Annu. Rev. Genomics Hum. Genet. 2000. 01:99–116

HOW MANY GENES CAN MAKE A CELL: The Minimal-Gene-Set Concept[a]

Eugene V. Koonin
National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894; e-mail: koonin@ncbi.nlm.nih.gov

Key Words comparative genomics, orthologs, nonorthologous gene displacement, genome evolution, transposon mutagenesis

■ **Abstract** Several theoretical and experimental studies have endeavored to derive the minimal set of genes that are necessary and sufficient to sustain a functioning cell under ideal conditions, that is, in the presence of unlimited amounts of all essential nutrients and in the absence of any adverse factors, including competition. A comparison of the first two completed bacterial genomes, those of the parasites *Haemophilus influenzae* and *Mycoplasma genitalium*, produced a version of the minimal gene set consisting of ∼250 genes. Very similar estimates were obtained by analyzing viable gene knockouts in *Bacillus subtilis, M. genitalium,* and *Mycoplasma pneumoniae*. With the accumulation and comparison of multiple complete genome sequences, it became clear that only ∼80 genes of the 250 in the original minimal gene set are represented by orthologs in all life forms. For ∼15% of the genes from the minimal gene set, viable knockouts were obtained in *M. genitalium*; unexpectedly, these included even some of the universal genes. Thus, some of the genes that were included in the first version of the minimal gene set, based on a limited genome comparison, could be, in fact, dispensable. The majority of these genes, however, are likely to encode essential functions but, in the course of evolution, are subject to nonorthologous gene displacement, that is, recruitment of unrelated or distantly related proteins for the same function. Further theoretical and experimental studies within the framework of the minimal-gene-set concept and the ultimate construction of a minimal genome are expected to advance our understanding of the basic principles of cell functioning by systematically detecting nonorthologous gene displacement and deciphering the roles of essential but functionally uncharacterized genes.

BACKGROUND AND HISTORY OF THE MINIMAL-GENE-SET CONCEPT

The numbers of genes in well-characterized genomes of cellular life forms range from as few as 480 in the parasitic bacterium *Mycoplasma genitalium* to

[a]The US government has the right to retain a nonexclusive, royalty-free licence in and to any copyright covering this paper.

∼100,000–150,000 in multicellular eukaryotes, such as humans (information on the completely sequenced genomes, including several complementary views of gene arrangement, can be found at http://www.ncbi.nlm.nih.gov/Entrez/Genome/org.html). Is it possible to combine comparative genomics with biochemical and molecular-genetic data to determine the minimal number of genes required to make a modern-type cell? Furthermore, what are our chances of generating a realistic list of genes that constitute such a minimal gene set? Here I explore these questions using a comparative analysis of 21 genomes of bacteria, archaea, and eukaryotes that have been completely sequenced to date and relevant experimental data.

The idea of a minimal gene set refers to the smallest possible group of genes that would be sufficient to sustain a functioning cellular life form under the most favorable conditions imaginable, that is, in the presence of a full complement of essential nutrients and in the absence of environmental stress (5, 14, 29, 32). Deriving such a minimal gene set and examining its features are of interest both to further our understanding of the basics of cell functioning and, in a more practical perspective, to define the subset of genes that are expected to be essential in most, if not all, species. Furthermore, minimal-gene-set reconstructions are, at least in principle, experimentally testable. A first-approximation, relatively straightforward test involves knocking out the genes from the minimal set and assessing the phenotype—generally, these genes are expected to be essential, although the possibility of functional redundancy should be considered. Direct testing requires actually constructing and manipulating the hypothetical minimal genome.

The upper bound of the minimal set is given by the number of genes in the smallest known genome, that of *M. genitalium*, which consists of 480 genes (10). The lower bound is suggested by salient features of any modern cell—the requirements for complete systems of translation, transcription, and replication as well as integral components of the cell membrane and minimal transport systems. A crude estimate indicates that these systems cannot be supported by <100 proteins. A remarkable experimental study that resulted in an estimation of the minimal genome size was published at the end of the pregenomic era. Itaya has shown that of 79 random gene disruptions in *Bacillus subtilis*, only 6 were lethal (16). Furthermore, even simultaneous insertions into 33 loci have produced a viable bacterium. These findings resulted in an estimate of 318–562 kb for the minimal genome, which, given the average size of ∼1 kb for a bacterial protein-coding gene, translates into 300–500 genes.

The sequencing, in 1995, of the first two complete genomes of cellular life forms, those of the parasitic bacteria *Haemophilus influenzae* (9) and *M. genitalium* (10), enabled a comparative genomic approach to the minimal-gene-set issue. This approach is based on two simple notions. (*a*) Cellular life forms are capable of importing a number of, if not all, metabolites and, accordingly, may dispense with the majority of metabolic enzymes; by contrast, cells, at least those of unicellular organisms, do not normally takeup proteins from the outside, and, therefore, all housekeeping proteins must be encoded in the genome. (*b*) Genes shared by multiple genomes are likely to be essential and therefore are good candidates for inclusion in the minimal gene set.

Generally, to apply the second notion meaningfully, one would need a number of complete genomes to compare. Any work in this direction based on the comparison of only a few genomes, let alone just two, necessarily would be preliminary, if not outright premature. The first two sequenced genomes, however, appeared to be particularly suitable for such a preliminary exercise of deriving a version of the minimal gene set, because they belong to phylogenetically distant groups of parasitic bacteria, each of which clearly has shed a number of genes in the process of its adaptation to the parasitic lifestyle. The gene losses have taken place subsequent to the divergence of these bacteria from their last common ancestor, in other words, independently; therefore, those common genes that remained, in principle, could be considered a good foundation for constructing a minimal gene set.

