ELSEVIER

Full text provided by www.sciencedirect.com
ScienceDirect

How do we compare hundreds of bacterial genomes?

Dawn Field, Gareth Wilson and Christopher van der Gast

The genomic revolution is fully upon us in 2006 and the pace of discovery is set to accelerate with the emergence of ultra-high-throughput sequencing technologies. Our complete genome collection of bacteria and archaea continues to grow in number and diversity, as genome sequencing is applied to an array of new problems, from the characterization of the pan-genome to the detection of mutation after experimentation and the exploration of microbial communities in unprecedented detail. The benefits of large-scale comparative genomic analyses are driving the community to think about how to manage our public collections of genomes in novel ways.

Addresses
Oxford Centre for Ecology and Hydrology, Mansfield Road,
Oxford OX1 3SR, UK

Corresponding author: Field, Dawn (dfield@ceh.ac.uk)

Current Opinion in Microbiology 2006, **9**:499–504

This review comes from a themed issue on
Genomics
Edited by David W Ussery and Timothy D Read

Available online 30th August 2006

1369-5274/$ – see front matter

DOI 10.1016/j.mib.2006.08.008

Introduction

Twenty years ago a student might have earned a doctorate from the sequencing and analysis of a single gene. Although not as routine as it might well be in the future, it is now feasible to require a student to earn a PhD by generating and analysing one or more genomes. Likewise, only ten years ago, most software for manipulating pieces of DNA (e.g. to find restriction sites for recombinant DNA studies) could not handle a sequence of 100 kb, the size of a large, but not giant, phage (http://giantvirus.org); yet now we have vast numbers of tools and databases for the manipulation of complete genomes [1].

The number of bacterial and archaeal genomes has grown exponentially in the past decade. It has doubled in the past two years and we now have more than 300 completed genomes from these two domains of life in public databases [2•]. We expect to have at least 1000 draft genomes within a year or two [3•]. Our capacity to generate and analyze genome sequences has grown at an astonishing rate, but we are on the cusp of yet another leap forward, with the advent of a new family of ultra-high-throughput, low-cost sequencing methods [4•,5,6••]. This revolution is creating new opportunities as well as challenges. High-quality comparative genomic analysis of hundreds of genomes depends on the availability of a framework with three parts: a large collection of genomes; adequate tools and analysis techniques; and relevant and pressing scientific questions. Here, we review key contributions, highlight challenges and speculate on what the future of large-scale comparative genomics will hold.

Our complete genome collection

Our ability to draw useful comparisons from hundreds of genomes depends on the taxonomic and ecological composition of our available genome collection. The number of complete genomes is continuing to grow into the hundreds (and thousands for viruses, organelles, and plasmids) and yet is still essentially a disjointed collection of isolates, which have been sequenced for a large variety of reasons. The most obvious reason is to mine genomes for the benefit of human health and wealth, and biases towards the sequencing of pathogens and organisms of economic consequence are clearly evident [7]. This bias is now being balanced by interest in isolates from the environment [8••], and the genome collection of the future will be vastly richer and more complex in terms of evolutionary and ecological diversity than the current one. New sequencing technologies are making possible the sequencing of random community DNA and single cells of bacteria, without the need for cloning or laboratory cultivation [4•,5,6••]. These technologies are set to revolutionize fields as diverse as microbial ecology [9] and human health, as, for example, researchers explore the metabolic capacity of microbial communities in the human gut [10,11••].

The number of closely related genomes is also set to increase, further compounding existing biases, but also opening up the possibility of detailed studies of the genomic evolution over even the smallest time scales. The concept of the pan-genome, or the fact that the gene pool of many microbial species is far larger than the number of genes found in any single genome, is one of the most important discoveries of the genomic era. A landmark study of seven *Streptococcus agalactiae* genomes (five of which were generated *de novo* for the purpose of this study) has proved that sampling more genomes would continue to reveal new genes — on average 33 per genome [12••]. A subsequent analysis of seven *Escherichia coli* genomes predicts that an average of 441 new genes will be provided by the sequencing of each new isolate [13•]. Mathematical modeling suggests that hundreds of genomes of other species will follow the same trend [14•]. The significance of the pan-genome is not yet understood, but it might be involved in niche adaptation [15].

