Genes Preferring Non-AUG Start Codons in Bacteria

Anne Gvozdjak\textsuperscript{1,2} and Manoj P. Samanta\textsuperscript{2,3,4}

1. Bellevue High School, 10416 SE Wolverine Way, Bellevue, WA 98004
2. Coding for Life Science, Redmond, WA 98053
3. Systemix Institute, Redmond, WA 98053
4. samanta@homolog.us

Abstract

Here we investigate translational regulation in bacteria by analyzing the distribution of start codons in fully assembled genomes. We report 36 genes (infC, rpoC, rnpA, etc.) showing a preference for non-AUG start codons in evolutionarily diverse phyla (“non-AUG genes”). Most of the non-AUG genes are functionally associated with translation, transcription or replication. In \textit{E. coli}, the percentage of essential genes among these 36 is significantly higher than among all genes. Furthermore, the functional distribution of these genes suggests that non-AUG start codons may be used to reduce gene expression during starvation conditions, possibly through translational autoregulation or IF3-mediated regulation.

Introduction

Understanding the regulation of gene expression at the transcriptional and translational levels is a key step in deciphering the information contained by genomes. Whereas inexpensive, high-throughput technologies (ChIP-seq, RNAseq) are helping in extensive experimental investigation of transcriptional regulation \cite{1,2}, the measurement of translational regulation using high-throughput mass spectroscopy \cite{3} is less widely adopted. Bioinformatics provides another avenue to leverage the rapidly falling cost of DNA sequencing to explore translational regulation. Prior to 2000, computational researchers identified a number of relevant patterns through the comparative analysis of gene sequences \cite{4,5}. With the availability of \textasciitilde 250,000 prokaryotic genomes, thousands of eukaryotic genomes, and additional metagenomic sequences, those patterns can now be studied at a significantly larger scale, and new patterns may also be identified.

Among the three phases of translation, namely initiation, elongation, and termination, initiation is the rate-limiting step \cite{6,7}. In prokaryotes, the Shine-Dalgarno sequence upstream of the start codon was considered to be the primary regulatory mechanism for initiation \cite{8,9}, but whole-genome sequencing changed this understanding. Leaderless genes, once considered rare \cite{9}, appear to be far more abundant \cite{10} and are argued to provide a more ancestral
initiation procedure [11]. Although the exact initiation mechanism in these genes is not understood, the nucleotides around the start codon likely play important roles.

The above realization brought renewed attention to start codons. Whole genome sequencing observed a widespread presence of GUG, UUG, and other non-AUG start codons in bacteria [12–15]. Irrespective of the start codon sequence, the same methionine-containing initiator tRNA always binds to it. Therefore, the differences in start codons may be used to regulate the rate of translation. Empirical studies have indeed shown that the rate of gene expression varies greatly depending on the start codon being used [16]. For instance, the well-studied gene infC uses an unconventional start codon in many evolutionarily diverse bacterial genomes to autoregulate its translation [17–20].

In this work, we looked for other genes showing a preference for non-AUG start codons by analyzing the distribution of start codons in fully assembled bacterial genomes from diverse phyla. Our analysis identified 36 genes (infC, rpoC, rmpA, etc.) with a preference for non-AUG start codons. Functionally, they are often associated with translation, transcription and replication. In E. coli, this group of genes carries a significantly larger proportion of essential genes than across all E. coli genes. Furthermore, based on data from existing literature, we argue that non-AUG start codons in these genes may help their regulation at the translational level, possibly as a response to nutrient stress.

**Results**

This analysis started with 210,130 publicly available prokaryotic genomes, from which all genomes with incomplete assemblies or annotations were excluded. From the remaining 11,904 genomes, the start codon nucleotide sequences of 3,818,377 gene occurrences were extracted. This raw data produced a preliminary distribution of start codon frequencies across bacterial phyla (Figure 1). Overall, about 85% of gene occurrences use AUG as their primary start codon, followed by GUG (10%) and UUG (< 5%). Actinobacteria has a substantially lower percentage of AUG start codons, in agreement with the prior studies [13,21].

