Orphans as taxonomically restricted and ecologically important genes

The abundance of orphan genes, or genes without known homologues, is amongst the greatest surprises uncovered by the sequencing of a large number of eukaryotic and bacterial genomes. It is therefore important to determine how the number of orphan genes will change as we sample new genomes. There are three possibilities. First, the number of orphans could continue to rise as we sample new genomes. Alternatively, orphan numbers could plateau in the future despite the sampling of novel taxa as has been suggested in the past (Siew & Fischer, 2003). Finally, the number could decrease by finding homes (gene families) for current orphans by improving our annotation methods and the sensitivity of our similarity searching algorithms (Skovgaard et al., 2001).

Here we examine these possibilities using data generated for a set of 122 bacterial species for which complete genomes are available. We have used these data to show that orphans are continuing to increase in number. This further emphasizes the importance of sequencing taxonomically diverse isolates (especially from environmental samples) to find novel predicted proteins. We suggest that orphans should now be classified as 'taxonomically restricted genes' (TRGs), as this concept seems more useful for advancing our knowledge of these sequences and their potential ecological significance.

Numbers of orphan genes in bacterial genomes

We examined the accumulation of bacterial orphans using the proteomes of the first 122 published bacterial species (Fig. 1a) and the decline in orphans over genomes sequenced as a percentage of total predicted proteins in these proteomes (Fig. 1b) (datasets D1 and D2 taken from the OrphanMine database; www.genomics.ceh.ac.uk/orphan_mine). These datasets were generated by comparison of each proteome to every other proteome using BLASTP with a cut-off of e^-3. D2 was generated by removing all predicted proteins smaller than 150 aa or containing any regions of low complexity (>0% calculated by SEG using default settings (Wootton & Federhen, 1993)) from D1. These orphans are predicted genes found in only one genome in this set of bacterial genomes and are only orphans with respect to this dataset (a small proportion of these genes do have matches in phage and plasmids and among bacteria without complete genome sequences – this observation will be discussed elsewhere). Fig. 1(a) shows that the number of these orphan bacterial genes is continuing to rise in a roughly linear fashion despite the large number of genomes sequenced, and this trend shows no signs of levelling off. In fact, the last 30 species included in this study provided 30% of the total orphans in our study (mean = 441 ± 643 for dataset D1; despite the large standard deviation all species contributed orphans).

With the availability of relatively few bacterial genome sequences, the addition of new species removed a large percentage of orphans (Siew & Fischer, 2003). However, as new species are added, the fall in the percentage of orphans slows and each new genome contributes very little to the decrease in orphans. In dataset D1, the percentage of orphans fell from 100 to 30% after the inclusion of the first 10 bacterial species; however, after 55 species the percentage is only down to 15%, and drops only 3% further to a value of 12% after 122 species.

Trend lines were fitted and used to predict orphan gene levels after the sampling of 200 species. For the more conservative dataset D2, the percentage of orphans after the inclusion of 122 species was 1.89% (6696 of 355 079 ORFs) and after 200 species, 1.16% (6751 of 582 000 ORFs). Therefore, although the percentage of orphans is falling, the actual number of orphans continues to rise, albeit very slowly. A similar pattern can be seen for D1 where 10% of all predicted coding regions in 200 species are predicted to be orphans. This is a far more significant percentage, but it is possible that this larger dataset contains genes which represent annotation artefacts (Skovgaard et al., 2001). However, it has been recently shown that A+T-rich, short proteins, which look like mis-annotated junk, may actually be derived from phage genomes by horizontal gene transfer (Daubin & Ochman, 2004).

These trends reveal several interesting points. First, given our current dataset for bacteria, it is not possible to make an estimate of the maximum number of orphans, as orphan growth does not show evidence of reaching a plateau. This conclusion is also supported by examining
the rate at which new protein families are discovered (Kunin et al., 2003).

Second, it appears that improved taxonomic sampling of distantly related genomes is continuing to reveal large numbers of orphans. These data suggest that the number of bacterial orphan genes will continue to increase for the foreseeable future as long as we continue to assay novel branches of the microbial tree of life. Therefore, although improved taxonomic sampling is reducing the overall percentage of orphans, it cannot be used to assign all orphans to known gene families. Furthermore, it is also likely that orphans will continue to be found in lineages that have already been heavily sampled (Hayashi et al., 2001; Perna et al., 2001).

Third, the number of currently known genes is undoubtedly a small proportion of the number of genes yet to be found as we sample more taxonomically and ecologically diverse species. It is well known that our selection of genomes for sequencing is highly biased. For example, nearly half of the species in this dataset are pathogens, and 76 of the 122 species examined here are from only two divisions, Proteobacteria and Firmicutes. Of these 122 species, 7 represent the only isolate from a division. These taxonomically unique species contribute approximately 13% of the total orphans in our dataset. It is therefore expected that our current databases are a significant underestimate of the number of new genes that might be sequenced in the future. Fortunately, there are now projects aimed at maximizing the taxonomic diversity of our current genome collection (Eisen & Fraser, 2003).

