doi:10.1093/nar/gkt610

Minimal genome encoding proteins with constrained amino acid repertoire

Olga Tsoy[1,2], Marina Yurieva[2,3], Andrey Kucharavy[3,4], Mary O'Reilly[5] and Arcady Mushegian[3,6,*,†]

[1]A.A.Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Bolshoi Karetny per. 19, Moscow, 127994, Russia, [2]Faculty of Bioengineering and Bioinformatics, Moscow State University, Vorobievy Gory 1-73, Moscow 119992, Russia, [3]Stowers Institute for Medical Research, 1000 E 50th Street, Kansas City, MO 64110, USA, [4]École Polytechnique, Palaiseau Cedex, 91128, France, [5]O'Reilly Science Art, LLC, P.O. Box 416 Cardiff, CA 92007, USA and [6]Department of Microbiology, Molecular Genetics and Immunology, Kansas University Medical Center, Kansas City, KS 66160, USA

Received April 7, 2013; Revised June 2, 2013; Accepted June 20, 2013

ABSTRACT

Minimal bacterial gene set comprises the genetic elements needed for survival of engineered bacterium on a rich medium. This set is estimated to include 300–350 protein-coding genes. One way of simplifying an organism with such a minimal genome even further is to constrain the amino acid content of its proteins. In this study, comparative genomics approaches and the results of gene knockout experiments were used to extrapolate the minimal gene set of mollicutes, and bioinformatics combined with the knowledge-based analysis of the structure-function relationships in these proteins and their orthologs, paralogs and analogs was applied to examine the challenges of completely replacing the rarest residue, cysteine. Among several known functions of cysteine residues, their roles in the active centers of the enzymes responsible for deoxyribonucleoside synthesis and transfer RNA modification appear to be crucial, as no alternative chemistry is known for these reactions. Thus, drastic reduction of the content of the rarest amino acid in a minimal proteome appears to be possible, but its complete elimination is challenging.

INTRODUCTION

Among the extracellularly living bacteria with sequenced genomes, a mollicute *Mycoplasma genitalium* has the smallest gene repertoire, with 487 protein-coding and 43 RNA-coding genes identified thus far (1). The *M. genitalium* genome was called 'minimal' when first sequenced (2), but it has been suggested that the minimal gene set may be more appropriately defined as the complement of genes necessary and sufficient for bacterial cell propagation on a defined medium; this may require fewer genes than are encoded by *M. genitalium* (3).

In addition to the set of genes, a minimal genome specification also should include their layout on the chromosome, the definition of regulatory intergenic regions and the account of other loci important for cell survival. Recent project of high-density transposon insertions in *Caulobacter crescentus* examined these genetic elements and suggested additional dimensions of genome minimization, by delineating the insertion-tolerating portions of the intergenic regions and identifying non-essential regions of essential proteins that may be removed to strip the members of the minimal gene set to their shortest functional versions (4).

In this study, we suggest yet another possible constraint on the minimal genome, namely, a restricted amino acid composition of the encoded proteins. We ask 'Can there be a minimal gene set that does not use a particular amino acid?' To do that, we first re-define the gene content of the possible minimal gene set, i.e. the list of gene products whose protein sequences have to be constrained, and then propose the ways to restrict or eliminate the usage of the rarest amino acid, cysteine.

Computational estimates of the minimal genome use comparative genomics, i.e. matching the lists of

*To whom correspondence should be addressed. Tel: +1 703 292 8528; Fax: +1 703 292 9061; Email: mushegian2@gmail.com
†The views expressed in this article are those of the author in his personal capacity and do not necessarily represent the view of the NSF or the United States.
Present address:
Arcady Mushegian, Division of Molecular and Cellular Biosciences, National Science Foundation, Arlington, VA 22230, USA.

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

orthologous genes in different species and defining the subsets of them that appear to persist in evolution. Augmentation of such lists by the candidate cellular or viral gene products may be required to assure that the pathways encoded by engineered minimal genome are functionally complete. The first computational reconstruction of a minimal bacterial genome was done by deriving the set of orthologous protein-coding genes shared by two completely sequenced genomes of distantly related bacteria *M. genitalium* and *Haemophilus influenzae* and supplementing this list with 'missing links', i.e. functions needed for pathway completeness but performed by distantly related or unrelated proteins in two 'parental' species (3). The amended list contained 256 genes, ~5% representing such 'missing links' (3,5). Subsequent genome comparisons demonstrated that at larger evolutionary distances and with inclusion of more species, the number of shared genes declines rapidly: only ~50 proteins, mostly involved in mRNA synthesis and its translation, are conserved in all cellular organisms without exception (6). This is because even the essential functions can be performed by non-orthologous genes in different species, as metabolism is constantly modified in evolution by gene loss, gene gain and pathway 'rewiring' (5–8). As a result, comparative computational predictions of minimal gene set tend to be underestimations because the complete set of functions performed by non-orthologous proteins is not known for any pair of species separated by a relatively long evolutionary distance.

Experimental analysis of gene essentiality has also been performed in different organisms. Expectedly, the number of essential protein-coding genes—as the first approximation, defined as those genes that cannot be knocked out—tends to be higher in eukaryotes with their more complex cell biology than in prokaryotes [878 genes in yeast *Saccharomyces cerevisiae*, compared with 300–650 protein-coding genes in bacteria, depending on species and the experimental protocol (7)]. In *M. genitalium*, 101 of 487 protein-coding genes can be individually disrupted without the loss of viability, whereas the remaining 386 cannot (1). Thus, for the modern-type bacterial cell, the lower bound of the number of essential protein-coding genes may be close to 300–350, if a rich medium is provided.