To investigate possible evolutionary patterns among various genes’ start codon frequencies, this analysis was further restricted to the genomes from 30 genuses covering various large bacterial phyla (Table 1). Restricting the analysis to genuses with high total gene occurrences reduced potential noise from annotation errors [21]. The set of chosen genuses included 10 gammaproteobacteria and firmicutes each, and 10 from the other remaining phyla. For each gene within a genus, a normalized value (“non-AUG ratio”) was introduced, measuring the ratio of its non-AUG start codon occurrences to its total occurrences within the genus. Genes with a higher “non-AUG ratio” across numerous genuses were considered more desirable. Therefore, a minimum cutoff of a 0.5 “non-AUG ratio” was implemented here.
Four genes (infC, rpoC, rmpA, hisS) were above this cutoff in at least half of the 30 genuses. It is encouraging to find infC at the top of this list. Not only is this gene involved in translation initiation itself, but its unconventional start codon has been the subject of extensive research [17–20]. Very little systematic research regarding the start codons has been conducted on the other three genes.

A total of 36 genes were above the cutoff in at least a third of the genuses (Table 2). They are referred to as “non-AUG genes” in the following text. To compare the phylogenetic distribution of start codon preferences, the percentage presence of various start codons for these genes in different genuses is displayed in Figure 2.

A number of interesting patterns emerge from Table 2 and Figure 2. The translation initiation factor infC maintains a non-AUG start codon in most genuses, but it is also joined by another initiation factor (infB) in a subset of the genuses. A similar pattern is observed among RNA polymerase components rpoC and rpoB.

Additionally, a large fraction of the genes are involved in translation, but they appear to play roles in only a subset of pathways. The genes rpsT, rpsN, and rpsC are all components of the 30S ribosomal subunit, whereas rlmH, rsmH, and rsmG are all methyltransferases targeting 16S rRNA, a component of the same subunit. Moreover, rimP is involved in the maturation of the same subunit. In contrast, no component of the 50S ribosomal subunit showed up, apart from typA assisting in its assembly at the low temperatures. The other translational genes in Table 2 are translation elongation factor tuf and ribosomal silencing genes rsfS, hflX, and rsgA.

Several other genes are involved in various aspects of transcription, translation and DNA replication. This set includes rmpA cleaving precursor tRNAs as part of the RNase P ribonuclease complex; hisS and argS ligating tRNAs with amino acids; grpE regulating protein folding; secF and secD assisting in protein export; pnp helping in mRNA degradation; nusB involved in transcription antitermination; dnaA taking part in chromosomal replication; and radA, ubrC, and recJ repairing DNA replication errors.

The remaining eight of the 36 non-AUG genes are not related to cellular information storage and processing pathways (i.e. translation, transcription, replication). They include atpF as an ATP synthase component; plsX involved in lipid metabolism; murE taking part in cell wall biogenesis; ispG performing isoprenoid biosynthesis; crcB reducing fluoride toxicity; pxpB catalyzing the cleavage of 5-oxoprolinate to L-glutamate; carA hydrolyzing glutamine to glutamine; and fdhD acting as a sulphur carrier necessary for formate dehydrogenation. Overall very few genes from the entire metabolic pathway was present in this set of 36 genes.

Among the non-AUG genes reported here, 47% (17 out of 36) are essential in E. coli, whereas deletion of 9 others lead to slowed growth. In comparison, only 4% of all E. coli genes are essential.
The non-AUG genes identified here yielded multiple members previously researched for their translational autoregulation. Indeed, genes such as rpsT, coding for the 30S ribosomal protein S20, and infC, coding for the bacterial translation initial factor IF3, have both been specifically investigated for the correlation between their use of a non-standard start codon and their translational autoregulation using negative feedback loops [17,22,23]. In both cases, replacing the start codon with AUG resulted in derepression and an increase in gene expression. The other genes known for autoregulation or IF3-mediated suppression of expression are rpoB, rpoC, dnaA, uvC, pnp, and recJ [24–27]. In the case of mPA, experimental evidence suggested potential translational autoregulation, although this was not further investigated [28].

Discussions

Early gene sequencing in bacteria showed the presence of non-AUG start codons in about 8-9% of all genes. This estimate increased substantially to 20-40% after complete bacterial genomes became available [12,13,21]. Researchers in both pre- and post-genomic eras attempted to understand the regulatory roles of these alternate start codons [16–18,29]. Here we discuss our results in the context of those early and recent reports.

