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,*,†]

8444–8451 *Nucleic Acids Research, 2013, Vol. 41, No. 18*
doi:10.1093/nar/gkt610

Introduction
Materials and Methods
Results and Discussion
Summary

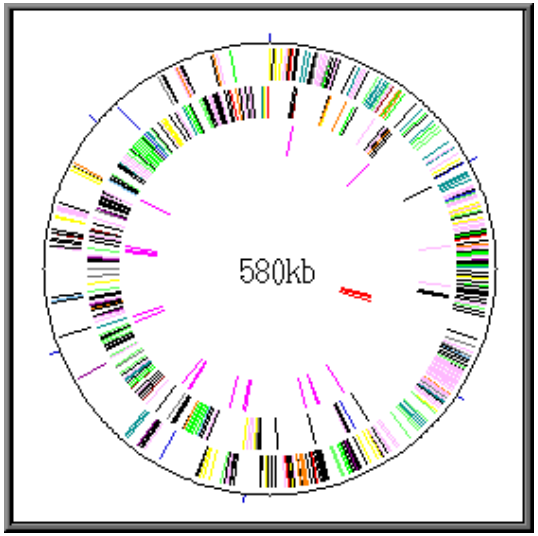
Introduction

Minimal bacterial gene set comprises the genetic elements needed for survival of engineered bacterium on a rich medium

Mycoplasma genitalium, *487 protein-coding and 43 RNA-coding genes*

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* .

layout on the chromosome
the definition of regulatory intergenic regions
and the account of other loci important



Introduction

Can there be a minimal gene set that does not use a particular amino acid?

In M. genitalium

101 of 487 protein-coding genes, individually disrupted without the loss of viability

386 cannot

lower bound of the number of essential protein-coding genes：

300–350,

if a rich medium is provided

Introduction

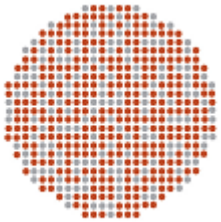
Can there be a minimal gene set that does not use a particular amino acid?

Two considerations

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

Second, an amino acid conserved all its orthologous proteins is most likely to be functionally indispensable

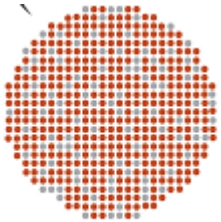
Materials and Methods



computational minimal genome 328

List of orthologous genes shared by the majority of Mollicutes , 25 genomes,

EdgeSearch algorithm.



experimental minimal genome 386

(2006) Essential genes of a minimal bacterium.

the derived minimal genome 439

union of the List 1 and List 2

Results and Discussion

Minimal genome with orthologs

345 COGs, 44 789 proteins found on average in 86 species
347 MOGs, 6903 proteins found in on average in 18 species

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

Proteins devoid of any amino acid are not common: most amino acids are present at least once in>95% of proteins in all lists

Rarest amino acids： **C,W**,H and internal M

Results and Discussion

Rarest amino acids: C

380 no C or can be substituted

76 proteins lack C in *M.genitalium*

248 *M. genitalium* proteins have C residues that are substituted in at least one mollicute, and 56 proteins, the cysteine-free orthologs are detected in more distantly related bacteria or archaea

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.

Results and Discussion

Rarest amino acids: C 59 no C-free orthologs

Have C-free paralogs

MG034: thymidine kinase, 1 conserved Cys ,paralogous eukaryotic thymidine kinases have none conserved Cys

MG174: ribosomal protein L36, has 3 Cys residues conserved in almost all organisms, The Cys-form of L36 in *Mesorhizobium loti* strain MAFF303099 has no Cys at all. (paralogs)

MG254: NAD-dependent DNA ligase, two fully conserved and two partially conserved cysteine residues, homologous entomopoxvirus lacks the Zn-finger domain.

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.

Replaced cys-free isofunctional analog

MG227: ThyA, 1 conserved Cys

The related Gram-positive bacteria, such as *Clostridiales,* however, encode non-homologous flavine-dependent thymidylate synthase ThyX (COG1351) that has only non-conserved cysteines

ThyA in minimal genome may be replaced by its cysteine-free isofunctional analog ThyX

Results and Discussion

Rarest amino acids: C 59 no C-free orthologs

Conserved cysteines have been experimentally mutated without loss of cell viability

MG431: triose phosphate isomerase
MG301: glyceraldehyde-3-phosphate dehydrogenase
MG271: subunit of dihydrolipoamide dehydrogenase

Dispensable in several *Mycoplasma* species

MG408: peptide methionine sulfoxide reduc-tases A
MG448: peptide methionine sulfoxide reduc-tases B
MG336: NifS
MG337: NifU

Results and Discussion

Rarest amino acids: C 59 no C-free orthologs

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 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 Cys in these protein families await further investigation.

