

REVIEWS AND
SYNTHESES

Ecological perspectives on the sequenced genome collection

Jennifer B. Hughes Martiny¹ and Dawn Field^{2*}

¹Department of Ecology and Evolutionary Biology, Brown University, Providence, RI 02912, USA

²Molecular Evolution and Bioinformatics Section, Oxford Centre for Ecology and Hydrology, Mansfield Road, Oxford OX1 3SR, UK

*Correspondence: E-mail: dfield@ceh.ac.uk

Abstract

Our complete genome collection is one of our most valuable biological resources. A key challenge for the future is the interpretation of these genomes from an ecological perspective. In this review, we discuss current work at this increasingly important interface. In particular, we review ongoing work aimed at developing high quality data sets that combine ecological, environmental, evolutionary and genomic information. Such data will help to identify biases in the sequence collection and facilitate future discoveries about the nature of ecological adaptation at the genome level. These efforts will be greatly enhanced by the contributions of ecologists.

Keywords

Bioinformatics, comparative method, extremophiles, genome architecture, genomics, host–parasite interactions, metadata, procaryotes, species interactions, viruses.

Ecology Letters (2005) 8: 1334–1345

INTRODUCTION

The first bacteriophage, a genome of only 5386 base pairs (bp), was sequenced by Sanger *et al.* (1977), and it took 18 more years until the first whole bacterial genome was sequenced (Fleischmann *et al.* 1995). As of this writing, though, there are more than 2500 complete genomes in the public domain from eucaryotes (> 30), procaryotes (> 200), organelles (> 500), plasmids (> 600) and viruses (> 1200), comprising over 11 billion bp (Field *et al.* 2005; Wheeler *et al.* 2005) and the pace of genome sequencing is only increasing in speed.

The value of this collection extends far beyond the value of individual sequences. An organism's genome sequence is certainly a tremendous resource to the biologists who happen to study that organism, not to mention the value of human and pathogen sequences to human health. In addition, many scientists reap unquantifiable benefits from the contributions of model organism sequences to studies that are applicable to a wide range of organisms. However, in this review we focus on the scientific value of the genome collection as a whole, for comparative analyses across a range of ecological diversity.

Insights from the growing field of comparative genomics have revealed new insights about genome evolution (e.g. Moran & Wernegreen 2000; Dobrindt & Hacker 2001; Tekaiia *et al.* 2002; Wernegreen 2002; Lynch & Conery 2003; Konstantinidis & Tiedje 2004). However, for the most part these studies do not consider the influence of an organism's

ecology. The relative rarity of large-scale comparative ecogenomic studies that explore the relationship between organismal ecology and genome features is largely due to the difficulties of collecting the data required to conduct such analysis. These difficulties are the limited knowledge of the ecology of many species and the distributed nature of such data when it is recorded. Both of these issues must be tackled through improved interactions between ecologists, computational biologists and bioinformaticians, and those sequencing genomes.

In this review, we first discuss the type of insights revealed by merging ecological and genomic data. We review ongoing work aimed at developing new databases and data standards that capture both ecological and genomic data. We then describe the generation of three data sets for comparative ecogenomic analyses and discuss some of the issues involved in collating such metadata (data that describes genome sequence data). Finally, we use these data sets to investigate the ecological biases in the current genome collection and to discuss the types of analyses that can be applied to ecogenomic data.

BIOLOGICAL RULES AND EXCEPTIONS

The value of examining large collections of genome sequences lies in the potential to detect universal rules about the structure of genomes. For example, the 'genome size vs. predicted coding sequence (CDS)' relationship is one of the most robust rules describing procaryotic

genomes. Genome size is tightly correlated with the number of CDSs in a genome (Doolittle 2002); in general, bacteria contain one gene every 1000 nucleotides.

Such rules provide predictive power. For practical purposes, the genome size–CDS rule can be used to estimate the number of predicted genes in an unsequenced genome. However, equally important is that rules allow the identification of exceptions and subsequent evaluation of their biological significance. *Mycobacterium leprae* is an oft cited example of an exception to this relationship. The species has far fewer genes than expected and is hypothesized to be undergoing a process of gene loss as it adapts to its host (Cole *et al.* 2001). In fact, when pseudogenes (inactive genes derived from ancestral active genes) are included, the genome has the expected number of CDSs. The large number of pseudogenes suggests an ongoing process of ecological adaptation involving reductive genome evolution.

Indeed, the factors that determine genome size only become discernable when considering the ecological context of the organisms. Highly host-adapted parasites have small genomes ranging roughly from 0.5 to 1 million bp while free-living species that live in relatively variable environments range from 5 to 9 million bp. Thus, genome size reflects the complexity of metabolic pathways that an organism requires. A recent study of 115 procaryotic genomes further revealed a relative enrichment of particular gene categories in species with large genomes compared to those with smaller genomes. The patterns in gene composition suggest that species with larger-sized genomes may be able to dominate in environments where resources are scarce but diverse (e.g. soil) and where there is minimal selection against slow growth (Konstantinidis & Tiedje 2004).

Such patterns only become discernable through the study of large numbers of genomes. These examples demonstrate that the examination of basic genomic features across large numbers of genomes can reveal striking insights into the genetic basis of ecological adaptation. These studies further suggest that increased efforts to examine genome collections from an ecological perspective will speed up the detection of additional rules and exceptions of genome features. Open-access databases to which ecologists contribute will enable further comparative analyses, especially as the diversity and number of complete genome sequences increases.

ENABLING LARGE-SCALE COMPARATIVE ECOGENOMIC STUDIES: GETTING ECOLOGISTS INVOLVED

Perhaps the most important bottleneck in comparative ecogenomic studies is the time required to generate and

collect sufficiently high quality metadata describing our complete genome collection. Therefore, a powerful and practical way to promote exploration of the interface between ecological lifestyle and genomic features is for the community to create open-access repositories of such data. A new generation of databases is starting to include ecological data, from habitat preferences to metabolic attributes. The database efforts mentioned below are actively soliciting expert contributions and offer a direct mechanism for ecologists to contribute to the quality and depth of the metadata associated with the complete genomes.

The primary international nucleotide databases are working to increase the amount of metadata, and in particular ecological metadata, collected from contributors of complete genomes. For example, the NCBI Genome Projects database now collects a wider range of data from contributors of genomes and maintains a list of genomic and ecological attributes about procaryotic genomes (<http://web.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj>). Perhaps the most advanced public effort is TIGR's Genome Properties database (Haft *et al.* 2005). This database is dedicated to the description of features of procaryotes and contains over 150 fields of information. Much of this information is generated automatically from bioinformatic analysis of the predicted gene content of these genomes, but ecological data is also curated from the literature.