A simple model of the translational machinery would argue that the alternate start codons are there to reduce the expression levels of the corresponding genes. It is easy to see how the tRNA-codon binding would weaken for the non-canonical start codons, given that the same methionine-containing initiator tRNA binds to all start codons. Empirical studies confirm that the rate of gene expression reduces with the use of non-AUG start codons [16]. Based on this model, the genes identified in this work should have a general need for reduced expression in diverse phyla, thus explaining their preference for non-AUG start codons in general.

However, this explanation has two inconsistencies [30]. First, in isolated examples of ribosomal proteins (e.g. rpsM), changing the start codon from GUG to AUG reduced translation by seven-fold. Second, the genes involved in translation are among the most abundant in a cell. Therefore, it does not appear that non-AUG start codons are used to maintain a general reduction of their expression. Below we offer a refined model that takes this functional difference of the non-AUG genes into account.

Based on observations of the preservation of many of its common components in all domains of life, translation appears to have evolved earlier than the separation of the domains. Translational initiation, on the other hand, is not shared among the domains. Therefore, it is likely that genes involved in translation had another shared mechanism of initiation prior to the evolution of the AUG-based system, and that they may continue to maintain it by using alternate start codons.

Indeed, there may actually be a need to maintain this earlier system of gene regulation. By observing the tightly-regulated expression of all members of the ribosome, Nomura argued that
those genes were regulated through negative feedback at the translational level [31]. Especially under environmental starvation conditions, cells need to reduce translation to survive. It is possible that non-AUG start codons play a role to maintain this feedback control. Rather than opting for the less-restrained AUG-based system, these translationally-related genes may have preferred to maintain their ancestral system of initiation in order to maximize their ability to be regulated.

Many researchers are trying to reconstruct minimal genomes in bacteria in different ways [32–34], primarily to understand the evolution of the universal genetic code and the translational machinery. If the above model is valid, finding genes maintaining non-AUG start codons may offer another way to identify the ancestral form of the translational apparatus.

The reliability of the results presented here depends entirely on the accuracy of gene prediction in the public databases. Here we highlight various sources of errors and biases in our primary data and the measures taken to mitigate their impacts. First, the distribution of assembled bacterial genomes in the public database is skewed heavily toward medically relevant species (E. coli, Salmonella, C. difficile, etc.). At the phylum level, gammaproteobacteria dominate the list for the same reason. Second, the quality of annotations varies between genuses. Annotations in the well-funded and well-studied bacteria like E. coli and B. subtilis are backed by extensive empirical data, whereas many other rarely studied species rely entirely on computational gene predictions. The prediction of translational start-sites is notoriously error prone [21], compelling other researchers to perform start codon analysis on a much smaller set of curated genomes [29]. Finally, the inconsistent formatting of gene names in various gff files contributes to the potential inaccuracies.

This analysis can be extended to many more microbial species given that there are currently over 250,000 genomes in the NCBI database. However, a number of technical challenges need to be addressed, or otherwise additional genomes will contribute more to noise than signal. These challenges fall into two categories: errors and biases. The errors are contributed by incorrect or incomplete assemblies (10% complete, 40% scaffold, and 50% contig level) and poor annotation quality for incomplete assemblies. Even the labeling of the organisms themselves may not be correct in all cases [35]. The biases appear from the distribution of available organisms toward medically relevant species.

**Methods**

A total of 210,130 publicly available prokaryotic genomes were downloaded from the NCBI database in FASTA format along with their corresponding GFF annotations. After excluding all genomes with incomplete assemblies or annotations, 11,904 genomes remained. All gene occurrences from these genomes were extracted based on their corresponding annotations. Genes either missing the gene name attribute or tagged with the pseudo attribute were removed. Only the first feature was considered for genes with multiple consecutive CDS
features. This analysis produced a total of 3,818,377 gene occurrences across 11,904 genomes for which the start codon nucleotide sequences were extracted. For each gene occurrence, its gene name, start codon, start position, end position, and source file name were captured in a separate table.