Based on these considerations, an attempt was made to construct a minimal gene set by comparing the *H. influenzae* and *M. genitalium* genomes (32; Figure 1). A detailed comparison of the protein sets from the two bacteria revealed 240 direct counterparts or likely orthologs (8). These genes, however, did not seem to add up to a viable minimal genome, because some of the metabolic pathways contained gaps that would preclude them from functioning in a theoretical minimal organism.

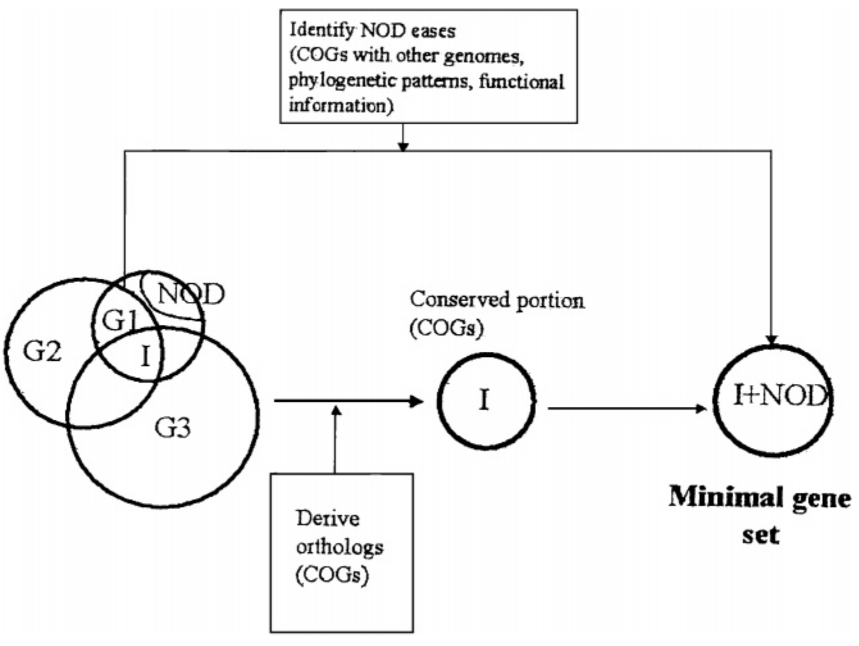

Figure 1 A generalized procedure for constructing a version of the minimal gene set. For simplicity, the schematic shows the derivation of a minimal gene set through a three-genome (*G1, G2, G3*) comparison. *I*, Intersection of the three genomes, which consists of orthologs (COGs); NOD, nonorthologous gene displacement.

To account for these gaps, one had to invoke nonorthologous gene displacement (NOD)—the situation when the same function is performed by unrelated or very distantly related and nonorthologous proteins (20). The *M. genitalium/H. influenzae* genome comparison produced a sketch of the minimal set of 256 genes that consisted mostly of orthologs, with NOD cases composing ~5% of these genes (32). In addition, this version of the minimal gene set was trimmed in a more arbitrary manner, namely by removing genes that appeared, at the time, to be specific for parasitic bacteria. Although undoubtedly just a crude approximation, the minimal gene set derived in this fashion appeared to correspond to a plausible minimalist bacterium. This organism would possess more or less complete systems for translation, transcription, and replication but would have all other cellular components, including the repair machinery, the set of molecular chaperones, the metabolic pathways, and particularly the signal transduction apparatus, reduced to a bare minimum.

THE CURRENT STATUS OF THE MINIMAL GENE SET

How does the version of the minimal gene set that was derived from the comparison between *M. genitalium* and *H. influenzae* withstand the test with new genome sequences? The first such tests have shown that ~90% of the genes from the minimal set were represented in the genome of a taxonomically distant bacterium, *Synechocystis* sp., but, in the first sequenced eukaryotic genome, that of the yeast *Saccharomyces cerevisiae*, orthologs of only 40% of the minimal set genes could be identified (19). We have the opportunity, 3 years and about 25 complete genomes later (Table 1), to assess the original version of the minimal gene set in a fairly comprehensive manner, and I do so here, by using the system of clusters of orthologous groups of proteins (COGs) from 21 complete genomes (22, 35, 36).

The COG approach is based on the notion that any group of at least three proteins from distant genomes that are more similar to each other than to any other proteins from the same genomes most probably belong to a family of orthologs (36). This notion is relevant even if the absolute level of sequence similarity between the proteins in question is relatively low; thus the COG approach accommodates both slow-evolving and fast-evolving genes. The procedure for constructing the COGs involves the detection of all triangles of genome-specific best hits from the complete matrix of pairwise comparisons between proteins encoded in the analyzed set of genomes and then merging those triangles that have a common side to form the complete orthologous families. In addition, a detailed, case-by-case analysis of each COG was performed to eliminate potential false positives and to add weakly conserved proteins that had been missed by the automatic procedure, but nevertheless appeared to be orthologous to the rest of the members of a particular COG. For the latter purpose, additional searches were performed using the PSI-BLAST program. The resulting protein families capture not only

TABLE 1 Coverage of completely-sequenced genomes by conserved families of orthologs

Species[a]	Number of genes: Total	Number of genes: In COGs (% of total)
Bacteria		
Aquifex aeolicus	1526	1265 (83%)
Thermotoga maritima	1852	1437 (78%)
Rickettsia prowazekii	834	632 (76%)
Mycoplasma genitalium	480	366 (76%)
Haemophilus influenzae	1694	1246 (74%)
Chlamydia trachomatis	895	612 (68%)
Treponema pallidum	1033	677 (66%)
Escherichia coli	4292	2752 (64%)
Bacillus subtilis	4100	2600 (63%)
Helicobacter pylori	1577	996 (63%)
Mycoplasma pneumoniae	678	408 (60%)
Chlamydia pneumoniae	1053	629 (60%)
Synechocystis sp.	3168	1883 (59%)
Borrelia burgdorferi	1256	656 (52%)
Archaea		
Archaeoglobus fulgidus	2411	1703 (71%)
Methanobacterium	1871	1319 (70%)
Methanococcus jannaschii	1747	1227 (70%)
Pyrococcus horikoshii	2072	1276 (62%)
Eukaryotes		
Saccharomyces cerevisiae	5932	2052 (35%)

[a]Within bacteria and archaea, the species are ordered by the percentage of genes included in clusters of orthologous groups of proteins (COGs).

one-to-one but also one-to-many and many-to-many orthologous relationships and hence clusters of orthologous groups of proteins.