For a defined set of protein products encoded by a particular organism, one can apply the ideas of orthology and functional analogy to devise the ways of eliminating most or all of instances of an amino acid from the set of these genes. Two considerations are of primary importance here. First, orthologs from closely related species often functionally complement each other, and the amino acid replacements between such orthologs are likely to preserve biological function. Thus, if an amino acid residue is found in a protein of interest but is missing from the homologous aligned position in one or more orthologs of this protein, this residue as a rule is not expected to be essential in that position. Second, an amino acid conserved all its orthologous proteins is most likely to be functionally indispensable; however, if a non-orthologous protein with the same function in other species lacks this amino acid, the protein in the minimal set can be replaced by such a protein, imitating naturally occurring non-orthologous gene displacement.

In this study, we have applied these ideas to the genome of *M. genitalium*, which has become the starting point for many kinds of bacterial genome minimization because of its small size and amenability to genetic manipulations.

MATERIALS AND METHODS

The starting list of orthologous genes shared by the majority of Mollicutes (List 1 or 'computational minimal genome') was constructed using the EdgeSearch algorithm (9), and the proteins encoded by the following 25 mollicute genomes: *Acholeplasma laidlawii* PG 8A, *Aster yellows witches broom phytoplasma, Mesoplasma florum* L1, *Mycoplasma agalactiae* PG2, *Mycoplasma arthritidis* 158L3 1, *Mycoplasma bovis* PG45, *Mycoplasma capricolum, Mycoplasma conjunctivae* HRC 581, *Mycoplasma crocodyli* MP145, *Mycoplasma fermentans* M64, *Mycoplasma gallisepticum* F, *M. genitalium* G37, *Mycoplasma hyopneumoniae* J, *Mycoplasma hyorhinis* HUB 1, *Mycoplasma leachii* PG50, *Mycoplasma mobile* 163 K, *Mycoplasma mycoides* capri, *Mycoplasma penetrans* HF 2, *Mycoplasma pneumoniae* M129, *Mycoplasma pulmonis* UAB CTIP, *Mycoplasma suis* str. Illinois, *Mycoplasma synoviae* 53, *Onion yellows phytoplasma* OY M, *Ureaplasma parvum* serovar 3 str. ATCC 27815, *Ureaplasma urealyticum* serovar 10 ATCC 33699. List 2 ('experimental minimal genome') was taken from (1). The union of the List 1 and List 2 that consists of 439 genes was the main subject of the study—List 3 or 'the derived minimal genome'. The gene content of List 1, List 2 and List 3 is given in Supplementary Table S4. A gene may be covered by none or several (in case of multiple domains) orthologous groups, or several genes might belong to the same orthologous group, e.g. if they are mollicute-specific gene duplications. Multiple alignments of orthologs from List 4 were produced by the MUSCLE program (10). The PSI-BLAST program (11) with inclusion cutoff (-h parameter) 0.01 and HHPred server (12) were used to search for distant homologs of proteins from List 1 that are not part of any mollicute orthologous group (MOG) or clusters of orthologous group (COG). Non-orthologous displacements by isofunctional proteins were identified by literature search, aided by the most recent catalog of the known displacements (13). Assignment of COGs to functional categories was taken from the COG (14) and EggNOG (15) databases.

RESULTS AND DISCUSSION

To design the strategy of removing a particular amino acid from the minimal proteome, we first re-estimated the minimal set of proteins in the smallest extracellularly living, genetically tractable bacterial species, a mollicute *M. genitalium*. We combined the data from comparative genomics and from the published results of gene knockout experiments, starting with the list of genes from *M. genitalium* that persist in more than a half of the completely sequenced species of *Mollicutes* (List 1 or 'computational minimal genome'; Figure 1 and 'Materials and Methods' section). List 1 comprises 328 of 487 *M. genitalium* protein-coding genes and is expected to track closely the functions needed for mycoplasma survival; it may be relatively little affected by nonorthologous gene displacement, which tends to be less frequent when more closely related species are compared.

The transposon mutagenesis experiments have suggested that 101 genes in *M. genitalium* can be disrupted (1). The remaining 386 experimentally determined essential genes make up List 2 ('experimental minimal genome'), which by construction is not sensitive to nonorthologous gene displacement. This list, however, may be an underestimation for its own reason, i.e. because single-gene knockout experiments do not take account of synthetic lethal effects (16,17). The union of our List 1 and List 2, consisting of 439 genes, may partially cancel out these distinct biases of the first two lists and constitutes List 3 or 'derived minimal genome' (Figure 1 and Supplementary Table S4).

The *M. genitalium* genes from List 3 together with their homologs from other species form List 4 ('minimal genome with orthologs'). These proteins belong to 345 COGs at NCBI (14) and 347 MOGs defined previously (9). The total number of proteins in these COGs is 44 789, found on average in 86 species of bacteria and archaea, and there are 6903 proteins in MOGs, found in on average in 18 species of mollicutes.

The frequencies of amino acids encoded by the derived minimal genome and the extent of their evolutionary conservation are shown in Table 1 and Supplementary Tables S1 and S2. At the level of the whole proteomes, proteins devoid of any amino acid are not common: most amino acids are present at least once in >95% of proteins in all lists; cysteine, tryptophan, non-initiatory methionine and histidine are the exceptions. These four are also the rarest amino acids among those that occupy the conserved positions in orthologous protein families, notwithstanding the fact that, at least for cysteine and tryptophan, the relatively small number of such positions accounts to a high proportion of all occurrences of these amino acids in proteins.