To break down the start codon frequencies phylogenetically, we ultimately concluded that it would be most insightful to categorize gene-start codon pairs based on the genus from which they were sequenced. For each gene within a genus, a normalized value (“non-AUG ratio”) was introduced, which measured the ratio of its number of non-AUG start codon occurrences to its total number of occurrences. Using this normalized value reduced biases from those genuses with a high number of sequenced genomes. Furthermore, to avoid bias from phyla with many sequenced genuses, this analysis was limited to the 30 genuses selected from proteobacteria, firmicutes, actinobacteria, planctomycetes and tenericutes. Only those genuses with high total gene occurrences were selected and this minimized noise from the genuses with few prior annotations.

The final 36 non-AUG genes are those who were present in at least two-thirds of the 30 genuses (at least 20), and who had a non-AUG ratio of more than 0.5 in at least a third of the 30 genuses (at least 10).

All code and data used in this work are publicly available from gitlab (https://gitlab.com/anne__g/start-codon-analysis).
Figure 1. A graphical representation of the preliminary distribution of start codon frequencies across bacterial genomes, categorized based on the phylum to which the genome belonged. The bars represent the ratio of the total number of genes using AUG start codons to the total number of genes within a phylum.

| Superphylum   | Phylum            | Genus                                                      |
|--------------|-------------------|------------------------------------------------------------|
| PVC Group    | Planctomycetes    | Planctomycetes                                            |
| Terrabacteria| Tenericutes       | Mycoplasma                                                 |
| Terrabacteria| Actinobacteria    | Corynebacterium, Mycobacterium, Streptomyces, Bifidobacterium |
| Terrabacteria| Firmicutes        | Bacillus, Lactobacillus, Listeria, Staphylococcus, Streptococcus, Paenibacillus, Clostridium, Brevibacillus, Clostridioides, Enterococcus |
| Genus Type | Genus | Examples |
|------------|-------|----------|
| Proteobacteria | Betaproteobacteria | Bordetella, Neisseria |
| Proteobacteria | Delta/Epsilon Subdivisions | Campylobacter, Helicobacter |
| Proteobacteria | Gammaproteobacteria | Mannheimia, Pasteurella, Serratia, Citrobacter, Enterobacter, Escherichia, Klebsiella, Salmonella, Xanthomonas, Pseudomonas |

**Table 1.** The list of the 30 genuses we considered in our research, organized by their superphylum and phylum.

| Gene | Detailed Function | Functional Category | Effect of Deletion in *E. coli* |
|------|-------------------|---------------------|---------------------------------|
| infC | translation initiation factor | translation | essential |
| rpoC | RNA polymerase subunit beta | transcription | essential |
| rnpA | RNase P protein component | translation | essential |
| hisS | Histidine-tRNA ligase | translation | essential |
| rimP | ribosome maturation factor | translation | reduced growth at high-temperature |
| recJ | ssDNA-specific exonuclease | replication | viable |
| atpF | ATP synthesis | other | viable |
| bipA (typA) | 50S ribosomal subunit assembly factor | translation | reduced growth |
| dnaA | Chromosomal replication | replication | essential |
| rpsT | 30S ribosomal subunit | translation | essential |
| rsfS | ribosomal silencing factor | translation | substantially reduced cell viability |
| uvrC | DNA damage repair | replication | viable |
| Gene   | Function                                        | Process       | Phenotype               |
|--------|-------------------------------------------------|---------------|-------------------------|
| hflX   | ribosome rescue factor                          | translation   | viable                  |
| crcB   | F(-) channel                                    | other         | essential               |
| rsgA   | ribosome small subunit-dependent GTPase A       | translation   | essential               |
| carA   | carbamoyl phosphate synthetase subunit alpha    | other         | viable                  |
| infB   | translation initiation factor IF-2t             | translation   | essential               |
| rlmH   | 23S rRNA m(3)psi1915 methyltransferase          | translation   | viable                  |
| argS   | arginine-tRNA ligase                            | translation   | essential               |
| pnp    | polynucleotide phosphorylase                   | transcription | reduced growth          |
| pxpB   | 5-oxoprolinase component B                     | other         | slows growth on minimal medium |
| tuf    | translation elongation factor                   | translation   | viable                  |
| secF   | Sec translocon accessory complex subunit        | protein export| disruption of secD and secF confers cold-sensitive growth |
| rpoB   | RNA polymerase subunit beta                     | transcription | essential               |
| secD   | Sec translocon accessory complex subunit        | protein export| disruption of secD and secF confers cold-sensitive growth |
| fdhD   | sulfurtransferase for molybdenum cofactor sulfuration | other        | exhibit defect in FDH-H activity |
| rsmH   | 16S rRNA m(4)C1402 methyltransferase            | translation   | viable                  |
| rsmG   | 16S rRNA m(7)G527 methyltransferase             | translation   | low level streptomycin resistance |
| murE   | UDP-N-acetylglucuronyl-L-alanyl-D-glucurate--2%2C6-diaminopimelate ligase | other         | essential               |
| rpsN   | 30S ribosomal subunit                           | translation   | essential               |
| nusB   | transcription antitermination protein           | transcription | essential               |
| Gene | Function Description | Relation | Essentiality |
|------|----------------------|----------|-------------|
| grpE | nucleotide exchange factor | protein folding | essential |
| ispG | (E)-4-hydroxy-3-methylbut-2-enyl-diphosphate synthase (flavodoxin) | other | essential |
| radA | DNA recombination protein | replication | reduced survival after ionising irradiation |
| plsX | putative phosphate acyltransferase | other | viable |
| rpsC | 30S ribosomal subunit | translation | essential |

**Table 2.** The top 36 “non-AUG genes” are presented with a) their corresponding functions, b) information on whether or not they are functionally related to translation or DNA replication, and c) whether they are identified as essential in *E. coli*.

**Figure 2.** A table comparing the start codon frequencies for each gene-genus combination. Genuses are organized phylogenetically; the top 36 non-AUG genes are ordered according to their tendency to use a non-standard start codon. Blank squares occurred where the gene was not present in the genus among the selected gff files.
References

1. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. Science. 2007;316: 1497–1502.

2. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science. 2008;320: 1344–1349.

3. Nilsson T, Mann M, Aebersold R, Yates JR 3rd, Bairoch A, Bergeron JJM. Mass spectrometry in high-throughput proteomics: ready for the big time. Nat Methods. 2010;7: 681–685.

4. Gold L, Pribnow D, Schneider T, Shinedling S, Singer BS, Stormo G. Translational initiation in prokaryotes. Annu Rev Microbiol. 1981;35: 365–403.

5. Kozak M. Initiation of translation in prokaryotes and eukaryotes. Gene. 1999. pp. 187–208. doi:10.1016/s0378-1119(99)00210-3

6. Gualerzi CO, Pon CL. Initiation of mRNA translation in bacteria: structural and dynamic aspects. Cellular and Molecular Life Sciences. 2015. pp. 4341–4367. doi:10.1007/s00018-015-2010-3

7. Rodnina MV. Translation in Prokaryotes. Cold Spring Harb Perspect Biol. 2018;10. doi:10.1101/cshperspect.a032664

8. Dalgarno L, Shine J. Ribosomal RNA. The Ribonucleic Acids. 1977. pp. 195–232. doi:10.1007/978-1-4612-6360-9_6

9. Kozak M. Regulation of translation via mRNA structure in prokaryotes and eukaryotes. Gene. 2005;361: 13–37.

10. Zheng X, Hu G-Q, She Z-S, Zhu H. Leaderless genes in bacteria: clue to the evolution of translation initiation mechanisms in prokaryotes. BMC Genomics. 2011. doi:10.1186/1471-2164-12-361

11. Benelli D, Londei P. Begin at the beginning: evolution of translational initiation. Res Microbiol. 2009;160: 493–501.

12. Blattner FR. The Complete Genome Sequence of Escherichia coli K-12. Science. 1997. pp. 1453–1462. doi:10.1126/science.277.5331.1453

13. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, et al. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature. 1998;393: 537–544.

14. Cao X, Slavoff SA. Non-AUG start codons: Expanding and regulating the small and alternative ORFeome. Exp Cell Res. 2020;391: 111973.

