Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access
Author manuscript
Angew Chem Int Ed Engl. Author manuscript; available in PMC 2017 April 18.

Published in final edited form as:
Angew Chem Int Ed Engl. 2016 April 18; 55(17): 5337–5341. doi:10.1002/anie.201511657.

Facile recoding of selenocysteine in nature

Dr. Takahito Mukai,
Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520 (USA)

Dr. Markus Englert,
Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520 (USA)

Dr. H. James Tripp,
Department of Energy Joint Genome Institute (DOE JGI), Walnut Creek, CA 94598 (USA)

Dr. Corwin Miller,
Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520 (USA)

Dr. Natalia N. Ivanova,
Department of Energy Joint Genome Institute (DOE JGI), Walnut Creek, CA 94598 (USA)

Dr. Edward M. Rubin,
Department of Energy Joint Genome Institute (DOE JGI), Walnut Creek, CA 94598 (USA)

Dr. Nikos C. Kyrpides, and
Department of Energy Joint Genome Institute (DOE JGI), Walnut Creek, CA 94598 (USA)

Prof. Dieter Söll
Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520 (USA). Department of Chemistry, Yale University, New Haven, CT 06520 (USA)

Dieter Söll: dieter.soll@yale.edu

Abstract

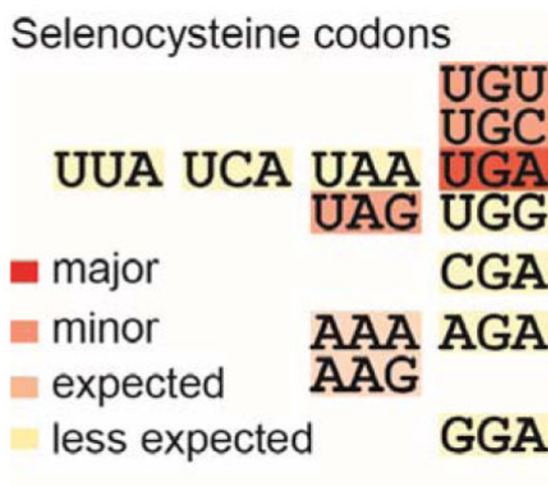
Selenocysteine (Sec or U) is encoded by UGA, a stop codon reassigned by a Sec-specific elongation factor and a distinctive RNA structure. To discover possible code variations in extant organisms we analyzed 6.4 trillion base pairs of metagenomic sequences and 24,903 microbial genomes for $tRNA^{Sec}$ species. As expected, UGA is the predominant Sec codon in use. We also found $tRNA^{Sec}$ species that recognize the stop codons UAG and UAA, and ten sense codons. Selenoprotein synthesis programmed by UAG in Geodermatophilus and Blastococcus, and by the Cys codon UGU in Aeromonas salmonicida was confirmed by metabolic labeling with ^{75}Se or mass spectrometry. Other $tRNA^{Sec}$ species with different anticodons enabled Escherichia coli to synthesize active formate dehydrogenase H, a selenoenzyme. This illustrates the ease by which the genetic code may evolve new coding schemes, possibly aiding organisms to adapt to changing

Correspondence to: Dieter Söll, `dieter.soll@yale.edu`.
Supporting information for this article is given via a link at the end of the document.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

environments. Our results reveal that the genetic code is much more flexible than previously thought.

Graphical Abstract



Keywords

genetic code; selenocysteine; sense codon recoding; metagenome; synthetic biology

The micronutrient selenium is present in proteins in the form of the versatile 21st amino acid, selenocysteine, in which the thiol moiety of cysteine (Cys) is replaced by a selenol group [1]. Selenoproteins are present in organisms from all domains of life [2]; such proteins are essential in mammalian cells [3], yet plants and fungi lack this amino acid. Sec is present in the active site of many redox enzymes [4]. The codon for Sec is UGA which is normally a translational stop signal [5]. During translation of selenoprotein mRNAs, UGA is recoded by the interaction of a specialized elongation factor SelB (in bacteria) with a downstream Sec insertion sequence [5–6]. Recently, a synthetic biology study succeeded in reassigning Sec to a large number of sense and stop codons in *Escherichia coli* [7] demonstrating that alterations to the genetic code can be tolerated. This prompted the question, whether deviations of the standard UGA Sec assignment may naturally occur.