The importance of a representative sample of genomes, especially from increased numbers of environmental bacteria, is underscored by the observation that the largest numbers of orphans are contributed by genomes that share one or more of the following characteristics: distant taxonomic relatedness, ecological uniqueness or large genome size. For example, Pirellula sp. 1, the first species belonging to the division Planctomycetes to be sequenced, produced 3576 orphan genes (49% of the total genes), despite being the 100th species to be sequenced. Leptospira interrogans, the third Spirochaetales species to be sequenced and 92nd species, contains 2138 orphan genes (45% of the total genes). This genome contains two chromosomes, and the species can survive as either a saprophyte or as a facultative parasite. It is believed that L. interrogans was originally an environmental bacterium that has subsequently emerged as an important

Fig. 1. The accumulation of bacterial orphans. For this analysis, data on the number of orphans in complete bacterial genomes was taken from the OrphanMine database (www.genomics.ceh.ac.uk/orphan_mine). Data set D1 represents all the orphans found in the bacterial genomes using BLASTP similarity searches and a cut-off threshold of e^-3 (corresponds to dataset D3 in the database). In addition we created a more conservative dataset (D2) in which all predicted proteins smaller than 150 aa containing any regions of low complexity were removed (corresponds to dataset D4 in the database). (a) Plot of the cumulative number of orphans versus non-orphans. The number of orphans in datasets D1 (■) and D2 (□) are plotted showing that the number of orphans is continuing to rise in a linear fashion. Each data point represents the addition of a complete genome sequence in chronological order of publication (n = 122 species). The two species contributing the largest numbers of orphans are shown. (b) Decline in the number of orphans in datasets D1 (■) and D2 (□) as a percentage of all predicted proteins. A power curve was fitted and the $R^2$ value is shown. An extrapolation of this curve is used to predict the percentage of orphans after 200 species have been sequenced (shown by the solid black line).
human pathogen (Ren et al., 2003). The ability to inhabit two different environments, in addition to its past as an environmental organism, could help to explain the presence of such a large number of orphan genes. The two species described above are the two biggest sources of bacterial orphan genes in this dataset.

**Classifying orphans as TRGs of potential ecological importance**

The cumulative number of orphans identified in complete bacterial genomes does not appear to be levelling off. This observation reflects both the small proportion of the total bacterial diversity sampled to date and the widespread occurrence of orphans in almost all bacterial taxa, with the exception of the very small genomes of intracellular parasites or endosymbionts. This suggests that, far from being non-coding ‘junk’ DNA, these orphan sequences may be taxon-specific genes that, because of their restricted taxonomic distributions, may play an important role in bacterial adaptation. Databases are continuing to grow in size and evidence is accumulating that orphans are often real genes (Daubin & Ochman, 2004) rather than annotation artefacts (Skovgaard et al., 2001). Therefore, we should stop referring to orphans as ‘mysterious’ and start classifying them more appropriately as biologically significant TRGs.

All genes are taxonomically restricted at some level. For example, any genes found in *Eubacteria* and not in *Archaeota* or *Eukaryotes* are TRGs at the domain level. Genes restricted to *Firmicutes* or *Proteobacteria* are TRGs at the division level. The orphan genes reported in this study are TRGs at the species level because isolates of 122 different species were included in the analysis. Orphans, defined as species- or strain-level TRGs, may be of special interest for their contributions to ecological adaptation. The concept of cataloguing genes that define (are restricted to) a given taxonomic group is already established (e.g. Graham et al., 2000) and we believe orphans firmly belong within this framework.

The availability of a large collection of complete prokaryotic genome sequences makes it possible to begin to explore in detail how the evolutionary diversification of gene content reflects the ecological needs and opportunities of different taxa. Surprisingly few bacterial genes are truly universal (Charlebois & Doolittle, 2004) and many hypothetical coding regions appear to be unique to a given family, genus or species. It is also well known that strains within a species can vary greatly in their shared gene content (Lan & Reeves, 2000). The study of these TRGs could reveal the genotypic basis of exclusive ecological adaptations. Furthermore, once the contributions of undersampling of bacterial lineages and computational errors in gene prediction and assignment to gene families (to be discussed elsewhere) have been removed from our current estimated number of orphans, the number of orphans found in many genomes will probably become experimentally tractable. Therefore, orphans, better defined as TRGs restricted to the species and strain levels, should be an important target of future study.

**G. A. Wilson,**1 **N. Bertrand,**1 **Y. Patel,**1 **J. B. Hughes,**2 **E. J. Feil**2 and **D. Field**1

1Molecular Evolution and Bioinformatics Section, CEH-Oxford, Mansfield Road, Oxford, OX1 3SR, UK
2Department of Ecology and Evolutionary Biology, Brown University, USA
3Department of Biology and Biochemistry, University of Bath, Claverton Down, Bath BA2 7AY, UK


DOI 10.1099/mic.0.28146-0