside the proteobacteria where it had previously been identified. The finding of many genes involved in phosphate uptake and utilisation of poly- and pyrophosphates is puzzling, as the marine environment is extremely phosphate-limited.

The challenge to analyse the complex communities of a nutrient-rich environment was taken up by Tringe and Rubin (2005). One sample that was analysed was derived from agricultural soil and three were from marine whale carcasses. First, rRNA libraries were generated by PCR to investigate the microbial diversity. The soil sample (DNA obtained from 5 g of surface clay loam from land that had been used for livestock) was extremely rich in species with at least 847 ribotypes detected representing over 12 phyla. The whale samples (two bone parts and one biofilm covering a whale carcass) were less diverse but still contained between 25 and 150 ribotypes. Although the assembly of sequences obtained from shotgun libraries was not possible, the genes that were identified on the sequenced library clones demonstrated that approximately half of the predicted proteins found similarities (homologs) in existing gene databases. Plotting the number of novel gene families against the amount of generated sequences suggested that, for the soil sample, few novel orthologues were found after sequencing 25 Mbp. The functions of predicted proteins from the sequences were naturally diverse, but for the soil sample, potassium channelling systems were overrepresented, whereas for the whale samples sodium ion exporters were abundant—which fit with the abundance of these two ions in the two environments, respectively.

The metagenomics analyses will continue to see databases expanding, with the interpretation and assembly of raw data becoming more complete. The human gastrointestinal tract, for example, is the target of a metagenomics sequencing project (Mongodin et al. 2005). It is apparent that each individual carries a large variety of microflora, probably acquired early in life (and which may have health

consequences even though these organisms are not pathogenic) as well as bacterial microheterogeneity that was not recognised previously. Against the common belief that Firmicutes and Bacteroides would be the most abundant microbes present in the human gut, it appears that Actinobacteria and Archaea may be more prominent (Mongodin et al. 2005). The intestinal microflora of obese mice differs considerably to that of lean animals, an observation in support of the view that the microbiota of mammals are good indicators (be it cause or effect) of their health status (Ley et al. 2005). There are clearly many microbial communities to be analysed and compared using metagenomics.

---

### **Application: computational vaccine development**

Vaccines remain an extremely important tool for controlling infectious diseases of humans and animals, although they are only available for about 10% of the microorganisms known to be harmful to humans (Lund et al. 2005). Traditional vaccines typically have incorporated whole live attenuated or killed microorganisms, but, particularly for use in humans, such vaccines now have limited application due to concerns about safety, efficacy and/or ease of production. Much recent work, therefore, has focused on developing vaccines composed of prominent immunogenic parts of microorganisms (subunit vaccines) or genes encoding these components (genetic vaccines, Ellis 1999). For bacterial vaccine discovery, these newer approaches have been greatly assisted by the recent availability of whole genomic sequence data and has allowed a new approach to vaccine development called “reverse vaccinology” (Rappuoli 2001).

In reverse vaccinology, bioinformatics tools are used to undertake comprehensive in silico screening of genomic sequence to identify genes encoding proteins that have desirable characteristics. The power of this process has increased as more and more genomic sequences that encode proteins of known function become available in the databases for comparative analysis. Targets for consideration for use in vaccines include genes encoding outer membrane proteins or lipoproteins, transmembrane domains or export signal peptides, and proteins with homologies to bacterial factors already known to be involved in virulence or pathogenicity. Surface-exposed or secreted proteins as well as virulence factors such as toxins or adhesive factors are likely to induce an immune response that may be protective (Zagursky and Russell 2001). In this way, large numbers of potential vaccine components can be identified from a whole (or partial) genome sequence. This approach was first taken for the human pathogen *Neisseria meningitidis* serogroup B, with 600 open reading frames (ORFs) of potential interest initially being identified (Pizza et al. 2000). Recombinant proteins from 350 ORFs were eventually produced and, after screening in for distribution in different serotypes, stability, immunogenicity and cross-protection, 15 were selected as potential subunit vaccine candidates. This same

approach to vaccine discovery is now being taken for a number of important human and animal pathogens (Serruto et al. 2004). Reverse vaccinology allows rapid identification of a large number of potential subunit vaccine candidates, many of which would not have been recognised by more traditional approaches. It is complemented by the use of microarrays to analyse gene expression and of proteomic approaches to study protein expression and distribution and can be focused further by the use of computer algorithms that scan and identify sequences encoding specific epitopes involved in immunogenicity (reviewed in Lund et al. 2002; see also, for a review, Theoretical Biology and Biophysics Group, Los Alamos National Laboratory [<http://www.hiv.lanl.gov/content/immunology/pdf/2002/1/Lund2002.pdf>]). These algorithms have been strengthened by the availability of full genomic sequences for many pathogens.

Methods for the three main types of epitopes targeting B cell, helper T lymphocyte and cytotoxic T lymphocyte have been made, and improved methods are constantly being developed. Thus, it is possible to take a genome sequence, use some predictors as described above and select potential peptide sequences for construction of vaccines. These vaccines can be either chemically synthesised peptide based or DNA based. With regards to peptides, these can be used directly or used to construct a “polytope”, which is a composite protein made from individual epitopes.

---

### **Intellectual property rights: who owns the genome sequence?**

This review started by giving the US patent numbers for the first two genomes sequenced. This final section will briefly discuss some of the issues facing researchers working with genomic data. At the time of writing, ten whole genome patents have been granted, with more patents being applied for (O’Malley et al. 2005). Some of these patents include the use of the sequence in silico and clearly raise a number of issues related to freedom to operate in research. In addition, the enforcement of the patents could be difficult, with many bioinformatic tools being developed in the public domain.

Another related difficulty has to do with using or analysing genome sequences before they are presented in scientific publications. Now that it is possible to sequence a bacterial genome in an afternoon and have a GenBank file a day or two later, the time gap between having the sequence publicly available and having the paper in print can be several years. Some public granting agencies have pushed hard for the data to be made available as soon as possible for people to search for their particular gene of interest. On the other hand, it is also understandable that the individuals who have actually sequenced the genomes need some lead time to analyse their data. With high-throughput bioinformatic techniques, it is possible, for example, for some groups to do in a few days what would take other groups months (or years) to complete.

A final problem has to do with obtaining basic information about the strain used for sequencing a genome. For example, what was the strain isolated from? What was the growth temperature or culture medium pH for the culture that the genomic DNA was derived from? What is the doubling time of this organism under these conditions? These are all important pieces of data, but they are often missing in genome publications. A recent “minimal information about a genome sequence” standard has been proposed (Field and Hughes 2005), which is in the same spirit as the MIAMI standard for microarray experiments.<sup>3</sup> In the future, it could well be that something resembling a GenBank file with additional biological information will be the “publication” for a bacterial genome sequence, as genome sequencing becomes ever cheaper and easier to perform. Overall, it is important that genome sequence information is released into the public domain in a timely manner so that global scientific progress can be maintained.

**Acknowledgements** DWU, PFH and TTB are supported by grants from the Danish Research Foundation. We are grateful to the Sanger Center for allowing prepublication access to the sequences for the *E. coli* 042 genome (the DNA sequence and annotation files were downloaded from the Sanger web site <http://www.sanger.ac.uk/>).

## References

- Abbott JC, Aanensen DM, Rutherford K, Butcher S, Spratt BG (2005) WebACT—an online companion for the Artemis Comparison Tool. *Bioinformatics* 21(18):3665–3666
- Acinas SG, Marcelino LA, Klepac-Ceraj V, Polz MF (2004) Divergence and redundancy of 16S rRNA sequences in genomes with multiple rrn operons. *J Bacteriol* 186(9):2629–2635
- Alain K, Querellou J, Lesongeur F, Pignet P, Crassous P, Raguene G, Cuff V, Cambon-Bonavita M-A (2002) *Caminibacter hydrogeniphilus* gen. nov., sp. nov., a novel thermophilic, hydrogen-oxidizing bacterium isolated from an East Pacific Rise hydrothermal vent. *Int J Syst Evol Microbiol* 52:1317–1323
- Alm EJ, Huang KH, Price MN, Koche RP, Keller K, Dubchak IL, Arkin AP (2005) The MicrobesOnline Web site for comparative genomics. *Genome Res* 15(7):1015–1022
- Alm RA, Trust TJ (1999) Analysis of the genetic diversity of *Helicobacter pylori*: the tale of two genomes. *J Mol Med* 77(12):834–846 (Review)
- Backhed F, Ley RE, Sonnenburg JL, Peterson DA, Gordon JI (2005) Host–bacterial mutualism in the human intestine. *Science* 307(5717):1915–1920
- Bendtsen JD, Binnewies TT, Hallin PF, Sicheritz-Ponten T, Ussery DW (2005a) Genome update: prediction of secreted proteins in 225 bacterial proteomes. *Microbiology* 151(Pt 6):1725–1727
- Bendtsen JD, Binnewies TT, Hallin PF, Ussery DW (2005b) Genome update: prediction of membrane proteins in prokaryotic genomes. *Microbiology* 151(Pt 7):2119–2121
- Binnewies TT, Hallin PF, Staerfeldt HH, Ussery DW (2004) Genome update: proteome comparisons. *Microbiology* 151(Pt 1):1–4
- Burrus V, Waldor MK (2004) Shaping bacterial genomes with integrative and conjugative elements. *Res Microbiol* 155(5):376–386
- Carattoli A (2001) Importance of integrons in the diffusion of resistance. *Vet Res* 32(3–4):243–259
- Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG, Parkhill J (2005) ACT: the Artemis Comparison Tool. *Bioinformatics* 21(16):3422–3423
- Chen S, Lesnik EA, Hall TA, Sampath R, Griffey RH, Ecker DJ, Blyn LB (2002) A bioinformatics based approach to discover small RNA genes in the *Escherichia coli* genome. *Biosystems* 65(2–3):157–177
- Dobrindt U, Hacker J (2001) Whole genome plasticity in pathogenic bacteria. *Curr Opin Microbiol* 5(4):550–557
- Dobrindt U, Hochhut B, Hentschel U, Hacker J (2004) Genomic islands in pathogenic and environmental microorganisms. *Nat Rev Microbiol* (2):414–424
- Doolittle WF (1999a) Lateral genomics. *Trends Cell Biol* 12(9):M5–M8
- Doolittle WF (1999b) Phylogenetic classification and the universal tree. *Science* 284(284):2124–2129
- Dufraigne C, Fertil B, Lespinats S, Giron A, Deschavanne P (2005) Detection and characterisation of horizontal transfers in prokaryotes using genomic signature. *Nucleic Acids Res* 33(33):e6
- Duponnois R, Ba AM, Mateille T (1999) Beneficial effects of *Enterobacter cloacae* and *Pseudomonas mendocina* for bio-control of *Meloidogyne incognita* with the endospore-forming bacterium *Oasteuria penetrans*. *Nematology* 1(1):95–101
- Ellis RW (1999) New technologies for making vaccines. *Vaccine* 17(13–14):1596–1604
- Falkow S (1975) Infectious multiple drug resistance. Pion Limited, London, England
- Fani R, Brilli M, Lio P (2005) The origin and evolution of operons: the piecewise building of the proteobacterial histidine operon. *J Mol Evol* 60(3):378–390
- Field D, Hughes J (2005) Cataloguing our current genome collection. *Microbiology* 151(Pt 4):1016–1019
- Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, McKenney K, Sutton G, FitzHugh W, Fields C, Gocayne JD, Scott J, Shirley R, Liu LI, Glodek A, Kelley JM, Weidman JF, Phillips CA, Spriggs T, Hedblom E, Cotton MD, Utterback TR, Hanna MC, Nguyen DT, Saudek DM, Brandon RC, Fine LD, Fritchman JL, Fuhrmann JL, Geoghagen NSM, Gnehm CL, McDonald LA, Small KV, Fraser CM, Smith HO, Venter JC (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269(5223):496–498, 507–512
- Fluit AC, Schmitz F-J (2004) Resistance integrons and super-integrons. *Clin Microbiol Infect* 10:272–288
- Fouts DE, Mongodin EF, Mandrell RE, Miller WG, Rasko DA, Ravel J, Brinkac LM, DeBoy RT, Parker CT, Daugherty SC, Dodson RJ, Durkin AS, Madupu R, Sullivan SA, Shetty JU, Ayodeji MA, Shvartsbeyn A, Schatz MC, Badger JH, Fraser CM, Nelson KE (2005) Major structural differences and novel potential virulence mechanisms from the genomes of multiple campylobacter species. *PLoS Biol* 3(1):e15
- Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM, Fritchman RD, Weidman JF, Small KV, Sandusky M, Fuhrmann J, Nguyen D, Utterback TR, Saudek DM, Phillips CA, Merrick JM, Tomb JF, Dougherty BA, Bott KF, Hu PC, Lucier TS, Peterson SN, Smith HO, Hutchison CA 3rd, Venter JC (1995) The minimal gene complement of *Mycoplasma genitalium*. *Science* 270(5235):397–403
- Fraser-Liggett CM (2005) Insights on biology and evolution from microbial genome sequencing. *Genome Res* 15:1603–1610
- Galun E (2003) Transposable elements: a guide to the perplexed and the novice. Kluwer Academic, Dordrecht, The Netherlands, pp 25–73
- Gil R, Latorre A, Moya A (2004) Bacterial endosymbionts of insects: insights from comparative genomics. *Environ Microbiol* 6(11):1109–1122
- Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D, Bibbs L, Eads J, Richardson TH, Noordewier M, Rappe MS, Short JM, Carrington JC, Mathur EJ (2005) Genome streamlining in a cosmopolitan oceanic bacterium. *Science* 309(5738):1242–1245

<sup>3</sup> <http://www.ucl.ac.uk/wibr/services/docs/miamiv1.doc>

- Goebel W, Gross R (2001) Intracellular survival strategies of mutualistic and parasitic prokaryotes. *Trends Microbiol* 9(6):267–273
- Goldmann DA, Klinger JD (1986) *Pseudomonas cepacia*: biology, mechanisms of virulence, epidemiology. *J Pediatr* 108(5 Pt 2):806–812
- Gottesman S (2005) Micros for microbes: non-coding regulatory RNAs in bacteria. *Trends Genet* 7:399–404
- Hallin PF, Ussery DW (2004) CBS genome atlas database: a dynamic storage for bioinformatic results and sequence data. *Bioinformatics* 20(18):3682–3686
- Hallin PF, Binnewies TT, Ussery DW (2004a) Genome update: chromosome atlases. *Microbiology* 150(Pt 10):3091–3093
- Hallin PF, Coenye T, Binnewies TT, Jarmer H, Saerfeldt HH, Ussery DW (2004b) Genome update: correlation of bacterial genomic properties. *Microbiology* 150(Pt 12):3899–3903
- Handelsman J (2004) Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev* 68:669–685
- Harrison A, Dyer DW, Gillaspay A, Ray WC, Mungur R, Carson MB, Zhong H, Gipson J, Gipson M, Johnson LS, Lewis L, Bakaletz LO, Munson RS Jr (2005) Genomic sequence of an otitis media isolate of nontypeable *Haemophilus influenzae*: comparative study with *H. influenzae* serotype d, strain KW20. *J Bacteriol* 187(13):4627–4636
- Hayashi T, Makino K, Ohnishi M, Kurokawa K, Ishii K, Yokoyama K, Han CG, Ohtsubo E, Nakayama K, Murata T, Tanaka M, Tobe T, Iida T, Takami H, Honda T, Sasakawa C, Ogasawara N, Yasunaga T, Kuhara S, Shiba T, Hattori M, Shinagawa H (2001) Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res* 8:11–22
- Holmes AJ, Gillings MR, Nield BS, Mabbutt BC, Nevalainen KM, Stokes HW (2003) The gene cassette metagenome is a basic resource for bacterial genome evolution. *Environ Microbiol* 5(5):383–394
- Horowitz NH (1945) On the evolution of biochemical synthesis. *Proc Natl Acad Sci U S A* 31:153–157
- Horowitz NH (1965) The evolution of biochemical synthesis—retrospect and prospect. In: Bryson V, Vogel HJ (eds) *Evolving genes and proteins*. Academic, New York, pp 15–23
- Itoh T, Takemoto K, Mori H, Gojobori T (1999) Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. *Mol Biol Evol* 3:332–346
- Jacob F, Monod J (1961) Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* 3:318–356
- Jacob F, Perrin D, Sanchez C, Monod J (1960) Operon: a group of genes with the expression coordinated by an operator. *C R Hebd Seances Acad Sci* 250:1727–1729
- Jaffe JD, Stange-Thomann N, Smith C, DeCaprio D, Fisher S, Butler J, Calvo S, Elkins T, FitzGerald MG, Hafez N, Kodira CD, Major J, Wang S, Wilkinson J, Nicol R, Nusbaum C, Birren B, Berg HC, Church GM (2004) The complete genome and proteome of *Mycoplasma mobile*. *Genome Res* 14(8):1447–1461
- Janga SC, Collado-Vides J, Moreno-Hagelsieb G (2005) Nebulon: a system for the inference of functional relationships of gene products from the rearrangement of predicted operons. *Nucleic Acids Res* 33(8):2521–2530
- Jores J, Rumer L, Wieler LH (2004) Impact of the locus of enterocyte effacement pathogenicity island on the evolution of pathogenic *Escherichia coli*. *Int J Med Microbiol* 294(2–3):103–113 (Review)
- Juhala RJ, Ford ME, Duda RL, Youlton A, Hatfull GF, Hendrix RW (2000) Genomic sequences of bacteriophages HK97 and HK022: pervasive genetic mosaicism in the lambdoid bacteriophages. *J Mol Biol* 299(1):27–51
- Kennedy SP, Ng WV, Salzberg SL, Hood L, DasSarma S (2001) Understanding the adaptation of *Halobacterium* species NRC-1 to its extreme environment through computational analysis of its genome sequence. *Genome Res* 11:1641–1650
- Kiil K, Binnewies TT, Sicheritz-Ponten T, Willenbrock H, Hallin PF, Wassenaar TM, Ussery DW (2005a) Genome update: sigma factors in 240 bacterial genomes. *Microbiology* 151(Pt 10):3147–3150
- Kiil K, Ferchaud JB, David C, Binnewies TT, Wu H, Sicheritz-Ponten T, Willenbrock H, Ussery DW (2005b) Genome update: distribution of two-component transduction systems in 250 bacterial genomes. *Microbiology* 151(Pt 11):3447–3452
- Kong H, Lin L-F, Porter N, Stickel S, Byrd D, Posfai J, Roberts RJ (2000) Functional analysis of putative restriction–modification system genes in the *Helicobacter pylori* J99 genome. *Nucleic Acids Res* 28:3216–3223
- Kummerfeld SK, Teichmann SA (2006) DBD: a transcription factor prediction database. *Nucleic Acids Res* 34(Database issue):D74–D81
- Kunin V, Goldovsky L, Darzentas N, Ouzounis CA (2005) The net of life: reconstructing the microbial phylogenetic network. *Genome Res* 15(7):954–959
- Kuwahara T, Yamashita A, Hirakawa H, Nakayama H, Toh H, Okada N, Kuhara S, Hattori M, Hayashi T, Ohnishi Y (2004) Genomic analysis of *Bacteroides fragilis* reveals extensive DNA inversions regulating cell surface adaptation. *Proc Natl Acad Sci U S A* 101(41):14919–14924
- Lawrence JG, Roth JR (1996) Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics* 143(4):1843–1860
- Lazcano A, Diaz-Villagomez E, Mills T, Oro J (1995) On the levels of enzymatic substrate specificity: implications for the early evolution of metabolic pathways. *Adv Space Res* 15(3):345–356
- Lewis M, Chang G, Horton NC, Kercher MA, Pace HC, Schumacher MA, Brennan RG, Lu P (1996) Crystal structure of the lactose operon repressor and its complexes with DNA and inducer. *Science* 271(5253):1247–1254
- Ley RE, Backhed F, Turnbaugh P, Lozupone CA, Knight RD, Gordon JI (2005) Obesity alters gut microbial ecology. *Proc Natl Acad Sci U S A* 102(31):11070–11075
- Lin L-F, Posfai J, Roberts RJ, Kong H (2001) Comparative genomics of the restriction–modification systems in *Helicobacter pylori*. *Proc Natl Acad Sci U S A* 98:2740–2745
- Lobner-Olesen A, Skovgaard O, Marinus MG (2005) Dam methylation: coordinating cellular processes. *Curr Opin Microbiol* 8(2):154–160
- Lund O, Nielsen M, Kesmir C, Christensen JK, Lundegaard C, Worning P, Brunak C (2002) Web-based tools for vaccine design. In: Korber BT, Brander C, Haynes BF, Koup R, Kuiken C, Moore JP, Walker BD, Watkins D (eds) *HIV molecular immunology*. Los Alamos, NM, pp 45–51
- Lund O, Nielsen M, Lundegaard C, Kesmit C, Brunak S (2005) *Immunological bioinformatics*. MIT, Cambridge, Massachusetts
- Lupski JR, Weinstock GM (1992) Short, interspersed repetitive DNA sequences in prokaryotic genomes. *J Bacteriol* 174(14):4525–4529
- Maas R (2004) Prereplicative purine methylation and postreplicative demethylation in each DNA duplication of the *Escherichia coli* replication cycle. *J Biol Chem* 279(49):51568–51573
- Mahillon J, Leonard C, Chandler M (1999) IS elements as constituents of bacterial genomes. *Res Microbiol* 150:675–687
- Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437(7057):376–380
- McClintock B (1950) The origin and behavior of mutable loci in maize. *Proc Natl Acad Sci U S A* 36(6):344–355
- McGillivray G, Tomaras AP, Rhodes ER, Actis LA (2005) Cloning and sequencing of a genomic island found in the Brazilian purpuric fever clone of *Haemophilus influenzae* biogroup aegyptius. *Infect Immun* 73(4):1927–1938

- Middendorf B, Hochhut B, Leipold K, Dobrindt U, Blum-Oehler G, Hacker J (2004) Instability of pathogenicity islands in uropathogenic *Escherichia coli* 536. *J Bacteriology* 186 (10):3086–3096
- Mongodin EF, Emerson JB, Nelson KE (2005) Microbial metagenomics. *Genome Biol* 6(10):347
- Mullis K, Faloona F, Scharf S, Saiki R, Horn G, Erlich H (1986) Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harb Symp Quant Biol* 51(Pt 1):263–273
- Nagy Z, Chandler M (2004) Regulation of transposition in bacteria. *Res Microbiol* 155:387–398
- Nishi T, Ikemura T, Kanaya S (2005) GeneLook: a novel ab initio gene identification system suitable for automated annotation of prokaryotic sequences. *Gene* 346:115–125
- Novikova N, De Boever P, Poddubko S, Deshevaya E, Polikarpov N, Rakova N, Coninx I, Mergeay M (2006) Survey of environmental biocoenosis on board the International Space Station. *Res Microbiol* 157(1):5–12
- Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial evolution. *Nature* 405:299–304
- Ohnishi M, Kurokawa K, Hayashi T (2001) Diversification of *Escherichia coli* genomes: are bacteriophages the major contributors? *Trends Microbiol* 9:481–485
- Okuda S, Katayama T, Kawashima S, Goto S, Kanehisa (2006) MODB: a database of operons accumulating known operons across multiple genomes. *Nucleic Acids Res* 34(Database issue):D358–362
- Olsen GJ, Lane DJ, Giovannoni SJ, Pace NR, Stahl DA (1986) Microbial ecology and evolution: a ribosomal RNA approach. *Annu Rev Microbiol* 40:337–365
- O'Malley MA, Bostanci A, Calvert J (2005) Whole-genome patenting. *Nat Rev Genet* 6(6):502–506
- Ortutay C, Gaspari Z, Toth G, Jager E, Vida G, Orosz L, Vellai T (2003) Speciation in *Chlamydia*: genome-wide phylogenetic analyses identified a reliable set of acquired genes. *J Mol Evol* 57:672–680
- Ou HY, Chen LL, Lonnen J, Chaudhuri RR, Thani AB, Smith R, Garton NJ, Hinton J, Pallen M, Barer MR, Rajakumar K (2006) A novel strategy for the identification of genomic islands by comparative analysis of the contents and contexts of tRNA sites in closely related bacteria. *Nucleic Acids Res* 34(1):e3
- Pal C, Hurst LD (2004) Evidence against the selfish operon theory. *Trends Genet* 20(6):232–234
- Parkhill J, Sebahia M, Preston A, Murphy LD, Thomson N, Harris DE, Holden MT, Churcher CM, Bentley SD, Mungall KL, Cerdeno-Tarraga AM, Temple L, James K, Harris B, Quail MA, Achtman M, Atkin R, Baker S, Basham D, Bason N, Cherevach I, Chillingworth T, Collins M, Cronin A, Davis P, Doggett J, Feltwell T, Goble A, Hamlin N, Hauser H, Holroyd S, Jagels K, Leather S, Moule S, Norberczak H, O'Neil S, Ormond D, Price C, Rabinowitsch E, Rutter S, Sanders M, Saunders D, Seeger K, Sharp S, Simmonds M, Skelton J, Squares R, Squares S, Stevens K, Unwin L, Whitehead S, Barrell BG, Maskell DJ (2003) Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. *Nat Genet* 35(1):32–40
- Paulsen IT, Banerjee L, Myers GSA, Nelson KE, Seshadri R, Read TD, Fouts DE, Eisen JA, Gill SR, Heidelberg JF, Tettelin H, Dodson RJ, Umayam L, Brinkac L, Beanan M, Daugherty S, DeBoy RT, Durkin S, Kolonay J, Madupu R, Nelson W, Vamathevan J, Tran B, Upton J, Hansen T, Shetty J, Khouri H, Utterback T, Radune D, Ketchum KA, Dougherty BA, Fraser CM (2003) Role of mobile DNA in the evolution of vancomycin-resistant *Enterococcus faecalis*. *Science* 299(5615):2071–2074
- Pedersen AG, Jensen LJ, Brunak S, Staerfeldt HH, Ussery DW (2000) A DNA structural atlas for *Escherichia coli*. *J Mol Biol* 299(4):907–930
- Pennisi E (2005) Biochemistry. Cut-rate genomes on the horizon? *Science* 309(5736):862
- Penyalver R, Lopez MM (1999) Cocolonization of the rhizosphere by pathogenic agrobacterium strains and nonpathogenic strains K84 and K1026, used for crown gall biocontrol. *Appl Environ Microbiol* 65(5):1936–1940
- Peters EDJ, Leverstein-Van Hall MA, Box ATA, Verhoef J, Fluit AC (2001) Novel gene cassettes and integrons. *Antimicrob Agents Chemother* 45(10):2961–2964
- Pizza M, Scarlato V, Masignani V, Giuliani MM, Arico B, Comanducci M, Jennings GT, Baldi L, Bartolini E, Capecchi B, Galeotti CL, Luzzi E, Manetti R, Marchetti E, Mora M, Nuti S, Ratti G, Santini L, Savino S, Scarselli M, Storni E, Zuo P, Broecker M, Hundt E, Knapp B, Blair E, Mason T, Tettelin H, Hood DW, Jeffries AC, Saunders NJ, Granoff DM, Venter JC, Moxon ER, Grandi G, Rappuoli R (2000) Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. *Science* 287:1816–1820
- Prescott L, Harvey JP, Klein DA (1999) *Microbiology*, 4th edn. McGraw-Hill, New York, USA
- Price MN, Huang KH, Alm EJ, Arkin AP (2005) A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Res* 33(3):880–892
- Rappuoli R (2001) Reverse vaccinology, a genome-based approach to vaccine development. *Vaccine* 19:2688–2691
- Rendulic S, Jagtap P, Rosinus A, Eppinger M, Baar C, Lanz C, Keller H, Lambert C, Evans KJ, Goesmann A, Meyer F, Sockett RE, Schuster SC (2004) A predator unmasked: life cycle of *Bdellovibrio bacteriovorus* from a genomic perspective. *Science* 303(5658):689–692
- Reznikoff WS (1992) The lactose operon-controlling elements: a complex paradigm. *Mol Microbiol* 6(17):2419–2422
- 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
- Roberts RJ, Vincze T, Psfai J, Macelis D (2005) REBASE—restriction enzymes and DNA methyl transferases. *Nucleic Acids Res* 33:D230–D232
- Rocha EPC, Danchin A, Viari A (1999) Functional and evolutionary role of long repeats in prokaryotes. *Res Microbiol* 150:725–733
- Rogozin IB, Makarova KS, Wolf YI, Koonin EV (2004) Computational approaches for the analysis of gene neighbourhoods in prokaryotic genomes. *Brief Bioinform* 5(2):131–149
- Rosenfeld JA, Sarkar IN, Planet PJ, Figurski DH, DeSalle R (2004) ORFcurator: molecular curation of genes and gene clusters in prokaryotic organisms. *Bioinformatics* 20(18):3462–3465
- Salgado H, Gama-Castro S, Peralta-Gil M, Diaz-Peredo E, Sanchez-Solano F, Santos-Zavaleta A, Martinez-Flores I, Jimenez-Jacinto V, Bonavides-Martinez C, Segura-Salazar J, Martinez-Antonio A, Collado-Vides J (2006a) RegulonDB (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res* 34(Database issue):D394–D397
- Salgado H, Santos-Zavaleta A, Gama-Castro S, Peralta-Gil M, Penaloza-Spinola MI, Martinez-Antonio A, Karp PD, Collado-Vides J (2006b) The comprehensive updated regulatory network of *Escherichia coli* K-12. *BMC Bioinformatics* 7(1):5
- Sanger F, Donelson JE, Coulson AR, Kossel H, Fischer D (1973) Use of DNA polymerase I primed by a synthetic oligonucleotide to determine a nucleotide sequence in phage  $\phi$ 1 DNA. *Proc Natl Acad Sci U S A* 70(4):1209–1213
- Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM, Smith M (1977) Nucleotide sequence of bacteriophage  $\phi$ 1 X174 DNA. *Nature* 265(5596):687–695