15. Tikole S, Sankararamakrishnan R. A survey of mRNA sequences with a non-AUG start
16. Hecht A, Glasgow J, Jaschke PR, Bawazer L, Munson MS, Cochran J, et al. Measurements of translation initiation from all 64 codons in E. coli. doi:10.1101/063800

17. Butler JS, Springer M, Grunberg-Manago M. AUU-to-AUG mutation in the initiator codon of the translation initiation factor IF3 abolishes translational autocontrol of its own gene (infC) in vivo. Proceedings of the National Academy of Sciences. 1987. pp. 4022–4025. doi:10.1073/pnas.84.12.4022

18. Brombach M, Pon CL. The unusual translational initiation codon AUU limits the expression of the infC (initiation factor IF3) gene of Escherichia coli. Mol Gen Genet. 1987;208: 94–100.

19. Muralikrishna P, Wickstrom E. Inducible high expression of the Escherichia coli infC gene subcloned behind a bacteriophage T7 promoter. Gene. 1989. pp. 369–374. doi:10.1016/0378-1119(89)90301-6

20. Gold L, Stormo G, Saunders R. Escherichia coli translational initiation factor IF3: a unique case of translational regulation. Proc Natl Acad Sci U S A. 1984;81: 7061–7065.

21. Villegas A, Kropinski AM. An analysis of initiation codon utilization in the Domain Bacteria - concerns about the quality of bacterial genome annotation. Microbiology. 2008;154: 2559–2661.

22. Parsons GD, Donly BC, Mackie GA. Mutations in the leader sequence and initiation codon of the gene for ribosomal protein S20 (rpsT) affect both translational efficiency and autoregulation. J Bacteriol. 1988;170: 2485–2492.

23. Wirth R, Littlechild J, Böck A. Ribosomal protein S20 purified under mild conditions almost completely inhibits its own translation. Molecular and General Genetics MGG. 1982. pp. 164–166. doi:10.1007/bf00333013

24. Braun RE, O'Day K, Wright A. Autoregulation of the DNA replication gene dnaA in E. coli K-12. Cell. 1985. pp. 159–169. doi:10.1016/0092-8674(85)90319-8

25. Dennis PP, Nene V, Glass RE. Autogenous posttranscriptional regulation of RNA polymerase beta and beta` subunit synthesis in Escherichia coli. Journal of Bacteriology. 1985. pp. 803–806. doi:10.1128/jb.161.2.803-806.1985

26. Jarrige AC, Mathy N, Portier C. PNPase autocontrols its expression by degrading a double-stranded structure in the pnp mRNA leader. EMBO J. 2001;20: 6845–6855.

27. Haggerty TJ, Lovett ST. IF3-mediated suppression of a GUA initiation codon mutation in the recJ gene of Escherichia coli. J Bacteriol. 1997;179: 6705–6713.

28. Vioque A, Arnez J, Altman S. Protein-RNA interactions in the RNase P holoenzyme from Escherichia coli. J Mol Biol. 1988;202: 835–848.

29. Belinky F, Rogozin IB, Koonin EV. Selection on start codons in prokaryotes and potential codon in RefSeq database. J Biomol Struct Dyn. 2006;24: 33–42.
compensatory nucleotide substitutions. Sci Rep. 2017;7: 12422.

30. La Teana A, Pon CL, Gualerzi CO. Translation of mRNAs with degenerate initiation triplet AUU displays high initiation factor 2 dependence and is subject to initiation factor 3 repression. Proc Natl Acad Sci U S A. 1993;90: 4161–4165.

31. Nomura M, Gourse R, Baughman G. Regulation of the synthesis of ribosomes and ribosomal components. Annu Rev Biochem. 1984;53: 75–117.

32. Grosjean H, Breton M, Sirand-Pugnet P, Tardy F, Thiaucourt F, Citti C, et al. Predicting the minimal translation apparatus: lessons from the reductive evolution of mollicutes. PLoS Genet. 2014;10: e1004363.

33. Gil R, Silva FJ, Peretó J, Moya A. Determination of the core of a minimal bacterial gene set. Microbiol Mol Biol Rev. 2004;68: 518–37, table of contents.

34. Hutchison CA 3rd, Chuang R-Y, Noskov VN, Assad-Garcia N, Deerinck TJ, Ellisman MH, et al. Design and synthesis of a minimal bacterial genome. Science. 2016;351: aad6253.

35. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics. 2019. doi:10.1093/bioinformatics/btz848