The current collection of COGs shows two striking and, in a sense, opposing trends that are relevant for the discussion of the minimal-gene-set concept (35; http://www.ncbi.nlm.nih.gov/COG). First, it is notable that 55%–83% of the proteins encoded in each of the bacterial and archaeal genomes belong to the COGs, which, it should be emphasized, by definition include representatives of at least three phylogenetically distant clades (Table 1). Thus a good majority of bacterial and archaeal proteins are, in fact, highly conserved in evolution. Second, most of the COGs comprise only a few clades, whereas ubiquitous or nearly ubiquitous COGs are a small minority. The composition of protein families can be conveniently described using the language of phylogenetic patterns, that is, the patterns of species that are represented or missing in a given COG

(36; http://www.ncbi.nlm.nih.gov/COG/palog?pytall=a). Similar approaches to the analysis of phylogenetic representation of protein families have been developed by two other groups (11, 33). Among the 2112 COGs that comprise the current collection, as many as 1234 unique patterns are seen, which emphasizes the evolutionary plasticity of the families. The predominant evolutionary explanations for this mosaicism include clade-specific gene loss and horizontal gene transfer—phenomena that are increasingly recognized as major evolutionary factors, at least in the prokaryotic world (2, 6, 7, 21, 25, 32). On many occasions, the appearance of clade-specific gene loss may be created by rapid evolution in some of the lineages.

The phylogenetic patterns for all members of the original minimal gene set were extracted from the respective COGs. The outcome of this reanalysis is, primarily, that the role of NOD in evolution is by far more fundamental than originally imagined. The status of the members of the original minimal gene set after reassessment was performed by the COG approach can be classified as shown in Table 2 [see also supplementary material on the *Annual Reviews* web site (http://www.Annual Review.org)]. Of the minimal-gene-set members, ∼30% proved to be truly universal—evidently, this group should coincide with the set of 80 ubiquitous COGs, and this is indeed the case, except that two of the universal COGs were missed in the original study for very different reasons. The clamp loader ATPase that is encoded by the *M. genitalium* gene *MG420* was not included, because the respective mycoplasmal protein was much shorter than its *H. influenzae* counterpart and, because of that, was not considered an ortholog. Subsequent genome comparisons, taken together with experimental data, indicate that the clamp loader is a ubiquitous and essential DNA polymerase subunit (28; http://www.ncbi.nlm.nih.gov/COG). In all likelihood, the *M. genitalium* sequence contains a frameshift, and the protein is a bona fide member of the minimal gene set. The second case is that of the MG046 protein, which, although it has a highly conserved ortholog in *H. influenzae*, has been excluded from the minimal gene set because its counterpart from *Pasteurella haemolytica* has been characterized as a sialoglycoprotease (26), a function that was considered parasite specific. Comparison of multiple sequenced genomes showed, however, that this protein is highly conserved in all of them; moreover, it has been shown to be essential in *Escherichia coli* and *B. subtilis* (3), and further sequence and structure analyses have led to the prediction that this is an intracellular protease that has chaperone activity (1). These examples demonstrate how comparative analysis of multiple genomes, supported by computational and experimental studies on individual protein families, can correct shortcomings of the preliminary studies and call for caution in applying straightforward biological reasoning.

A slightly smaller group of minimal-set members are those that are conserved in all or nearly all bacteria whose genomes have been sequenced; some of these proteins are also represented in eukaryotes and/or in a subset of the archaea (Table 2). It appears most likely that the majority of these genes encode essential functions, but

TABLE 2 Phylogenetic patterns in the full COG[a] collection and in the minimal gene set classified by functional classes of proteins

Functional class of proteins	Phylogenetic pattern (number and %)					
	Full COG collection			Minimal gene set		
	Universal	Conserved in bacteria	Scattered[b]	Universal	Conserved in bacteria	Scattered[b]
Translation, ribosome structure, and biogenesis	53 (30%)	33 (18%)	93 (52%)	53 (57%)	33 (36%)	6 (7%)
Transcription	4 (5%)	2 (2%)	78 (93%)	4 (50%)	2 (25%)	2 (25%)
Replication, recombination, repair	5 (4%)	15 (12%)	108 (84%)	5 (17%)	15 (54%)	8 (29%)
Metabolism	9 (1%)	7 (1%)	686 (98%)	9 (12%)	7 (9%)	62 (79%)
Cellular processes: chaperone functions, secretion, cell division, cell wall biogenesis	9 (2%)	12 (3%)	374 (95%)	9 (28%)	12 (38%)	10 (34%)
Miscellaneous	1 (0%)	6 (1%)	637 (99%)	1 (6%)	6 (38%)	9 (56%)
Total	81 (4%)	75 (4%)	1953 (92%)	81 (32%)	75 (30%)	97 (38%)

[a]COG, cluster of orthologous groups of proteins.

[b]Defined in this case as missing at least one bacterial species; thus, within the full COG collection, this category includes proteins that are universally conserved in archaea and eukaryotes.

archaea and bacteria or archaea and eukaryotes have evolved different, in many cases evolutionarily unrelated implementations of these functions (see below). In other words, this category of proteins is a major manifestation of NOD at a deep level of evolutionary divergence (ancient NOD cases; see next section for specific examples).

