COMPARATIVE GENOMICS, MINIMAL GENE-SETS AND THE LAST UNIVERSAL COMMON ANCESTOR

Eugene V. Koonin

Comparative genomics, using computational and experimental methods, enables the identification of a minimal set of genes that is necessary and sufficient for sustaining a functional cell. For most essential cellular functions, two or more unrelated or distantly related proteins have evolved; only about 60 proteins, primarily those involved in translation, are common to all cellular life. The reconstruction of ancestral life-forms is based on the principle of evolutionary parsimony, but the size and composition of the reconstructed ancestral gene-repertoires depend on relative rates of gene loss and horizontal gene-transfer. The present estimate suggests a simple last universal common ancestor with only 500–600 genes.

ESSENTIAL GENE
A gene for which knockout is lethal under certain conditions.

Comparative genomics started in earnest when two bacterial genome sequences were completed[1,2]. Although comparisons of many viral and partial cellular genomes had been carried out for years, the completion of the bacterial genomes added a new dimension and a new level of excitement[3,4]. Even the simplest cells that are known differ from viruses because they are autonomous. A complex array of functional systems, including those for translation, transcription, replication, membrane transport and energy conversion, is indispensable for maintaining cellular integrity[5]. When reverse-engineering a complex machine, one basic goal is to draw up a list of essential parts. Having a list of ESSENTIAL GENES might eventually enable a biological engineer to manipulate a cellular system to perform desirable functions[6]. The concept of a minimal gene-set for cellular life[7–10] originated from these straightforward ideas: the functional parts of a living cell are protein and RNA molecules, and the instructions for making these parts are encoded in genes.

Until recently, searching for the minimal gene-set was the domain of computational comparative genomics. Molecular-genetic studies addressed the problem indirectly because, even with today's impressive technologies, the task of engineering a minimal cell remains prohibitively difficult. However, this area of research received a huge boost (and underwent a corresponding surge in popularity) when J. Craig Venter announced that his new genomic research institute would achieve this goal within the next 3 years[11].

The gene complement of prokaryotes and eukaryotes varies from ~500 to ~10,000 genes in prokaryotes and ~2,000 to ~35,000 genes in eukaryotes. The smallest genomes sequenced so far are those of parasites that are dependent on their hosts — for example *Mycoplasma genitalium*[2], a bacterial pathogen, and *Encephalitozoon cuniculi*[12], a microsporidian. Smaller genomes will probably be found; for example, a tiny, symbiotic archaeal organism (tentatively assigned to a new phylum, Nanoarchaeota), seems to have a genome of only ~300 kilobases[13], and the sequence of this genome is eagerly anticipated. The smallest sequenced genome of an autotrophic organism — that of the hyperthermophilic bacterium *Aquifex aeolicus* — has fewer than 1,600 genes[14]. So the minimal gene-set required for autonomous cellular life must be remarkably small.

Essential genes and minimal gene-sets

Analyses of minimal gene-sets aim to define, by comparative and experimental approaches, the repertoire of genes that is necessary and sufficient to support cellular

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, NIH Building 38A, 8600 Rockville Pike, Bethesda, Maryland 20894, USA.
e-mail: koonin@ncbi.nlm.nih.gov
doi:10.1038/nrmicro751

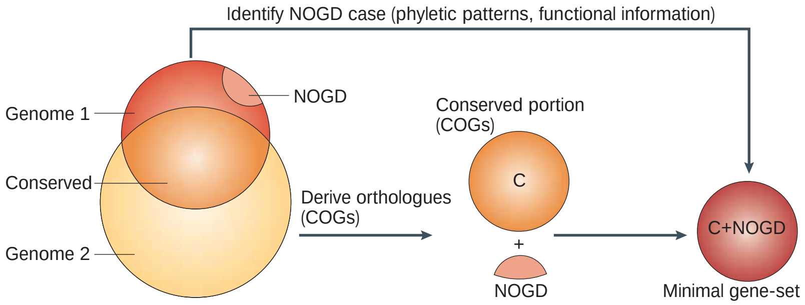
Identify NOGD case (phyletic patterns, functional information)
Genome 1
NOGD
Conserved
Genome 2
Derive orthologues (COGs)
Conserved portion (COGs)
C
+
NOGD
C+NOGD
Minimal gene-set

Figure 1 | **How to derive minimal gene-sets by genome comparison: a scheme.** Genomes 1 and 2 are arbitrary designations for two compared genomes — for example, those of *Haemophilus influenzae* and *Mycoplasma genitalium*[7]. 'C' indicates the conserved (shared) portion of genes. The non-orthologous gene displacement (NOGD) cases are arbitrarily put into the smaller genome. COGs, clusters of orthologous groups of proteins. Modified with permission from REF. 10 © (2000) Annual Reviews.

life. However, the phrase 'minimal gene-set' in itself makes no sense. It is only meaningful when associated with a defined set of environmental conditions[10,15]. So the number of minimal gene-sets could be enormous, because the minimal complement of genes required for growth on a medium that contains all the amino acids except for cysteine will differ from that required on a medium that lacks lysine. Conceivably, the absolute minimal set of genes will correspond to the most favourable conditions possible, in which all essential nutrients are provided and there are no environmental stress factors. Under these conditions, an organism might be expected to shed many genes, sticking to the bare essentials. *M. genitalium*, with ~480 genes, is a good case in point, and this genome was dubbed 'minimal' at the time of its sequencing[2]. However, there is no proof of its actual minimality and, moreover, a parasite might have extra genes for interaction with its host — for example those that encode adhesins, proteins that allow the bacterium to stick to the surface of host cells[16].

In a minimal gene-set, each gene is essential for the survival of the cell, at least under the conditions for which the minimal gene-set was defined. Straightforward, even if less than complete, experimental validation of the minimal gene-set is therefore possible, as discussed in the subsequent sections of this review.

Minimal gene-sets: the comparative-genomics approach. When the first two bacterial genome sequences were completed, a straightforward concept came to the fore: genes that are shared by distantly related organisms are likely to be essential and a catalogue of these genes might comprise a minimal gene-set for cellular life[7]. Things are, of course, not that simple, and the colloquial phrase 'shared genes' must be redefined in (semi)formal terms. Shared genes can be defined as ORTHOLOGUES[17–19]. When two genomes are examined, it is apparent that orthologues are often duplicated after the speciation event that led to the divergence of the species that are being compared. This results in more complex situations, in which two or more genes in one genome are co-orthologous to one gene in the other genome[20]. Technical difficulties notwithstanding, sets of (probable) orthologues for two or more compared genomes can be determined by straightforward computational procedures. These are based on the identification of groups of genes from different genomes, the sequences of which are more similar to each other than they are to other sequences from the group of genomes analysed[21,22]. When this was done with the first two bacterial genomes sequenced, those

Table 1 | Some cases of non-orthologous gene displacement within the minimal gene-set*