Sequencing now offers a realistic alternative to the use of comparative genomic hybridization (CGH) microarrays, which, unlike genome sequencing, cannot detect novel genes. In fact, genome sequencing is currently being used as a tool for unraveling the molecular basis of phenotypic differences among strains that diverged by as little as 200 generations [4•]. Strikingly, a recent study of *Myxococcus xanthus* has recently revealed a single mutation in the huge 9.14 Mb genome of an evolved strain [16] that can transform an evolutionary cheater to a social cooperator [17].

A new chapter in comparative genomics is also about to be written as we move from using metagenomics to assay natural microbial communities, to the ambitious and perhaps more useful practice of 'community whole genome sequencing'. For example, whereas the human gut metagenome has been explored using metagenomic approaches [11••], the Human Gut Microbiome Initiative (HGMI) aims to produce deep draft genomes of 100 intestinal species (http://www.genome.gov/Pages/Research/Sequencing/SeqProposals/HGMISeq.pdf). This approach enables the study of the total genes involved in producing the metabolic capacity of a community, the levels of redundancy, the rates of horizontal transfer based on confirmed proximity of species, and the role of the pan-genome in bacterial adaptation.

Such community-level studies serve as a stark reminder that we have only just begun to sample the natural microbial diversity [7]. Table 1 lists the members of a simple community found in sump-tanks in an engineering workshop [18]. This low-diversity community (47 species in total) contains many pathogens and is dominated by proteobacteria (the most sampled bacterial division with respect to genome sequencing studies). Although these are the two most significant biases in our current genome collection, [7] this community still contains a significant proportion of as-of-yet unsampled genera and species, and a few species of particular interest (e.g. *E. coli*). This would be a fascinating community to sequence because it provides further evidence that the species–area relationship applies to microbes [18,19]. Using this community, for example, it would be possible to test for a corresponding 'gene–area' relationship, and, if it exists, to characterize it.

Requirements for the comparison of hundreds of genomes

The informatics associated with annotating or analyzing a hundred, or even one thousand, genomes instead of one might seem daunting to many, but to a growing number of researchers with adequate resources it is vastly preferable because of the power of comparative methods [3•]. Handling genomic data is becoming increasingly easy, because the software, databases and analysis tools that have emerged over the past years become tried-and-tested, new scaled-up resources are developed and we gain an improved community infrastructure for enabling the analysis of hundreds of genomes [1].

Certainly in the case of annotation, quality improves with the number of relevant genomes available for comparison, especially with respect to accurate gene prediction [20•]. Likewise, many biological patterns only become 'visible' through comparative genomic approaches, such as gene fusions [21], pseudogenes [22] or the non-coding RNAs of the 'RNome' [23]. This, of course, also extends to the study of orphan and lineage-specific genes, which can only be properly characterized in light of many related sequences. Take, for example, the fact that genomes with the largest numbers of orphans are often those that are the most evolutionarily or ecologically distinct [24,25]. Even for well-characterized genomes, the large number of uncharacterized and orphan genes can be a major bottleneck, for example in the study of pathogenesis [26].

Computational studies of hundreds of genomes depend on high quality annotations that must be directly comparable. However, genomic annotations, when deposited into the databases (DDBJ [http://www.ddbj.nig.ac.jp/; Japan], EMBL [http://www.ebi.ac.uk/embl/; UK] and GenBank [http://www.ncbi.nlm.nih.gov/Genbank/index.html; USA]) of the International Nucleotide Sequence Database Collaboration (INSDC) vary in their level of detail, choice of terms and language and the exact types of features reported (e.g. proteins, tRNAs, ribosomal operons and repeats) [27]. The solution to this problem is the development of new databases that combine and standardize information from a variety of sources and apply uniform reannotation techniques [28,29•,30•].

However, the standardization of *in silico* resources is not enough. We must also improve annotations through empirical work on specific loci [31••]. For example, an increasing number of annotations are being validated by transcriptomic [32•] or proteogenomic experiments that verify the expression, start and stop positions of proteins [32•]. To support the integration of empirical annotations with submitted public genome sequences, the INSDC has developed a Third Party Annotation project that collects peer-reviewed reannotations of genomic sequences from anyone in the community [33].