We focused our analysis on the rarest amino acid, cysteine. All told, 76 proteins in *M. genitalium* lack cysteine, 248 *M. genitalium* proteins have Cys residues that are substituted in at least one mollicute, and for 56 proteins, the cysteine-free orthologs are detected in more distantly related bacteria or archaea (Figure 1 and Tables 1 and 2). This gives the blueprint for orthology-directed cysteine substitutions in 380 proteins of 439 from the derived minimal genome (Table 2 and Supplementary Table S4).

The remaining 59 proteins have no cysteine-free orthologs, but three of them, i.e. MG034, MG174 and MG254, have cysteine-free paralogs. The ribosomal protein L36 (MG174) has three cysteine residues conserved in almost all organisms, but several Proteobacteria contain two paralogous forms: the Cys+ form with all three conserved cysteine residues and Cys− form, in which the cysteines are partially substituted. The Cys− form of L36 in *Mesorhizobium loti* strain MAFF303099 has no cysteines at all (18). Thymidine kinase MG034 has one conserved cysteine, whereas paralogous eukaryotic thymidine kinases have none. The third protein in this category, MG254, is NAD-dependent DNA ligase. Orthologous bacterial ligases have two fully conserved and two partially conserved cysteine residues, found close to each other in the predicted Zn-finger domain. A homologous NAD-dependent ligase from an entomopoxvirus, however, lacks the Zn-finger domain as well as the C-terminal BRCT domain; yet, it is active *in vivo* and seals the nicked substrates *in vitro* (19). We predict that a functional form of bacterial ligase without the Zn-finger domain (but probably with BRCT domain) may be engineered to provide the nick sealing and perhaps gap-repair functions for the DNA of the minimal genome.

One essential protein, i.e. folate-dependent flavine-independent thymidylate synthase ThyA (MG227), has a conserved Cys residue and no known cysteine-free homologs. The related Gram-positive bacteria, such as *Clostridiales*, however, encode non-homologous flavine-dependent thymidylate synthase ThyX (COG1351) that has only non-conserved cysteines, whereas the ThyX homologs from more distant *Oceanithermus profundus* and *Marinithermus hydrothermalis* have no cysteines at all. Thus, ThyA in minimal genome may be replaced by its cysteine-free isofunctional analog ThyX.

In three proteins (triose phosphate isomerase MG431, glyceraldehyde-3-phosphate dehydrogenase MG301 and a subunit of dihydrolipoamide dehydrogenase MG271), conserved cysteines have been experimentally mutated without loss of cell viability. Cysteine replacement in eukaryotic ortholog of MG431 results in a 4- to 6-fold decrease in K_{cat} and mild increase in K_m (20,21), and in ortholog of MG301 from *Bacillus stearothermophilus*, cysteine replacement gives a considerable decrease in K_{cat} but the wild-type level K_m (22). A more complicated case of conditional viability may be represented by MG271 ortholog in *Ralstonia eutropha*, where cysteine-free mutant is apparently being rescued by change of expression of genes in the central metabolism (23). *M. genitalium* contains only a subset of these rescuing proteins, including glycolytic enzymes; further experimentation is needed to establish whether this partial gene complement is sufficient to rescue the mutation in MG271 itself.

Two more proteins, peptide methionine sulfoxide reductases A and B (MG408, MG448), both involved in protein repair, are not essential in yeast and mammalian cells, are missing from some bacteria and archaea and are believed to be dispensable in several *Mycoplasma* species (24). Two gene products, NifS (MG336) and NifU (MG337) catalyze Fe-S cluster formation in metalloproteins and were shown to be dispensable for axenic growth of *M. agalactiae* (25).

In the case of 33 proteins, the role of conserved cysteine residues remains unknown. Thirty of these are hypothetical proteins missing from a large subset of mollicute genomes, suggesting that, even if cysteine residues are required for function in some of these proteins, permissive cultivation conditions may perhaps be found to