Function/activity	*M. genitalium* gene	NOGD		Comment
		Version 1 (*M. genitalium*)/ phyletic pattern/COG	Version 2/phyletic pattern /COG‡	
Lysyl-tRNA synthetase	MG136	Class II lysyl-tRNA synthetase; most bacteria and eukaryotes, crenarchaeota; COG1190	Class I lysyl-tRNA synthetase; euryarchaeota, spirochetes, most α-proteobacteria; COG1384	The two lysyl-tRNA synthetases are unrelated — a rare case of perfectly complementary phyletic patterns
Glycyl-tRNA synthetase	MG251	One-subunit glycyl-tRNA synthetase; archaea, eukaryotes, mycoplasma, actinobacteria, *Deinococcus*; COG0423	Two-subunit glycyl-tRNA synthetase; most bacteria; COG0751/0752	The α-subunit of the bacterial synthetase enzyme is distantly related to the archaeal/eukaryotic one, but they are not orthologues
Cysteinyl-tRNA synthetase	MG253	Class I aminoacyl-tRNA synthetase; most bacteria, archaea and eukaryotes; COG0215	Unidentified; archaeal methanogens	The identity of cysteinyl-synthetase in methanogenic archaea remains unknown
DNA-dependent DNA polymerase main catalytic subunit	MG261	Bacterial DNA polymerase III (class C); all bacteria; COG0587	Class B DNA polymerase; all archaea and eukaryotes, some γ-proteobacteria, *Nostoc*; COG0417	The two classes of catalytic polymerase seem to be unrelated. The class B polymerase in bacteria is implicated in repair
Thymidylate synthase	MG227	Folate-dependent thymidylate synthase; most eukaryotes, most bacteria, archaeal methnogens; COG207	Flavin-dependent thymidylate synthase; most archaea, actinobacteria, cyanobacteria, *Rickettsia*, *Chlamydia*; COG1351	The two classes of synthase are unrelated

*See REFS 7,10. ‡An update; not all of these are present in *H. influenzae*. COGs, clusters of orthologous groups of proteins; NOGD, non-orthologous gene replacement; tRNA, transfer RNA.

Table 2 | The ubiquitous genes

Functions	Number of genes	Comments
Translation and associated functions		
Ribosomal proteins	30	–
Aminoacyl-transfer-RNA synthetases	15	–
Translation factors	6	Two are predicted to function as translation factors on the basis of their domain composition and genomic context, but this remains to be validated experimentally
Enzymes involved in RNA and protein modficiation	3	One is predicted to be an RNA-modification enzyme on the basis of its domain composition and genomic context, but this remains to be validated experimentally
Signal-recognition-particle components involved in secretion	3	–
Molecular chaperone/protease	1	Predicted function; role in translation possible on the basis of genomic context
Transcription		
RNA-polymerase subunits	2	–
Replication/repair		
DNA-polymerase subunit, exonuclease, topoisomerase	3	–
Total	63	–

of *Haemophilus influenzae* and *M. genitalium*, an important problem became apparent. For a substantial number of essential functions, different organisms use genes that are not orthologues and, in some cases, are not even homologues[23]. These genes are missed by the procedures for orthologue identification and must instead be detected using other approaches, namely the examination of the set of probable orthologues for missing essential functions. This approach was applied to the *M. genitalium*/*H. influenzae* comparison[7], and the set of orthologues was supplemented with versions of the missing, but apparently essential, genes from one of the compared genomes (the *M. genitalium* versions were chosen more or less arbitrarily, simply because this genome is smaller and therefore seems to be 'closer' to the minimal genome; FIG. 1).

The existence of two or more distinct (distantly related or non-homologous) sets of orthologues that are responsible for the same function in different organisms is called NON-ORTHOLOGOUS GENE DISPLACEMENT (NOGD)[23]. The extent of apparent NOGD between *M. genitalium* and *H. influenzae* is limited to a maximum of 15 genes in the reconstructed 256-gene absolute minimal set of orthologues[7,10]. However, subsequent, wider genome comparisons have shown that NOGD occurs with most essential genes, including some of the central components of the translation, transcription and, especially, replication machineries[10,24]. Genes that comprise an NOGD set often show partially complementary PHYLETIC PATTERNS; a few striking examples of NOGD in the 256-gene minimal gene-set derived from the *M. genitalium*/*H. influenzae* comparison are given in TABLE 1. With ~100 genomes sequenced, only ~60 genes remain ubiquitous; most of these genes are translation-system components, and a few are basal components of the transcription system (TABLE 2). These findings show that it is, in fact, more appropriate to speak of a minimal set of essential functional niches (given specified conditions) rather than of minimal sets of genes. These functional niches differ in their evolutionary/structural redundancy (propensity for NOGD), and two or more distinct solutions have evolved for most of them. This creates enormous combinatorial possibilities for constructing (theoretically, but eventually perhaps, experimentally) numerous versions of minimal gene-sets, even for the same set of conditions. The original minimal gene-set constructed from the comparison of *M. genitalium* and *H. influenzae* might approximate the set of essential functional niches. Clearly, however, this is just one of the numerous possible versions of the absolute minimal gene-set for cellular life. A recent comparative analysis of the accumulated sequenced genomes of endosymbiotic bacteria revealed a shared set of 313 genes, of which only 179 were present in *M. genitalium*[25]. This might be another incarnation of the minimal set of functions, and the limited overlap with *M. genitalium* reinforces the importance of NOGD.

Minimal gene-sets: experimental approaches. An experimental attempt to determine the number of essential genes and, by inference, the approximate size of the minimal gene-set, was undertaken by Itaya even before the advent of comparative genomics[26]. He generated 79 random gene-knockouts in the bacterium *Bacillus subtilis* and found that only 6 of them were lethal. Furthermore, simultaneous knockout of several of the remaining 73 genes did not kill the bacteria, suggesting that SYNTHETIC LETHALS are relatively rare. This early experimental analysis, although it was done on a limited scale, indicated that the minimal gene-set derived from the

ORTHOLOGUES
Homologous genes in different species that originate from the same ancestral gene in the last common ancestor of the species compared.

NON-ORTHOLOGOUS GENE DISPLACEMENT
Displacement of a gene responsible for a particular biological function in a certain set of species by a non-orthologous (unrelated or paralogous) gene in a different set of species.

PHYLETIC PATTERN
The pattern of presence or absence (representation by orthologues) of a gene in different lineages across the species tree.

Table 3 | Essential genes identified by genome-wide gene inactivation in bacteria and eukaryotes*

Species	Number of protein-coding genes	Number of genes analysed	Number of essential genes	Number of essential genes extrapolated to complete genome	Gene-inactivation method	References
M. genitalium/ M. pneumoniae	480	All (random mutagenesis)	351 without insert	265–380 (55–79%)	Transposon-insertion mutagenesis	31
B. subtilis‡	4,118	3,613	192	271 (6.6%)	Plasmid-insertion mutagenesis	33
H. influenzae	1,714	1,272	478	670 (38%)	Antisense-mediated gene inactivation	32
E. coli	4,275	3,746	620	708 (17%)	Transposon-insertion mutagenesis	15
S. cerevisiae	~6,000	5,916	1,105	1,124 (~19%)	Precise deletion by mitotic recombination	70
C. elegans	~20,000	16,757	929	1,080 (~5.4%)	Inactivation by RNA interference	71

*All experiments assayed the survival of the respective organisms on rich media. ‡Combined with previously published data.

B. subtilis genome, which contains a total of ~4,100 genes, might consist of ~300 genes. This number is remarkably close to 256, the minimal set derived by computational genomics based on the *M. genitalium*/ *H. influenzae* comparison[7].

Of course, the prescient work of Itaya produced gene numbers without revealing the identity of the genes in the minimal set. The actual experimental engineering of a cell containing a minimal gene-set is a formidable task that is beyond the routine technologies available today, although the required improvements might be within reach[27]. The crucial first step, which is technically feasible, if challenging, is the identification of complete sets of essential genes. Genetic methods used for this type of analysis include transposon-insertion mutagenesis[28], plasmid-insertion mutagenesis[29] and the inactivation of genes using antisense RNAs[30]. Genome-wide analyses of gene knockouts produced using these approaches have been reported for several bacteria and two eukaryotes (TABLE 3). Although, for technical reasons, none of these studies succeeded in mutagenizing all the genes in the respective genomes, more than 50% of genes have been disrupted in each case, which is sufficient for reliable extrapolations.