Recognition of the benefits of expert curation and the value of having a large collection of genomes with which to create an 'annotation and analysis environment' are driving a paradigm shift in the way we build annotation tools [34] and maintain genomic databases [3•]. The SEED database (http://theseed.uchicago.edu/FIG/index.cgi) aims to use expert curation of various pathways and traits to annotate the first 1000 genomes [3•]. Thus,

Table 1

Species composition of a sump-tank community[a].

Species				C		CU		O		Total	
Class	**Family**	**Genus**	**Species**	**G**	**S**	**G**	**S**	**G**	**S**	**G**	**S**
Actinobacteria	Microbacteriaceae	*Rathayibacter*	*rathayi*								
	Micrococcaceae	*Arthrobacter*	*atrocyaneus*					2		2	
			aurescens					2	1	2	1
		Kocuria	*rosea*					1		1	
		Micrococcus	*luteus*					1		1	
			lylae					1		1	
Bacilli	Bacillaceae	*Bacillus*	*psychrosaccharolyticus*	9		5		42		56	
Alphaproteobacteria	Brucellaceae	*Ochrobactrum*	*anthropi*					1	1	1	1
Betaproteobacteria	Alcaligenaceae	*Achromobacter*	*xylosoxidans*								
		Acidovorax	*avenae*					3	1	3	1
			delafieldii					3		3	
			facilis					3		3	
	Neisseriaceae	*Neisseria*	*mucosa*	2		4		5		11	
	Oxalobacteraceae	*Janthinobacterium*	*lividum*					1		1	
Gammaproteobacteria	Altermonadaceae	*Shewanella*	*putrefaciens*	1		2		18	4	21	4
	Enterobacteriaceae	*Cedecea*	*davisae*								
		Citrobacter	*freundii*					2		2	
			koseri					2	1	2	1
		Enterobacter	*aerogenes*			1		2		3	
			amnigenus			1		2		3	
			cloacae			1	1	2		3	1
			hormaechei			1		2		3	
			intermedius			1		2		3	
			pyrinus			1		2		3	
		Escherichia	*coli*	5	5	4	4	28	26	37	35
		Klebsiella	*planticola*			1		3		4	
			pneumoniae			1	1	3	2	4	3
		Kluyvera	*ascorbata*								
			cryocrescens								
		Leclercia	*adecarboxylata*								
		Morganella	*morganii*								
		Pantoea	*agglomerans*			1		2		3	
		Pectobacterium	*chrysanthemi*					1	1	1	1
		Proteus	*vulgaris*			1		1		2	
		Salmonella	*typhimurium*	5	1	2		16	4	23	5
		Serratia	*marcescens*			1	1	1		2	1
			odorifera			1		1		2	
	Moraxellaceae	*Acinetobacter*	*johnsonii*	1		1		4		6	
	Pseudomonadaceae	*Pseudomonas*	*aeruginosa*	7	1	3	1	13	6	23	8
			alcaligenes	7		3		13		23	
			balearica	7		3		13		23	
			mendocina	7		3		13	1	23	1
			pseudoalcaligenes	7		3		13		23	
			putida	7	1	3		13	4	23	5
			stutzeri	7		3		13	1	23	1
	Vibrionaceae	*Vibrio*	*parahaemolyticus*	4	1	1		18		23	1
	Xanthomonadaceae	*Stenotrophomonas*	*maltophilia*			1	1	2	1	3	2

[a] Each species in a well-characterized environment [18,unpublished data] was compared with the GOLD [37•] to determine the number of complete published (C), complete unpublished (CU), and ongoing (O) genome projects at the genus (G) and species (S) level. This shows that 7 of the 47 species in this community are not represented at the genus level, and 23 are represented at the genus level by at least one genome project but not at the species level.

users of this approach have the best possible annotations available for downstream use, for example to detect duplicated or missing genes within known pathways [35] or to detect novel differences in metabolic capacity between two or more genomic datasets [36•].