The GenomeMine database contains information about all complete genomes (eucaryotes, procaryotes, viruses, plasmids and organelles) and hopes to provide a community resource for the archiving of comparative ecogenomic data sets (Field *et al.* 2005). Authors can contribute existing data sets to this database where they can be viewed in several formats and downloaded thus promoting their exchange and re-use. For example, the GenomeMine data set of the ecological attributes of procaryotic genomes (described below) has recently been downloaded and incorporated into the SEED database, a project to manually annotate 1000 genomes from different taxa (<http://theseed.uchicago.edu/FIG/index.cgi>). This contextual information is already being used for further bioinformatic analyses. For example, PSORT, a computer program for the prediction of protein localization sites in cells used by the SEED project, requires input on the source of the proteins being analysed (e.g. whether a bacterium is Gram-negative or Gram-positive) to generate with highest quality predictions (Nakai & Horton 1999).

The primary source of ecological metadata for the above projects is Bergey's Manual (Garrity 2001). This procaryotic reference is currently available only in print, but its Editor, George Garrity, has recently defined a core set of more than 300 fields of information about all taxa (in SGML format). If such a dataset were to become electronically available it

would surely become the gold standard for ecogenomics work of the future.

The increased interest in genomic metadata has also prompted calls for the establishment of improved data standards for both complete genomes (Field & Hughes 2005) and partial sequence data sets (<http://www.jgi.doe.gov/16s/>). For example, the isolate list from the Moore Foundation's Marine Microbe Genome Sequencing project is exemplary in containing details of the location (e.g. depth), isolation method, primary citation and 16S sequence for each isolate (<http://www.moore.org/microgenome/>). To describe partial sequences better, the DOE Joint Genome Institute (JGI) is currently conducting a survey soliciting expert help from a range of scientists, especially microbial ecologists. The JGI aims to formulate a minimum set of standard sequence-associated information fields for partial environmental sequences (16S clone libraries and metagenomic studies). Descriptions will include up to 20 fields of information depending on the habitat sampled. The importance of such data sets is underscored by the rapid growth of 16S (Cole *et al.* 2005) and metagenomic data sets (Venter *et al.* 2004; Cowan *et al.* 2005; Edwards & Rohwer 2005; Tringe *et al.* 2005).

GENERATING ECOLOGICAL METADATA FOR THE GENOME COLLECTION

For the sake of reviewing the biases in the genome collection and performing comparative ecogenomic analyses, we generated three data sets. These data sets have been archived in the GenomeMine and are freely available for viewing and downloading (Field *et al.* 2005). The genomes in these data sets are summarized in Table 1 and were taken from the NCBI Entrez genomes database (Wheeler *et al.* 2004). They represent over 2500 of the first genomes published for eucaryotes, procaryotes, organelles, plasmids and viruses. The total number of complete but unpublished genome sequences is significantly larger, but we have restricted our list to published genomes for two reasons. First, these genomes are freely accessible to everyone in the research community. Second, unpublished genome sequences are largely unannotated, and without access to pre-existing annotation we could not easily summarize the basic features of those genomes.

The first data set, 'Genomic Features', summarizes genomic features derived directly from the genome annotations downloaded from GenBank. This was performed for

Table 1 Tally of published genomes as of March 2004 (eucaryotes by June 2004)

Group	Amount sequenced			Number of 'species'		No. of phyla/divisions	
	No. genomes	Total (Mb)	No. CDSs	Described	Estimated	Described	Sequenced
Eucaryotes*	24	11 016	401 468	1.8×10^6	12×10^6	82	11
Archaea†	17	38	38 818	217		4	3
Bacteria†	133	436	404 733	5007	$4.5\text{--}1000 \times 10^6$	23	14
Viruses‡	1299	29	29 457	1550	400 000	72	60
Organelles§							
Mitochondria	479	10	7259	–	12×10^6	–	–
Plastids	39	5.3	4158	–	720 000	–	–
Plasmids¶							
Eucaryotic	32	0.16	97	–	–	–	–
Procaryotic	576	24	23 486	1000	14×10^6	–	–

The term 'species' is used very loosely as it must be defined differently across taxa. The number of genomes reported is the number of strains; therefore, the number of species sequenced is lower. Some of the smaller genomes such as the plasmids are not annotated, so the actual number of predicted coding sequences (CDSs) is probably higher.

*Described and estimated total eucaryote diversity are working estimates proposed by the Global Biodiversity Assessment (Hammond 1995). Eucaryotic phyla are based on Margulis & Schwartz (1998).

†The estimate of procaryotic species richness (Archaea and Bacteria together) is the range between conservative estimates from Curtis *et al.* (2002) and Dykhuizen (1998) respectively. The number of described bacterial and archaeal phyla are based on Bergey's Manual (Garrity 2001), plus two new archaeal phyla recently proposed (Barns *et al.* 1996; Huber *et al.* 2002).

‡This includes phage (210 strains) and viroids. Virus diversity estimates are from The Committee's Seventh Report on Virus Taxonomy (Regenmortel *et al.* 2000). Number of viral families (rather than phyla) are based on the International Committee on Taxonomy of Viruses. (<http://www.ncbi.nlm.nih.gov/ICTVdb/index.htm>) (Regenmortel *et al.* 2000).

§Mitochondrial diversity is estimated as the number of eucaryotes. Some eucaryotes lack mitochondria, however, excluding them does not change the rounded, working estimate of eucaryotes. The diversity of plastids (including chloroplasts and cyanelles) is estimated as the number of plants plus the number of eucaryotic algae (Hammond 1995).

¶A very conservative estimate of the diversity of procaryotic plasmids is the estimated number of procaryote species times the average number of plasmids sequenced per procaryotic host sequenced in the NCBI database (3.2 plasmids per host).

all genomes except eucaryotes, for which genomic features were curated from primary genome reports. The second data set, 'Curated Genomic Features and Ecological Information', focuses on procaryotes, a group for which we have a good deal of ecological information and now over 220 complete genomes. The data set summarizes genomic features that could not be calculated directly from the published genome annotation, for example, whether a bacterial genome had one or two chromosomes or whether it contained plasmids. The data set also includes 20 fields of ecological information collated from primary species description publications and Bergey's Manual (Holt 1984–1989; Holt *et al.* 1994; Garrity 2001). Below we discuss eight of these variables: (i) primary habitat or the major habitat thought to be used for growth; (ii) optimal growth temperature, almost always determined by laboratory culture studies; (iii) whether the strain is an extremophile, i.e. if its optimal growth temperature was > 48C (thermophile) or > 80C (hyperthermophile) or if it preferred acidic (pH < 4), alkali (pH > 8.5) or high salinity (> 15%) environments (Madigan 2000); (iv) if the organism is primarily free-living or host-associated, and if host-associated, whether the host is obligate (needed for long-term persistence) or not; (v) for a host-associated strain, its primary interaction with the host (pathogen, mutualist, commensal or predator); (vi) for a host-associated strain, whether the organism lives within the host's cells (is intracellular); (vii) primary energy source (organic, inorganic or light); and (viii) primary carbon source (organic or inorganic).

The third data set, 'Genomic Interactions,' records the primary host of all of the organisms whose genomes have been sequenced. For the viruses, we classify the hosts into general taxonomic categories. We also identify potential interactions between the genomes that are already sequenced, or what we call here 'genome pairs.' Thus, if a sequenced bacteriophage primarily infects one of the bacteria already sequenced, then this interaction counts as one genome pair. A genome can participate in multiple genome pairs; for instance, 25 different plasmids are sequenced from multiple strains of *Lactobacillus lactis*. This example highlights the fact that these interactions are 'potential' interactions; the particular specimens or strains sequenced may have never interacted with the other genome.