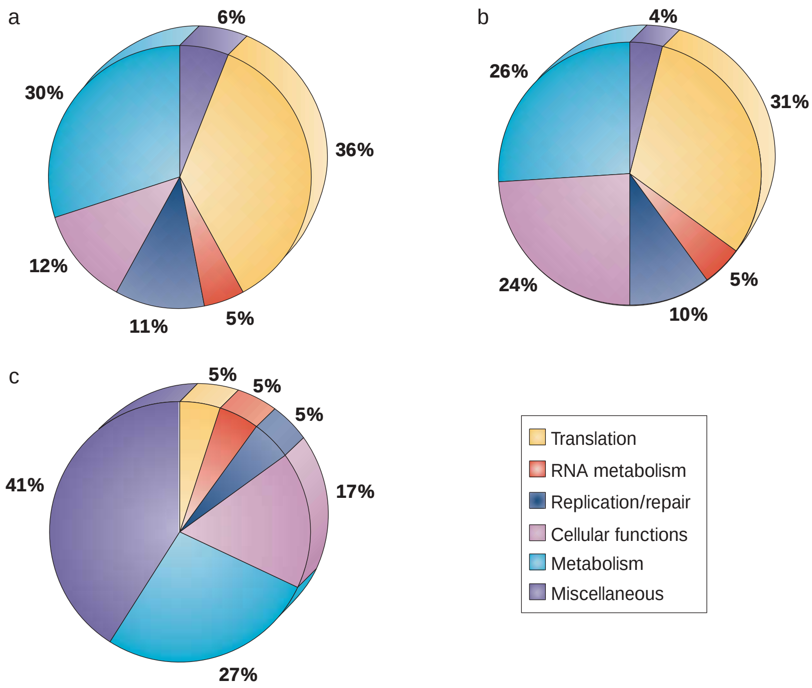

Figure 2 | **Protein functions encoded in the mimimal gene-set.** A rough, functional breakdown of (**a**) the minimal gene-set predicted from genome comparison of *Haemophilus influenzae* and *Mycoplasma genitalium*, (**b**) the set of essential genes of *Bacillus subtilis* and (**c**) the complete set of conserved genes (COGs (clusters of orthologous groups of proteins)). Cellular functions include molecular chaperones, proteins involved in cell division, proteins involved in membrane biogenesis and components of the cell envelope. The 'miscellaneous' category includes uncharacterized proteins.

These approaches yielded minimal gene-set numbers that are compatible with comparative-genomic reconstructions. Very few genes from the computationally derived minimal set were found to be dispensable, and among these, some are thought to be essential on obvious biological grounds — for example, certain aminoacyl-transfer-RNA synthetases[10,31]. It remains to be determined whether these results reflect artefacts of the knockout strategy or unexpected functional redundancy. The systematic knockout of *H. influenzae* genes[32] and *Escherichia coli* genes[15] consistently produced much larger sets of essential genes than the analogous experiments in *B. subtilis*[33] and *M. genitalium*[31] or the comparative-genomics approach (TABLE 3). The reasons for this remain unclear; one possibility is that the double membrane of Gram-negative bacteria (such as *H. influenzae* and *E. coli*) might require substantially more protein components for secretion, transport and regulation than the single membrane of Gram-positive bacteria. There are even more essential genes in eukaryotes (TABLE 3). Whatever the exact explanations, the large numerical differences between the sets of essential genes for different organisms emphasize the conditional nature of the minimal gene-set.

The computationally derived minimal gene-set and the experimentally determined sets of essential genes have similar functional features, which are distinct from

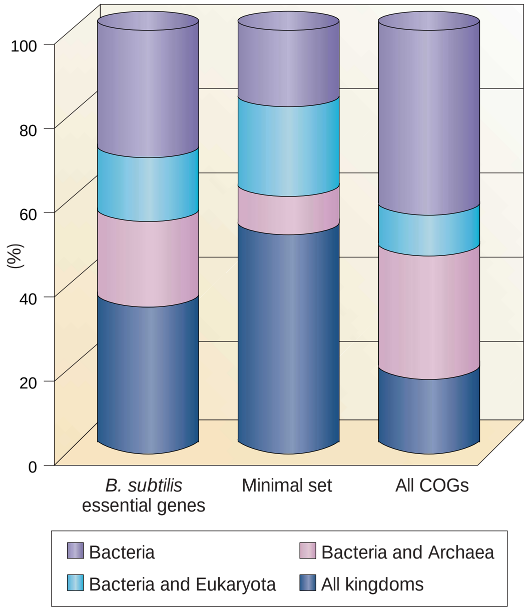

Figure 3 | **Gene conservation (phyletic spread) in the set of essential genes of *Bacillus subtilis,* the computational minimal gene-set, and the complete set of conserved genes.** A protein was assigned to one of the conservation classes if it was represented in at least one species of the respective primary kingdom. COGs, clusters of orthologous groups of proteins.

those of the general population of conserved microbial genes (the latter are represented in the database of clusters of orthologous groups (COGs) of proteins[21,34,35]). Both types of putative minimal gene-set are substantially enriched in genes that encode components of genetic-information processing systems, and contain relatively few genes for metabolic enzymes and a very small fraction of functionally uncharacterized genes (FIG. 2). It seems that we are already aware of most essential cellular functions, but we still have a long way to go before we understand the specific functions that are essential for different microbes, let alone more complex organisms.

Both computational and experimental approaches to minimal gene-set analysis have their pitfalls. Computational strategies are based primarily on the identification of probable orthologues in genomes of distantly related species and, secondarily, on filling other essential functional niches with NOGD candidates (FIG. 1). The catch is in the latter stage: in spite of the optimistic conclusion described above regarding our knowledge of essential functions, it seems almost certain that some cases of NOGD remain undetected and, accordingly, a minimal gene-set derived by a comparative-genomic computational approach is likely to be an underestimate. By contrast, genome-scale knockout experiments tend to score as essential genes those whose knockout slows down, but does not abrogate, the organism's growth. Because of this, the experimental approach can overestimate the minimal set of essential genes, perhaps substantially[33]. Conversely, in all genome-wide studies, genes are knocked out one by one and, although the number of synthetic lethals is unlikely to be large, this is bound to result in the set of essential genes missing some *bona fide* essential functions.

The first attempt at 'genome minimization' in bacteria has already been reported[36]. Yu and co-workers developed a method to combine multiple gene-deletions in one chromosome and used it to eliminate ~10% of the *E. coli* genome without detectable phenotypic effects using a rich growth medium. It seems that playing with minimal gene-sets might become an interesting avenue for research in the future, especially as the combinatorial space of possible minimal sets is explored and the compatibility (or incompatiblity) of essential genes from different, phylogenetically distinct sources, including those that do not occur together in nature due to NOGD (because they are 'on different sides' of NOGD gene-sets; TABLE 1), is revealed. Major technical advances in genome engineering are required before experimental research in this direction starts in earnest.

When considering experimental approaches to microbial-genome 'minimization', it is hard to refrain from making an analogy to the classic experiments of Sol Spiegelman and co-workers on *in vitro* Darwinian evolution of bacteriophage RNA[37]. Under the conditions in which the phage genome did not require anything except replication signals, the outcome of its evolution was rapid and spectacular: the selected variants lost >90% of the genome. Bacteria do not evolve as rapidly as phages, but it might prove valuable to devise a similar experiment using carefully designed selective conditions.