The need to capture expert curations extends to top level information describing genomes. For example, the Genomes Online Database (GOLD; http://www.genomesonline.org/), the 'gold' standard for tracking ongoing and completed genome projects, has recently integrated into its version 2.0 a large amount of curated metadata from the literature [37•]. Further, this information is available for download and is imported into the Integrated Microbial Genomes system [30•]. The need for more metadata to describe collections of genomes and the benefits of this, especially in the realm of eco-genomic studies, has recently been reviewed [38••]. A call for improved capture of additional information at the time of submission of genomes to the INSDC [39] has led to the formation of the Genomic Standards Consortium (GSC; http://gensc.sf.net). This international group is working together towards the creation of a richer set of descriptions of complete genomes and metagenomes [40]. Two key drivers for the extension of metadata capture, in addition to the large number of environmental genomes and metagenomes, are the fact that far more genomes in the future will only be completed to draft stage and the emergence of new sequencing technologies.

What questions can be addressed through the comparison of hundreds of genomes?

With a good infrastructure supporting the analysis of large collections of genomes, the number of computational studies that can be imagined becomes boundless. The true power of large-scale comparative genomic studies lies in their ability to identify and characterize biological trends (or even rules) that explain particular phenomena or that highlight interesting exceptions [7]. For example, with a coding capacity of only 51%, the genome of *Sodalis glossinidius*, which is evolving from a free-living bacterium to a mutualist endosymbiont, is a classic exception to the rule that gene density is conserved across bacteria from the smallest to the largest genome sizes [22].

The comparative approach can be extended to explore and characterize any pattern that is widely shared among microbes, and improves in power with the number of genomes included. Such patterns include the distribution of a range of structural features [41•], the global characteristics of proteomes [42], the abundances of repetitive sequences [43], the relative abundances of specific types of genes, such as two-component systems [44], or the speed and action of particular processes, such as gene movement and loss of synteny [45•]. Large-scale computational studies can also be used to search for relationships between genomic features and ecology [7,29•], reconstruct evolutionary relationships among genomes [46•,47••], and explore the concept of a bacterial species [48••]. The comparative approach yields fundamental insights into the function and evolution of genomes, but can also lead to practical results. For example, understanding interactions between phage and bacteria through comparative genomic studies has use in engineering widespread phage protection for industrially important bacteria used in bioprocessing activities (e.g. fermentation) [49].

Conclusions and future outlook

The analysis of microbial genomes is continuing to shed light on our fundamental understanding of microbiology [50•]. Although comparative genomic studies of hundreds of genomes are still relatively rare compared with comparative genomic studies of particular groups of bacteria, they are rapidly increasing in number. Closing the gap between our ability to generate vast quantities of data using computational methods and our ability to ensure resulting annotation and analyses of the highest quality (especially through curation) will be a major goal of the next decade. If the community can continue to provide critical stewardship of its complete genome collection [40], this will open the doors to ever more powerful comparative genomic studies, especially in the future, as whole genome collections from natural microbial communities and evolutionary time-series studies become available.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Field D, Feil E, Wilson G: **Databases and software for the comparisons of prokaryotic genomes**. *Microbiology* 2005, **151**:2125-2132.

2. • Binnewies TT, Motro Y, Hallin PF, Lund O, David Dunn D, La T, Hampson DJ, Bellgard M, Wassenaar TM, Ussery DW: **Ten years of bacterial genome sequencing: 10 comparative-genomics-based discoveries**. *Funct Integr Genomics* 2006, **6**:165-185.

A rich overview of the outcomes of comparative genomic studies over the past years.

3. • Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, de Crecy-Lagard V, Diaz N, Disz T, Edwards R *et al.*: **The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes**. *Nucleic Acids Res* 2005, **33**:5691-5702.

This database represents a new chapter in the way in which we annotate genomes. Annotations in the SEED database are the result of expert curations of single 'subsystems' (a pathway or particular function) across all available genomes.

4. • Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM: **Accurate multiplex polony sequencing of an evolved bacterial genome**. *Science* 2005, **309**:1728-1732.

Proof of principle of one of the ultra-high-throughput sequencing technologies applied to the sequencing of a strain of *E. coli*, which was shown to diverge from its last sequenced ancestor by a mere 200 generations.

5. Clarke SC: **Pyrosequencing: nucleotide sequencing technology with bacterial genotyping applications**. *Expert Rev Mol Diagn* 2005, **5**:947-953.