A computational study scanning several trillion base pairs of metagenomic data revealed a large number of stop codon reassignments in bacteria and bacteriophages [8]. This inspired us to perform a comprehensive search of the available metagenomic and microbial genomic sequence data for anticodon variants of the typical $tRNA^{Sec}_{UCA}$, the longest tRNA [9] with a tertiary structure quite different from that of canonical tRNAs [10]. A BLAST search of the tRNADB-CE database [11] revealed four $tRNA^{Sec}_{CUA}$ sequences, suggesting UAG codon recognition. Searching all public microbial genomes in the National Center for Biotechnology Information (NCBI) and all assembled metagenome data in the Integrated Microbial Genomes (IMG) system [12] yielded a $tRNA^{Sec}_{GCA}$ group, indicating UGC codon recognition. We then developed a general computational pipeline that scanned ~6.4 Tb of

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

unassembled short reads, ~180 Gb of assembled contigs (> 2 kb), and 24,903 microbial genomes in IMG (Figure 1A). The results affirmed UGA as the predominant Sec codon. In addition, 12 different $tRNA^{Sec}$ anticodon variants capable of recognizing the stop codons UAG and UAA, and 10 sense codons were discovered (Figure 1A). Further sequence validations (see Supplementary Information) ascertained these $tRNA^{Sec}$ variants not to be sequencing artifacts.

We grouped these non-canonical $tRNA^{Sec}$ species by anticodon type, sequence, and structural similarity (Figure 1A, and Supplementary Information). The largest group (anticodon CUA) contains 366 nearly identical $tRNA^{Sec}$ sequences from the actinobacterial *Geodermatophilaceae* family [13] (Figure 1A, 1B left panel), while the other 3 $tRNA^{Sec}_{CUA}$ species may be of rhizosphere bacterial origin (Figures S1, S2). The $tRNA^{Sec}_{CUA}$ species amounted to 3% of the total $tRNA^{Sec}$ species found in a soil metagenome (3300001205).

The next group contains $tRNA^{Sec}$ species (with anticodon GCA able to decode Cys) from *Betaproteobacteria* and termite gut symbionts (Figure 1A, Figure S3). Two *Aeromonas salmonicida* genomes contain a $tRNA^{Sec}_{ACA}$, able to decode the UGU Cys codon (Figure 1B, middle panel). The other 16 members of this group may originate from *Chloroflexi* (Figure S4) and ocean bacteria (Figure S5). The 2 $tRNA^{Sec}_{UUU/CUU}$ species (recognizing AAA/AAG Lys codons) may derive from the *Solirubrobacterales* (Figure S6). Additional $tRNA^{Sec}$ variants able to recognize the stop codon UAA and 6 sense codons (CGA, AGA, GGA, UUA, UCA, UGG) were also found (Figure S7).

We then wanted to confirm the coding properties of these non-canonical $tRNA^{Sec}$ species. For proof of selenoprotein synthesis three strategies were possible: (i) metabolic labeling of the organisms with ^{75}Se, (ii) replacing parts of the *E. coli* selenoprotein synthesis machinery with genes and $tRNA^{Sec}$ from our genomic or metagenomic findings, and (iii) replacing *E. coli* $tRNA^{Sec}$ with the newly discovered $tRNA^{Sec}$ species. In the last two strategies *E. coli* formate dehydrogenase H ($FDH_{H,}$ encoded by the *fdhF* gene) would serve as reporter [7].

To confirm UAG-directed Sec incorporation we grew *Geodermatophilus obscurus* G-20 and *Blastococcus saxobsidens* cells in the presence of [^{75}Se]selenite and detected radiolabeled selenoproteins of 140 and 50 kDa size. The genome sequences predict formate dehydrogenases (FDHs) (Figure S8, Tables S1 & S2) and UGSC-motif proteins [14], (Figures S1, S8, Table S3). This was the first indication for UAG read-through by Sec to form the expected FDH (FdxG) and UGSC-motif protein products (Figure 2A). Crude cell extracts were then resolved by SDS-PAGE; the proteins in the gel slices corresponding to 140 and 50 kDa were trypsinized for subsequent liquid chromatography (LC) coupled with tandem mass spectrometry (LC-MS/MS) analysis. The 140 kDa gel slices from *G. obscurus* and *B. saxobsidens* harbored full-length FdxG, while the 50 kDa gel slice from *G. obscurus* contained a UGSC-motif protein (Figure 2A & Table S4).