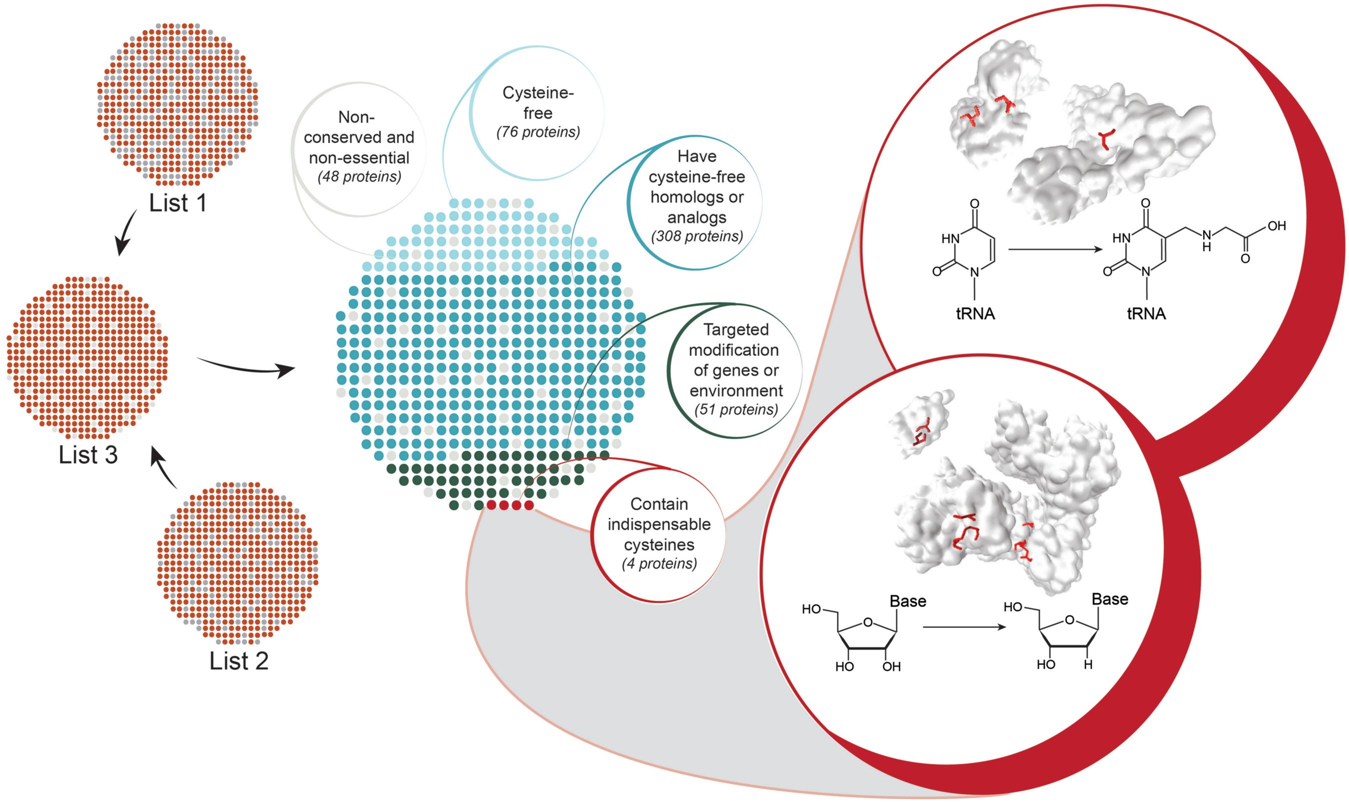
List 1
List 2
List 3
Non-conserved and non-essential (48 proteins)
Cysteine-free (76 proteins)
Have cysteine-free homologs or analogs (308 proteins)
Targeted modification of genes or environment (51 proteins)
Contain indispensable cysteines (4 proteins)
tRNA
tRNA
Base
Base
HO
HO
OH
H

Figure 1. Computational design of minimal genome with reduced content of the rarest amino acid, cysteine.

Table 1. The percentages of proteins lacking each amino acid

Amino acid	In all COGs	In minimal genome	Amino acid	In all COGs	In minimal genome	Amino acid	In all COGs	In minimal genome
A	0.38	0.13	I	0.59	0.29	Q	2.26	1.06
C	21.00	22.30	K	1.60	0.39	R	0.84	0.30
D	1.19	0.63	L	0.14	0.14	S	0.34	0.21
E	0.96	0.47	initiatory M	0.39	0.17	T	0.67	0.31
F	1.70	1.17	internal M	6.66	2.40	V	0.39	0.09
G	0.58	0.10	N	2.15	0.81	W	16.71	22.50
H	6.35	4.39	P	1.69	0.81	Y	3.50	3.04

compensate for deletion of these genes. Two components of a phosphotransferase system (PtsH MG069 and PtsG MG429) and a subunit of ATP synthase (AtpA MG401) are relatively well-studied, but the roles of conserved cysteines in these protein families await further investigation.

In seven proteins (MG019, MG052, MG106, MG110, MG375, MG421 and MG498), the function of conserved cysteines is to coordinate divalent metal cations (Zn or, in the case of MG106, Fe). In the best-studied case, cytidine deaminase MG052, the coordinating site with three cysteines in a closely related homolog of *Bacillus subtilis* can be remodeled with the aid of histidine substitutions followed by additional mutation of nearby arginine to glutamine to restore the capacity to bind Zn and partially restore catalytic activity (26). More generally, the database-wide analysis of metal-binding sites in proteins with known spatial structure (27) reveals that in the majority of Zn-binding motifs, the metal chelation is performed not by cysteines, but by a combination of aspartic acid, histidine and/or glutamine residues (Supplementary Table S5), giving rise to a testable hypothesis that many of the Zn-binding sites in proteins might be rebuilt to coordinate the catalytic metal ions without participation of the Cys residues.