The last universal common ancestor

In principle, a minimal gene-set is a purely functional concept that does not explicitly incorporate evolution. In practice, however, the comparative-genomic approach to minimal gene-set derivation is based on the key evolutionary notion of orthology, and the resulting sets of genes should approximate those of ancestral life-forms[7]. By contrast, experimental approaches that identify essential genes do not have an evolutionary basis at all. It is a crucial, if not unexpected, observation that essential genes tend to be highly evolutionarily conserved, in terms of both the rate of sequence evolution[38] and, particularly, in terms of wide phyletic spread[15,33,38]. FIGURE 3 illustrates the predominance of highly conserved (widespread) genes in both the computationally derived minimal gene-set and in the set of essential genes of *B. subtilis* compared with the complete COG collection.

Analysis of the evolutionary conservation of sets of essential genes or members of minimal gene-sets indicates that they are relevant for the reconstruction of the gene sets of ancestral life-forms, another conceptually straightforward application of comparative genomics. There is a considerable history of conceptual thinking and comparative analysis aimed at the reconstruction of the last universal common ancestor (LUCA); a thorough representation of these can be found, for example, in a special issue of the *Journal of Molecular Evolution* edited by Antonio Lazcano and Patrick Forterre[39]. Furthermore, Carl Woese argued in

SYNTHETIC LETHALS
Genes for which simultaneous knockout is lethal, whereas individual knockouts are viable.

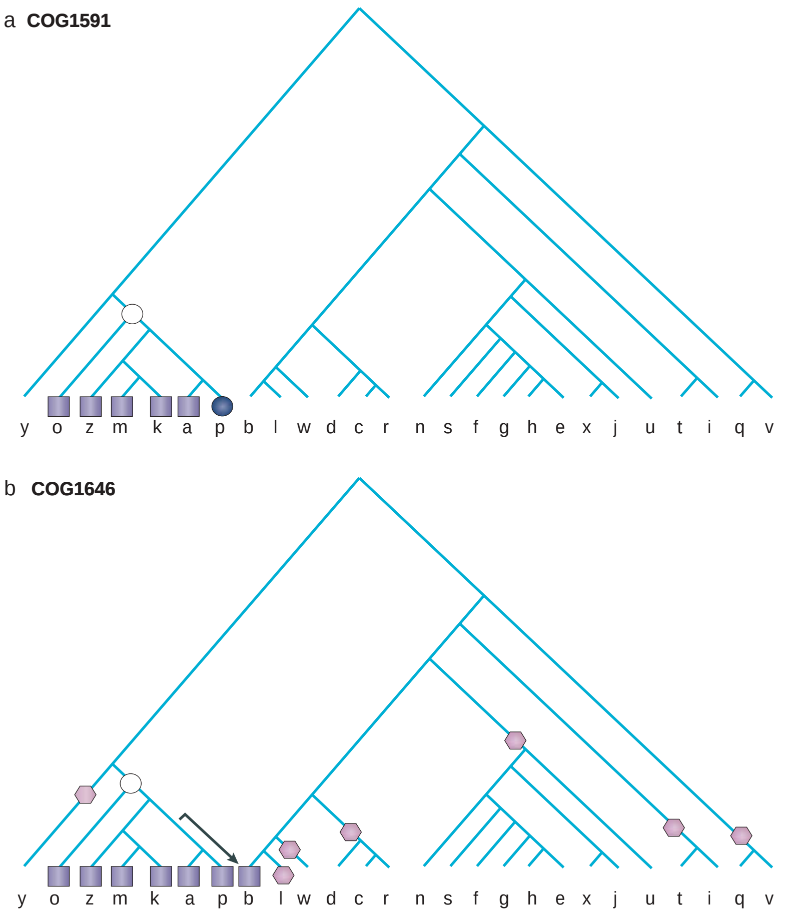
a COG1591
y o z m k a p b l w d c r n s f g h e x j u t i q v
b COG1646
y o z m k a p b l w d c r n s f g h e x j u t i q v

Figure 4 | **Parsimonious evolutionary scenarios for two genes. a** | COG1591 (an archaeal Holliday-junction resolvase): an archaeal gene that is lost in only one lineage. **b** | COG1646 (a predicted TIM (triose phosphate isomerase) -barrel enzyme): an archaeal gene that was apparently transferred to a single bacterial lineage by horizontal gene-transfer (HGT). The purple boxes indicate the presence of the given gene in the respective lineage; white circles indicate the point of emergence of the given gene (COG); blue circles indicate lineage-specific gene loss; the arrow in **b** indicates HGT; pink hexagons indicate gene losses that would need have to have occurred to explain the phyletic pattern of COG1636 if HGT was substantially less frequent than gene loss. The topology of the species tree was based on the analysis of concatenated alignments of ribosomal proteins[52]. Species are designated using a one-letter code. Pairs of related species designated by the same letter were treated in all analyses as a single entity. Eukaryotes: y, *Saccharomyces cerevisiae*. Archaea: a, *Archaeoglobus fulgidus*; k, *Pyrococcus horikoshii*; m, *Methanococcus jannaschii* and *Methanothermobacter thermoautotrophicus*; o, *Halobacterium* sp.; p, *Thermoplasma acidophilum*; z, *Aeropyrum pernix*. Bacteria: b, *Bacillus subtilis*; c, *Synechocystis* sp.; d, *Deinococcus radiodurans*; e, *Escherichia coli*; f, *Pseudomonas aeruginosa*; g, *Vibrio cholerae*; h, *Haemophilus influenzae*; i, *Chlamydia trachomatis* and *Chlamydophila pneumoniae*; j, *Mesorhizobium loti*; l, *Lactococcus lactis* and *Streptococcus pyogenes*; n, *Neisseria meningitidis*; q, *Aquifex aeolicus*; r, *Mycobacterium tuberculosis*; s, *Xylella fastidiosa*; t, *Treponema pallidum* and *Borrelia burgdorferi*; u, *Helicobacter pylori* and *Campylobacter jejuni*; v, *Thermotoga maritima*; w, *Mycoplasma genitalium* and *Mycoplasma pneumoniae*; x, *Rickettsia prowazekii*.

SPECIES TREE
A phylogenetic tree that represents evolutionary relationships between species as a whole, as opposed to phylogenetic trees for individual genes.

an influential article that a single LUCA might never have existed; instead, he proposed that extant life-forms evolved from a population of diverse organisms that exchanged their genes at an extremely high rate[40]. A critical discussion of this entire body of work would require a separate review. Here, I consider only some recent algorithmic approaches to the reconstruction of LUCA's gene set and their connection with the minimal gene-set for cellular life.

The starting material for such reconstructions consists of the phyletic patterns of orthologous gene-sets and a phylogenetic tree of the species analysed (the species tree)[41–43]. Both types of data are affected markedly by the evolutionary processes of lineage-specific gene loss and horizontal gene-transfer (HGT)[24,44–46], the scale of which only became apparent through recent comparative analyses of multiple prokaryotic genomes. The phyletic patterns of orthologous gene-sets (as represented, for example, by the COGs) show an extremely high degree of scatter, which indicates contributions from gene loss and HGT to the evolution of most of the COGs[42]. It has been suggested that HGT might have been so rampant that the feasibility and meaningfulness of a species tree would become dubious[44,46–49], although an argument was also made that the extent of HGT might have been overestimated, mainly due to artefacts of phylogenetic methods[50].

The results of several recent attempts to use information on a genomic scale for phylogenetic analyses indicates that there might be a phylogenetic signal in the sequences of large ensembles of genes, particularly those that do not seem to be subject to extensive HGT — for example, genes coding for ribosomal proteins[51–55]. So, the concept of a species tree might survive the challenge from comparative genomics, although it will inevitably have to be downgraded to a representation of a central trend in genome evolution, rather than a full depiction of this process.

Assuming that a species tree makes sense, phyletic patterns of orthologous-gene clusters can be mapped onto the branches of the tree. By using one of the implementations of the EVOLUTIONARY PARSIMONY principle[56], we can reconstruct the evolutionary scenario that includes the smallest number of elementary events (the most parsimonious scenario). The elementary events in the evolution of genes are the emergence of a gene, gene loss and HGT. FIGURE 4 shows the most parsimonious scenarios for two simple COGs. The scenario shown in FIG. 4a clearly requires just one gene-loss event, whereas the scenario shown in FIG. 4b requires one HGT event. However, this is the most parsimonious scenario only if we assume that gene loss and HGT are equally likely. By contrast, if HGT were much less frequent than gene loss, the most parsimonious scenario for COG1646 in FIG. 4b would consist of seven losses, and the COG would have been inferred to descend from LUCA. The main problem with these reconstructions is that we have no reliable estimate of the actual relative rates (probabilities) of gene loss and HGT. It is often hypothesized that gene loss is (much) more common because it is mechanistically easier and because it has occurred *en masse* in many parasites[57]. However, there is no hard evidence to support this notion, and it is suspected that a lot of HGT might go unnoticed[46]. Given these uncertainties, the score for an evolutionary scenario can be reasonably calculated as $S = l + gh$, where l is the number of losses, h is the number of HGT events (plus the event of the initial emergence of the given gene; one for each COG), and g is the 'gain penalty'[41,42]. The scenario with the lowest score is the

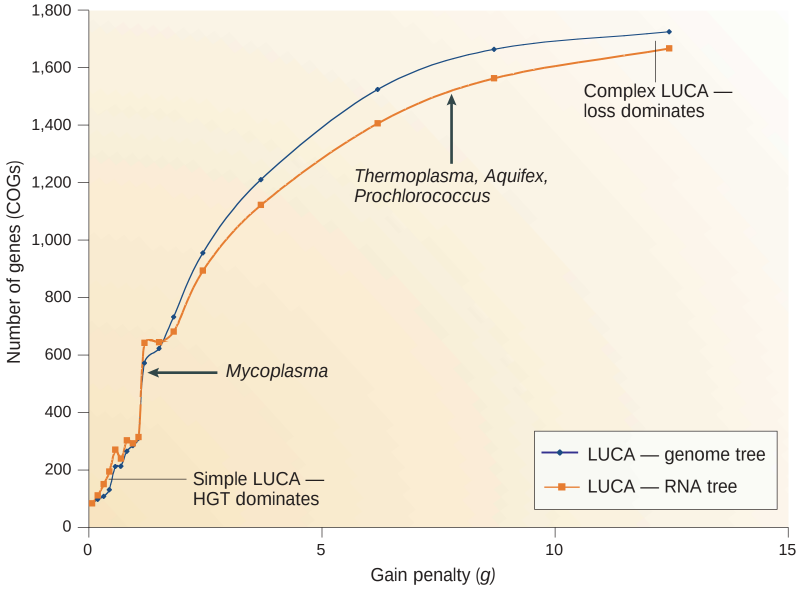

Figure 5 | **Dependence of the number of genes in a reconstructed last universal common ancestor on the relative rates of gene loss and horizontal gene transfer.** The sizes of the smallest extant genome (*Mycoplasma genitalium*) and the smallest genomes of free-living prokaryotes (*Aquifex aeolicus, Prochlorococcus marinus* and *Thermoplasma acidophilum*) are indicated for comparison. The genome tree was constructed on the basis of concatenated analysis of ribosomal proteins[52] and the ribosomal-RNA tree is described in REF. 69. COGs, clusters of orthologous groups of proteins; *g*, gain penalty; HGT, horizontal gene-transfer; LUCA, last universal common ancestor.

EVOLUTIONARY PARSIMONY
A methodological approach in evolutionary biology that aims to explain an observed distribution of character states (for example, the phyletic pattern of a gene in a species tree) by postulating the minimal number of events in the course of evolution that could have led to that distribution.

most parsimonious one. If $g >> 1$, the contribution of HGT is negligible; by contrast, if $g < 1$, the scenario is dominated by HGT.

The most parsimonious scenario for each gene (COG) shows either the presence or absence of the given gene in each of the internal (ancestral) nodes of the species tree, and therefore contributes to the reconstruction of the gene complements of these ancestors, including LUCA (FIG. 4). The combination of the scenarios for all COGs gives a conservative (because some ancestral genes might have been lost in all extant species with sequenced genomes) approximation of the most likely gene repertoire for each of the ancestors. Clearly, the size and composition of the reconstructed ancestral gene-sets depend critically on the value of the *g* parameter — the relative rates of gene loss and HGT. At high *g* values, when HGT is practically disallowed, genes with scattered phyletic patterns (for example, see FIG. 4b) will be assigned to ancestors and, in many cases, to LUCA, indicating that LUCA was a complex organism. By contrast, low *g* values indicate a simple LUCA with far fewer genes. FIGURE 5 shows the strong dependence of the number of genes in the reconstructed LUCA gene-set on the *g* value. The reconstructed ancestral gene-sets depend not only on the relative rates of gene loss and HGT but also on the topology of the species tree that is used for the reconstruction. However, examination of the gene sets for LUCA obtained with alternative species-tree topologies showed that the effect of these differences is not dramatic, at least quantitatively (as long as reasonable tree topologies are considered)[42].

Without knowing the true ratio (or ratios, because these might differ in different phylogenetic lineages) of the frequencies of gene loss and HGT, how can we choose the optimal *g* value for the evolutionary reconstruction? A definitive solution is probably out of our reach, but a crude one can be obtained by reverting to the minimal gene-set approach or, more precisely, examination of the state of essential functional niches in the reconstructed gene-set of LUCA. It is possible to trace how these niches are filled with the increasing size of the reconstructed LUCA or, to use technical jargon, with an increasing *g* value. This type of analysis becomes particularly convincing when, as well as individual functions, the completion of entire metabolic pathways is examined, as illustrated in FIG. 6. A systematic survey showed that most, if not all, known essential pathways are filled up with genes at $g = 1$ — when one assumes that gene loss and HGT are equally common, and ~600 genes are assigned to LUCA[42]. Of course, the possibility that the number of essential functional niches is underestimated in these reconstructions should always be considered. This would lead to a more complex LUCA, and functional arguments for a greater complexity of LUCA have indeed been made[58]. With that caveat, the results seem to be compatible with the notion of a relatively simple LUCA, with a genome that is considerably smaller than those of any known extant free-living prokaryotes.

Of course, the approaches to ancestral gene-set reconstruction described here are over-simplified in more than one important respect. First, only phyletic patterns and the species tree, but not the phylogenetic trees for individual genes, are taken into account. When such phylogenetic trees are analysed, some genes that are extremely widespread and are assigned to LUCA by phyletic-pattern analysis might actually have had a later origin and been disseminated by HGT; this has been noticed, in particular, for several ribosomal-protein genes that are generally considered to be refractory to HGT[59–61]. Such cases of hidden HGT could further reduce the number of genes traced back to LUCA. Second, it is assumed that the same *g* value is valid for the entire history of microbial life, from LUCA to the present day, and this is certainly an oversimplification. On one hand, it is likely that HGT between closely related species occurs at higher rates than that between distant lineages[46]. Conversely, in the early days of microbial evolution, at the time LUCA existed and shortly thereafter, HGT might have been much more rampant than it was after the main phylogenetic lineages were fully established[62]. These opposing trends would add uncertainty to our estimates of the ancestral gene-sets; narrowing the margin of error will require much more comparative-genomic analysis.

Minimal gene-sets and LUCA

The phrases 'minimal genome' and 'minimal gene-set' sound appealing and encroach on the fundamentals of life but, in reality, we learn relatively little from a computationally derived, hypothetical minimal gene-set or from an experimentally defined set of essential

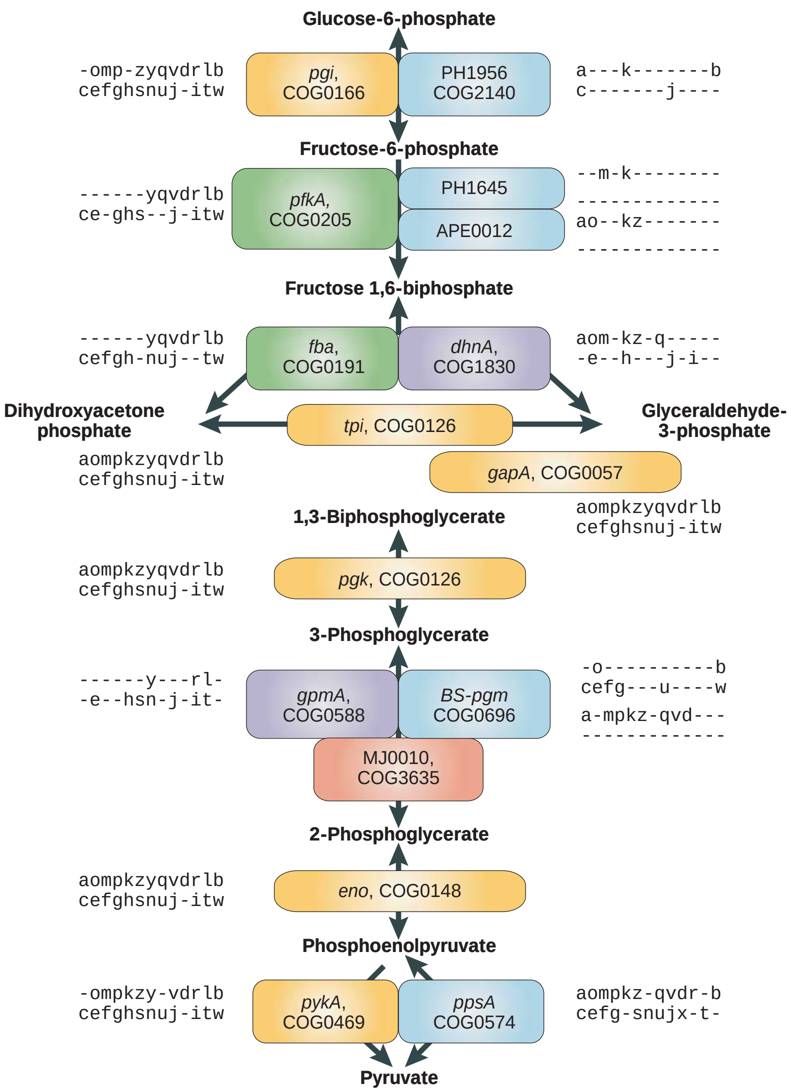

Figure 6 | **Essential functions for different versions of the last universal common ancestor: glycolysis and gluconeogenesis.** Enzyme names are accompanied by COG (clusters of orthologous groups of proteins) numbers and gene names (from *Escherichia coli*, unless indicated otherwise as follows: APE, *Aeropyrum pernix*; BS, *Bacillus subtilis*; PH, *Pyrococcus horikoshii*). Phyletic patterns are shown using the following species abbreviations: Eukaryotes: y, *Saccharomyces cerevisiae*. Archaea: a, *Archaeoglobus fulgidus*; k, *Pyrococcus horikoshii*; m, *Methanococcus jannaschii* and *Methanothermobacter thermoautotrophicus*; o, *Halobacterium* sp.; p, *Thermoplasma acidophilum*; z, *Aeropyrum pernix*. Bacteria: b, *Bacillus subtilis*; c, *Synechocystis* sp.; d, *Deinococcus radiodurans*; e, *Escherichia coli*; f, *Pseudomonas aeruginosa*; g, *Vibrio cholerae*; h, *Haemophilus influenzae*; i, *Chlamydia trachomatis* and *Chlamydophila pneumoniae*; j, *Mesorhizobium loti*; l, *Lactococcus lactis* and *Streptococcus pyogenes*; n, *Neisseria meningitidis*; q, *Aquifex aeolicus*; r, *Mycobacterium tuberculosis*; s, *Xylella fastidiosa*; t, *Treponema pallidum* and *Borrelia burgdorferi*; u, *Helicobacter pylori* and *Campylobacter jejuni*; v, *Thermotoga maritima*; w, *Mycoplasma genitalium* and *Mycoplasma pneumoniae*; x, *Rickettsia prowazekii*. Pairs of related species designated by the same letter were treated in all analyses as a single entity. The COGs that first appear in different reconstructed versions of the last universal common ancestor (LUCA; designated according to the *g* value (gain penalty) are colour-coded as follows: LUCA0.9, yellow; LUCA1.0, green; LUCA2.0, purple; LUCA3.0, red; COGs that do not appear in LUCA, blue). Multiple COGs corresponding to the same reaction are cases of non-orthologous gene displacement. *dhnA*, DhnA-type fructose 1,6-bisphosphate aldolase, and related enzymes; *eno*, enolase; *fba*, fructose/tagatose bisphosphate aldolase; *gapA*, glyceraldehyde-3-phosphate dehydrogenase A/erythrose-4-phosphate dehydrogenase; *gpmA*, phosphoglycerate mutase 1; MJ, predicted phosphoglycerate mutase family; *pfkA*, phosphofructokinase A; *pgi*, glucose-6-phosphate isomerase; *pgk*, phosphoglycerate kinase; *pgm*, phosphoglyceromutase; *ppsA*, phosphoenolpyruvate synthase A/pyruvate phosphate dikinase; *pykA*, pyruvate kinase A; *tpi*, triose phosphate isomerase. Modified with permission from REF. 42 © (2003) BioMed Central.

genes. The important realization that came from this type of analysis is the remarkable evolutionary plasticity of even the central, essential biological functions. Only a tiny group of genes (nearly all of them associated with translation and transcription) is truly ubiquitous among living things[63] (TABLE 2). Although a detailed examination of the reconstructed gene-sets is not possible here, it is interesting to note that the reconstructed gene-set of LUCA (within a broad range of g values) lacks some of the main components of the DNA-replication machinery (such as the replicative polymerase and helicase), in agreement with the hypothesis that LUCA might not have had a modern-type DNA genome and replication system[64,65]. An alternative hypothesis is that LUCA did have a DNA-replication system, but this ancestral system was obliterated by NOGD in one of the principal branches of life, probably bacteria[66–68]. This remains a distinct possibility but, at present, the scheme for a mixed RNA–DNA replication system of LUCA, with a genome distributed among multiple RNA segments, seems to be the most parsimonious reconstruction[64].

There seems to be a limited number of indispensable cellular functions, but the number of unique realizations of the minimal gene-set for cellular life is likely to be astronomically large. Construction of minimal gene-sets allows a researcher to systematically identify cases of NOGD for subsequent experimental analysis, which is crucial for uncovering the diversity of essential cellular systems. In the (perhaps not so distant) future, when experimental manipulations with genome-scale gene-sets become as routine as working with recombinant plasmids is today, theoretically derived minimal gene-sets might have an inportant role in attaining a new level of understanding of the cell. At that time, it should become clear whether organisms with minimal genomes might be good starting material for biotechnological designs, as suggested recently by J. Craig Venter and others[11].

The immortal dictum of Theodosius Dobzhansky states that "Nothing makes sense in biology except in the light of evolution". Minimal gene-sets certainly don't. These constructs, whether theoretical or experimental, are intimately linked to the problem of the reconstruction of ancestral genomes. I think Dobzhansky's saying could be extended to point out that "Nothing about (at least prokaryotic) evolution makes sense except in the light of horizontal gene-transfer and lineage-specific gene-loss" (taken loosely from a recent paper by Gogarten, Doolittle and Lawrence[46]). The apparent preponderance of HGT makes evolutionary reconstruction both challenging and interesting. Knowledge of the parameters of the evolutionary process, particularly the relative rates of gene loss and HGT, is crucial for enabling parsimony-type algorithms to produce realistic reconstructions of ancestral gene-sets. Examination of the population of the minimal set of functional niches in different versions of reconstructed ancestors (for example, LUCA) helps in determining the optimal evolutionary parameters, thereby linking the functional and evolutionary aspects of minimal gene-sets.

On the basis of the latest reconstructions, I tend to favour a simple LUCA that had a small number of genes and certain qualitative differences from modern cells — for example, in the replication system. However, the jury is definitely still out on this issue. In the future, it is easy to imagine not only experimental manipulation of minimal genomes, but also a 'Jurassic Park' of cellular evolution, with the experimental study of various theoretically reconstructed ancestral forms.

1. Fleischmann, R. D. *et al*. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512 (1995).
The first bacterial genome sequenced.
2. Fraser, C. M. *et al.* The minimal gene complement of *Mycoplasma genitalium*. *Science* **270**, 397–403 (1995).
The second bacterial genome sequenced, and still the smallest.
3. Fraser, C. M., Eisen, J. A. & Salzberg, S. L. Microbial genome sequencing. *Nature* **406**, 799–803 (2000).
4. Koonin, E. V., Aravind, L. & Kondrashov, A. S. The impact of comparative genomics on our understanding of evolution. *Cell* **101**, 573–576 (2000).
5. Alberts, B. *et al. Molecular Biology of the Cell* (Garland Science, New York, 2002).
6. Gerstein, M. & Hegyi, H. Comparing genomes in terms of protein structure: surveys of a finite parts list. *FEMS Microbiol. Rev.* **22**, 277–304 (1998).
7. Mushegian, A. R. & Koonin, E. V. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc. Natl Acad. Sci. USA* **93**, 10268–10273 (1996).
The first attempt to derive a minimal gene-set using a comparative-genomic computational approach (comparing the gene sets of *H. influenzae* and *M. genitalium*, the only two bacterial genomes sequenced at the time).
8. Maniloff, J. The minimal cell genome: 'on being the right size'. *Proc. Natl Acad. Sci. USA* **93**, 10004–10006 (1996).
9. Mushegian, A. The minimal genome concept. *Curr. Opin. Genet. Dev.* **9**, 709–714 (1999).
10. Koonin, E. V. How many genes can make a cell: the minimal-gene-set concept. *Annu. Rev. Genomics Hum. Genet.* **1**, 99–116 (2000).
11. Zimmer, C. Tinker, tailor: can Venter stitch together a genome from scratch? *Science* **299**, 1006–1007 (2003).
The closest so far to a scientific publication on the brave new project of minimal-genome construction.
12. Katinka, M. D. *et al.* Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature* **414**, 450–453 (2001).
13. Huber, H. *et al.* A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. *Nature* **417**, 63–67 (2002).
14. Deckert, G. *et al.* The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. *Nature* **392**, 353–358 (1998).
15. Gerdes, S. Y. *et al.* Experimental determination and system-level analysis of essential genes in *Escherichia coli* MG1655. *J. Bacteriol.* **185**, 5673–5684.
A thorough experimental and theoretical analysis of the essential genes of *E. coli.*
16. Rottem, S. Interaction of mycoplasmas with host cells. *Physiol. Rev.* **83**, 417–432 (2003).
17. Pauling, L. & Zuckerkandl, E. Chemical paleogenetics. Molecular 'restoration studies' of extinct forms of life. *Acta Chemica Scandinavica* **17**, S9–S16 (1963).
18. Fitch, W. M. Distinguishing homologous from analogous proteins. *Systematic Zoology* **19**, 99–106 (1970).
19. Fitch, W. M. Homology: a personal view on some of the problems. *Trends Genet*. **16**, 227–231 (2000).
20. Sonnhammer, E. L. & Koonin, E. V. Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet.* **18**, 619–620 (2002).
21. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. *Science* **278**, 631–637 (1997).
22. Huynen, M. A. & Bork, P. Measuring genome evolution. *Proc. Natl Acad. Sci. USA* **95**, 5849–5856 (1998).
23. Koonin, E. V., Mushegian, A. R. & Bork, P. Non-orthologous gene displacement. *Trends Genet.* **12**, 334–336 (1996).
24. Koonin, E. V. & Galperin, M. Y. *Sequence — Evolution — Function. Computational Approaches in Comparative Genomics* (Kluwer Academic, New York, 2002).
25. Gil, R. *et al.* The genome sequence of *Blochmannia floridanus*: comparative analysis of reduced genomes. *Proc. Natl Acad. Sci. USA* **100**, 9388–9393 (2003).
26. Itaya, M. An estimation of minimal genome size required for life. *FEBS Lett.* **362**, 257–260 (1995).
A prescient attempt to estimate the minimal genome size in the pre-genomic era. The estimate comes uncannily close to those based on computational and experimental analysis of complete genomes.
27. Venter, J. C., Levy, S., Stockwell, T., Remington, K. & Halpern, A. Massive parallelism, randomness and genomic advances. *Nature Genet.* **33** (Suppl.), 219–227 (2003).
28. Judson, N. & Mekalanos, J. J. Transposon-based approaches to identify essential bacterial genes. *Trends Microbiol.* **8**, 521–526 (2000).
29. Vagner, V., Dervyn, E. & Ehrlich, S. D. A vector for systematic gene inactivation in *Bacillus subtilis*. *Microbiology* **144**, 3097–3104 (1998).
30. Ji, Y., Woodnutt, G., Rosenberg, M. & Burnham, M. K. Identification of essential genes in *Staphylococcus aureus* using inducible antisense RNA. *Methods Enzymol.* **358**, 123–128 (2002).
31. Hutchison, C. A. *et al.* Global transposon mutagenesis and a minimal *Mycoplasma* genome. *Science* **286**, 2165–2169 (1999).
The first attempt to identify essential genes at the whole-genome level.
32. Akerley, B. J. *et al.* A genome-scale analysis for identification of genes required for growth or survival of *Haemophilus influenzae*. *Proc. Natl Acad. Sci. USA* **99**, 966–971 (2002).
33. Kobayashi, K. *et al.* Essential *Bacillus subtilis* genes. *Proc. Natl Acad. Sci. USA* **100**, 4678–4683 (2003).
34. Tatusov, R. L. *et al.* The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* **29**, 22–28 (2001).
35. Tatusov, R. L. *et al.* The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 2003 Sep 11; [epub ahead of print].
36. Yu, B. J. *et al.* Minimization of the *Escherichia coli* genome using a Tn5-targeted Cre/*loxP* excision system. *Nature Biotechnol.* **20**, 1018–1023 (2002).
37. Mills, D. R., Peterson, R. L. & Spiegelman, S. An extracellular Darwinian experiment with a self-duplicating nucleic acid molecule. *Proc. Natl Acad. Sci. USA* **58**, 217–224 (1967).
38. Jordan, I. K., Rogozin, I. B., Wolf, Y. I. & Koonin, E. V. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res.* **12**, 962–968 (2002).
39. Lazcano, A. & Forterre, P. The molecular search for the last common ancestor. *J. Mol. Evol.* **49**, 411–412 (1999).
Introduction to a special issue on the last universal common ancestor, which provides an excellent overview of the state of this field at the end of the twentieth century.
40. Woese, C. The universal ancestor. *Proc. Natl Acad. Sci. USA* **95**, 6854–6859 (1998).
A profound discussion of the nature of the last universal common ancestor. The two principal ideas are that the last universal common ancestor did not comprise a unique species, but rather a community of organisms that engaged in rampant gene exchange, and that the different cellular systems 'crystallized' asynchronously during the early evolution of life.
41. Snel, B., Bork, P. & Huynen, M. A. Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Res.* **12**, 17–25 (2002).
The first earnest attempt to construct evolutionary scenarios on the basis of genome comparisons, taking into account gene loss and HGT.
42. Mirkin, B. G., Fenner, T. I., Galperin, M. Y. & Koonin, E. V. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol. Biol.* **3** [online], (cited 22 Sept. 2003), <http://www.biomedcentral.com/1471-2148/3/2> (2003).
A detailed analysis of parsimony algorithms for reconstruction of ancestral life forms and an attempt to use the feedback from examination of essential functional niches to adjust the parameters of the algorithms — the relative rates of gene loss and HGT.
43. Kunin, V. & Ouzounis, C. A. The balance of driving forces during genome evolution in prokaryotes. *Genome Res.* **13**, 1589–1594 (2003).
44. Doolittle, W. F. Phylogenetic classification and the universal tree. *Science* **284**, 2124–2129 (1999).
45. Doolittle, W. F. Lateral genomics. *Trends Cell Biol.* **9**, M5–M8 (1999).
46. Gogarten, J. P., Doolittle, W. F. & Lawrence, J. G. Prokaryotic evolution in light of gene transfer. *Mol. Biol. Evol.* **19**, 2226–2238 (2002).
A veritable manifesto for HGT. Makes the case for numerous instances of hidden HGT.
47. Doolittle, W. F. Uprooting the tree of life. *Sci. Am.* **282**, 90–95 (2000).
48. Pennisi, E. Genome data shake tree of life. *Science* **280**, 672–674 (1998).
49. Pennisi, E. Is it time to uproot the tree of life? *Science* **284**, 1305–1307 (1999).
50. Kurland, C. G., Canback, B. & Berg, O. G. Horizontal gene transfer: a critical view. *Proc. Natl Acad. Sci. USA* **100**, 9658–9662 (2003).
A useful counterpoint to reference 46. Makes the argument that numerous apparent cases of HGT are artefacts.
51. Clarke, G. D., Beiko, R. G., Ragan, M. A. & Charlebois, R. L. Inferring genome trees by using a filter to eliminate phylogenetically discordant sequences and a distance matrix based on mean normalized BLASTP scores. *J. Bacteriol.* **184**, 2072–2080 (2002).
52. Wolf, Y. I., Rogozin, I. B., Grishin, N. V., Tatusov, R. L. & Koonin, E. V. Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evol. Biol.* **1** [online], (cited 22 Sept. 2003), < http://www.biomedcentral.com/1471-2148/1/8> (2003).
53. Wolf, Y. I., Rogozin, I. B., Grishin, N. V. & Koonin, E. V. Genome trees and the tree of life. *Trends Genet.* **18**, 472–479 (2002).
54. Korbel, J. O., Snel, B., Huynen, M. A. & Bork, P. SHOT: a web server for the construction of genome phylogenies. *Trends Genet.* **18**, 158–162 (2002).
55. Daubin, V., Gouy, M. & Perriere, G. A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. *Genome Res.* **12**, 1080–1090 (2002).
56. Nei, M. & Kumar, S. *Molecular Evolution and Phylogenetics* (Oxford Univ. Press, Oxford, 2001).
57. Moran, N. A. Microbial minimalism: genome reduction in bacterial pathogens. *Cell* **108**, 583–586 (2002).
58. Glansdorff, N. About the last common ancestor, the universal life-tree and lateral gene transfer: a reappraisal. *Mol. Microbiol.* **38**, 177–185 (2000).
59. Brochier, C., Philippe, H. & Moreira, D. The evolutionary history of ribosomal protein RpS14: horizontal gene transfer at the heart of the ribosome. *Trends Genet.* **16**, 529–533 (2000).
60. Matte-Tailliez, O., Brochier, C., Forterre, P. & Philippe, H. Archaeal phylogeny based on ribosomal proteins. *Mol. Biol. Evol.* **19**, 631–639 (2002).
61. Makarova, K. S., Ponomarev, V. A. & Koonin, E. V. Two C or not two C: recurrent disruption of Zn-ribbons, gene duplication, lineage–specific gene loss, and horizontal gene transfer in evolution of bacterial ribosomal proteins. *Genome Biology* **2** [online], (cited 11 Sept. 2003), <http://genomebiology.com/2001/2/9/research/0033> (2001).
62. Woese, C. R. On the evolution of cells. *Proc. Natl Acad. Sci. USA* **99**, 8742–8747 (2002).
63. Harris, J. K., Kelley, S. T., Spiegelman, G. B. & Pace, N. R. The genetic core of the universal ancestor. *Genome Res.* **13**, 407–412 (2003).
64. Leipe, D. D., Aravind, L. & Koonin, E. V. Did DNA replication evolve twice independently? *Nucleic Acids Res.* **27**, 3389–3401 (1999).
65. Forterre, P. The origin of DNA genomes and DNA replication proteins. *Curr. Opin. Microbiol.* **5**, 525–532 (2002).
66. Forterre, P. Displacement of cellular proteins by functional analogues from plasmids or viruses could

explain puzzling phylogenies of many DNA informational proteins. *Mol. Microbiol.* **33**, 457–465 (1999).

67. Delaye, L., Vazquez, H. & Lazcano, A. in *First Step in the Origin of Life in the Universe* (ed. Chela-Flores, J.) 223–230 (Kluwer Academic, Amsterdam, 2001).
68. Dworkin, J. P., Lazcano, A. & Miller, S. L. The roads to and from the RNA world. *J. Theor. Biol.* **222**, 127–134 (2003).
69. Olsen, G. J., Woese, C. R. & Overbeek, R. The winds of (evolutionary) change: breathing new life into microbiology. *J. Bacteriol.* **176**, 1–6 (1994).
70. Giaever, G. *et al.* Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**, 387–391 (2002).
71. Kamath, R. S. *et al.* Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature* **421**, 231–237 (2003).

Acknowledgements

I gratefully acknowledge my intellectual debt to A. Mushegian (minimal gene-set analysis) and B. Mirkin (reconstruction of evolutionary scenarios) and constructive discussions with F. Doolittle and P. Forterre on the nature of the last universal common ancestor. I thank A. Osterman for useful discussions on minimal gene-sets and for providing me with his data prior to publication.

Online links

FURTHER INFORMATION
Kyoto Encyclopedia of Genes and Genomes: http://www.genome.ad.jp/kegg/
NCBI COGs database: http://www.ncbi.nlm.nih.gov/COG/
NCBI Entrez Genome database: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome
The Tree of Life Web Project: http://www.tolweb.org/tree/phylogeny.html
TIGR Genome Projects Database: http://www.tigr.org/tdb/
Access to this interactive links box is free online.