Like *E. coli*, the *G. obscurus* genome [13b] encodes a Sec incorporation machinery consisting of the *selA*, *selB*, *selC* and *selD* genes. In bacteria, SelD produces the Se donor selenophosphate, SelA converts Ser-$tRNA^{Sec}$ (the tRNA is the product of the *selC* gene) to Sec-$tRNA^{Sec}$, and SelB carries Sec-$tRNA^{Sec}$ to the ribosome in a SECIS-dependent manner.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

To test their functionality, the *E. coli* s*elABC fdhF* strain ME6 was complemented with the *G. obscurus selABC* genes. The product of the *E. coli fdhF* gene is the selenoenzyme FDH_H (Table S1) whose activity requires Sec at position 140 [15], if replaced by Cys the activity drops 300-fold [16]. FDH_H is readily detected by the reduction of benzyl viologen resulting in a purple color [17]. To serve as a reporter, the plasmid-encoded *E. coli fdhF* gene transformed into strain ME6 was modified to have a TAG codon in position 140, followed by a *G. obscurus*-type SECIS element leading to an FDH_H variant with two amino acid changes (see Figure 2B). Expression of the *G. obscurus selABC* genes in this modified strain produced active FDH_H (Figure 2B). These data confirm that in this *E. coli* FDH_H variant the UAG_{140} codon is recoded to Sec.

UGU (Cys) recoding was confirmed in *Aeromonas salmonicida* subsp. *pectinolytica* 34mel, the type strain of the γ-proteobacterium *A. salmonicida* subspecies [18]. Unlike other *Aeromonas* species the *pectinolytica* subspecies and strain Y577 [19] pair $tRNA^{Sec}{}_{ACA}$ with a UGU Cys codon in *fdhF*. In addition to the *in vivo* FDH activity in *pectinolytica* cells (Figure 2C), their anaerobic metabolic labeling with ^{75}Se produced radioactive Sec-containing FDH_H (Figure 2D). We confirmed Sec incorporation encoded by UGU_{140} in FDH_H by overexpressing the protein from a plasmid (Figure 2E) and LC-MS/MS analysis (Table S4). The recoded selenopeptide (LC retention time 24.82) had the correct mass (Figure S9) and the appropriate secondary fragmentation pattern (Figure 2F). The co-eluting Cys- and Sec-peptides were detected through their different masses with ion intensities of ~100:1, respectively (Figure S9A). The mass peaks of the Sec-peptides were absent in protein samples obtained from *pectinolytica* cells expressing a *fdhF* variant lacking the SECIS element (Table S4). Thus, the Cys_{140} codon of the *fdhF* gene is translated as Sec in a SECIS-dependent manner in *A. pectinolytica* 34mel.

UGC Cys→Sec recoding by $tRNA^{Sec}{}_{GCA}$ may be a common trait in the *Burkholderiales.* In some long metagenomic *Burkholderiales* contig the *selC* gene is flanked by *selB* and *selA* genes, and the *selABC* operon is located next to a formate dehydrogenase (*fdoG*) gene whose active site UGC Cys codon is followed by a putative SECIS element (Figures 1B right panel & 2G). The *E. coli selABC* strain was complemented with the *Burkholderiales* contig *selAB* and *selC*-opal variant, and an *E. coli fdhF* variant harboring the *Burkholderiales* contig SECIS element. This strain produced active FDH_H (purple color) (Figure 2H, the 2nd row), but an inactive SECIS element (with a G25C mutation) did not form FDH_H and the cells were colorless (Figure 2H, the 4th row). In combination with the *Burkholderiales* contig *selC*, we changed UGA_{140} to UGC for the chimeric *fdhF* variants that carried functional or inactive SECIS elements. As the FDH_H Cys_{140} enzyme produced a purple color (Figure 2H, the 1st & 3rd rows), ^{75}Se-labeling was used to demonstrate that the functional SECIS element led to a clear signal (Figure 2H). Thus, the *Burkholderiales* contig *selA*, *selB*, $tRNA^{Sec}{}_{GCA}$, and SECIS enabled UGC-recoding in *E. coli.*