Coordinate divalent metal cations?

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).

The metal chelation is performed not by cysteines, but by a combination of aspartic acid, histidine and/or glutamine residues giving rise to a testable hypothesis that many of the Zn-binding sites in proteins might be rebuilt to coordinate the catalytic metal ions

The distribution of Zn binding site types in PDB

Type	Number	Proportion (%)
CXXX	141	4
CCXX	198	6
CCCX	203	6
CCCC	370	12
other	253	8
HHHX	571	18
YYY	1493	46

Abbreviation: X – any amino acid residue; Y – D, H or E

Results and Discussion

Rarest amino acids: C

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

The remaining eight proteins with absolutely conserved cysteines **focus one molecular function, namely, the redox potential of the thiol group.**

MG124:(thioredoxin) and MG102(thioredoxin reductase) are dispensable in *E.coli* when the medium is supplemented with glutathione, cysteine and methionine, probably because glutaredoxin takes over as the source of redox equivalents.

MG127:Glutaredoxin-like protein, may play this role in the minimal genome (indispensable)

MG231: The large subunit of class I ribonucleoside-diphosphate reductase.

Five conserved Cys, all of which are essential for activity, and C386 is essential and conserved among all three classes of ribonucleotide reductases. **This residue is responsible for the thyil radical formation. (indispensable)**

Results and Discussion

Rarest amino acids: **C**

The remaining eight proteins with absolutely conserved cysteines

focus one molecular function, namely, the redox potential of the thiol group.

Four enzymes are involved in modification of bases within transfer RNA (tRNA), also requiring the thiol groups of conserved Cys.

MG372:ThiI, ThiI ortholog is missing from three mycoplasma genomes and from recently characterized *Candidatus Riesia pediculicola* , which has a drastically reduced repertoire of tRNA modifications and not essential in *E. coli* .

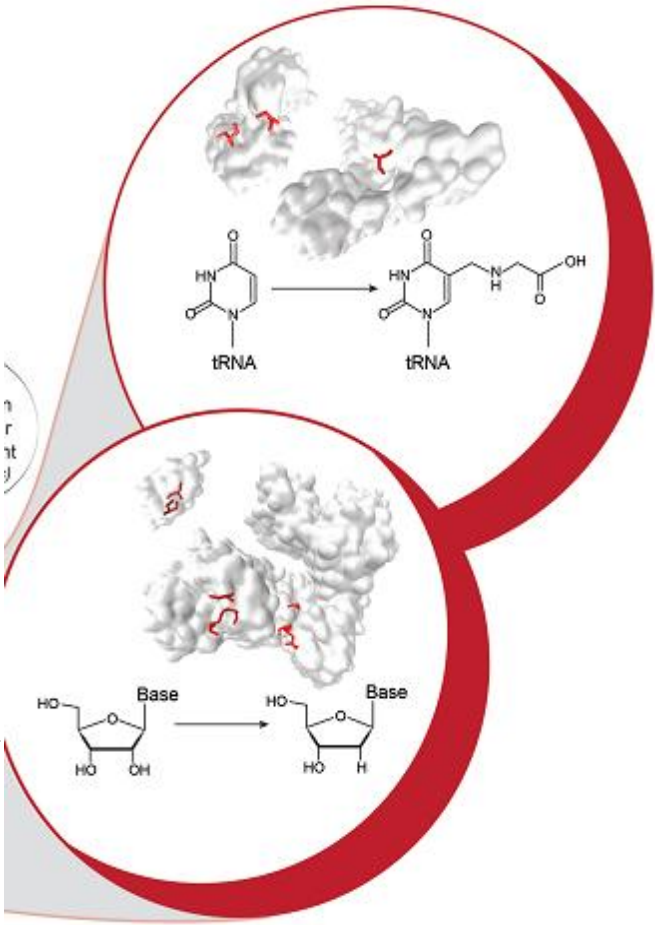
MG295:MnmA, is dispensable in *M.genitalium*, in agreement with the data that mutant *Salmonella enterica* lacking thiolated U34 is viable.

MG008,MG379(MnmE,GidA): a modification that appears to be essential for decoding codons with the wobble base.

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 . (indispensable)

Summary

- **4 indispensable proteins with Cys residues** MG008, MG127, MG231 and MG379
  all of which are involved in ribonucleoside modification
- A drastic reduction of cysteine content in *M. genitaliumis*, within reach of synthetic biology. Cys 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.
- Elimination of cysteine residues makes obsolete the genes whose products are involved in cysteine metabolism.
- The work toward redesigning and minimizing other bacterial genomes. Should the restriction of amino acid usage become desirable in these cases, the general strategy outlined here will be applicable



Summary

- Our derivation of a proteome with restricted use of cysteine, nevertheless, has evolutionary implications.

- 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.

Thank you