- Schmidt H, Hensel M (2004) Pathogenicity islands in bacterial pathogenesis. *Clin Microbiol Rev* 17(1):14–56
- Schneider G, Dobrindt U, Bruggemann H, Nagy G, Janke B, Blum-Oehler G, Buchrieser C, Gottschalk G, Emody L, Hacker J (2004) The pathogenicity island-associated K15 capsule determinant exhibits a novel genetic structure and correlates with virulence in uropathogenic *Escherichia coli* strain 536. *Infect Immun* 72(10):5993–6001
- Serruto D, Adu-Bobie J, Capecchi B, Rappuoli R, Pizza M, Masignani V (2004) Biotechnology and vaccines: application of functional genomics to *Neisseria meningitidis* and other bacterial pathogens. *J Biotechnol* 113:15–32
- Sharp PM, Li WH (1987) The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* 15(3):1281–1295
- Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 309(5741):1728–1732
- Shimizu T, Ohtani K, Hirakawa H, Ohshima K, Yamashita A, Shiba T, Ogasawara N, Hattori M, Kuhara, Hayashi H (2002) Complete genome sequence of *Clostridium perfringens*, an anaerobic flesh-eater. *Proc Natl Acad Sci U S A* 99(2):996–1001
- Skovgaard M, Jensen LJ, Brunak S, Ussery D, Krogh A (2001) On the total number of genes and their length distribution in complete microbial genomes. *Trends Genet* 17(8):425–428
- Stahl FW, Murray NE (1966) The evolution of gene clusters and genetic circularity in microorganisms. *Genetics* 53(3):569–576
- Starlinger P, Saedler H (1976) IS-elements in microorganisms. *Curr Top Microbiol Immunol* 75:111–152
- Talarico S, Cave MD, Marrs CF, Foxman B, Zhang L, Yang Z (2005) Variation of the *Mycobacterium tuberculosis* PE\_PGRS 33 gene among clinical isolates. *J Clin Microbiol* 43(10):4954–4960
- Taoka M, Yamauchi Y, Shinkawa T, Kaji H, Motohashi W, Nakayama H, Takahashi N, Isobe T (2004) Only a small subset of the horizontally transferred chromosomal genes in *Escherichia coli* are translated into proteins. *Mol Cell Proteomics* 3(8):780–787
- Tobes R, Ramos JL (2005) REP code: defining bacterial identity in extragenic space. *Environ Microbiol* 7(2):225–228
- Toh H, Weiss BL, Perkin SA, Yamashita A, Oshima K, Hattori M, Aksoy S (2006) Massive genome erosion and functional adaptations provide insights into the symbiotic lifestyle of *Sodalis glossinidius* in the tsetse host. *Genome Res* 16:149–156
- Tomb JF, White O, Kerlavage AR, Clayton RA, Sutton GG, Fleischmann RD, Ketchum KA, Klenk HP, Gill S, Dougherty BA, Nelson K, Quackenbush J, Zhou L, Kirkness EF, Peterson S, Loftus B, Richardson D, Dodson R, Khalak HG, Glodek A, McKenney K, Fitzgerald LM, Lee N, Adams MD, Hickey EK, Berg DE, Gocayne JD, Utterback TR, Peterson JD, Kelley JM, Cotton MD, Weidman JM, Fujii C, Bowman C, Watthey L, Wallin E, Hayes WS, Borodovsky M, Karp PD, Smith HO, Fraser CM, Venter JC (1997) The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* 388(6642):539–547
- Torsvik V, Salte K, Sorheim R, Goksoyr J (1990) Comparison of phenotypic diversity and DNA heterogeneity in a population of soil bacteria. *Appl Environ Microbiol* 56:776–781
- Tringe SG, Rubin EM (2005) Metagenomics: DNA sequencing of environmental samples. *Nat Rev Genet* 6(11):805–814
- Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428(6978):37–43
- Ussery DW, Hallin PF (2004a) Genome update: AT content in sequenced prokaryotic genomes. *Microbiology* 150(Pt 4):749–752
- Ussery DW, Hallin PF (2004b) Genome update: length distributions of sequenced prokaryotic genomes. *Microbiology* 150(Pt 3):513–516
- Ussery DW, Binnewies TT, Gouveia-Oliveira R, Jarmer H, Hallin PF (2004a) Genome update: DNA repeats in bacterial genomes. *Microbiology* 150(Pt 11):3519–3521
- Ussery DW, Hallin PF, Lagesen K, Coenye T (2004b) Genome update: rRNAs in sequenced microbial genomes. *Microbiology* 150(Pt 5):1113–1115
- Ussery DW, Hallin PF, Lagesen K, Wassenaar TM (2004c) Genome update: tRNAs in sequenced microbial genomes. *Microbiology* 150(Pt 6):1603–1606
- Ussery DW, Tindbaek N, Hallin PF (2004d) Genome update: promoter profiles. *Microbiology* 150(Pt 9):2791–2793
- Vallenet D, Labarre L, Rouy Z, Barbe V, Bocs S, Cruveiller S, Lajus A, Pascal G, Scarpelli C, Medigue C (2006) MaGe: a microbial genome annotation system supported by synteny results. *Nucleic Acids Res* 34(1):53–65
- van Belkum A, Scherer S, van Alphen L, Verbrugh H (1998) Short sequence DNA repeats in prokaryotic genomes. *Microbiol Mol Biol Rev* 62(2):275–293
- van der Meer JR, Sentchilo V (2003) Genomic islands and the evolution of catabolic pathways in bacteria. *Curr Opin Biotechnol* 14:248–254
- Van Domselaar GH, Stothard P, Shrivastava S, Cruz JA, Guo A, Dong X, Lu P, Szafron D, Greiner R, Wishart DS (2005) BASys: a web server for automated bacterial genome annotation. *Nucleic Acids Res* 33(Web Server issue):W455–W459
- Venter JC, 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 YH, Smith HO (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304(5667):66–74
- Vezi A, Campanaro S, D'Angelo M, Simonato F, Vitulo N, Lauro FM, Cestaro A, Malacrida G, Simonati B, Cannata N, Romualdi C, Bartlett DH, Valle G (2005) Life at depth: *Photobacterium profundum* genome sequence and expression analysis. *Science* 307(5714):1459–1461
- Willenbrock H, Binnewies TT, Hallin PF, Ussery DW (2005) Genome update: 2D clustering of bacterial genomes. *Microbiology* 151(Pt 2):333–336
- Worning P, Jensen LJ, Nelson KE, Brunak S, Ussery DW (2000) Structural analysis of DNA sequence: evidence for lateral gene transfer in *Thermotoga maritima*. *Nucleic Acids Res* 28(3):706–709
- Worning P, Jensen LJ, Hallin PF, Stærfeldt H-H, Ussery DW (2006) Origin of replication in circular prokaryotic chromosomes. *Environ Microbiol* (In press)
- Yan F, Polk DB (2004) Commensal bacteria in the gut: learning who our friends are. *Curr Opin Gastroenterol* 20(6):565–571
- Zagursky RJ, Russell D (2001) Bioinformatics: use in bacterial vaccine discovery. *Biotechniques* 31:636–659
- Zhang R, Zhang CT (2004) A systematic method to identify genomic islands and its applications in analyzing the genomes of *Corynebacterium glutamicum* and *Vibrio vulnificus* CMCP6 chromosome I. *Bioinformatics* 20(5):612–622
- Zheng Y, Anton BP, Roberts RJ, Kasif S (2005) Phylogenetic detection of conserved gene clusters in microbial genomes. *BMC Bioinformatics* 6:243
- Zubrzycki IZ (2004) Analysis of the products of genes encompassed by the theoretically predicted pathogenicity islands of *Mycobacterium tuberculosis* and *Mycobacterium bovis*. *Proteins: Struct, Funct, Bioinf* 54:563–568