The metagenomic $tRNA^{Sec}$ variants that recognize other stop and sense codons were also tested for Sec reassignment. We selected one representative metagenomic $tRNA^{Sec}$ species for each anticodon type and expressed them in an *E. coli* s*elC fdhF* strain, together with the *E. coli fdhF* variants that carry the proper cognate codons at position 140 [7]. Surprisingly, all but $tRNA^{Sec}{}_{UCC}$ of the tested $tRNA^{Sec}$ species recoded the respective

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

codons for Sec, as they supported the expression of active FDH_H in their host *E. coli* cells (Figure S10). It should be mentioned that GGA was also poorly recoded in our earlier Sec recoding strategy [7]. The different recoding efficiencies may result from distortions of the ideal SECIS element structure by the nature of the upstream codon [20]. In light of these results we believe that these $tRNA^{Sec}$ species may be used for recoding sense codons in the organisms they originate from.

What about eukaryotic organisms? Although we found 9 $tRNA^{Sec}$ variants of algal origin (2 are shown in Figure S7), they need further validation, because they are almost identical to canonical $tRNA^{Sec}$ species. A similar search of 92 mammalian genomes (215 Gbp) and of the *Drosophila melanogaster* genome (139 Mbp) showed no exception to the use of UGA as the Sec codon. Whether this is related to the necessity of selenoproteins in high-level redox signaling pathways [21] or due to the sophisticated backup systems [22] remains to be investigated. However, in the lower eukaryote *Euplotes crassus* UGA serves both as a Cys and also as a SECIS-dependent Sec codon [23].

Natural reassignment of sense codons has not been seen in bacteria. But it is known in mitochondrial genomes, where a particular codon lost its original assignment and now leads to insertion of another amino acid [24]. Our case here is different; Sec insertion is mediated by a SECIS element and thus gives rise to dual use of the codon for another amino acid (through pairing with $tRNA^{Sec}$ variants carrying the proper anticodon).

What might account for this facile recoding to Sec? It is pertinent to note that Sec incorporation is different from that of all other amino acids; it is facilitated by its own 'orthogonal' system [5] consisting of a different elongation factor (SelB), a required SECIS RNA element, a structurally unusual tRNA ($tRNA^{Sec}$) [10], a dual meaning stop codon (UGA), and the use of release factor 2. Therefore, Sec recoding events may not have as general an effect on the protein translation machinery as one might expect from recoding canonical sense codons [25]. The ease of Cys to Sec recoding may be a consequence of the often desirable properties of selenoenzymes and selenoproteins with novel redox functions and increased enzyme activity [4, 26], while still allowing the expression of useful Cys-proteins and Cys-enzymes. Our finding of facile Sec recoding also opens our minds to the possible existence of other coding schemes. It also underscores the limitations of the current computational programs to predict selenoproteins from genome sequences, as these algorithms rest on UGA as the sole Sec codon.

Overall our approach provides new evidence of a limited but unequivocal plasticity of the genetic code whose secrets still lie hidden in the majority of unsequenced organisms.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Hans Aerni and Jesse Rinehart for advice on LC-MS/MS and Jean Kanyo (Yale University) for the dedicated efforts on the MS analyses. We also thank Andreas Brune, Filipa Gody-Vitorino, Hans-Peter Klenk, Ryan Lynch, Katherine McMahon, Daniel Marcus, William Mohn, Len Pennacchio, and Ameet Pinto for permission to

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

use unpublished sequence data produced through the DOE-JGI's community sequencing program. We are grateful to Patrick O'Donoghue, Oscar Vargas-Rodriguez, Jiqiang Ling for enlightened discussions and Daniel Drell and Robert Stack for encouragement. This work was supported by grants from the National Institute for General Medical Sciences (GM22854 to D.S.) and from the Division of Chemical Sciences, Geosciences, and Biosciences, Office of Basic Energy Sciences of the Department of Energy (DE-FG02-98ER20311 to D.S.; for funding the genetic experiments). The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, was supported under Contract No. DE-AC02-05CH11231.