More than one third of the proteins from the original minimal gene set show a less consistent phyletic distribution (Table 2). Nearly one half of these, however (37 of 92), are missing in only one or two bacterial clades and so are highly conserved genes, even if they are not ubiquitous. It seems most likely that these genes correspond to critical functions, with NOD, at least among bacteria, being an exception. Some of the remaining genes might be NOD cases that have evolved a patchy phylogenetic pattern as a result of horizontal gene transfers, whereas others indeed are likely to be nonessential and do not belong to a true minimal gene set for cellular life. Distinguishing between these situations may be possible only by examination of the specific information on the biological functions of the respective proteins (see next section).

The predominant phylogenetic patterns are significantly different for different functional categories of proteins included in the minimal gene set; predictably, the regularities seen here are the same as those observed in the full COG collection (Table 2). The ubiquitous proteins are mostly components of the translation machinery and RNA polymerase subunits; very few are scattered among other functional categories (examples include the HSP60 chaperonin and glycine hydroxymethyltransferase). The replication-recombination-repair systems are consistently conserved among bacteria, but ubiquitous proteins are in the minority (see below). By contrast, among the metabolic functions, scattered phyletic distribution is prevalent (Table 2), which indicates both the wide spread of NOD and the loss of pathways in many organisms.

On the whole, it appears that the approach of constructing a minimal gene set by comparing the genomes of just two species, which are, however, phylogenetically distant parasites, has survived the test of multiple-genome comparison reasonably well. Indeed, this set is significantly enriched in universal and highly conserved proteins and includes a relatively small number of proteins with a scattered phyletic distribution, compared with an analogous breakdown of the full set of COGs (Table 2).

A recent major extension of minimal-gene-set studies has involved global, transposon-mediated knockout mutagenesis of *M. genitalium* and *M. pneumoniae* genes (14). Viable mutants with disruptive insertions have been obtained for 129 distinct mycoplasmal genes; an estimate based on a Poisson distribution of transposon insertion sites among genes indicates that the actual number of nonessential genes should be between 180 and 215. The upper bound on the number of nonessential genes obtained by this approach suggests a minimal gene set of 265 genes, which is remarkably close to the size of the set produced by the comparative genomic approach (30). A case-by-case examination of the list of viable knockouts shows that 38 of the genes included in the computer-derived minimal gene set have been proved nonessential (14; see supplementary material at http://www.sciencemag.org/feature/data/1042937.shl). Had the 250-gene minimal set been drawn at random from the 480 genes of *M. genitalium* and given the 129 nonessential genes identified, one would expect 67 hits to fall into the minimal gene set. Thus this set is clearly enriched in essential genes. Nevertheless, the significant number of viable disruptions within the theoretical minimal gene set is somewhat unexpected and could suggest that evolutionary conservation of a gene does not automatically translate into it being essential under any conditions. Among the 38 hits into the minimal gene set, 16 are into genes with a scattered phyletic distribution, 15 are into genes conserved in all bacteria, and 7 are into universal genes. For the first of these groups of genes, the results of comparative genomic analysis converge with those of global mutagenesis in indicating that the inclusion of these genes in the minimal set simply reflected the limited nature of the original genome comparison. The viability of the disruptions of conserved, particularly universal genes is, however, perplexing. Certainly, as indicated by Hutchison and coworkers (14), nonessentiality of a gene under laboratory conditions, in the absence of competition, is not a particularly good measure of its real-life importance. For example, the disruption of the ubiquitous gene for the GroEL chaperonin may not be immediately lethal, as indicated by the mutagenesis results, but the disadvantage under any limiting conditions is expected to be devastating. Similarly, it can be rationalized that disruption of the genes coding for the components of the UvrABC excinuclease, a repair enzyme present in all bacteria, as well as RecA, a ubiquitous enzyme involved in recombination and repair, does not kill the cell, but a cell with practically no capacity for DNA repair clearly is not facing a bright future. Still, for some of the ubiquitous genes, viable disruptions of which have been reported, one is hard pressed to imagine a mechanism for the cells' survival. Cases in point are isoleucyl- and tyrosyl-tRNA synthetases. An outlandish possibility might be considered that these genes are rendered dispensable by the low level of mischarge of the respective tRNAs. It cannot be ruled out, however, that there are some unrecognized problems with the method of mutagenesis used, which result in leakage for some of the mutants.

In general, the results of the massive knockout mutagenesis of mycoplasmal genes lead to an important, even if in retrospect not particularly unexpected, conclusion. The evolutionary conservation of genes that is revealed by the comparative genomic approach reflects exactly what this approach was designed to reflect, namely the critical importance of the conserved genes for species evolution. The genes in question frequently prove to be essential under all tested conditions, but one cannot expect this correlation to be strict. Accordingly, it seems that the comparative genomic methodology is more applicable to deriving a minimal set of genes that are sufficient to sustain a robust evolutionary trajectory, for example that of the ultimate parasitism typical of the mycoplasmas (34), rather than just to support a cell under artificially favorable conditions.

NONORTHOLOGOUS GENE DISPLACEMENT—A MINIMAL GENE SET OR A MINIMAL SET OF FUNCTIONS?