All these considerations allow us to propose either robust or tentative strategies for eliminating cysteines or whole cysteine-containing proteins, from *M. genitalium* genome, altogether getting rid of >1200 cysteine residues (Tables 1 and 2 and Supplementary Table S4). The remaining eight proteins with absolutely conserved cysteines put in sharp focus one molecular function, namely, the redox potential of the thiol group. Two proteins in the minimal set, thioredoxin MG124 and thioredoxin reductase MG102, are dispensable in *Escherichia coli* when the

Table 2. Reducing cysteine content of proteins with different functions within minimal genome

Function	All[a]	List 1[b]			List 2[b]			List 3[b]			No Cys[a]	Orthologs without Cys[c]		Other ways of Cys removal[c]		Indispensable[c]	
C: Energy production	22	20	18	54	17	15	42	21	19	55	2	15	39	4	16	0	0
D: Cell division, chromosome partitioning	5	3	3	13	4	3	12	5	4	16	1	4	16	0	0	0	0
E: Amino acid transport and metabolism	13	10	10	41	11	11	41	12	12	45	0	11	42	1	3	0	0
F: Nucleotide transport and metabolism	24	23	21	83	21	19	76	24	22	85	2	15	55	3	11	4	19
G: Carbohydrate transport and metabolism	27	21	20	65	20	20	61	23	22	69	1	18	57	4	12	0	0
H: Coenzyme transport and metabolism	12	10	10	38	10	10	40	12	12	48	0	11	45	1	3	0	0
I: Lipid transport and metabolism	9	6	5	13	6	4	10	8	6	15	2	5	13	1	2	0	0
J: Translation, ribosome biogenesis	109	98	67	281	99	67	274	104	72	298	32	66	273	6	25	0	0
K: Transcription	15	13	11	46	12	9	43	14	11	46	3	11	46	0	0	0	0
L: Replication, recombination and repair	40	35	32	173	29	27	141	38	35	178	3	32	154	3	24	0	0
M: Cell wall/membrane/envelope biogenesis	12	6	6	25	9	9	40	12	12	54	0	11	49	1	5	0	0
O: Protein modification/turnover, chaperones	20	14	12	37	14	13	44	18	16	50	2	11	29	5	21	0	0
P: Inorganic ion transport and metabolism	19	17	15	54	14	12	45	18	16	56	2	16	54	1	2	0	0
R: General (molecular) function	37	29	26	109	27	25	101	34	31	123	3	30	119	1	5	0	0
S: Conserved protein, unknown function	16	9	6	19	13	11	32	15	12	35	3	12	35	0	0	0	0
T: Signal transduction	2	2	2	6	2	2	6	2	2	6	0	2	6	0	0	0	0
U: Intracellular trafficking, secretion	7	4	3	9	7	5	11	7	5	12	2	5	12	0	0	0	0
V: Defense mechanisms	8	5	5	22	6	6	19	7	7	28	0	6	19	1	9	0	0
Unknown non-conserved	90	3	2	11	65	47	127	65	47	127	18	23	59	24	68	0	0
Total	487	328	274	1099	386	315	1165	439	363	1346	76	304	1122	56	206	4	19

[a] The number indicates the counts of proteins in each functional category.

[b] Three columns represent the total of all proteins, only Cys-containing proteins and the count of Cys residues in these proteins within each functional category.

[c] Two numbers indicate Cys-containing proteins and the count of Cys residues in these proteins within each functional category.

medium is supplemented with glutathione, cysteine and methionine, probably because glutaredoxin takes over as the source of redox equivalents (28,29). Glutaredoxin-like protein MG127 may play this role in the minimal genome. Its main client in the minimal context appears to be ribonucleoside-diphosphate reductase, the enzyme that converts ribonucleosides to 2′-deoxyribonucleosides. The large subunit of class I ribonucleoside-diphosphate reductase in *M. genitalium* (MG231) itself has five conserved cysteines, all of which are essential for activity, and one of these five (C386) is essential and conserved among all three classes of ribonucleotide reductases. This residue is responsible for the thyil radical formation (30). Thus, MG127 and MG231 may be parts of an indispensable Cys-containing subsystem of the minimal genome.

Four enzymes are involved in modification of bases within transfer RNA (tRNA), also requiring the thiol groups of conserved cysteines. ThiI (MG372) uses free cysteine as a sulfur donor to convert uridine 8 into thiouridine in several tRNAs, but ThiI ortholog is missing from three mycoplasma genomes and from recently characterized *Candidatus Riesia pediculicola*, which has a drastically reduced repertoire of tRNA modifications; ThiI is also not essential in *E. coli* (31). MnmA (MG295) is involved in nucleoside thiolation in position 2 of a subset of tRNAs but is dispensable in *M. genitalium* (32), in agreement with the data that mutant *Salmonella enterica* lacking thiolated U34 is viable (33). In contrast, MnmE and GidA (MG008, MG379) add carboxymethylaminomethyl to the position 5 of U34, a modification that appears to be essential for decoding codons with the wobble base (31). Both MnmE and GidA contain pairs of conserved cysteine residues, and one cysteine in each pair is crucial for base modification, in both cases initiating the activation of a carbon atom in the pyrimidine ring for nucleophilic attack (35,36).

These results leave us with a group of four functionally coherent, but structurally diverse, proteins with invariant cysteines—MG008, MG127, MG231 and MG379—all of which are involved in ribonucleoside modification (Figure 1). The need to perform these molecular functions appears to be a formidable obstacle in the project of eliminating genetically encoded cysteine from a minimal bacterial genome. A drastic reduction of cysteine content in *M. genitalium* is, however, within reach of synthetic biology.

In summary, cysteine residue incorporated into proteins accomplishes many functions in the cell, but our study suggests that not all of them are equally important in small bacterial genomes. Indeed, such role of cysteine as protein stabilization by disulphide bonds is rare in bacterial cytoplasmic proteins (36), whereas chelation of divalent cations like Zn++ appears less relevant or more modifiable in the minimal-genome context than participation of Cys as a strong nucleophile in the active centers of enzymes and engaging in redox reactions with various electron donors, including other cysteines.

We also note that if elimination of cysteine residues was successful after all, this would make obsolete the genes whose products are involved in cysteine metabolism. In a small metabolically streamlined genome of *M. genitalium*, there is just one such protein-coding gene, i.e. cysteinyl-tRNA synthetase MG253, as well as an RNA-coding gene for cognate tRNA-Cys. If these genetic elements are no longer in the picture, the UGU and UGC codons would become available for reassignment by synthetic biologists, concerned with the projects of evolving the genetic code and, in the exact opposite of the current work, addition to the repertoire of naturally encoded amino acids. Another direction of synthetic biology that may benefit from the knowledge of cysteine dispensability is design of physiologically active cysteine-free versions of proteins, which can be then genetically derivatized by cysteine in specific positions suitable for affinity labeling. Among the homologs of the proteins from the minimal set, such strategy has been applied to investigate the E1 component of pyruvate dehydrogenase from *E. coli* (37).

In this work, we examined the challenges of eliminating the rarest amino acid, cysteine, from a minimal bacterial genome devised on the chassis of existing small genome of a mollicute, *M. genitalium*. Analysis of the requirements for amino acid residues in proteins allowed us to formulate a number of testable mechanistic hypotheses concerning known and novel functions of these residues. We feel that either success or failure of computational, and eventually laboratory, approaches to eliminate rare residues, such as cysteine (or the second-rarest residue, tryptophan—see Supplementary Table S3 for the preliminary data on Trp distribution in our lists), perhaps will be better tractable and more informative than replacing a common amino acid by its similar, such as, for example, leucine–isoleucine swap.

Practical implementation of any kind of genome minimization faces a major challenge in the form of negative epistasis/synthetic sickness, when milder effects of single genetic changes are aggravated when several of these changes are combined in one genome. Far from being specific to designing a cysteine-free proteome, these problems are inherent in any experimental scheme of genome reduction. In the case of whole-gene knockouts, computational methods of data analysis are emerging to predict the unobserved phenotypes of double mutants on the basis of the known effects of single mutations (38,39). The mutational fitness landscape of genomes at the amino acid resolution, in contrast, is just beginning to be studied, and small genomes with constrained amino acid repertoires may be particularly suitable models for this kind of analysis. Of course, it is also possible that some double knockouts may have higher fitness than one or both single-knockout parents, as in the cases of toxin–antitoxin or restriction–modification systems in bacteria (40) or, perhaps more generally, when two gene products have an opposite effect on the accumulation of a toxic intermediate.

The minimal genome derived from *M. genitalium* will have the imprimatur of a mollicute at all levels of organization, from nucleotide frequencies and codon usage to protein sequences and metabolic pathways to cell biology. The work toward redesigning and minimizing other bacterial genomes is also under way (40), technology suitable for minimization of yeast genome has been developed (41), and the first attempts on minimization of fission yeast genome have been described recently (42). Should the restriction of amino acid usage become desirable in these cases, the general strategy outlined here will be applicable there as well, though of course specific solutions will differ depending on the proteins encoded by these other genetic backgrounds.

The streamlined genomes discussed in this work are the objects of synthetic biology, designed by re-engineering of the existing bacteria and not expected to represent any past stage of biological evolution. Our derivation of a proteome with restricted use of cysteine, nevertheless, has evolutionary implications. Indeed, the rarity of amino acids in minimal proteome correlates, albeit imperfectly, with the order in which the amino acids are thought to have become genetically encoded: Cys, His, Met and Trp are all thought to be the most recent additions to the canonical set, based on many lines of evidence (42). Thus, minimal organisms with proteins depleted of Cys or other amino acids may serve as models of the functional limitations experienced by the cells at the recent stages of evolution of the genetic code. In particular, the implausibility of obtaining deoxyribonucleosides without Cys residues within ribonucleoside reductases may provide the relative timing for a major evolutionary transition, i.e. emergence of DNA in the world dominated by RNA genomes. This leap appears to have been unlikely before the incorporation of genetically encoded cysteines into proteins. Interestingly, ribonucleoside reductase is consistently placed in the last universal common ancestor of living organisms by most efforts of reconstructing ancestral gene content [reviewed in (43)].

Modification of the minimal proteome that restricts or completely abolishes the usage of one or more amino acids is akin to a lipogram, i.e. a text in a human language that avoids using one of the letters in the alphabet (44–46). A connection between protein lipograms and enzyme evolution has been also explored in a different context (47).

We conclude this article with lipogrammatically rewritten abstract of itself.

Minimal gene set for the lifestyle of a prokaryote is a list of genes needed for survival of an engineered *M. genitalium* that has all nutrients for survival provided in the medium. The estimates suggest that this set may have 300–350 protein-enabling genes. May these minimal proteins be made of a limited monomer repertoire? The genome eyeballing method is used here to extrapolate the minimal gene set of the pliable-skinned Gram-positive-like prokaryotes and to engineer out the rarest residue, i.e. the One that Has a Thiol Group. Nearly full, but not final, elimination of That Residue seems possible. Notably, the main deal-breaker is the essentiality of the thiol groups in four proteins that form deoxyribose from ribose and modify tRNA.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors are grateful to Dmitry Fomenko and Valerie de Crécy-Lagard for critical reading of the manuscript.

FUNDING

Stowers Institute for Medical Research [Intramural Research Award to A.M.]; École Polytechnique Foundation [scholarship to A.K.]. Funding for open access charge: Stowers Institute for Medical Research.

Conflict of interest statement. None declared.

REFERENCES

1. Glass,J.I., Assad-Garcia,N., Alperovich,N., Yooseph,S., Lewis,M.R., Maruf,M., Hutchison,C.A. 3rd, Smith,H.O. and Venter,J.C. (2006) Essential genes of a minimal bacterium. *Proc. Natl Acad. Sci. USA*, **103**, 425–430.
2. Fraser,C.M., Gocayne,J.D., White,O., Adams,M.D., Clayton,R.A., Fleischmann,R.D., Bult,C.J., Kerlavage,A.R., Sutton,G., Kelley,J.M. *et al.* (1995) The minimal gene complement of *Mycoplasma genitalium*. *Science*, **270**, 397–403.
3. Mushegian,A.R. and Koonin,E.V. (1996) A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc. Natl Acad. Sci. USA*, **93**, 10268–10273.
4. Christen,B., Abeliuk,E., Collier,J.M., Kalogeraki,V.S., Passarelli,B., Coller,J.A., Fero,M.J., McAdams,H.H. and Shapiro,L. (2011) The essential genome of a bacterium. *Mol. Syst. Biol.*, **7**, 528.
5. Mushegian,A. (1999) The minimal genome concept. *Curr. Opin. Genet. Dev.*, **9**, 709–714.
6. Charlebois,R.L. and Doolittle,W.F. (2004) Computing prokaryotic gene ubiquity: rescuing the core from extinction. *Genome Res.*, **14**, 2469–2477.
7. Juhas,M., Eberl,L. and Glass,J.I. (2011) Essence of life: essential genes of minimal genomes. *Trends Cell Biol.*, **21**, 562–568.
8. Koonin,E.V. (2003) Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat. Rev. Microbiol.*, **1**, 127–136.
9. Kristensen,D.M., Kannan,L., Coleman,M.K., Wolf,Y.I., Sorokin,A., Koonin,E.V. and Mushegian,A.R. (2010) A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches. *Bioinformatics*, **26**, 1481–1487.
10. Edgar,R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.
11. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
12. Söding,J., Biegert,A. and Lupas,A.N. (2005) The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.*, **33**, W244–W248.
13. Omelchenko,M.V., Galperin,M.Y., Wolf,Y.I. and Koonin,E.V. (2010) Non-homologous isofunctional enzymes: a systematic analysis of alternative solutions in enzyme evolution. *Biol. Direct.*, **5**, 31.
14. Tatusov,R.L., Natale,D.A., Garkavtsev,I.V., Tatusova,T.A., Shankavaram,U.T., Rao,B.S., Kiryutin,B., Galperin,M.Y., Fedorova,N.D. and Koonin,E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, **29**, 22–28.
15. Powell,S., Szklarczyk,D., Trachana,K., Roth,A., Kuhn,M., Muller,J., Arnold,R., Rattei,T., Letunic,I., Doerks,T. *et al.* (2012) eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res.*, **40**, D284–D289.