References

1. Cone JE, Del Rio RM, Davis JN, Stadtman TC. Proc Natl Acad Sci USA. 1976; 73:2659–2663. [PubMed: 1066676]
2. Ambrogelly A, Palioura S, Söll D. Nat Chem Biol. 2007; 3:29–35. [PubMed: 17173027]
3. Kryukov GV, Castellano S, Novoselov SV, Lobanov AV, Zehtab O, Guigo R, Gladyshev VN. Science. 2003; 300:1439–1443. [PubMed: 12775843]
4. Metanis N, Hilvert D. Curr Opin Chem Biol. 2014; 22:27–34. [PubMed: 25261915]
5. Böck, A.; Thanbichler, M.; Rother, M.; Resch, A. Aminoacyl-tRNA Synthetases. Ibba, M.; Francklyn, CS.; Cusack, S., editors. Landes Bioscience; Georgetown, TX: 2005. p. 320-327.
6. Heider J, Baron C, Böck A. EMBO J. 1992; 11:3759–3766. [PubMed: 1396569]
7. Bröcker MJ, Ho JM, Church GM, Söll D, O'Donoghue P. Angew Chem Int Ed Engl. 2014; 53:319–323. [PubMed: 24511637]
8. Ivanova NN, Schwientek P, Tripp HJ, Rinke C, Pati A, Huntemann M, Visel A, Woyke T, Kyrpides NC, Rubin EM. Science. 2014; 344:909–913. [PubMed: 24855270]
9. Schön A, Böck A, Ott G, Sprinzl M, Söll D. Nucleic Acids Res. 1989; 17:7159–7165. [PubMed: 2529478]
10. a) Palioura S, Sherrer RL, Steitz TA, Söll D, Simonovic M. Science. 2009; 325:321–325. [PubMed: 19608919] b) Itoh Y, Chiba S, Sekine S, Yokoyama S. Nucleic Acids Res. 2009; 37:6259–6268. [PubMed: 19692584] c) Itoh Y, Sekine S, Suetsugu S, Yokoyama S. Nucleic Acids Res. 2013; 41:6729–6738. [PubMed: 23649835]
11. Abe T, Inokuchi H, Yamada Y, Muto A, Iwasaki Y, Ikemura T. Front Genet. 2014; 5:114. [PubMed: 24822057]
12. Markowitz VM, Chen IM, Chu K, Szeto E, Palaniappan K, Pillay M, Ratner A, Huang J, Pagani I, Tringe S, Huntemann M, Billis K, Varghese N, Tennessen K, Mavromatis K, Pati A, Ivanova NN, Kyrpides NC. Nucleic Acids Res. 2014; 42:D568–573. [PubMed: 24136997]
13. a) Luedemann GM. J Bacteriol. 1968; 96:1848–1858. [PubMed: 5726312] b) Ivanova N, et al. Stand Genomic Sci. 2010; 2:158–167. [PubMed: 21304698]
14. Zhang Y, Gladyshev VN. PLoS Genet. 2008; 4:e1000095. [PubMed: 18551170]
15. Zinoni F, Birkmann A, Leinfelder W, Böck A. Proc Natl Acad Sci USA. 1987; 84:3156–3160. [PubMed: 3033637]
16. Axley MJ, Böck A, Stadtman TC. Proc Natl Acad Sci USA. 1991; 88:8450–8454. [PubMed: 1924303]
17. Mandrand-Berthelot MA, Wee MYK, Haddock BA. FEMS Microbiology Lett. 1978; 4:37–40.
18. Pavan ME, Pavan EE, Lopez NI, Levin L, Pettinari MJ. Genome Announc. 2013; 1 pii: e00675–00613.
19. Vincent AT, Trudel MV, Freschi L, Nagar V, Gagne-Thivierge C, Levesque RC, Charette SJ. BMC Genomics. 2016; 17:44. [PubMed: 26753691]
20. Yoshizawa S, Rasubala L, Ose T, Kohda D, Fourmy D, Maenaka K. Nat Struct Mol Biol. 2005; 12:198–203. [PubMed: 15665870]
21. Lu J, Holmgren A. Free Radic Biol Med. 2014; 66:75–87. [PubMed: 23899494]
22. a) Lin HC, Ho SC, Chen YY, Khoo KH, Hsu PH, Yen HC. Science. 2015; 349:91–95. [PubMed: 26138980] b) Turanov AA, Everley RA, Hybsier S, Renko K, Schomburg L, Gygi SP, Hatfield DL, Gladyshev VN. PLoS One. 2015; 10:e0140353. [PubMed: 26452064]
23. Turanov AA, Lobanov AV, Fomenko DE, Morrison HG, Sogin ML, Klobutcher LA, Hatfield DL, Gladyshev VN. Science. 2009; 323:259–261. [PubMed: 19131629]