6. •• Edwards RA, Rodriguez-Brito B, Wegley L, Haynes M, Breitbart M, Peterson DM, Saar MO, Alexander S, Alexander EC Jr, Rohwer F: **Using pyrosequencing to shed light on deep mine microbial ecology under extreme hydrogeologic conditions**. *BMC Genomics* 2006, **7**:57.

The first application of pyrosequencing technology for the characterization of the metagenomes of two microbial communities.

7. Martiny JBH, Field D: **Ecological perspectives on our complete genome collection**. *Ecol Lett* 2005, **8**:1334-1345.

8. •• Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu DY, Paulsen I, Nelson KE, Nelson W *et al.*: **Environmental genome shotgun sequencing of the Sargasso Sea**. *Science* 2004, **304**:66-74.

Landmark study that opened the doors to the investigation of natural microbial communities using shotgun sequencing. It includes attempted reconstructions of genomes from dominant species.

9. Xu J: **Microbial ecology in the age of genomics and metagenomics: concepts, tools, and recent advances**. *Mol Ecol* 2006, **15**:1713-1731.

10. Furrie E: **A molecular revolution in the study of intestinal microflora**. *Gut* 2006, **55**:141-143.

11. •• Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM, Nelson KE: **Metagenomic analysis of the human distal gut microbiome**. *Science* 2006, **312**:1355-1359.

As these authors state, the human intestinal microbiota is composed of 10^{13} to 10^{14} microorganisms whose collective genome (or 'microbiome') contains at least 100 times as many genes as our own genome. Metagenomic characterization of the gut community, as well as comparison to our bacterial genome collection, shows that humans are a 'super-organism' made up of eukaryotic and bacterial metabolic potential that is enriched for the production of amino acids, xenobiotics, methanogenesis, vitamins and other metabolites.

12. •• Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS *et al.*: **Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial 'pan-genome'**. *Proc Natl Acad Sci USA* 2005, **102**:13950-13955.

In a comparison of seven *S. agalactiae* genomes it was found that each new strain contributed on average 33 new genes, thus firmly establishing the concept of the pan-genome.

13. • Chen SL, Hung CS, Xu J, Reigstad CS, Magrini V, Sabo A, Blasiar D, Bieri T, Meyer RR, Ozersky P *et al.*: **Identification of genes subject to positive selection in uropathogenic strains of *Escherichia coli*: a comparative genomics approach**. *Proc Natl Acad Sci USA* 2006, **103**:5977-5982.

Good illustration of the power of comparative genomics to detect genes under selection. By analyzing seven *E. coli* genomes, it was also confirmed that the pan-genome of this species is vast — on average each new isolate contributes 441 new genes.

14. • Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R: **The microbial pan-genome**. *Curr Opin Genet Dev* 2005, **15**:589-594.

Useful review of the pan-genome concept and the mathematical models for calculating the extent of the pan-genome for different species on the basis of the availability of multiple genomes.

15. Purdy A, Rohwer F, Edwards R, Azam F, Bartlett DH: **A glimpse into the expanded genome content of *Vibrio cholerae* through identification of genes present in environmental strains**. *J Bacteriol* 2005, **187**:2992-3001.

16. Velicer GJ, Raddatz G, Keller H, Deiss S, Lanz C, Dinkelacker I, Schuster SC: **Comprehensive mutation identification in an evolved bacterial cooperator and its cheating ancestor**. *Proc Natl Acad Sci USA* 2006, **103**:8107-8112.

17. Fiegna F, Yu YT, Kadam SV, Velicer GJ: **Evolution of an obligate social cheater to a superior cooperator**. *Nature* 2006, **441**:310-314.

18. van der Gast CJ, Lilley AK, Ager D, Thompson IP: **Island size and bacterial diversity in an archipelago of engineering machines**. *Environ Microbiol* 2005, **7**:1220-1226.

19. Bell T, Ager D, Song JI, Newman JA, Thompson IP, Lilley AK, van der Gast CJ: **Larger islands house more bacterial taxa**. *Science* 2005, **308**:1884.

20. • Powell BC, Hutchison CA III: **Similarity-based gene detection: using COGs to find evolutionarily-conserved ORFs**. *BMC Bioinformatics* 2006, **7**:31.