Probably the most notable change in our thinking about the minimal-gene-set concept, brought about by the comparison of multiple genomes, is the much greater extent of NOD than originally perceived. Indeed, only ~30% of the members of the original minimal gene set belong to ubiquitous protein families, which suggests that many of the remaining proteins from this set are responsible for critical functions but are subject to NOD. Examination of the known biological context of the 55 members of the minimal gene set that showed a scattered phyletic distribution (an analysis that inevitably includes a degree of arbitrariness), along with the transposon mutagenesis data, suggested that 20–30 of them are likely to be nonessential and simply should be removed from a more robust version of a minimal gene set (Table 3; see supplementary material at http://www.AnnualReview.org). The remaining proteins in this category are likely NOD cases. Moreover, whenever a protein is found in all bacteria but not in archaea and eukaryotes, or in bacteria and eukaryotes but not in archaea, NOD appears most probable. In some of the apparent NOD cases, both alternative solutions for the same functional niche are known, whereas in others, one of them remains to be identified (Table 4). Proteins that compose a NOD pair tend to display complementary phyletic patterns (Table 4). Although this complementarity may not be perfect, because it is common for some organisms to encode both members of a NOD pair, this feature can be used to predict previously undetected NOD cases (Table 4; EV Koonin & MY Galperin, unpublished data).

There is no functional category of proteins or functional system within the minimal gene set that would be immune to NOD, but in some systems it is a relatively rare exception, whereas in others the majority of functions are not performed by orthologs in all organisms. The translation machinery is by far more uniform in all life forms than any other functional system, but, even here, several notable examples of NOD are seen (Table 4). The cases of glutamine and asparagine activation for protein synthesis are particularly interesting, because these amino acids are linked to the cognate tRNAs by two completely different mechanisms—either via the corresponding aminoacyl-tRNA synthetases or via the transamidation mechanisms (Table 4). In these cases, as in some others, NOD is manifested as a one-to-many, rather than a one-to-one, relationship—a single gene is displaced by three unrelated genes whose products provide the same function as a complex (15, 17). The most striking display of NOD is seen in the DNA replication system, where the principal components in bacteria are not orthologous and, in some cases, appear to be unrelated to those in archaea/eukaryotes (24; Table 4).

NOD can be interpreted in a broader sense, with entire systems and pathways displacing others for a particular general role. Thus, glycolysis, or at least its lower part leading from trioses to phosphoenopyruvate, is nearly universal and,

TABLE 3 Members of the original minimal gene set that show scattered phylogenetic patterns and are predicted to be dispensable (examples)

M. genitalium gene	Function/activity	Phylogenetic pattern[a]	Transposon insertion knockout reported?
MG012	Ribosomal protein S6 modification Enzyme (glutaminyl transferase)	`am-k---ce--h--gp-----`	No
MG104	Exoribonuclease (RNase B family)	`----yqvceb-hujgp-lin-`	No
MG346	rRNA methylase (SpoU class)	`-------cebrh--gp--in-`	Yes
MG278	Guanosine polyphosphate Pyrophosphohydrolase (SpoT)	`-----qvcebrhujgpo----`	Yes
MG097	Uracil DNA glycosylase	`----y---ebrhujgpo-in-`	No
MG262.1	Formamidopyrimidine-DNA glycosylase	`-------cebrh--gp-----`	No
MG408	Peptide methionine sulfoxide reductase	`--t-yqvcebrhujgp-l---`	Yes
MG448	Conserved domain frequently associated with peptide Methionine sulfoxide reductase	`--t-yqvcebrhujgp-l---`	No
MG049	Purine-nucleoside phosphorylase	`--------eb-huj-------`	Yes
MG052	Cytidine deaminase	`----y-v-ebrh--gpo----`	Yes
MG127	Arsenate reductase	`--------eb-h--gp-----`	No
MG033	Glycerol uptake facilitator	`a-t-y-vceb-h--gpo----`	Yes
MG038	Glycerol kinase	`a---yqvcebrh--gpo----`	No
MG299	Phosphotransacetylase	`------vcebrh--gpol--x`	Yes

[a]Pattern notation: a letter indicates presence of the respective species in the given cluster of orthologous groups of proteins, and a hyphen in the corresponding position indicates its absence. Species abbreviations: a, *Archeoglobus fulgidus*; m, *Methanococcus jannaschii*, t, *Methanobacterium thermoautotrophicum*; k, *Pysococcus horikoshii*; y, *S. cerevisiae*; q, *Aquifex aeolicus*; v, *Thermotoga maritima*; c, *Synechocystis* sp.; e, *E. coli*; b, *B. subtilis*; r, *Mycobacterium tuberculosis*; h, *H. influenzae*; u, *Helicobacter pylori*; j, *H. pylori* J strain; g, *M. genitalium*; p, *M. pneumoniae*; o, *Borrelia burgdorferi*; l, *Treponema pallidum*; I, *Chlamydia trachomatis*; n, *C. pneumoniae*; x, *Rickettsia prowazekii*.

TABLE 4 Nonorthologous gene displacement (NOD) within the minimal gene set (examples)

Function/activity	*M. genitalium* gene	Organisms with NOD (representative)	Phylogenetic pattern[a]	Comment
Lysyl-tRNA synthetase	*MG136*		`----yqvcebrhujgp-lin-`	Most bacteria an eukaryotes encode class II Lysyl-tRNA synthetase, whereas archaea, spirochetes, and rickettsiae, possess class I enzyme; the two enzymes are unrelated.
		Archaea, spirochetes, Rickettsia (MJ0359–*M. jannaschii*)	`amtk-----------ol--x`	
Glycyl-tRNA synthetase	*MG251*		`amtky-----r---gpol---`	Mycoplasma, spirochetes, and mycobacteria encode an archaeal/eukaryotic-type, one-subunit glycyl-tRNA synthetase, whereas most bacteria possess a distinct, 2-subunit enzyme. The α-subunit is distantly related to the archaeal/eukaryotic, but they are not orthologs.
		Most bacteria (GlyQ, Glys–*E. coli*)	`-----qvceb-huj----inx`	
Cysteinyl-tRNA synthetase	*MG253*		`a--kyqvcebrhujgpolinx`	In the methanogenic archaea, *M. jannaschaii* and *M. thermoautotrophicum*, the function of cysteinyl-tRNA synthetase is fulfilled by prolyl-tRNA synthetase, which in these organisms is a bifunctional enzyme (34a).
		Archaeal methanogens—*M. Jannaschii*, *M. thermoautotrophicum*	`-mt------------------`	
Glutamine activation for translation	*MG098*, *MG099*, *MG100*		`amtkyqvc-br-ujgpolinx`[b]	γ-Proteobacteria and eukaryotes possess an aminoacyl-tRNA synthetase for glutamine, whereas most bacteria and archaea use the transamidation mechanism (29).
		γ-Proteobacteria, eukaryotes (GlnS—*E. coli*)	`----y---e--h---------`	
DNA-dependent DNA polymerase main catalytic subunit	*MG261*[c]		`-----qvcebrhujgpolinx`	The main catalytic subunits of the DNA polymerase in bacteria and in archaea/eukaryotes appear to be unrelated; archaea possess an additional DNA polymerase not seen in eukaryotes or bacteria.
		Archaea, eukaryotes (MJ0885, MJ1630—*M. jannaschii*)	`amtky---e------------` `amtk-----------------`	
ATPase involved in DNA replication initiation (Dna-A)	*MG469*		`-----qvcebrhujgpolinx`	The ATPases involved in replication initiation in bacteria and in archaea-eukaryotes are distantly related but not orthologous (24). Not

		Archaea, eukaryotes (MTH1412—*M. thermoautotrophicum*)	`a-tky----------------`	the nearly complementary phylogenetic patterns (although NOD is likely for *M. jannaschii*).
Ribonuclease HII (family II)	*MG199*		`-----q---b----gp--in-`	A complex case of NOD—most bacteria have both a typical RNase HII (family I) and RNase HI. By contrast, mycoplasmas encode only a distinct form of RNase HII [family II; 4], whereas archaea possess only the typical RNase HII.
		Ribonuclease HII, family I: archaea, eukaryotes, most bacteria (RnhB—*E. coli*)	`amtkyqvcebrhuj--o-inx`	
		Ribonuclease HI: eukaryotes, most bacteria (RnhA—*E. coli*)	`----y--cebrhuj---l--x`	
Holliday junction resolvase endonuclease subunit	*MG291.1*		`-----qvcebrhujgp--inx`	Another complex case of NOD. RuvC and its functionally analogous but unrelated archaeal counterpart have been characterized experimentally. The mycoplasmas and *B. subtilis* do not encode orthologs of either of these but, along with most other bacteria, possess a protein that is distantly related to RuvC (L. Aravind and EV Koonin, unpublished data) and could function as an alternative resolvase.
		Archaea (MJ0947 — *M. jannaschii*)	`amtk-----------------`	
		Most bacteria (RuvC—*E. coli*)	`------vce-rhuj---linx`	
Fructose biphosphate aldolase	*MG023*		`----yqvcebrhujgpol---`	A classic case of NOD in an essential step of glycolysis. The two aldolases are distantly related but not orthologous. Note that the phylogenetic patterns are nearly complementary, but *A. aeolicus* and *E. coli* possess both aldolases.
		Archaea, *Chlamydia*, a few other bacteria (DhnA—*E. coli*)	`amtk-q--e---------in-`	

[a]The pattern notation and species abbreviations are the same as in Table 3.

[b]B subunit. Other subunits are missing in some species.

[c]MG261 is the ortholog of the replicative polymerase of *H. influenzae*, but in *M. genitalium* this protein is likely to be involved in repair, whereas MG031, a distinct form of DNA polymerase III represented only in gram-positive bacteria, is the likely replicative polymerase (14,18).

being present in the mycoplasmas, is an integral part of the minimal gene set. This pathway is, however, completely missing in *Rickettsia prowazekii*, which instead possess the tricarboxylic-acid cycle; the latter may be considered a displacement of glycolysis as the central loop of energy metabolism. In the same vein, no metabolite transport systems are ubiquitous, with unique solutions for metabolite intake in different organisms.

The prevalence of NOD suggests a shift in perspective on the entire concept of the minimal gene set. It seems that a more general and hence more robust idea is a minimal set of functional niches, most of which can be filled by proteins that belong to two or more distinct families of orthologs. A conserved core of functions with a single, ubiquitous solution certainly exists. The list of proteins included in this group is expected to further shrink with the accumulation of diverse genome sequences. There is, however, little doubt at this stage that a significant group of key proteins, probably at least 50, are truly universal, including several translation factors, the majority of aminoacyl-tRNA synthetases, and core RNA polymerase.

EXTENSION AND DEVELOPMENT OF THE MINIMAL-GENE-SET CONCEPT

With the evolution of the minimal-gene-set concept towards the more inclusive notion of a minimal set of functions, it is becoming clear that the task of delineating the minimal subset of the genome of, for example, *M. genitalium* by purely computational means is less straightforward than initially perceived, but also perhaps less generally important. The concept itself, however, may have considerable heuristic value. As mentioned above, constructing a minimal gene set makes no sense without explicitly defining the conditions under which the respective "minimal organism" should be expected to survive. Construction and analysis of minimal gene sets for different conditions could be a useful approach to predicting subsets of genes that are specifically required for life in the respective niches, for example, for thermophily. The conserved portions of minimal sets for different lifestyles are easy to identify using the tools for phyletic pattern analysis that are associated with the COG system (Table 5). The challenge lies in delineating the NOD cases to supplement these conserved sets of proteins, which requires careful computational analyses and biological reasoning and is beyond the scope of this review, but even examination of the conserved portions is of some interest. It shows, for example, that the gene set shared by all autotrophs whose genomes have been sequenced includes a large number of metabolic functions, indicating that these diverse organisms share a significant repertoire of biochemical pathways (Table 5).

What about the original quest for a "minimum minimorum" gene set for cellular life? Taken together, examination of the phylogenetic patterns, the transposon knockout data, and minimalist biochemical reasoning suggest that, in principle, a cell could be supported even by a considerably smaller number of proteins than the originally proposed 250. In addition to the proteins from the original minimal

TABLE 5 Conserved portions of hypothetical minimal gene sets for different lifestyles

	Organism's lifestyle[a] (number of shared proteins/COGs)			
Functional class of proteins	**Free-living organisms**	**Autotrophs**	**Chemoautotrophs**	**Thermophiles**
Translation, ribosome structure and biogenesis	57	64	68	62
Transcription	4	9	11	6
Replication recombination, repair	9	14	18	14
Metabolism	62	152	171	89
Cellular processes: chaperone functions, secretion, cell division, cell wall biogenesis	17	48	50	30
Miscellaneous	13	46	74	44
Total	158	320	379	237

[a]The following species whose genomes have been completely sequenced were included in this analysis: Free-living organisms—*A. fulgidus*, *M. jannaschii*, *M. thermoautotrophicum*, *P. horikoshii*, *S. cerevisiae*, *A. aeolicus*, *T. maritima*, *Synechocystis* sp., *E. coli*, *B. subtilis*; autotrophs—*A. fulgidus*, *M. jannaschii*, *M. thermoautotrophicum*, *A. aeolicus*, *Synechocystis* sp.; chemoautotrophs—*A. fulgidus*, *M. jannaschii*, *M. thermoautotrophicum*, *A. aeolicus*; thermophiles—*A. fulgidus*, *M. jannaschii*, *M. thermoautotrophicum*, *D. horikoshii*, *A. aeolicus*, and *T. maritima*.

set that turned out to be poorly conserved and, in all likelihood, dispensable, a considerable number of conserved proteins also could be tentatively subtracted. These include all repair systems; some of the remaining metabolic enzymes, transport systems, and cell wall components; and even some ribosomal proteins. The result would be a bare-bones set of ∼150 genes with the basal systems for translation, transcription, and replication; intermediate metabolism essentially reduced to glycolysis; a primitive transport system characterized by a broad specificity; and no cell wall [see supplementary material (http://www.AnnualReview.org)]. It remains questionable, of course, whether such a minimalist cell could survive under any realistic conditions. Although many conserved genes are individually dispensable, there are, so far, too few data on the effect of their simultaneous deletion, and the relatively small number of viable knockouts actually obtained in the recent large-scale experiment (14) calls for caution in interpretations. It is possible that concomitant removal of too many conserved and hence important genes would result in such a drop in fitness that, although theoretically possible, a bare-bones minimal cell could never be constructed in practice.

Comparative genomic approaches to the minimal-gene-set issue are straightforward but involve inherent uncertainties in that some of the widespread genes could still be dispensable, whereas identification of NOD can be ambiguous. This points to the importance of experimental approaches. With all of the advances of genomic engineering, however, the goal of actually constructing and manipulating a minimal genome still appears to be a major technical challenge. The recent global knockout studies, although an impressive scale-up of the analogous early experiments, still include only individual gene disruptions. Combining these in a single genome remains to be achieved, and, given that to actually define a minimal genome requires a number of trials, it seems that we are at least a few years away from a practicable minimal-genome technology. In principle, one could imagine a radically different approach based on selection of fast-replicating bacterial clones on a rich medium, perhaps starting from a strain with enhanced recombinational capabilities. This approach could be an attractive strategy modeled on the classic experiments of Spiegelman and coworkers with RNA bacteriophage genomes (27), but it is unclear whether such an approach could be implemented with bacteria on an acceptable timescale. In any case, the goal of experimentally constructing a minimal cell seems worth pursuing because not only will it help in verifying comparative genomic results and, accordingly, enhance our understanding of evolution, but a minimal cell also could provide a valuable model system for probing the principles of cell functioning.

The final issue to be tackled is the relevance, or lack thereof, of the minimal-gene-set concept to the reconstruction of ancestral genomes (31). The comparative procedure used to derive a hypothetical minimal gene set (Figure 1) has not been designed to retrace the actual course of evolution. Nevertheless, this shortcoming does not seem to justify a sweeping conclusion that the entire concept is evolutionarily irrelevant, as recently claimed (23). Not only are the universal genes a likely heritage of the "last universal common ancestor," but the identified cases of NOD can be at least tentatively mapped to specific stages of life's evolution, thus helping in the reconstruction of ancestral genomes.

CONCLUSIONS

Accumulation of multiple genome sequences provides ample material for computational approaches to minimal-genome construction. Comparative analyses of these genomes show that the majority of genes originally included in the minimal gene set derived by a comparison of the *H. influenzae* and *M. genitalium* genomes are either universal or at least conserved in all bacteria, whereas a minority show a scattered phyletic distribution. These results lead to a re-evaluation of the minimal-gene-set concept to accommodate a greater-than-originally-perceived contribution of nonorthologous gene displacement. It seems to be more appropriate to consider not a rigid minimal gene set but rather a minimal set of functional niches, some of which are occupied by members of the same orthologous family in all organisms but the majority of which allow at least two distinct solutions. Further development of the notion of the minimal gene set in which minimal gene sets are constructed for different conditions and lifestyles, for example thermophily or chemoautotrophy, seems to be a fruitful research direction.

Global knockout mutagenesis of the mycoplasmal genes, aimed at delineating a minimal gene set, has resulted in estimates that are very similar to those produced by original comparative genomic analysis but has also shown that even some of the universal or highly conserved genes can be dispensable. These results could indicate that even absolute evolutionary conservation does not automatically entail indispensability of a gene under any conditions, but their definitive interpretation requires further experiments. Actual experimental construction of a minimal genome may not be attainable in the nearest future but appears to be a goal worth pursuing.

ACKNOWLEDGMENTS

I am grateful to Arcady Mushegian for his collaboration during the development of the minimal-gene-set approach and for numerous helpful conversations and to Clyde A Hutchinson III for critical reading of the manuscript.

Visit the Annual Reviews home page at www.AnnualReviews.org

LITERATURE CITED

1. Aravind L, Koonin EV. 1999. Gleaning nontrivial structural, functional and evolutionary information about proteins by iterative database searches. *J. Mol. Biol.* 287:1023–40
2. Aravind L, Tatusov RL, Wolf YI, Walker DR, Koonin EV. 1998. Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles. *Trends Genet.* 14:442–44
3. Arigoni F, Talabot F, Peitsch M, Edgerton MD, Meldrum E. 1998. A genome-based approach for the identification of essential bacterial genes. *Nat. Biotechnol.* 16:851–56
4. Bellgard MI, Gojobori T. 1999. Identification of a ribonuclease H gene in both *Mycoplasma genitalium* and *Mycoplasma pneumoniae* by a new method for exhaustive identification of ORFs in the complete genome sequences. *FEBS Lett.* 445:6–8
5. Cho MK, Magnus D, Caplan AL, McGee D. 1999. Policy forum: genetics—ethical considerations in synthesizing a minimal genome. *Science* 286:2087–90
6. Doolittle WF. 1999. Lateral genomics. *Trends Cell Biol.* 9:M5–M8
7. Doolittle WF. 1999. Phylogenetic classification and the universal tree. *Science* 284:2124–29
8. Fitch WM. 1970. Distinguishing homologous from analogous proteins. *Syst. Zool.* 19:99–106
9. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, et al. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269:496–512
10. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, et al. 1995. The minimal gene complement of *Mycoplasma genitalium*. *Science* 270:397–403
11. Gaasterland T, Ragan MA. 1998. Microbial genescapes: phyletic and functional patterns of ORF distribution among prokaryotes. *Microb. Comp. Genomics* 3:199–217
12. Deleted in proof
13. Deleted in proof
14. Hutchison CA, Peterson SN, Gil SR, Cline RT, White O, et al. 1999. Global transposon mutagenesis and a minimal *Mycoplasma* genome. *Science* 286:2165–69
15. Ibba M, Curnow AW, Soll D. 1997. Aminoacyl-tRNA synthesis: divergent routes to a common goal. *Trends Biochem. Sci.* 22:39–42
16. Itaya M. 1995. An estimation of minimal genome size required for life. *FEBS Lett.* 362:257–60
17. Koonin EV, Aravind L. 1998. Genomics: re-evaluation of translation machinery evolution. *Curr. Biol.* 8:R266–69
18. Deleted in proof
19. Koonin EV, Mushegian AR. 1996. Complete genome sequences of cellular life

forms: glimpses of theoretical evolutionary genomics. *Curr. Opin. Genet. Dev.* 6:757–62
20. Koonin EV, Mushegian AR, Bork P. 1996. Non-orthologous gene displacement. *Trends Genet.* 12:334–36
21. Koonin EV, Mushegian AR, Galperin MY, Walker DR. 1997. Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea. *Mol. Microbiol.* 25:619–37
22. Koonin EV, Tatusov RL, Galperin MY. 1998. Beyond complete genomes: from sequence to structure and function. *Curr. Opin. Struct. Biol.* 8:355–63
23. Kyrpides N, Overbeek R, Ouzounis C. 1999. Universal protein families and the functional content of the last universal common ancestor. *J. Mol. Evol.* 49:413–23
24. Leipe DD, Aravind L, Koonin EV. 1999. Did DNA replication evolve twice independently? *Nucleic Acids Res.* 27:3389–401
25. Logsdon JM, Faguy DM. 1999. Thermotoga heats up lateral gene transfer. *Curr. Biol.* 9:R747–51
26. Mellors A, Lo RY. 1995. O-sialoglycoprotease from Pasteurella haemolytica. *Methods Enzymol.* 248:728–40
27. Mills DR, Peterson RL, Spiegelman S. 1967. An extracellular Darwinian experiment with a self-duplicating nucleic acid molecule. *Proc. Natl. Acad. Sci. USA* 58:217–24
28. Mossi R, Hubscher U. 1998. Clamping down on clamps and clamp loaders—the eukaryotic replication factor C. *Eur. J. Biochem.* 254:209–16
29. Mushegian A. 1999. The minimal genome concept. *Curr. Opin. Genet. Dev.* 9:709–14
30. Mushegian AR, Koonin EV. 1996. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc. Natl. Acad. Sci. USA* 93:10268–73
31. Mushegian AR, Koonin EV. 1998. A minimal gene complement for cellular life and reconstruction of primitive life forms by analysis of complete bacterial genomes. In *Bacterial Genomes. Physical Structure and Analysis*, ed. FJ De Bruijn, JR Lupski, GM Weinstock, pp. 478–88. New York: Chapman & Hall
32. Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, et al. 1999. Evidence for lateral gene transfer between archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature* 399:323–29
33. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. USA* 96:4285–88
34. Razin S, Yogev D, Naot Y. 1998. Molecular biology and pathogenicity of mycoplasmas. *Microbiol. Mol. Biol. Rev.* 62:1094–156
34a. Stathopoulos C, Li T, Longman R, Vothknecht UC, Becker HD, et al. 2000. One polypeptide with two aminoacyl-tRNA synthetase activities. *Scienci* 287: 479–82
35. Tatusov RL, Galperin MY, Natale DA, Koonin EV. 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* 28:33–36
36. Tatusov RL, Koonin EV, Lipman DJ. 1997. A genomic perspective on protein families. *Science* 278:631–37

Annual Review of Genomics and Human Genetics
Volume 1, 2000

CONTENTS