24. a) Yokobori S. Viva Origino. 2014; 42:47–49.b) Ling J, O'Donoghue P, Söll D. Nat Rev Microbiol. 2015; 13:707–721. [PubMed: 26411296]
25. a) Ho JM, Reynolds NM, Rivera K, Connolly M, Guo LT, Ling J, Pappin DJ, Church GM, Söll D. ACS Synth Biol. 201510.1021/acssynbio.1025b00197b) Mukai T, Yamaguchi A, Ohtake K, Takahashi M, Hayashi A, Iraha F, Kira S, Yanagisawa T, Yokoyama S, Hoshi H, Kobayashi T, Sakamoto K. Nucleic Acids Res. 2015; 43:8111–8122. [PubMed: 26240376]
26. a) Beld J, Woycechowsky KJ, Hilvert D. Oxidative Folding of Peptides and Proteins, The Royal Society of Chemistry. 2009:253–273.b) Snider GW, Ruggles E, Khan N, Hondal RJ. Biochemistry. 2013; 52:5472–5481. [PubMed: 23865454]

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

Author Manuscript

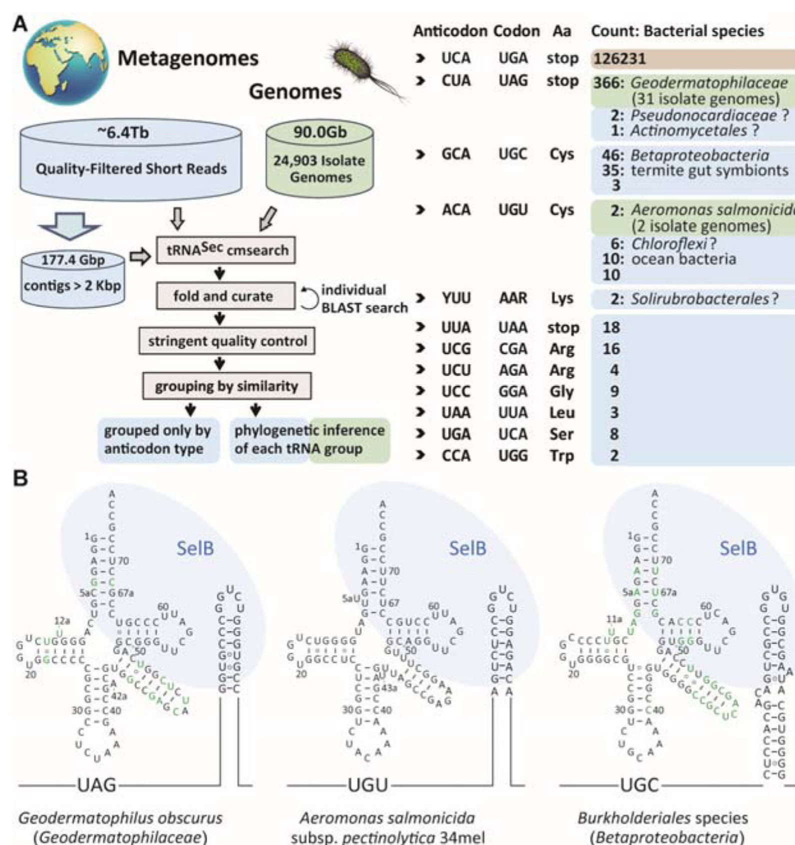


Figure 1. Non-canonical selenocysteine assignments in nature

(**A**) The $tRNA^{Sec}$ search pipeline and the manually curated output. The non-canonical $tRNA^{Sec}$ sequences are grouped by codon recognition; their numbers are given with their (putative) bacterial origins from phylogenetic inference. Green color indicates results from whole genomes, while blue represents results from only metagenomic data. For comparison the number of canonical $tRNA^{Sec}_{UCA}$ sequences is shown (most of them were not curated). “Y” and “R” denote “C or U” and “A or G”, respectively. **(B)** Inferred cloverleaf structures of non-canonical $tRNA^{Sec}$ species. The nucleotide polymorphism among the same $tRNA^{Sec}$ group is indicated with green letters. The Sec codons and the SECIS elements of formate dehydrogenase mRNAs are shown.