16. Hartman,J.L. 4th, Garvik,B. and Hartwell,L. (2001) Principles for the buffering of genetic variation. *Science*, **291**, 1001–1004.
17. Nichols,R.J., Sen,S., Choo,Y.J., Beltrao,P., Zietek,M., Chaba,R., Lee,S., Kazmierczak,K.M., Lee,K.J., Wong,A. *et al.* (2011) Phenotypic landscape of a bacterial cell. *Cell*, **144**, 143–156.
18. Makarova,K.S., Ponomarev,V.A. and Koonin,E.V. (2001) 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 Biol.*, **2**, RESEARCH0033.
19. Sriskanda,V., Moyer,R.W. and Shuman,S. (2001) NAD+-dependent DNA ligase encoded by a eukaryotic virus. *J. Biol. Chem.*, **276**, 36100–36109.
20. Samanta,M., Banerjee,M., Murthy,M.R.N., Balaram,H. and Balaram,P. (2011) Probing the role of the fully conserved Cys126 in triosephosphate isomerase by site-specific mutagenesis—distal effects on dimer stability. *FEBS J.*, **278**, 1932–1943.
21. González-Mondragón,E., Zubillaga,R.A., Saavedra,E., Chánez-Cárdenas,M.E., Pérez-Montfort,R. and Hernández-Arana,A. (2004) Conserved cysteine 126 in triosephosphate isomerase is required not for enzymatic activity but for proper folding and stability. *Biochemistry*, **43**, 3255–3263.
22. Boschi-Muller,S. and Branlant,G. (1999) The active site of phosphorylating glyceraldehyde-3-phosphate dehydrogenase is not designed to increase the nucleophilicity of a serine residue. *Arch. Biochem. Biophys.*, **363**, 259–266.
23. Raberg,M., Bechmann,J., Brandt,U., Schlüter,J., Uischner,B., Voigt,B., Hecker,M. and Steinbüchel,A. (2011) Versatile metabolic adaptations of *Ralstonia eutropha* H16 to a loss of PdhL, the E3 component of the pyruvate dehydrogenase complex. *Appl. Environ. Microbiol.*, **77**, 2254–2263.
24. Le,D.T., Lee,B.C., Marino,S.M., Zhang,Y., Fomenko,D.E., Kaya,A., Hacioglu,E., Kwak,G.H., Koc,A., Kim,H.Y. *et al.* (2009) Functional analysis of free methionine-R-sulfoxide reductase from *Saccharomyces cerevisiae*. *J. Biol. Chem.*, **284**, 4354–4364.
25. Baranowski,E., Guiral,S., Sagné,E., Skapski,A. and Citti,C. (2010) Critical role of dispensable genes in *Mycoplasma agalactiae* interaction with mammalian cells. *Infect. Immun.*, **78**, 1542–1551.
26. Johansson,E., Neuhard,J., Willemoës,M. and Larsen,S. (2004) Structural, kinetic, and mutational studies of the zinc ion environment in tetrameric cytidine deaminase. *Biochemistry*, **43**, 6020–6029.
27. Hsin,K., Sheng,Y., Harding,M.M., Taylor,P. and Walkinshaw,M.D. (2008) MESPEUS: a database of the geometry of metal sites in proteins. *J. Appl. Crystallogr.*, **41**, 963–968.
28. Holmgren,A., Ohlsson,I. and Grankvist,M.L. Thioredoxin from *Escherichia coli*. Radioimmunological and enzymatic determinations in wild type cells and mutants defective in phage T7 DNA replication. *J. Biol. Chem.*, **253**, 430–436.
29. Russel,M. and Holmgren,A. (1988) Construction and characterization of glutaredoxin-negative mutants of *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **85**, 990–994.
30. Holmgren,A. and Sengupta,R. (2010) The use of thiols by ribonucleotide reductase. *Free Radic. Biol. Med.*, **49**, 1617–1628.
31. de Crécy-Lagard,V., Marck,C. and Grosjean,H. (2012) Decoding in *Candidatus Riesia pediculicola*, close to a minimal tRNA modification set? *Trends Cell Molec. Biol.*, **7**, 12–39.
32. de Crécy-Lagard,V., Marck,C., Brochier-Armanet,C. and Grosjean,H. (2007) Comparative RNomics and modomics in *Mollicutes*: prediction of gene function and evolutionary implications. *IUBMB Life*, **59**, 634–658.
33. Nilsson,K., Lundgren,H.K., Hagervall,T.G. and Björk,G.R. (2002) The cysteine desulfurase IscS is required for synthesis of all five thiolated nucleosides present in tRNA from *Salmonella enterica* serovar *typhimurium*. *J. Bacteriol.*, **184**, 6830–6835.
34. Yim,L., Martínez-Vicente,M., Villarroya,M., Aguado,C., Knecht,E. and Armengod,M.E. (2003) The GTPase activity and C-terminal cysteine of the *Escherichia coli* MnmE protein are essential for its tRNA modifying function. *J. Biol. Chem.*, **278**, 28378–28387.
35. Osawa,T., Ito,K., Inanaga,H., Nureki,O., Tomita,K. and Numata,T. (2009) Conserved cysteine residues of GidA are essential for biogenesis of 5-carboxymethylaminomethyluridine at tRNA anticodon. *Structure*, **17**, 713–724.
36. Tan,J.T. and Bardwell,J.C.A. (2004) Key players involved in bacterial disulfide-bond formation. *Chembiochem*, **5**, 1479–1487.
37. Kale,S., Ulas,G., Song,J., Brudvig,G.W., Furey,W. and Jordan,F. (2008) Efficient coupling of catalysis and dynamics in the E1 component of *Escherichia coli* pyruvate dehydrogenase multienzyme complex. *Proc. Natl Acad. Sci. USA*, **105**, 1158–1163.
38. Babu,M., Gagarinova,A. and Emili,A. (2011) Array-based synthetic genetic screens to map bacterial pathways and functional networks in *Escherichia coli*. *Methods Mol. Biol.*, **781**, 99–126.
39. Baryshnikova,A., Costanzo,M., Kim,Y., Ding,H., Koh,J., Toufighi,K., Youn,J.Y., Ou,J., San Luis,B.J., Bandyopadhyay,S. *et al.* (2010) Quantitative analysis of fitness and genetic interactions in yeast on a genome scale. *Nat. Methods.*, **7**, 1017–1024.
40. Acevedo-Rocha,C.G., Fang,G., Schmidt,M., Ussery,D.W. and Danchin,A. (2013) From essential to persistent genes: a functional approach to constructing synthetic life. *Trends Genet.*, **29**, 273–279.
41. Dymond,J.S., Richardson,S.M., Coombes,C.E., Babatz,T., Muller,H., Annaluru,N., Blake,W.J., Schwerzmann,J.W., Dai,J., Lindstrom,D.L. *et al.* (2011) Synthetic chromosome arms function in yeast and generate phenotypic diversity by design. *Nature*, **477**, 471–476.
42. Sasaki,M., Kumagai,H., Takegawa,K. and Tohda,H. (2013) Characterization of genome-reduced fission yeast strains. *Nucleic Acids Res.*, **41**, 5382–5399.
43. Trifonov,E.N. (2000) Consensus temporal order of amino acids and evolution of the triplet code. *Gene*, **261**, 139–151.
44. Mushegian,A. (2008) Gene content of LUCA, the last universal common ancestor. *Front Biosci.*, **13**, 4657–4666.
45. Perec,G. (1969) *La disparition*. Gallimard, Paris.
46. Perec,G. and Adair,G. (2005) *A Void*, Vol. 9. Verba Mundi, Jaffrey, NH.
47. Dunn,M. (2002) *Ella Minnow Pea: A Novel in Letters*. Anchor, New York.
48. White,H.B. 3rd and Dhurjati,P. (2006) Evolution of protein lipograms: a bioinformatics problem. *Biochem. Mol. Biol. Educ.*, **34**, 262–266.