This paper describes the use of clusters of orthologous genes to detect and correct unannotated genes in complete genomes.

21. Suhre K, Claverie J.M., Fusion DB: **A database for in-depth analysis of prokaryotic gene fusion events**. *Nucleic Acids Res* 2004, **32**:D273-D276.

22. Toh H, Weiss BL, Perkin SA, Yamashita A, Oshima K, Hattori M, Aksoy S: **Massive genome erosion and functional adaptations provide insights into the symbiotic lifestyle of *Sodalis glossinidius* in the tsetse host**. *Genome Res* 2006, **16**:149-156.

23. Pichon C, Felden B: **Small RNA genes expressed from *Staphylococcus aureus* genomic and pathogenicity islands with specific expression among pathogenic strains**. *Proc Natl Acad Sci USA* 2005, **102**:14249-14254.

24. Wilson GA, Bertrand N, Patel Y, Hughes JB, Feil EJ, Field D: **Orphans as taxonomically restricted and ecologically important genes**. *Microbiology* 2005, **151**:2499-2501.

25. Fukuchi S, Nishikawa K: **Estimation of the number of authentic orphan genes in bacterial genomes**. *DNA Res* 2004, **11**:219-231, 311–313.

26. Raskin DM, Seshadri R, Pukatzki SU, Mekalanos JJ: **Bacterial genomics and pathogen evolution**. *Cell* 2006, **124**:703-714.

27. Ussery DW, Hallin PF: **Genome update: annotation quality in sequenced microbial genomes**. *Microbiology* 2004, **150**:2015-2017.

28. Sterk P, Kersey PJ, Apweiler R: **Genome reviews: standardising content and representation of information about complete genomes**. *OMICS* 2006, **10**:114-118.

29. • Haft DH, Selengut JD, Brinkac LM, Zafar N, White O: **Genome properties: a system for the investigation of prokaryotic genetic content for microbiology, genome annotation and comparative genomics**. *Bioinformatics* 2005, **21**:293-306.

A database of 'assertions' about genomes derived from automated analysis of pathways and a variety of other traits. Curated information from the literature is also included.

30. • Markowitz VM, Korzeniewski F, Palaniappan K, Szeto E, Werner G, Padki A, Zhao X, Dubchak I, Hugenholtz P, Anderson I *et al.*: **The integrated microbial genomes (IMG) system**. *Nucleic Acids Res* 2006, **34**:D344-D348.

The Joint Genome Institute's IMG system (http://img.jgi.doe.gov/) is a major effort to provide an integrative environment that facilitates the genomic analysis of isolate organisms on a comparative level. Data can be organized by, for example, phylogeny, phenotypic or ecotypic properties.

31. •• Roberts RJ, Karp P, Kasif S, Linn S, Buckley MR: *An Experimental Approach to Genome Annotation. Critical Issues Colloquia Report*. Washington, DC: American Society for Microbiology; 2005.

Call for the systematic annotation of hypothetical and orphan genes using empirical methods.

32. • Kolker E, Picone AF, Galperin MY, Romine MF, Higdon R, Makarova KS, Kolker N, Anderson GA, Qiu X, Auberry KJ *et al.*: **Global profiling of *Shewanella oneidensis* MR-1: expression of hypothetical genes and improved functional annotations**. *Proc Natl Acad Sci USA* 2005, **102**:2099-2104.

Excellent demonstration of the use of transcriptomic approaches to the validation of genomic annotation.

33. Cochrane G, Bates K, Apweiler R, Tateno Y, Mashima J, Kosuge T, Mizrachi IK, Schafer S, Fetchko M: **Evidence standards in experimental and inferential INSDC Third Party Annotation data**. *OMICS* 2006, **10**:105-113.

34. Vallenet D, Labarre L, Rouy Z, Barbe V, Bocs S, Cruveiller S, Lajus A, Pascal G, Scarpelli C, Medigue C: **MaGe: a microbial genome annotation system supported by synteny results**. *Nucleic Acids Res* 2006, **34**:53-65.

35. Ye Y, Osterman A, Overbeek R, Godzik A: **Automatic detection of subsystem/pathway variants in genome analysis**. *Bioinformatics* 2005, **21**:i478-i486.

36. • Rodriguez-Brito B, Rohwer F, Edwards RA: **An application of statistics to comparative metagenomic**. *BMC Bioinformatics* 2006, **7**:162.

Statistical method for testing the significance of metabolic differences between two datasets for protein–function assignments.

37. • Liolios K, Tavernarakis N, Hugenholtz P, Kyrpides NC: **The genomes on line database (GOLD) v. 2: a monitor of genome projects worldwide**. *Nucleic Acids Res* 2006, **34**:D332-D334.

This updated version of GOLD contains a wide range of searchable, viewable and downloadable descriptive information about genomes curated from the literature.

38. •• Zhang K, Martiny A, Reppas NB, Barry KW, Malek J, Chisholm SW, Church GM: **Sequencing genomes from single cells by polymerase cloning**. *Nat Biotechnol* 2006, **24**:680-686.

Demonstration that genome sequencing can be accomplished from single cells without the need for cultivation.

39. Field D, Hughes J: **Cataloguing our current genome collection**. *Microbiology* 2005, **151**:1016-1019.

40. Field D, Morrison N, Selengut JD, Sterk P: **Cataloguing our Current Genome Collection II**. *OMICS* 2006, **10**:100-104.

41. Hallin PF, Ussery DW: **CBS genome atlas database: a dynamic**
• **storage for bioinformatic results and sequence data**. *Bioinformatics* 2004, **20**:3682-3686.

Content continues to accumulate in this genomic database from the authors of the 'Genome Update' monthly column in *Microbiology*.

42. Knight CG, Kassen R, Hebestreit H, Rainey PB: **Global analysis of predicted proteomes: functional adaptation of physical properties**. *Proc Natl Acad Sci USA* 2004, **101**:8390-8395.

43. Ussery DW, Binnewies TT, Gouveia-Oliveira R, Jarmer H, Hallin PF: **Genome update: DNA repeats in bacterial genomes**. *Microbiology* 2004, **150**:3519-3521.

44. Kiil K, Ferchaud JB, David C, Binnewies TT, Wu H, Sicheritz-Ponten T, Willenbrock H, Ussery DW: **Genome update: distribution of two-component transduction systems in 250 bacterial genomes**. *Microbiology* 2005, **151**:3447-3452.

45. Rocha EP: **Inference and analysis of the relative stability of**
• **bacterial chromosomes**. *Mol Biol Evo* 2006, **23**:513-522.

Study of the loss of synteny in bacterial chromosomes.

46. Konstantinidis KT, Tiedje JM: **Towards a genome-based**
• **taxonomy for prokaryotes**. *J Bacteriol* 2005, **187**:6258-6264.

Comparisons of shared gene content show that taxonomic classifications of these genomes might need to be revised in light of whole proteome information.

47. Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P:
•• **Toward automatic reconstruction of a highly resolved tree of life**. *Science* 2006, **311**:1283-1287.

Genomes are an excellent source of phylogenetic markers. This study uses 31 orthologous loci, pruned for any cases of horizontal gene transfer from 191 genomes, to build the tree of life across all three domains.

48. Konstantinidis KT, Tiedje JM: **Genomic insights that advance the**
•• **species definition for prokaryotes**. *Proc Natl Acad Sci USA* 2005, **102**:2567-2572.

The concept of a species is notoriously difficult to apply to bacteria. This study compared the gene content of 70 closely related and fully sequenced bacterial genomes. The results suggest that species boundaries do appear to exist, but are more conservative than current species definitions, and are better drawn on the basis of both ecology on shared gene content (evolutionary distance).

49. Sturino JM, Klaenhammer TR: **Engineered bacteriophage-defence systems in bioprocessing**. *Nat Rev Microbiol* 2006, **4**:395-404.

50. Ward N, Fraser CM: **How genomics has affected the concept of**
• **microbiology**. *Curr Opin Microbiol* 2005, **8**:564-571.

An excellent overview of how the comparison of microbial genomes with the genomes from fungi and viruses (such as the mimivirus, which has a genome size more than double than that of the smallest bacteria and contains tRNA genes) is changing our concept of what it means to be a bacterium.