ISSUES SURROUNDING THE ANALYSIS OF ECOGENOMIC DATA

The collection of these three data sets highlights some general issues associated with the analysis of ecogenomic metadata. A key issue is the quality of the ecological data itself. We found it difficult to curate a complete set of values for some of the variables in the ecological data set. Some of these values may be known, but we did not find them in the literature reports.

However, erroneous data is perhaps a more worrisome issue. In particular, isolates of the same named species can have very different phenotypic traits, and the ecological data reported for one strain may not apply to the sequenced strain.

Likewise, the classification of organisms by finite ecological categories is not always straightforward. For example, we call *Escherichia coli* an obligate, host-associated bacterium. We label it 'host associated' because although it is found in the environment outside of a host, its primary habitat is the digestive tract of vertebrates. We call it 'obligate' because its long-term persistence, even in the environment, is probably dependent on reproduction in vertebrate hosts. In contrast to *E. coli*, we call *Agrobacterium tumefaciens*, a plant pathogen that reproduces well in the soil without a host, a 'non-obligate, host-associated' bacterium. These detailed decisions are somewhat arbitrary, and such examples underscore the importance of expert curation, either by ecologists that work on the organism or by those close to the sequencing project.

Finally, when performing comparative analyses we must consider phylogenetic relatedness, otherwise the relationships among species' traits can be misinterpreted because of uneven phylogenetic sampling (Harvey & Pagel 1991; Hurst & Merchant 2001). A suite of sophisticated methods have been developed for this purpose (e.g. Purvis & Rambaut 1995; Harvey & Rambaut 2000; Giannini 2003; Quader *et al.* 2004). With our current understanding of the tree of life, these methods are likely to prove useful within specific clades of Bacteria. However, deep evolutionary relationships among eucaryotic phyla and bacterial divisions are highly debated (Doolittle 1999; Wolf *et al.* 2002; Rokas *et al.* 2003). Furthermore, virus taxonomy, largely based on the past classification of morphological features, is in flux as it is updated on the basis of molecular features (Büchen-Osmond 2003), and plasmid 'taxonomy', for example, as captured within the NCBI database, is only based on host (procaryotic) taxonomy (Wheeler *et al.* 2004).

Even in the absence of a complete phylogeny, however, comparative analyses are still possible when the basic grouping of organisms are understood (Graffen 1989). For example, to begin to control for the problem of uneven phylogenetic sampling, we carried out analyses on procaryotes (below) with only one species per genus represented. Where the patterns observed seem to be driven by one group of related taxa, further analyses can be carried out without the group. Still, until the apparent effect of phylogeny can be fully taken into account, such results should be considered preliminary.

ECOLOGICAL BIASES IN THE SEQUENCED GENOME COLLECTION

While it has been repeatedly noted that the sequenced genome collection is evolutionarily biased, less attention has

been paid to ecological biases. It is essential to understand such biases as they can also influence the outcome of comparative analyses and limit the applicability of the results. In the next three sections we use the GenomeMine data sets to investigate the biases in the current genome collection with respect to taxonomic diversity, types of interactions between genomes, and metabolic and habitat diversity.

Taxonomic bias

The number of genomes sequenced now, and in the foreseeable future, is miniscule compared with total diversity on the planet (Table 1). Sequencing of genomes, while often aimed at assaying novel genetic diversity (Eisen & Fraser 2003), has been significantly biased by four key factors: the use of cultivated isolates (Floyd *et al.* 2005), the ease of sequencing smaller genomes, the tremendous interest in human pathogens, and the selection of eucaryotes that are experimental models of human biology. At a broad scale, the ecological biases of the sequenced genome collection are related to biases in the taxonomic and evolutionary diversity sampled. For instance, interest in disease combined with their small genome size, has made viruses the best represented group in the genome collection. Indeed, 50% of all sequenced genomes are those of viruses, and over three-quarters (60 of 72) of all known viral families have been sequenced (Table 1).

The next most abundant genomes in the collection are plasmids and organelles (608 and 518 respectively; Table 1). Only 5% of the plasmids sequenced are eucaryotic, but this skew is mostly a reflection of the small number of eucaryotes sequenced rather than an indication of their relative diversity. Within the organelles, many more mitochondria (479) than plastids (39) have been sequenced. Because of their use in phylogenetic reconstruction, mitochondria, more than any other group, have been most systematically sequenced to maximize taxonomic diversity (Pollock *et al.* 2000). Still, the majority of mitochondrial genomes come from vertebrates and more specifically mammals, whose diversity (*c.* 4700 species) makes up a small fraction of the *c.* 12 million eucaryotic species that contain mitochondria.

The phylogenetic representation of the procaryotic genomes is quite good because of recent attention to evolutionary biases (Eisen 2000), even though we have sequenced perhaps < 0.000015% of all procaryotic species (Table 1). For instance, there are representatives sequenced from 17 of the 27 described bacterial and archaeal phyla (Table 1). The missing evolutionary diversity is due to phyla (alternatively called divisions) that have not yet been cultured. However, this limitation may be mitigated soon with the recent advent of new techniques to sequence whole

genomes (Tyson *et al.* 2004; Venter *et al.* 2004). There are also likely undiscovered procaryotic phyla. Half of the archaeal phyla (two of four) have been reported just in the last decade (Barns *et al.* 1996; Huber *et al.* 2002), and there are as many as 26 'candidate phyla' of procaryotes known only from 16S rDNA sequences (Rappe & Giovannoni 2003) suggesting that our picture of the extent of procaryotic life on Earth remains far from complete.

In contrast, a higher but still tiny (0.0002%), proportion of eucaryotic species have been sequenced than procaryotic species, simply because eucaryotic diversity is relatively low (Table 1). However, phylogenetic representation is not as broad relative to the procaryotes. Only 11 of 82 eucaryotic phyla are represented, because of the difficulty in sequencing large eucaryotic genomes. In summary, although the proportion sequenced of the total diversity of procaryotes, plasmids and viruses is miniscule, phylogenetic representation is much higher among these small genomes than among the eucaryotes and their organelles.

Metabolic and habitat diversity

The metabolism and habitat of an organism are intertwined components of ecological diversity. Almost two-thirds of the first 150 procaryotes whose genomes are sequenced are host-associated, and all of these are chemoorganotrophs, i.e. they use organic compounds for energy and carbon sources. In contrast, 53 free-living procaryotes have been sequenced from a wide variety of habitats representing a variety of metabolic types (Table 2). Still, most of the free-living procaryotes sequenced are chemoorganotrophs (Table 2). Furthermore, the metabolism of the species sequenced is not randomly distributed among habitats. For instance, the majority of marine procaryotes sequenced (excluding thermal vent organisms) are phototrophs (use light as their energy source). In contrast, the procaryotes sequenced from soils are overwhelmingly represented by chemoorganotrophs. However, at this point the true distribution of metabolic types among habitats is not known, and thus it is unclear whether these patterns represent a biased view of procaryotic diversity.

Another characteristic of the free-living procaryotes whose genomes are sequenced is that more than 40% are extremophiles or thermophiles. This bias is especially evident in the marine procaryotes, where 58% of those sequenced are thermo- or extremophiles (Table 2). Organisms that live in extreme environments provide a scale for the range of biochemical diversity on Earth and are often useful for industry (Madigan 2000; Rothschild & Mancinelli 2001). However, it is unlikely that the majority of procaryotic diversity on the planet falls within these categories. It is also notable that the extremophiles sequenced so far include alkaliphiles, acidophiles, halophiles and hyperthermophiles,

Table 2 Primary habitat vs. primary metabolic type of the sequenced free-living procaryotes

Primary habitat	<i>n</i>	<i>n</i> (%) thermo- and extremophiles	Primary metabolism			
			Chemoorganotrophs	Chemolithotrophs	Phototrophs	Mixotrophs/ flexible*
<i>Aquatic</i>						
Marine thermal vents	10	10 (100)	5	2		2
Other marine	9	1 (11)	2		6	1
Freshwater	4	1 (25)	1	1	1	1
Salt lakes	1	1 (100)	1			
Hot springs (terrestrial)	4	4 (100)	1	1	2	
<i>Terrestrial</i>						
Soil, rocks	19	2 (11)	15	1	1	2
Solfatara fields	4	4 (100)	2	2		
Plants	2		2			
Totals†	53	22 (42)	29	7	10	6

The number and per cent of the strains that are thermophiles or extremophiles are also reported. Metabolic types are defined by energy sources (organic, inorganic or light) and carbon sources (inorganic or organic) following Madigan *et al.* (1997).

*For convenience, we combine mixotrophs and organisms that are flexible in using organic and inorganic carbon for energy and carbon sources into one column.

†We exclude *Nanoarchaeum equitans* from the metabolism categories as its metabolism is unknown.

but no psychrophiles (cold-loving species). In fact, the lowest optimal temperatures of the free-living strains are around 22–24 °C, those of the marine strains *Pirellula* (also called *Rhodopirellula*) and *Prochlorococcus*.

A frequency distribution of the optimal temperatures of the sequenced, free-living procaryotes shows that the metabolic types are somewhat scattered among temperature classes (Fig. 1a). Chemoorganotrophs are represented across the entire range of temperatures. In contrast, few non-thermophilic chemolithotrophs and few thermophilic phototrophs have been sequenced. There is also a skew in terms of the temperature range of how Archaea and Bacteria are sampled. Most of the high temperature genomes are Archaea (Fig. 1b), whereas there is a complete absence of low temperature Archaea. While it was once thought that the Archaea were primarily extremophiles, it is now known that they inhabit a wide range of habitats including the open ocean (DeLong 1998).

As with metabolism and general habitat types, not enough is known about procaryotic diversity to quantify the extent of the temperature bias within the genome collection. We guess that extremophiles and chemoorganotrophs are overrepresented and that there is poor phylogenetic representation across temperatures. Psychrophiles, at least are clearly underrepresented.

Ecological interactions

A growing area of comparative ecogenomics research is how ecological interactions affect genome structure. For

instance, the genomes of bacterial endosymbionts of insects display many similar features, including low G + C content (Wernegreen 2002), a bias towards deletions over insertions (Mira *et al.* 2001), and signatures of host ecology on functional genes (Zientz *et al.* 2001).

The host-associated species (as opposed to free-living species) in the genome collection interact in many ways with a variety of host taxa. We define interaction type by whether the outcome is primarily negative (pathogenic), neutral (commensalistic) or positive (mutualistic). The group with the best representation of different interaction types is the procaryotes, and their representation is, not surprisingly, highly skewed. Half of the human-associated sequenced procaryotes are pathogens (Table 3). A less obvious bias is that this skew varies greatly by host type. For instance, five of six of the insect-associated sequences are mutualists. Moreover, these mutualists are obligatorily intracellular, whereas only 8% of the plant-associated and 16% of the human-associated procaryotes are intracellular (Table 3). While there may be true differences in the relative abundance of interaction types among taxa, so far the differences among the sequenced collection are likely driven by human interests.

Although viruses do not vary in terms of interaction type (they are usually negative, but for potential exceptions see Levin & Lenski 1985), those sequenced infect a broad array of hosts (Fig. 2). Most of the sequenced viruses infect eucaryotes rather than procaryotes, although based on host diversity, bacteriophage and archaeophage are probably orders of magnitude more diverse. Because of

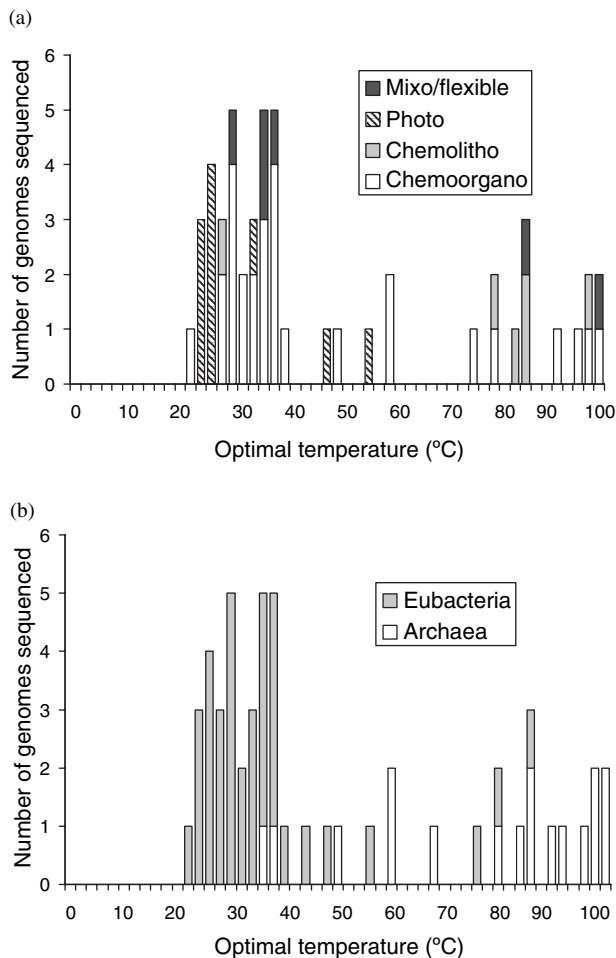


Figure 1 Optimal temperatures of the sequenced, free-living prokaryotes by (a) metabolic type (see category explanations in Table 2) and (b) domain.

their agricultural importance, plant viruses make up the largest host group accounting for 40% of all the sequenced viruses. The majority of vertebrate viruses sequenced infect

humans (201); e.g. 86 viral genomes are human papillomavirus.

One of the most exciting and underutilized aspects of the genome collection is the sequencing of over 700 'genome pairs,' pairs of genomes from organisms that potentially interact in the environment. Table 4 reports nine taxonomic classes of genome pairs, where both partners in the interaction are fully sequenced. The evolutionary breadth of these interactions varies enormously across different pairs. For instance, the breadth of interactions between bacteria and eucaryotes is narrow; 76 sequenced bacteria interact with only two different sequenced eucaryotes (mice and humans). The phage and prokaryote pairs are broader; 131 phages have been sequenced that infect 27 sequenced prokaryotic hosts. Still of these phage, 40 infect *E. coli*, followed by 15 that infect *Lactococcus lactis*. The broadest range of interactions is represented by the plasmid–bacteria pairs; 292 plasmids interact with 62 different bacteria.

ECOGENOMIC ANALYSIS OF PROCARYOTES

Despite ecological biases, the current genome collection in conjunction with ecological metadata can address innumerable questions about the link between an organism's ecology and genomic features. As discussed above, it appears that an organism's lifestyle influences its genome size, but there is also growing evidence that it shapes the form and content of other genomic features. For example, in a study of 56 genomes, amino acid usage varied between mesophiles, thermophiles and hyperthermophiles (Tekaija *et al.* 2002). Likewise, in a recent study of 90 prokaryotic genomes, codon bias signatures distinguished between thermophiles and mesophiles and groups of organisms with different respiratory characteristics (aerobic, anaerobic, facultative aerobic and facultative anaerobic) (Carbone *et al.* 2005). The elucidation of such relationships is critical not only for understanding why genomes are constructed as they are, but

Host	No. of strains	No. (%) intracellular	No. of pathogens	No. of mutualists	No. of commensals
Vertebrates					
Humans	70	13 (16)	58		12
Only non-humans	9	1 (11)	6		3
Plants	12	1 (8)	9	3	
Insects	6	5 (83)	1	5	
Totals	97	20	74	8	15

Table 3 Sequenced prokaryotic strains whose primary habitat is a eucaryotic host

If the primary habitat of a strain is humans, then it is not included in any other host category even if it can interact with other vertebrates or insect vectors. The number of strains that are obligatory intracellular is also reported (followed by the percentage intracellular) and total strain number is categorized by interaction type (pathogen, mutualist or commensal). This table does not include opportunistic pathogens; 13 prokaryote genomes sequenced are opportunistic pathogens of humans, but are otherwise free-living organisms.

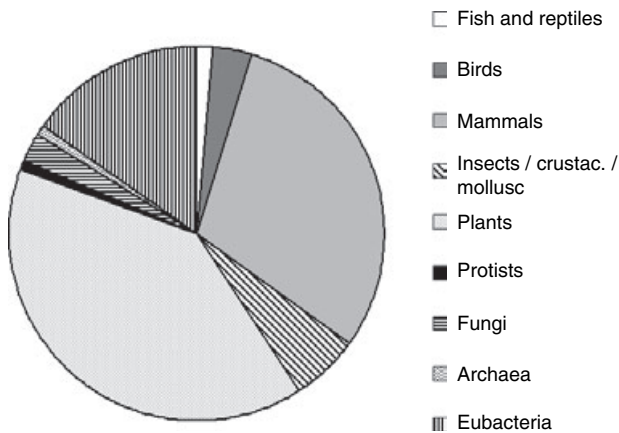


Figure 2 The distribution of host taxa of the sequenced viruses ($n = 1299$).

Table 4 The number of 'genome pairs' that are sequenced and publicly available

Host	Host associate				
	Eucaryotes	Bacteria	Viruses	Organelles	Plasmids
Eucaryotes	3/4	2/76	10/250	12/16	2/3
Bacteria	–	–	25/129	–	62/292
Archaea	–	–	2/2	–	3/4

The first number in the cell is the number of hosts that interact with any organism of that type. The second number is the number of host-associated organisms that interact with any of the sequenced host types (counting multiple strains of a species). For instance, the '3/4' in the cell for eucaryotic host and eucaryotic parasite counts the interaction between *Anopheles gambiae* and *Plasmodium falciparum*, *Homo sapiens* and *P. falciparum*, *H. sapiens* and *Encephalitozoon cuniculi*, *H. sapiens* and *Cryptosporidium parvum*, and *Mus musculus* and *Plasmodium yoelii* for a total of three hosts and four parasites.

for using genome sequences to infer aspects of organismal lifestyle.

In this section, we use the GenomeMine data sets we generated to further demonstrate and discuss the types of ecogenomic analyses that can be undertaken. We focus on four genomic features: (i) ORF number; (ii) rRNA gene copy number; (iii) tRNA gene number; and (iv) G + C content. In all the analyses, we use only one species per genus (the first alphabetically), narrowing the data set to 84 prokaryotic species.

As discussed above, free-living bacteria contain more CDSs (and have larger genome sizes) than obligate host-associated bacteria. The GenomeMine data set confirms this relationship, but we also find that non-obligate, host-

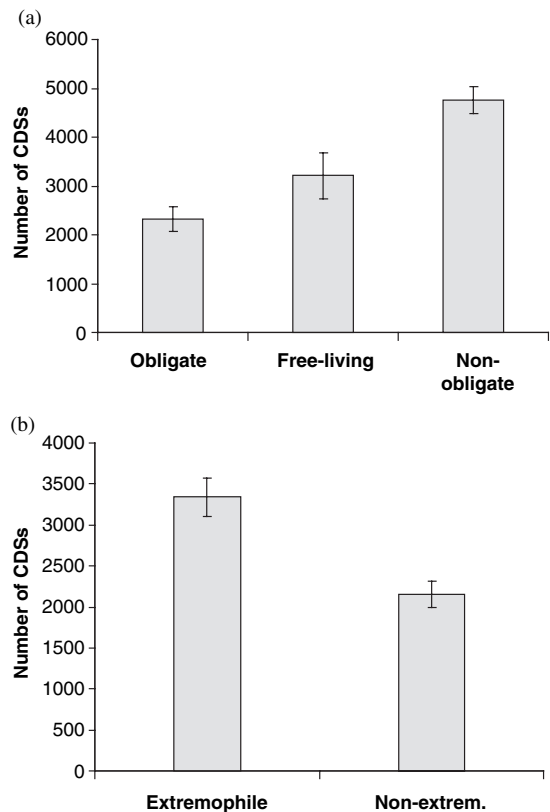


Figure 3 The average number of coding regions for: (a) obligate host-associated, free-living and non-obligate host-associated species; (b) free-living extremophiles and non-extremophiles. Error bars represent 1 SE. Only one species per genus is used in the analyses (total $n = 84$).

associated bacteria contain more CDSs than free-living species (Kruskal–Wallis test, $\chi^2 = 17.57$; $n = 84$; $P < 0.001$; Fig. 3a). This result suggests that the ability to be both free-living and host-associated may require more genes than being solely free-living. Among the free-living prokaryotes, extremophiles have significantly more CDSs than other species (Mann–Whitney $Z = -2.40$, $n = 84$, $P = 0.016$; Fig. 3b). This pattern may be driven by an evolutionary constraint of Archaea compared with Bacteria, as all but one of the non-extremophile genera is Bacteria.

Our data also reveal that rRNA gene copies in prokaryotes are lower for free-living species compared with both obligate and non-obligate host-associated species (Kruskal–Wallis test; $\chi^2 = 7.61$; $n = 84$; $P = 0.022$). Ribosomal RNA (rRNA) genes are transcribed to form a key structural component of ribosomes, which translates messenger RNA. It is generally thought that rRNA copy number in prokaryotes is related to an organism's growth rate. In a laboratory study, Klappenbach *et al.* (2000) demonstrated

that, within a community of soil organisms, there is a positive correlation between rRNA copy number and an organism's ability to grow quickly on a pulse of resources. If the same reasoning applies here, then host-associated organisms may be better adapted to respond to resource pulses than free-living species.

At the same time, we found no difference in the number of transfer RNA (tRNA) genes between free-living or host-associated species. Transfer RNAs translate messenger RNA into a protein sequence. There is evidence in both bacteria and eucaryotes that tRNA copy number is related to intracellular tRNA abundance and that this abundance is related to growth rate (e.g. Dong *et al.* 1999; Duret 2000; Rocha 2004). Rocha (2004) also found that fast-growing bacteria have more overall tRNAs, but relatively fewer tRNA anticodon species compared with slow-growing bacteria. Fewer tRNA types helps to optimize translation through a dependence on a smaller number of amino acids and the codons that code for them apparently allow faster growth. Finally, we found that the number of tRNAs and rRNAs in genomes are both significantly positively correlated with genome size (Spearman correlation; $\rho = 0.411$ $P < 0.001$ and $\rho = 0.651$, $P < 0.001$ respectively). Altogether, the results suggest that while genome size (and the ecological factors that influence it) may play a large role in determining the number of rRNAs and tRNAs in a genome, the exact numbers of these genes may be under somewhat different selective pressures.

Organisms vary widely in the GC content (the per cent guanine and cytosine) of their genomes. This variation is related to phylogeny (Muto & Osawa 1987) and appears to be related to some aspects of an organism's ecology (Haft *et al.* 2005). For instance, intracellular mutualists have relatively low GC content (Moran & Wernegreen 2000), and organisms living at high temperatures may have relatively high GC content, at least for some genes (Hurst & Merchant 2001). GC content can also be used as an indicator of lateral gene transfer, where some genes may have been subject to different mutational processes at different times (Lawrence & Ochman 1998). It has also recently been noted that the GC content of double-stranded DNA phages is positively correlated with genome size, although the reasons for this pattern remain unclear (Pedulla *et al.* 2003).

A correlation of GC content among genome pairs can also provide a signature of genome interactions. A positive relationship between the GC content of plasmid–procaryote and phage–procaryote pairs signals tight evolutionary relationships (Gibbs & Primrose 1976). Our updated analyses based on pairs of complete genomes draws the same conclusion, although we find even tighter correlations than previously reported. The per cent GC content of a procaryotic host is highly positively correlated with per cent

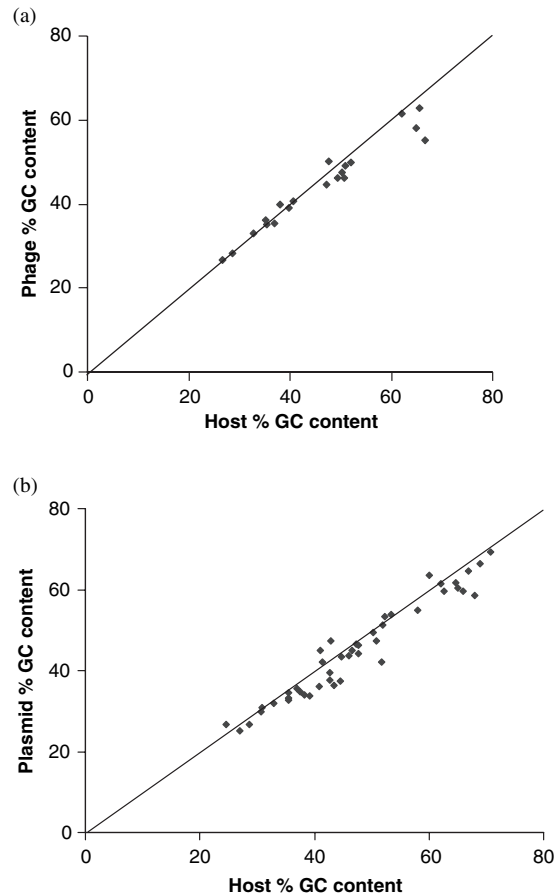


Figure 4 The relationship between host vs. (a) phage and (b) plasmid per cent GC content. If multiple phages or plasmids are sequenced for a given host, then GC content is given as the average of the phages or plasmids. Only one host per genus is used in the analyses. The line $x = y$ is drawn for reference. The regression line for (a) is $y = 0.83x + 6.25$ ($n = 19$; $r^2 = 0.949$) and for (b) is $y = 0.94x + 0.72$ ($n = 43$; $r^2 = 0.940$) (not shown).

GC content of its phage ($r^2 = 0.949$; Fig. 4a). A similarly strong positive correlation holds for plasmid GC content and that of their hosts ($r^2 = 0.940$; Fig. 4b). The strong plasmid–host relationship may be partly due to sequencing the plasmids that are found in the particular host isolates that are also sequenced, but it would not explain the tight correlation of the phage and their hosts, which are sequenced separately. The results suggest that one may be able to identify if a phage or plasmid host has gone through a recent host range expansion.

The analyses in this section serve to demonstrate the types of questions one can ask with the combination of genetic, evolutionary and ecological data for every fully sequenced genome. There is certainly far greater scope to extend these analyses. For instance, one could incorporate

more detailed genomic features such as numbers of genes in particular functional categories (Konstantinidis & Tiedje 2004) as well as more refined sets of ecological information such as the particular habitat from which the sequenced isolates were collected.

CONCLUSIONS

The current genome collection samples a small and biased fraction of all extant biodiversity. In the future, it will be beneficial to sequence organisms from under-sampled ecological niches if we are to capture a broader range of the diversity of life on Earth. While the phylogenetic diversity already captured by the current collection is encouraging, the ecological diversity represented appears more skewed. Of special importance is the sequencing of non-pathogenic relatives of the many pathogens in our current genome collection. For example, such genomes are revealing the surprising fact that some genes thought to be classic virulence determinants actually evolved in free-living relatives, causing a re-evaluation of both the genetic basis of pathogenicity and the nature of the genes involved in adaptation to preferred niche (Horn *et al.* 2004; Ruby *et al.* 2005). In addition, a broader range of metabolic types from within different habitat types should be sequenced. Equally important is the sampling of novel environments and the vast wealth of uncultured isolates now accessible through the advancement of metagenomic approaches (Venter *et al.* 2004; Cowan *et al.* 2005; Edwards & Rohwer 2005; Tringe *et al.* 2005). Finally, sequencing organisms that potentially interact with already sequenced organisms will further promote research on how ecological interactions influence genomic features.

Despite bias limitations, the current genome collection represents a considerable amount of taxonomic and ecological diversity, making it possible to address the influence of ecological lifestyle on genomic features. Complimented by ecological metadata, the analyses described here demonstrate the types of insights that comparative analyses of sequenced genomes can provide.

This review also highlights the challenges involved in the collection of a standardized set of basic ecological information about already sequenced organisms. While many genomic features can be automatically extracted from genome sequences (although not without its own difficulties), ecological data must be painstakingly collected from a variety of literature sources. Perhaps the most conspicuous piece of data absent from the vast majority of genome reports is any information about the geographical location and specific habitat from which the sequenced specimen was collected. This information may be well known for many of the sequences, but has not been consistently published. Hopefully this will be rectified in the future

through the establishment of new reporting standards (Field & Hughes 2005).

We propose that ecologists, researchers comfortable with comparative methodologies and in possession of knowledge at the organismal, community and ecosystem levels, are in a special position to make a unique and important contribution to the development and interpretation of ecogenomic information. Ecologists can contribute to decisions about which organisms to sequence, relevant data fields to incorporate into databases, standardization of methods for data collection, and interpretation of comparative analyses. Without this expert contribution, such standards, databases and analyses risk being poorly formulated, incomplete or worse, incorrect. Likewise, ecologists are already bringing new perspectives to the burgeoning field of genomic analysis (Feder & Mitchell-Olds 2003; Vinogradov 2004; Brookfield 2005). Further collaborations with bioinformaticians and those generating new genome sequences will certainly provide innovative insights into ecological adaptation at the genomic level.

ACKNOWLEDGEMENTS

JBHM was supported by Brown University. DF was supported by the Oxford Centre for Ecology and Hydrology. Many thanks to Gareth Wilson for helping to curate data for the bacterial ecology data set, and to Melissa Lage, Jay Lennon, Adam Martiny, Kathryn Maurer and Marc Tatar for discussion and critical reading of the manuscript.

REFERENCES

- Barns, S.M., Delwiche, C.F., Palmer, J.D. & Pace, N.R. (1996). Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. *Proc. Natl. Acad. Sci. USA*, 93, 9188–9193.
- Brookfield, J.F. (2005). The ecology of the genome – mobile DNA elements and their hosts. *Nat. Rev. Genet.*, 6, 128–136.
- Büchen-Osmond, C. (2003). The universal virus database ICTVdB. *Comput. Sci. Eng.*, 5, 16–25.
- Carbone, A., Kepes, F. & Zinovyev, A. (2005). Codon bias signatures, organization of microorganisms in codon space, and lifestyle. *Mol. Biol. Evol.*, 22, 547–561.
- Cole, S.T., Eiglmeier, K., Parkhill, J., James, K.D., Thomson, N.R., Wheeler, P.R. *et al.* (2001). Massive gene decay in the leprosy bacillus. *Nature*, 409, 1007–1011.
- Cole, J.R., Chai, B., Farris, R.J., Wang, Q., Kulam, S.A., McGarrell, D.M. *et al.* (2005). The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Res.*, 33, D294–D296.
- Cowan, D., Meyer, Q., Stafford, W., Muyanga, S., Cameron, R. & Wittwer, P. (2005). Metagenomic gene discovery: past, present and future. *Trends Biotechnol.*, 23, 321–329.
- Curtis, T.P., Sloan, W.T. & Scannell, J.W. (2002). Estimating prokaryotic diversity and its limits. *Proc. Natl. Acad. Sci. USA*, 99, 10494–10499.

- DeLong, E.F. (1998). Everything in moderation: Archaea as 'non-extremophiles'. *Curr. Opin. Genet. Dev.*, 8, 649–654.
- Dobrindt, U. & Hacker, J. (2001). Whole genome plasticity in pathogenic bacteria. *Curr. Opin. Microbiol.*, 4, 550–557.
- Dong, H.J., Nilsson, L. & Kurland, C.G. (1999). Co-variation of tRNA abundance and codon usage in *Escherichia coli* at different growth rates. *J. Mol. Biol.*, 260, 649–663.
- Doolittle, W.F. (1999). Phylogenetic classification and the universal tree. *Science*, 284, 2124–2128.
- Doolittle, R.F. (2002). Biodiversity: microbial genomes multiply. *Nature*, 416, 697–700.
- Duret, L. (2000). tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of highly expressed genes. *Trends Genet.*, 16, 287–289.
- Dykhuizen, D.E. (1998). Santa Rosalia revisited: why are there so many species of bacteria? *Antonie Van Leeuwenhoek*, 73, 25–33.
- Edwards, R.A. & Rohwer, F. (2005). Viral metagenomics. *Nat. Rev. Microbiol.*, 3, 504–510.
- Eisen, J.A. (2000). Assessing evolutionary relationships among microbes from whole-genome analysis. *Curr. Opin. Microbiol.*, 3, 475–480.
- Eisen, J.A. & Fraser, C.M. (2003). Phylogenomics: intersection of evolution and genomics. *Science*, 300, 1706–1707.
- Feder, M.E. & Mitchell-Olds, T. (2003). Evolutionary and ecological functional genomics. *Nat. Rev.*, 4, 649–655.
- Field, D. & Hughes, J. (2005). Cataloguing our current genome collection. *Microbiology*, 151, 1016–1019.
- Field, D., Hughes, J. & Gray, T. (2005). *The GenomeMine database*. <http://www.genomics.ceh.ac.uk/GMINE/>.
- Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R. *et al.* (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, 269, 496–512.
- Floyd, M.M., Tang, J., Kane, M. & Emerson, D. (2005). Captured diversity in a culture collection: case study of the geographic and habitat distributions of environmental isolates held at the American type culture collection. *Appl. Environ. Microbiol.*, 71, 2813–2823.
- Garrity, G.M. (2001). *Bergey's Manual of Systematic Bacteriology*, 2nd edn. Springer-Verlag, New York, NY.
- Giannini, N.P. (2003). Canonical phylogenetic ordination. *Syst. Biol.*, 52, 684–695.
- Gibbs, A. & Primrose, S. (1976). A correlation between the genome compositions of bacteriophages and their hosts. *Intervirology*, 7, 351–355.
- Grafen, A. (1989). The phylogenetic regression. *Philos. Trans. R. Soc. B*, 326, 119–157.
- Haft, D.H., Selengut, J.D., Brinkac, L.M., Zafar, N. & White, O. (2005). Genome properties: a system for the investigation of prokaryotic genetic content for microbiology, genome annotation and comparative genomics. *Bioinformatics*, 21, 293–306.
- Hammond, P.M. (1995). The current magnitude of biodiversity. In: *Global Biodiversity Assessment* (ed. Heywood, V.H.). Cambridge University Press, Cambridge, UK, pp. 113–138.
- Harvey, P.H. & Pagel, M.D. (1991). *The Comparative Method in Evolutionary Biology*. Oxford University Press, Oxford.
- Harvey, P.H. & Rambaut, A. (2000). Comparative analyses for adaptive radiations. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.*, 355, 1599–1605.
- Holt, G. (1984–1989). *Bergey's Manual of Systematic Bacteriology*, 1st edn. Lippincott Williams & Wilkins, Baltimore, MD.
- Holt, J.G., Krieg, N.R., Sneath, P.H., Staley, J.T. & Williams, S.T. (1994). *Bergey's Manual of Determinative Bacteriology*. Williams & Wilkins, Baltimore, MD.
- Horn, M., Collingro, A., Schmitz-Esser, S., Beier, C.L., Purkhold, U., Fartmann, B. *et al.* (2004). Illuminating the evolutionary history of chlamydiae. *Science*, 304, 728–730.
- Huber, H., Rachel, M.J.H.R., Fuchs, T., Wimmer, V.C. & Stetter, K.O. (2002). A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. *Nature*, 417, 63–67.
- Hurst, L.D. & Merchant, A.R. (2001). High guanine-cytosine content is not an adaptation to high temperature: a comparative analyses amongst prokaryotes. *Proc. R. Soc. Lond. B*, 268, 493–497.
- Klappenbach, J.A., Dunbar, J.M. & Schmidt, T.M. (2000). rRNA operon copy number reflects ecological strategies of bacteria. *Appl. Environ. Microbiol.*, 66, 1328–1333.
- Konstantinidis, K.T. & Tiedje, J.M. (2004). Trends between gene content and genome size in prokaryotic species with larger genomes. *Proc. Natl. Acad. Sci. USA*, 101, 3160–3165.
- Lawrence, J.G. & Ochman, H. (1998). Molecular archaeology of the *Escherichia coli* genome. *Proc. Natl. Acad. Sci. USA*, 95, 9413–9417.
- Levin, B.R. & Lenski, R.E. (1985). Bacteria and phage: a model system for the study of the ecology and co-evolution of hosts and parasites. In: *Ecology and Genetics of Host-Parasite Interactions* (eds Rollinson, D. & Anderson, R.M.). Academic Press, London, pp. 227–242.
- Lynch, M. & Conery, J.S. (2003). The origins of genome complexity. *Science*, 302, 1401–1404.
- Madigan, M.T. (2000). Extremeophilic bacteria and microbial diversity. *Ann. Mo. Bot. Gard.*, 87, 3–12.
- Madigan, M.T., Martinko, J.M. & Parker, J. (1997). *Brock Biology of Microorganisms*, 8th edn. Prentice Hall, Upper Saddle River, NJ.
- Margulis, L. & Schwartz, K.V. (1998). *Five Kingdoms: An Illustrated Guide to the Phyla of Life on Earth*, 3rd edn. Freeman, New York, NY.
- Mira, A., Ochman, H. & Moran, N.A. (2001). Deletional bias and the evolution of bacterial genomes. *Trends Genet.*, 17, 589–596.
- Moran, N.A. & Wernegreen, J.J. (2000). Lifestyle evolution in symbiotic bacteria: insights from genomics. *Trends Ecol. Evol.*, 15, 321–326.
- Muto, A. & Osawa, S. (1987). The guanine and cytosine content of genomic DNA and bacterial evolution. *Proc. Natl. Acad. Sci. USA*, 84, 166–169.
- Nakai, K. & Horton, P. (1999). PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem. Sci.*, 24, 34–36.
- Pedulla, M.L., Ford, M.E., Houtz, J.M., Karthikeyan, T., Wadsworth, C., Lewis, J.A. *et al.* (2003). Origins of highly mosaic mycobacteriophage genomes. *Cell*, 113, 171–182.
- Pollock, D.D., Eisen, J.A., Doggett, N.A. & Cummings, M.P. (2000). A case for evolutionary genomics and the comprehensive examination of sequence biodiversity. *Mol. Biol. Evol.*, 17, 1776–1788.
- Purvis, A. & Rambaut, A. (1995). Comparative analyses by independent contrasts (CAIC): an Apple Macintosh application for analysing comparative data. *Comput. Appl. Biosci.*, 11, 247–251.
- Quader, S., Isvaran, K., Hale, R.E., Miner, B.G. & Seavy, N.E. (2004). Nonlinear relationships and phylogenetically independent contrasts. *J. Evol. Biol.*, 17, 709–715.

- Rappe, M.S. & Giovannoni, S.J. (2003). The uncultured microbial majority. *Annu. Rev. Microbiol.*, 57, 369–394.
- Regenmortel, M.H.V.V., Fauquet, C.M., Bishop, D.H.L., Carstens, E.B., Estes, M.K., Lemon, S.M. *et al.* (2000). *Virus Taxonomy: the Classification and Nomenclature of Viruses. The Seventh Report of the International Committee on Taxonomy of Viruses*. Academic Press, San Diego, CA.
- Rocha, E.P. (2004). Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. *Genome Res.*, 14, 2279–2286.
- Rokas, A., Williams, B.L., King, N. & Carroll, S.B. (2003). Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, 425, 798–804.
- Rothschild, L.J. & Mancinelli, R.L. (2001). Life in extreme environments. *Nature*, 409, 1092–1101.
- Ruby, E.G., Urbanowski, M., Campbell, J., Dunn, A., Faini, M., Gunsalus, R. *et al.* (2005). Complete genome sequence of *Vibrio fischeri*: a symbiotic bacterium with pathogenic congeners. *Proc. Natl. Acad. Sci. USA*, 102, 3004–3009.
- Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, C.A. *et al.* (1977). Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*, 265, 687–695.
- Tekaia, F., Yeramian, E. & Dujon, B. (2002). Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: a global picture with correspondence analysis. *Gene*, 297, 51–60.
- Tringe, S.G., von Mering, C., Kobayashi, A., Salamov, A.A., Chen, K., Chang, H.W. *et al.* (2005). Comparative metagenomics of microbial communities. *Science*, 308, 554–557.
- Tyson, G., Chapman, J., Hugenholtz, P., Allen, E., Ram, R., Richardson, P. *et al.* (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428, 37–43.
- Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A. *et al.* (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, 304, 66–74.
- Vinogradov, A.E. (2004). Genome size and extinction risk in vertebrates. *Proc. Biol. Sci.*, 271, 1701–1705.
- Wernegreen, J.J. (2002). Genome evolution in bacterial endosymbionts of insects. *Nat. Rev. Genet.*, 3, 850–861.
- Wheeler, D.L., Church, D.M., Edgar, R., Federhen, S., Helmberg, W., Madden, T.L. *et al.* (2004). Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res.*, 32 Database Issue, D35–D40.
- Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Church, D.M. *et al.* (2005). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, 33 Database Issue, D39–D45.
- Wolf, Y.I., Rogozin, I.B., Grishin, N.V. & Koonin, E.V. (2002). Genome trees and the tree of life. *Trends Genet.*, 18, 472–479.
- Zientz, E., Silva, F.J. & Gross, R. (2001). Genome interdependence in insect-bacterium symbioses. *Genome Biol.*, 2, 1032.1–1032.6.

Editor, Paul Rainey

Manuscript received 26 July 2005

Manuscript accepted 2 September 2005