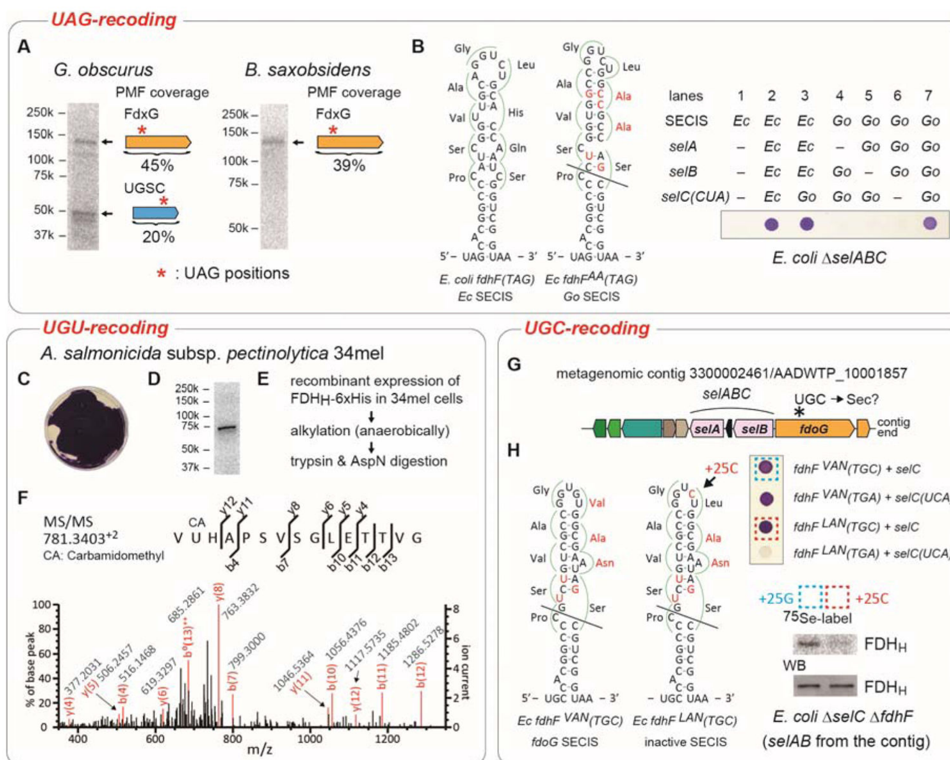


Figure 2. Recoding of UAG and cysteine codons to selenocysteine

(**A**) Metabolic ^{75}Se labeling of *G. obscurus* and *B. saxobsidens* cells. Crude extracts were resolved by SDS-PAGE, and their putative selenoproteins were visualized by PhosphorImager analysis. The results of peptide mass fingerprinting (PMF) analyses of the proteins in the excised gel bands are shown to the right of the bands. (**B**) FDH_H expression in *E. coli selABC fdhF* cells with the *G. obscurus selABC* genes and a chimeric *fdhF(140TAG)* gene variant having a *G. obscurus* SECIS element with a few nucleotide modifications shown in red. The *selC(CUA)* genes express $tRNA^{Sec}_{CUA}$. The expressed selenoprotein FDH_H reduced benzyl viologen, resulting in a purple color. (**C**) FDH_H activity of *A. salmonicida* subsp. *pectinolytica* 34mel cells. (**D**) Metabolic ^{75}Se labeling of the FDH_H of 34mel. (**E**) The procedure of sample preparation for the LC-MS/MS analysis of the FDH_H selenoprotein. (**F**) PMF confirms Sec incorporation at codon 140 in the recombinant FDH_H. **(G)** $tRNA^{Sec}_{GCA}$ gene (*selC*) locus in a metagenomic contig. (**H**) *In vivo* FDH_H assays in *E. coli selABC fdhF* cells with the *selABC* genes of the metagenomic contig and a chimeric *fdhF(140TGC)* gene carrying the contig's SECIS element with a few nucleotide modifications shown in red. The two transformed strains boxed were metabolically labeled with ^{75}Se, and the radioactive FDH_H proteins were analyzed by SDS-PAGE and autoradiography or western blotting (WB).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript