Open access • Posted Content • DOI:10.1101/614032

tRNAscan-SE 2.0: Improved Detection and Functional Classification of Transfer RNA Genes — Source link

Patricia P. Chan, Brian Y. Lin, Allysia J. Mak, Todd M. Lowe

Institutions: University of California, Santa Cruz

Published on: 30 Apr 2019 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: Gene prediction and Pseudogene

tRNAscan-SE 2.0: Improved Detection and Functional Classification of Transfer RNA Genes

Patricia P. Chan[†], Brian Y. Lin[†], Allysia J. Mak, and Todd M. Lowe*

Department of Biomolecular Engineering, University of California Santa Cruz, CA, 95064, USA

* To whom correspondence should be addressed. Tel: +1 831 459 1511; Fax: +1 831 459 4829; Email: lowe@soe.ucsc.edu

[†] These authors contributed equally to the paper. The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

ABSTRACT

tRNAscan-SE has been widely used for whole-genome transfer RNA gene prediction for nearly two decades. With the increased availability of new genomes, a vastly larger training set has enabled creation of nearly one hundred specialized isotype-specific models, greatly improving tRNAscan-SE's ability to identify and classify both typical and atypical tRNAs. We employ a new multi-model annotation strategy where predicted tRNAs are scored against a full set of isotype-specific covariance models. A post-filtering feature also better identifies tRNA-derived SINEs that are abundant in many eukaryotic genomes, and provides a "high confidence" tRNA gene set which improves upon prior pseudogene prediction. These new enhancements of tRNAscan-SE will provide researchers more accurate detection and more comprehensive annotation for tRNA genes.

INTRODUCTION

Transfer RNAs (tRNAs) are ubiquitous in all living organisms as the key translator of the genetic code into proteins. tRNAscan-SE (1) is the most widely used tool for identifying and annotating tRNA genes in genomes. With over eight thousand citations, its users include RNA biologists, sequencing centers, database annotators, and other basic researchers. To increase the ease of use for scientists who may not have the expertise to work with UNIX-based software, the tRNAscan-SE On-line website (2,3) provides quick, in-depth tRNA analysis. tRNAs predicted using tRNAscan-SE are available in the Genomic tRNA Database (GtRNAdb) (4,5) for thousands of genomes, enabling the research community to browse high-quality tRNA collections across all three domains of life.

The original tRNAscan-SE implementation pioneered the large-scale use of covariance models (CMs) (6) to annotate RNA genes in genomes, predating the invaluable Rfam database (7). By training on structurally aligned members of the same RNA family, covariance models capture RNA conservation via stochastic context-free grammars that are able to integrate both primary sequence and secondary structure information. Despite of the intensive computational requirements, covariance models yield unparalleled sensitivity and specificity in finding tRNAs and many other structured RNAs. Any given sequence can be searched for tRNAs by alignment to a tRNA covariance model. Depending on the training set of tRNAs used to construct the covariance model, the search can be

tailored for general detection of any tRNA sequence, or for more specialized searches to detect tRNAs with clade-specific features (e.g. eukaryotic cytosolic-type tRNAs) or specific tRNA types (e.g. initiator methionine tRNAs). Thus, tRNAscan-SE can easily be "tuned" to different types of tRNAs, limited only by high quality training set alignments. This flexibility enabled clade-specific search modes for eukaryotic, bacterial, archaeal, or organellar tRNAs in previous versions of tRNAscanSE; the same flexibility provides a framework for powerful new clade and isotype-specific search modes, limited only by tRNA sequence training sets. To reduce the computational load required for CM search and alignment across entire genomes, the original tRNAscan-SE used two fast and sensitive algorithms as a first-pass screen to identify putative tRNAs (8,9). The program then aligned only the putative tRNAs to CMs and identified key information such as the isotype and intron boundaries. The initial general and domain-specific CMs were built using an alignment of 1,415 tRNAs extracted from the gold-standard Sprinzl database (10,11), while CM construction and alignment was performed using COVE (6).

Since its first implementation, numerous other tRNA detection and classification methods have been developed, including ARAGORN (12) which detects both tRNA genes and tmRNA genes; DOGMA (13), ARWEN (14), and MITOS (15) which are designed for annotating tRNAs in various types of organellar genomes, TFAM (16) that classifies tRNAs based on log-odds profiles built from covariance models; tRNAfinder (17); a rule-based program that detects tRNAs through secondary structure; and SPLITS (18) which is designed to find split and intron-containing tRNAs in microbial genomes. All of these methods were designed to either improve upon or complement tRNAscan-SE, and notably, many depend on tRNAscan-SE's core detection software.

Although tRNAscan-SE has remained a reliable and easily accessible tool for tRNA detection over the past two decades, it has been in need of a major revision to incorporate new algorithms, new data, and new strategies for improved performance and functional prediction accuracy. Here, we describe the latest version of tRNAscan-SE that has enhancements including (1) improved covariance model search technology in the form ofintegration of Infernal 1.1 covariance model search software (19); (2) updated search models leveraging a more broadly representative diversity of tRNA genes from thousands of newly sequenced genomes (Table 1); (3) better functional classification of tRNAs, based on comparative information from a full suite of isotype-specific tRNA covariance models, and (4) a new "high confidence" filter to identify eukaryotic tRNAs that are the most likely to be used in protein translation.

MATERIAL AND METHODS

tRNAscan-SE prediction results

Predicted tRNA genes using tRNAscan-SE 2.0 are currently available in the Genomic tRNA Database (GtRNAdb) (5). The genome assemblies used in this study include 4041 bacteria, 216 archaea, and 181 eukaryotes. While these are not exhaustive analyses of all available genomes (which increases every day), it does constitute a good representation of high quality, substantially complete genomes.

Bacterial, archaeal and 72 fungal eukaryotic genomes were obtained from NCBI GenBank, and 101 large eukaryotic genomes were obtained from the UCSC Genome Browser (21) and JGI Phytozome (22). The prediction results of specific genomes highlighted in this manuscript are human (GRCh37/hg19 and GRCh38/hg38), mouse (GRCm38/mm10), cat (Felis_catus_9.0), *Saccharomyces cerevisiae* S288c (GCA_000146045.2/R64), *Escherichia coli* str. K-12 substr. MG1655 (GCA_000005845.2 / ASM584v2), and *Pyrococcus furiosus* DSM 3638 (GCA_000007305.1 / ASM730v1).

tRNA search modes

In default search mode, the new tRNAscan-SE 2.0 uses Infernal 1.1 (19) as the state-of-the-art sequence search engine to find and score tRNA genes (Figure 1). The original tRNAscan-SE 1.0 used tRNAscan (8) and software implementing an algorithm from Pavesi and colleagues (9) as the sensitive first-pass candidate-gathering searches, and Infernal's forerunner, COVE (6), as the high-specificity tRNA filter. This 1.0 search mode is still available in 2.0 as the "legacy mode" (-L) for researchers who wish to make backward version comparisons. The Infernal software implements profile stochastic context-free grammars, also known as "covariance models" because of their ability to detect covariation in conserved RNA secondary structures. Covariance models can be created to identify members of any RNA gene family based on structurally aligned, trusted examples which serve as training sets. tRNAscan-SE 2.0 (Figure 1) employs a combination of 76 different covariance models (described below) for identifying and classifying the many different types and biological sources of tRNAs in two phases. In step one, models trained on all tRNA isotypes from species in each domain of life (Eukayota, Bacteria, or Archaea) are used to maximize sensitivity for predicting different types of tRNA genes. The incorporation of the accelerated profile hidden Markov model (HMM) methods used in HMMER3 (23,24) and the constrained CM alignment algorithms (19,25) in Infernal provide multiple levels of filtering as part of sequence homology searches. In the default setting, tRNAscan-SE 2.0 adopts the mid-level strictness (cmsearch option: –mid) with a low-score cutoff (10 bits) to replace the first-pass filters in tRNAscan-SE 1.3 for high sensitivity and fast performance. Then, for the second-pass specificity scan, it uses Infernal without the HMM filter against the first-pass candidates with extra flanking sequences and a default score threshold of 20 bits. For users who need to obtain maximum search sensitivity and can accept a longer processing time, we also include the use of Infernal without HMM filter (--max option) as an addition single-pass search option. To further improve identification of slightly shortened tRNA genes, the new algorithm also makes use of the truncated hit detection feature in Infernal to annotate tRNA predictions that are possibly truncated at either or both ends of the sequence. After initial tRNA gene prediction, the task of isotype classification is performed by comparing each anticodon prediction (based purely on its position in the anticodon loop of the tRNA secondary structure) to scores against a suite of isotype-specific covariance models, a strategy similar to TFAM (16). Now, alongside the predicted anticodon, the highest scoring isotype-specific model is also reported in the output details; any disagreement between the two functional prediction methods can be noted and more closely inspected by the user.

Default tRNA prediction score threshold

The default score cutoff for tRNA predictions in tRNAscan-SE 1.3 using COVE (6) is 20 bits. To assess if this threshold can be applied to the new version with Infernal (19), we generated virtual genomes for *E. coli* K12, *Halobacterium sp.* NRC-1, *Saccharomyces cerevisiae* S288C, and *Homo sapiens* (representing a broad diversity of different GC contents and phylogenetic clades) by using a 5^{th} order Markov chain to retain the base frequency of the original genome (Supplementary Table S1). tRNAscan-SE 2.0 default search mode and score cutoff 10 bits was used to search 260 virtual genomes. The highest scoring hit in this negative control sequence set scored 19.8 bits, and just 125 other hits were found with scores ranging between 10-20 bits. For reference, 32,390 tRNAs were identified in the equivalent number of copies of real genomes using a score cutoff of 20 bits or greater. Based on these results, we did not alter the original 20 bit default score threshold used in prior versions of tRNAscan-SE.

Domain-specific covariance models

Using existing public predictions in GtRNAdb (4) and additional predictions from tRNAscan-SE 1.3, we assembled three sets of domain-specific genomic tRNA sequences from a total of over 4000 genomes, representing a broad diversity of eukaryotes, bacteria, and archaea (Table 1, Supplementary Figure S1). Before using these tRNA sequences as training sets for building domain-specific covariance models, multiple filtering steps were taken to ensure optimal results. To avoid the inclusion of common tRNA-derived repetitive elements that exist in many eukaryotic genomes, especially mammals (30-32), we first selected only the eukaryotic tRNAs with a COVE score greater than 50 bits (a threshold reflecting more conserved, canonical tRNA features). We then selected only the top 50 scoring tRNAs for each isotype per organism to avoid overrepresentation of highly conserved tRNA-like repetitive elements found in some species (for example, elephant shark has over 9,500 $tRNA^{Ala}$ scoring over 50 bits). For the bacterial tRNA training set, genes that have long self-splicing introns were excluded to eliminate large alignment gaps in model creation that resulted in greatly increased search time for these rare cases. Similarly for archaea, pre-processing of sequence training sets was necessary. Some species within the phyla Crenarchaeota and Thaumarchaeota contain a number of tRNAs that are known to have multiple non-canonical introns (4,5,18,33). Atypical tRNAs such as *trans*-spliced tRNAs and circularized permuted tRNAs have also previously been discovered in crenarchaea and nanoarchaea (34-36). To accommodate these special archaeal features without sacrificing performance, both mature tRNA sequences (without introns) and selected atypical genes with multiple introns at different locations were included in the archaeal tRNA training sets. As a last step, anticodons of all tRNA sequences in all training sets were replaced with NNN and were aligned to the corresponding original domain-specific tRNA covariance models using Infernal (19). The resulting alignments were then used to generate the new set of domain-specific tRNA covariance models with Infernal.

Isotype-specific covariance models

The isotype-specific covariance models for the three phylogenetic domains were built iteratively through two rounds of training to account for sequence features specific to each tRNA isotype

(Supplementary Figure S1). In the first round, the filtered tRNA genes used for building the domain-specific covariance models were divided into groups according to the tRNA isotypes determined by the anticodon sequence. For archaeal models, tRNA sequences of genes with noncanonical introns were "pre-spliced" to only include their mature sequences. The sequences of each isotype group were then aligned to the tRNA covariance model of the corresponding domain using Infernal (19). "Intermediate" covariance models for each isotype were built using the resulting alignments. In a second round of model-building, the original training set from each domain was scored against the new intermediate covariance models. The sequences were then grouped according to isotype, this time based on which isotype-specific covariance model yielded the highest score for each sequence (regardless of anticodon sequence). These revised isotype-classified sequence groups were re-aligned to the corresponding intermediate models to build the final isotype-specific covariance models.

Covariance models for methionine tRNAs and isoleucine tRNAs decoding AUA

In eukayotes and archaea, initiator methionine tRNA ($tRNA^{iMet}$) and elongator methionine tRNA ($tRNA^{Met}$) have distinct sequence features and functions but contain the same anticodon, CAU. Similarly, N-formylmethionine tRNA ($tRNA^{fMet}$) in bacteria also contains the same CAU codon as the structurally distinct elongator methionine. At the same time, $tRNA^{Ile2}$ which decodes the isoleucine AUA codon is *also* encoded by tRNA genes containing a CAU anticodon. However, these special $tRNA^{Ile2}$ are post-transcriptionally modified on their wobble bases (lysidine in bacteria and agmatidine in archaea), effectively giving them isoleucine-specific UAU anticodons. The strategy of prior versions of tRNAscan-SE to identify tRNAs based on their anticodon failed to separate these three functionally distinct tRNAs.

In order to develop accurate covariance models that represent these structurally and functionally different tRNAs, we applied the above two-round training method with carefully selected training sets. For eukaryptes, the sequences of $tRNA^{iMet}$ and $tRNA^{Met}$ were selected from the original 1415 tRNAs used for training tRNAscan-SE 1.3. For bacteria, we collected the sequences for $tRNA^{fMet}$, $tRNA^{Met}$, and $tRNA^{Ile2}$ from 234 genomes where these tRNAs were classified (37). For archaea, we curated the sequences based on known identity elements of these three different tRNAs (38). These sequences were aligned to the corresponding domain-specific covariance models for generating the first-round intermediate covariance models followed by the second-round training step, producing the final covariance models.

Selenocysteine tRNA modelling

Selenocysteine tRNAs ($tRNA^{SeC}$) have secondary structures and sequence lengths that differ from other tRNA isotypes, and are only a component of the protein translation system of some species that employ selenoproteins. While eukaryotic and archaeal $tRNA^{SeC}$ have a 9-bp acceptor stem and a 4-bp T-arm (9/4 fold) (39-42), bacteria have an 8-bp acceptor stem and a 5-bp T-arm (8/5 fold) (43,44). To build covariance models for these special cases, we curated the sequences of $tRNA^{SeC}$ from genomes

where selenoproteins were previously identified and removed those that do not retain published canonical features: the special secondary structure of $tRNA^{SeC}$, a UCA anticodon, and G at position 73. The collected sequences (65 eukaryotic (45-56), 61 bacterial (57), and 13 archaeal (42,58-61)) were aligned to the original $tRNA^{SeC}$ covariance models from tRNAscan-SE 1.3, with manual inspection and adjustments. The resulting alignments were used to build the domain-specific $tRNA^{SeC}$ covariance models.

Identification of non-canonical introns in archaeal tRNAs

tRNA candidates predicted with archaeal search model are analyzed for non-canonical introns (Supplementary Figure S2). Two covariance models that include the bulge-helix-bulge (BHB) secondary structure were built with manually curated non-canonical tRNA introns from (1) crenarchaea and euryarchaea, and (2) thaumarchaea, respectively. The tRNA candidates are extended with 60 nucleotides of flanking sequences in order to identify additional introns located near the 5' and/or 3' ends, and scanned with the BHB covariance models. Predicted non-canonical introns are confirmed when the score of the predicted mature tRNA with detected intron(s) removed is higher than the unspliced form. Some tRNAs contain two or three introns, with introns located in such closely proximity that one intron must be removed before a second BHB motif can be detected in the second intron; thus, multiple iterations of the intron search are required, and end when the final tRNA prediction score is higher than the previous iteration, or the length of the predicted mature tRNA is less than 70 nucleotides (the typical minimum length of archaeal tRNAs). Final reported scores in tRNAscan-SE 2.0 outputs are based on the predicted mature tRNAs.

Custom search configurations

To increase the flexibility of tRNAscan-SE 2.0, we include a configuration file that accompanies the software. This file contains default parameters such as score thresholds for the first and second-pass scans, file locations of covariance models, and legacy search mode settings. Advanced users can make changes to the settings as appropriate for their research needs. To further extend its capability, an extra search option was implemented in tRNAscan-SE 2.0 that allows researchers to use alternate covariance models specified in the configuration file for tRNA searching. If multiple alternate covariance models are included, the top scoring one at each overlapping locus will be reported. This new feature enables the use of custom-built covariance models that may better detect tRNAs with atypical unique features.

Post-filtering of potential tRNA-derived repetitive elements and classification of high confidence tRNAs

EukHighConfidenceFilter in tRNAscan-SE 2.0 is a post-scan filtering program for better distinguishing potential tRNA-derived repetitive elements in eukaryotic genomes from "real" tRNAs that function in protein translation. Three filtering stages are involved in the classification (Supplementary Figure S3). First, tRNA predictions that are considered as possible pseudogenes based on criteria defined in tRNAscan-SE (an overall score below 55 bits with the primary sequence score below 10 bits or the

secondary structure score below 5 bits, as used in the original version of tRNAscan-SE) are excluded from the high confidence set. Second, predictions with an isotype-specific model score below 70 bits, overall score below 50 bits, or secondary structure score below 10 bits are filtered from the high confidence set. Finally, if there are more than 40 predicted hits remaining for an isotype anticodon, a dynamic score threshold increases, starting at 71 bits, increasing one bit iteratively until the number of predictions is not over 40, or the score threshold reaches 95 bits. The score thresholds in stages two and three were manually determined by comparing score distributions of predictions between eukaryotic genomes with large number of tRNA-derived repetitive elements and those that do not have high portion of "false positive" predictions (Supplementary Figure S4). The remaining tRNA predictions are considered as part of the high confidence set if they have a consistent isotype prediction (inferred from anticodon versus the highest scoring isotype-specific model) and they have an "expected" anticodon -- based on known decoding strategies of synonymous codons in eukaryotes, 15 anticodons are not used (63).

RESULTS

The use of Infernal increases flexibility and speed

One of the most substantial improvements to tRNAscan-SE 2.0 is the incorporation of Infernal 1.1 (19) that allows us to improve tRNA prediction quality with comparable performance and sensitivity as earlier versions of tRNAscan-SE (see Methods). The original two-pass search strategy of tRNAscan-SE utilized a rule-based detection method (8) and static, unchanging eukaryotic weight matrices (9) to increase the speed of tRNAscan-SE. The predicted candidates were further searched for primary sequence and secondary structure similarities using the first generation of covariance models (6) built for tRNAs. These effective but inflexible and computationally expensive methods proved difficult to adapt for detection and classification of specialized sub-classes of tRNAs, such as distinguishing between initiator methionine tRNAs and elongator methionine tRNAs. By contrast, Infernal makes use of modern computer architecture for parallelizing hidden Markov model dynamic programming algorithms (24), enabling fast similarity search and development of better-trained covariance models that include tRNA features from over 4000 genomes across the three domains of life (Table 1).

Although tRNAscan-SE 2.0 does not provide significant speed advancement when predicting tRNAs in organisms with small genomes such as *E. coli* and *Saccharomyces cerevisiae* (budding yeast), the total run time for large eukaryotic genomes improves as the number of predicted genes increases. Searching the human and mouse genomes using the new version with ten computing cores in our test server (dual 10-core/20-thread 2.3GHz processors with 128GB RAM) reduces the processing time by 40%. The speed improvement is even more extreme with the use of the maximum sensitivity mode. When the tRNAscan-SE was first released, it was estimated that searching the human genome with covariance-model-only analysis would take nine-CPU years. Although the computer technology has been significantly advanced over the years, the covariance model searching program, COVE (6), used by tRNAscan-SE 1.3 was not designed to take advantage of the available multi-core parallel processors in our test, resulting in an estimate of over 3 months to complete the

search of the human genome in our test server (dual 10-core/20-thread processors with 128GB RAM). Whereas, using the new version of tRNAscan-SE with Infernal (19), the processing time for the human genome will be within four days and even less than 15 minutes for *E. coli*, making the covariance-model-only search feasible to be used by researchers who would like to maximize the sensitivity.

When comparing the detected tRNA genes in microorganisms, we found that the ones in *Saccharomyces cerevisiae* (budding yeast) and *Pyrococcus furiosus* (archaea) are the same using both versions of tRNAscan-SE. Yet, an extra predicted tRNA gene was identified by the new version in *Escherichia coli*. With the score cutoff as 20 bits (see Methods), this very low-scoring prediction (29.6 bits) with undetermined anticodon and isotype calls for investigation. In fact, we found that it is part of gene b2621, a tmRNA that is known to have tRNA-liike properties. Moreover, out of the 500+ predicted cytosolic tRNAs in the human genome, tRNAscan-SE 2.0 detects 56 more candidates but misses 49 originally predicted by the previous version. Of which, 57% and 73% respectively are classified as possible pseudogenes. The highest score of extra candidate is 47.6 bits whereas that of the missing one is 34.44 bits, both lower than the minimum score of a high-confidence eukaryotic tRNA (see Methods) that functions in translation. We further studied the sequencing results of those non-pseudogenes from ARM-Seq (64) and DASHR (65) and found that almost none of them shows significant expression level in multiple samples and conditions with only a few exceptions. For example, a predicted $tRNA^{Tyr(GTA)}$ gene with a score of 22.8 bits has low level of abundance across the precursor tRNA region in all the ARM-Seq samples (Supplementary Figure S5) (64). Uniform coverage is observed between the mature tRNA region and the 12-nt intron, suggesting that this candidate may not be processed as a typical tRNA. This shows increased sensitivity of the newly built covariance models and tRNAscan-SE 2.0 without losing the specificity to determine real tRNA hits. While there seems to be more differences in the number of detected tRNAs in other mammalian genomes such as mouse and cat, most of those have a low score especially due to the abundance of tRNA-derived repetitive elements, which will be discussed later in more details.

Enabling discrimination between initiator methionine, elongator methionine, and tRNA-Ile2

When tRNAscan-SE was designed almost two decade ago, only a limited number of tRNA genes were identified and available in the 1993 Sprinzl database (10), which were not sufficient to be used as a training set for creating robust isotype-specific covariance models. With thousands of new genomes in public databases, we estimated that there would be sufficient tRNA gene diversity to train specialized models for different functional classes of tRNAs. Sequence-based positive and negative determinants are used by tRNA aminoacyl synthetases to establish tRNA identity, and have been characterized in a number of model species (66). Previously, the original version of tRNAscan-SE used the anticodon exclusively to predict isotype because anticodon-isotype pairings are highly conserved and the anticodon is easily detected in a tRNA gene. However, this method does not provide the capability to distinguish between the three different tRNA types with anticodon CAT in genomic sequences: initiator methionine/N-formylmethionine tRNA ($tRNA^{iMet}$ and $tRNA^{fMet}$), elongator methionine tRNA ($tRNA^{Met}$), and isoleucine tRNA decoding AUA codon ($tRNA^{Ile2}$) in bacteria and

archaea. By creating specialized models for these sub-groups of tRNAs in each domain of life, we were able to distinguish them with high certainty.

When comparing the isotype-specific covariance model scores of the predicted tRNAs in a total of almost 4,500 genomes across Eukaryota, Bacteria, and Archaea domains, we found that the three $tRNA^{iMet/fMet}$, $tRNA^{Met}$, and $tRNA^{Ile2}$ form distinct clusters (Figure 2) with $tRNA^{iMet/fMet}$ cluster conspicuously separated from the other two tRNA groups in bacteria and archaea. This reflects that the consensus sequence of $tRNA^{iMet/fMet}$ is relatively less similar to those of $tRNA^{Met}$, and $tRNA^{Ile2}$ (38). We checked the number of identified tRNAs in each studied genome and found that 3.5% (6 of 173) eukaryotes and 3.6% (144 of 4034) bacteria miss at least one of the three AUA-anticodon tRNAs. Of the 6 eukaryotic genomes that belong to three different fungal genera, all of them have $tRNA^{iMet}$ mis-annotated as $tRNA^{Met}$ because iMet covariance model has the second highest score, suggesting that these tRNA sequences are relatively atypical when comparing to the consensus and may need further studies on the differences. On the other hand, missing tRNA genes in bacterial genomes could be a result from the insertion of self-splicing group I introns that tRNAscan-SE was not designed to detect. For example, $tRNA^{Ile2}$ was not detected in two *Burkholderia pseudomallei* strains (NAU35A-3 and TSV 48) but was found in reference strain K96243. Close inspection with manual sequence alignments and RNA family similarity search (7,19) shows that group I introns of 7,726bp and 7,730 bp respectively exist between positions 31 and 32 of $tRNA^{Ile2}$ in the two strains, causing the failure to correctly identify the tRNA genes.

Isotype-specific covariance models improve functional annotation of tRNAs

Besides distinguishing tRNA genes with anticodon CAT, isotype-specific covariance models can also help better classify predicted genes. For example, previous studies identified “chimeric” tRNAs that have identity elements recognized by one type of tRNA synthetase, but with altered anticodon corresponding to mRNA codon of a different amino acid (67,68). Although the biological significance of chimeric tRNAs is not well understood, it is valuable to identify a conflict between the anticodon and other structural features that may be related to protein “recoding” events.

Among the genomes we studied, over 95% of the typical predicted tRNA genes have anticodons that match with the highest scoring isotype-specific model (Supplementary Figure S6). When inspecting some of the cases with isotype-anticodon disagreement, we noticed that the uncertainty may be caused by various reasons or have different possible influence in translational events. In human, two high-scoring predictions, tRNA-Val-AAC-6-1 and tRNA-Leu-CAA-5-1 (69.9 bits and 66.5 bits respectively), are inconsistent with the isotype-specific model. tRNA-Val-AAC-6-1 has an anticodon for encoding Valine but scores much better with the $tRNA^{Ala}$ model (88.1 bits vs 45.0 bits). tRNA-Leu-CAA-5-1 is an exemplar chimeric tRNA that scores better with the $tRNA^{Met}$ model than $tRNA^{Leu}$ model (98.4 bits vs 2.6 bits). Its secondary structure (Figure 3B) shows the lack of a long variable arm, a typical character of a type II tRNA as the other $tRNA^{Leu(CAA)}$ (Figure 3C), and has only six nucleotides different from tRNA-Met-CAT-3-1 (Figure 3A, Supplementary Figure S7A). The gene is conserved in most of the primates. However, only human, chimp, and gorilla have A_{36} while the

other ancestral genomes have T_{36} that would transcribe into a $tRNA^{Met(CAU)}$ (Supplementary Figure S7B). This suggests that the mutation was acquired relatively recent and it is interesting to understand its significance in the human genome.

When studying the tRNAscan-SE results in archaea, we noticed *Methanobrevibacter ruminantium* has a $tRNA^{Arg}$ with anticodon ACG that is not found in its closest sequenced relatives, *Methanobrevibacter smittii* and *Methanobrevibacter sp.* AbM4. This is unusual, as A_{34} in $tRNA^{Arg}$ is generally found in bacteria and eukaryotes but not archaea that use G_{34}-containing tRNAs to decode pyrimidine-ending codons (63). The isotype prediction also disagrees with the anticodon with the highest scoring isotype as $tRNA^{Trp}$. In the previous genome analysis, over 13% (294 out of 2217) of the coding genes in *M. ruminantium* were identified to be originated from other species including bacteria and eukaryotes (72). Horizontal gene transfer commonly occurs in microbes that share the same or similar habitats (73,74). Although research have been focused on protein coding genes, we hypothesized that it is possible for non-coding RNA genes transferring between organisms. Applying the bacterial and eukaryotic models, the predicted $tRNA^{Arg(ACG)}$ scores 53.4 bits and 55.1 bits respectively, compared with 37.6 bits using the archaeal model. In addition, the predicted isotype is consistent with the anticodon, suggesting that this gene may be transferred from a species in another domain of life.

New process helps identify archaeal tRNA genes with noncanonical introns

Some tRNAs in eukaryotes and archaea have introns that are removed by tRNA splicing endonuclease during maturation. Although the majority of the archaeal tRNA introns are located one nucleotide downstream of the anticodon (position 37/38), some have been found at seemingly random, “noncanonical” positions in the tRNA genes (5,18,75,76). *Pyrobaculum calidifontis* has the highest number of introns (71 introns in 46 tRNA genes) found in complete genomes and an archaeal tRNA can have up to three introns. These noncanonical introns have presented a challenge for predicting the archaeal tRNA genes correctly. The introns preserve a general bulge-helix-bulge (BHB) secondary structure (75) that can be modelled using Infernal (19) for similarity search. In the previous version of tRNAscan-SE, we included a search routine and a covariance model to detect the noncanonical introns in archaea. The model was built using the known intron sequences at the time mostly identified in crenarchaea. Due to slower performance of the previous Infernal versions, we only made the routine as an optional feature to avoid the significant increase of default search time. In addition, the original covariance model cannot effectively detect the noncanonical tRNA introns in recently sequenced genomes that are more distantly related from those used as the training data. Therefore, we have redesigned the search process (see methods; Supplementary Figure S2) by including two covariance models: one further optimized from the existing model, and the other one newly trained with introns in thaumarchaea. Both models were built with the latest release of Infernal with the performance improvement that makes the process feasible to be included in the default archaeal search mode.

Among the 216 archaeal genomes studied, we found 1,527 canonical and 667 noncanonical introns in a total of 10,334 predicted tRNA genes (Supplementary Table S2). Previous study results and manual inspection show that our process has a low error rate, with 2.9% of the noncanonical introns misisng in the search due to low similarity to the consensus, and four noncanonical introns located at the anticodon loop misannotated as canonical introns. Although almost 70% of the identified introns are canonical, a small number of clades illustrate the opposite. As described in previous studies, a lot of tRNA genes in Thermoproteales have been known to harbour multiple introns (18,33). In our analysis, we found an average of one intron per 0.84 tRNA gene in Thermoproteales as compared with the overall 1-to-0.21 ratio. In addition, two-third of these introns are located at noncanonical positions of the tRNA genes while Euryarchaeota, the phylum with the least amount of noncanonical introns, have over 95% introns at canonical position. Similar to Thermoproteales, genomes in Thaumarchaeota only possess a majority of noncanonical introns. Although, in total, only about 2% of the tRNA genes have two or more introns, the 16 tRNA genes with three introns belong to Thermoproteales, which is consistent with the overall high number of introns identified in this clade.

A post-filtering tool distinguish high confidence predictions from tRNA-derived repetitive elements

Short interspersed repeated elements (SINEs) that are derived from tRNAs have conserved RNA polymerase III promoters internal of tRNA genes but not necessary the typical cloverleaf secondary structure (30-32). This causes the covariance model analysis being able to identify them in a lower score than true tRNA genes but still above the default score threshold (20 bits) as described in the original version of tRNAscan-SE (1). The tRNA-derived SINEs are numerous in mammalian genomes and some other large eukaryotes except primates (30,77-80), resulting in huge number of predictions that may not be accurate. For example, the cat genome has over 403,500 tRNA predictions (the largest amount in the studied genomes) while the rat genome that has been previously reported with many tRNA pseudogenes due to repetitive elements (77) has over 211,000 predictions in the latest assembly (Table 2). Although tRNAscan-SE classifies over 80% of the predictions in these mammals as pseudogenes, the remaining still exceeds our expectation of true tRNA genes in a genome given that there are only about 600 human tRNA gene predictions. When comparing the non-pseudogene prediction score distributions between primates and other mammals, we noticed that the median scores in mammals such as cow and armadillo are significantly lower than those in primates like human (Supplementary Figure S4). In addition, plants like maize that is known to contain repetitive elements also have lower prediction median scores. By checking against the repetitive elements annotated in mouse (one of the most-studied model organism) (79), we found that the low-scoring non-pseudogene tRNA predictions are mostly part of the B1 or B2 repeats. However, due to the different evolutionary age of the repetitive elements that leads to various mutation rates, SINE-origin predictions in marine mammals like minke whale and dolphin tend to have relatively higher tRNA (domain-specific) and isotype-specific scores but retain low secondary structure scores. We therefore developed a filtering tool that can be optionally applied to the tRNAscan-SE results for

better classifying the real tRNA genes. The tool assesses the predictions with a combination of domain-specific, isotype-specific, and secondary structure scores in two filtering stages on top of the pseudogene classification (see Methods), and determines the “high confidence” set of genes that are most likely to be functioned in the translation process. A small number of the predictions that have high scores but atypical features such as unexpected anticodons are separately marked for further investigation. In our study, the high confidence set remains below 1,000 tRNA genes in most genomes (Table 2), which provides researchers a stricter, more conservative set to better focus their experimental efforts.

DISCUSSION

With the improvement of technology, processes and methods that used to take very long execution time have become practically possible. The employment of multi-threaded Infernal v1.1 (19) has allowed us to eliminate the use of the two pre-filters in the original version of tRNAscan-SE without sacrificing performance. The multi-model strategy applied in tRNAscan-SE 2.0 also provides additional annotations through the isotype-specific covariance model classification to better identify functional ambiguity as well as atypical tRNA genes that are worthy of further investigation. The increased availability of genomes in clades that were not previously studied may reveal new trends of sequence features. Although the small number of genes with isotype uncertainty may be resulted from special scenarios demonstrated with examples above, unexpected features in some poorly represented clades may cause inaccurate classification with the current models which were mostly trained with tRNA genes in well-studied clades. This issue could possibly be addressed with new sets of clade-specific models developed with additional analyses.

Previously, three types of “interrupted” but functional tRNA genes have been identified: (1) genes with one or more introns, (2) *trans*-spliced tRNA genes, and (3) circularized permuted tRNA genes (18,34-36,76,81-84). Using tRNAscan-SE 2.0, researchers can now detect archaeal tRNA genes with noncanonical tRNA introns in addition to those with canonical introns. However, tRNAscan-SE was not designed to detect self-splicing group I introns found in cyanobacterial tRNA genes (81,82) due to the computational demands of aligning very large RNA structures to covariance models. During our analysis, we noticed that group I introns also exist in tRNA genes of other bacterial clades such as Proteobacteria, causing misannotation or failure to identify those genes. In addition to the Rfam model (7), a new set of covariance models has recently been built based on the group I introns identified in archaea (85). The need for detecting these missing genes and improvements to covariance model search software have motivated new work to the detect tRNA genes with group I introns in the next major software release. We also plan to add a process to detect the *trans*-spliced and permuted tRNA genes, but the rareness of these special classes make it a lower priority among other features that could have greater biological significance.

Since the initial release of tRNAscan-SE, the tool has been part of standard routines for annotating genomes decoded at national genome centers. While knowledge of basic tRNA function has been long known, the discovery of alternate regulatory functions and tRNA-derived small RNAs

has stimulated new interest in understanding the regulation and processing of this ancient RNA family. Together with experimental analyses, tRNAscan-SE serves as a key tool for expanding the world of tRNA biology.

ACKNOWLEDGEMENT

We would like to thank Aaron Cozen for his valuable feedback during the development of tRNAscan-SE 2.0.

FUNDING

This work was supported by a grant from the National Human Genome Research Institute, National Institutes of Health [R01HG006753 to T.L.].

TABLES

Table 1. Diversity of genomes used as training sets for tRNA covariance model creation. The tRNAs in the genomes were grouped by domains for building the domain-specific covariance models. For tRNAscan-SE 2.0, tRNA sequences from genomes of each domain were further grouped into different isotypes for the generation of isotype-specific models.

Domain	Models in tRNAscan-SE 1.3		Models in tRNAscan-SE 2.0	
	No. of genera	No. of genomes	No. of genera	No. of genomes
Eukaryota	88	115	110	155
Bacteria	23	33	647	4,016
Archaea	13	18	75	182
Total	124	166	838	4,285

Table 2. tRNA predictions and post-filtered high confidence set in eukaryotic genomes with numerous repetitive elements. Top ten genomes with large amount of raw predictions are shown in comparison with human and mouse. High confidence predictions are determined as a result of the three-stage post-filtering process. The values in the table represent the number of gene predictions at each category.

Genome	All tRNA predictions	tRNAscan-SE Predicted pseudogenes	Secondary post-filtered predictions	Tertiary post-filtered predictions	High confidence set
Cat	403,590	392,312	10,552	126	549
Cow	263,431	232,442	29,617	562	593
Sheep	256,819	225,920	29.710	420	534
Armadillo	227,726	154,522	72,576	89	462
Minke whale	212,492	194,977	13,916	2,879	428
Rat	211,167	198,126	12,605	51	364
Panda	187,373	183,250	3,548	143	398
Ferret	182,506	179,919	2,167	40	360
Dolphin	172,909	157,354	12,694	2,236	368
Squirrel	165,970	146,252	19,221	85	349
Human	596	85	91	0	417
Mouse	40,912	36,415	4,087	2	401

FIGURE LEGENDS

Figure 1. Schematic diagram of tRNAscan-SE 2.0 search algorithm. Three pathways were developed for cytosolic tRNA search modes with the addition of the mitochondrial tRNA search mode. The default method employs Infernal 1.1 (19) with newly built covariance models for similarity search while the legacy search remains the same as tRNAscan-SE 1.3 (1) for backward compatibility.

Figure 2. Isotype-specific covariance model score comparison between tRNAs with anticodon CAU. Dots represent individual tRNAs of initiator methionine/N-formylmethionine ($tRNA^{iMet/fMet}$), elongator methionine ($tRNA^{Met}$), and isoleucine decoding AUA codon ($tRNA^{Ile2}$). Each tRNA was scanned with the isotype-specific covariance models of the corresponding domain. The axis of the plots shows the bit scores of the tRNA gene scanned with the isotype-specific covariance models for eukayotes, bacteria, and archaea.

Figure 3. Isotype uncertainty in human tRNA-Leu-CAA-5-1. The primary sequence and the secondary structure comparison of (A) tRNA-Met-CAT-3-1, (B) tRNA-Leu-CAA-5-1, and (C) tRNA-Leu-CAA-1-1 show that tRNA-Leu-CAA-5-1 is more similar to a $tRNA^{Met}$ than a $tRNA^{Leu}$ even though it has an anticodon decoding leucine. The bases highlighted in orange and green represent the differences between the respective sequence and tRNA-Leu-CAA-5-1.

REFERENCES

1. Lowe, T.M. and Eddy, S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res*, **25**, 955-964.

2. Schattner, P., Brooks, A.N. and Lowe, T.M. (2005) The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res*, **33**, W686-689.

3. Lowe, T.M. and Chan, P.P. tRNAscan-SE On-line: Integrating Search and Context for Analysis of Transfer RNA Genes. *(Submitted)*.

4. Chan, P.P. and Lowe, T.M. (2009) GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res*, **37**, D93-97.

5. Chan, P.P. and Lowe, T.M. (2016) GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Res*, **44**, D184-189.

6. Eddy, S.R. and Durbin, R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res*, **22**, 2079-2088.

7. Kalvari, I., Argasinska, J., Quinones-Olvera, N., Nawrocki, E.P., Rivas, E., Eddy, S.R., Bateman, A., Finn, R.D. and Petrov, A.I. (2018) Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res*, **46**, D335-D342.

8. Fichant, G.A. and Burks, C. (1991) Identifying potential tRNA genes in genomic DNA sequences. *J Mol Biol*, **220**, 659-671.

9. Pavesi, A., Conterio, F., Bolchi, A., Dieci, G. and Ottonello, S. (1994) Identification of new eukaryotic tRNA genes in genomic DNA databases by a multistep weight matrix analysis of transcriptional control regions. *Nucleic Acids Res*, **22**, 1247-1256.

10. Steinberg, S., Misch, A. and Sprinzl, M. (1993) Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res*, **21**, 3011-3015.

11. Juhling, F., Morl, M., Hartmann, R.K., Sprinzl, M., Stadler, P.F. and Putz, J. (2009) tRNAdb 2009: compilation of tRNA sequences and tRNA genes. *Nucleic Acids Res*, **37**, D159-162.

12. Laslett, D. and Canback, B. (2004) ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res*, **32**, 11-16.

13. Wyman, S.K., Jansen, R.K. and Boore, J.L. (2004) Automatic annotation of organellar genomes with DOGMA. *Bioinformatics*, **20**, 3252-3255.

14. Laslett, D. and Canback, B. (2008) ARWEN: a program to detect tRNA genes in metazoan mitochondrial nucleotide sequences. *Bioinformatics*, **24**, 172-175.

15. Bernt, M., Donath, A., Juhling, F., Externbrink, F., Florentz, C., Fritzsch, G., Putz, J., Middendorf, M. and Stadler, P.F. (2013) MITOS: improved de novo metazoan mitochondrial genome annotation. *Mol Phylogenet Evol*, **69**, 313-319.

16. Ardell, D.H. and Andersson, S.G. (2006) TFAM detects co-evolution of tRNA identity rules with lateral transfer of histidyl-tRNA synthetase. *Nucleic Acids Res*, **34**, 893-904.

17. Kinouchi, M. and Kurokawa, K. (2006) tRNAfinder: A software system to find all tRNA genes in the DNA sequence based on the cloverleaf secondary structure. *J Computer Aided Chem.*, **7**, 116-126.

18. Sugahara, J., Kikuta, K., Fujishima, K., Yachie, N., Tomita, M. and Kanai, A. (2008) Comprehensive analysis of archaeal tRNA genes reveals rapid increase of tRNA introns in the order thermoproteales. *Mol Biol Evol*, **25**, 2709-2716.

19. Nawrocki, E.P. and Eddy, S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**, 2933-2935.

20. Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Ostell, J., Pruitt, K.D. and Sayers, E.W. (2018) GenBank. *Nucleic Acids Res*, **46**, D41-D47.

21. Casper, J., Zweig, A.S., Villarreal, C., Tyner, C., Speir, M.L., Rosenbloom, K.R., Raney, B.J., Lee, C.M., Lee, B.T., Karolchik, D. *et al.* (2018) The UCSC Genome Browser database: 2018 update. *Nucleic Acids Res*, **46**, D762-D769.

22. Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N. *et al.* (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res*, **40**, D1178-1186.

23. Eddy, S.R. (2008) A probabilistic model of local sequence alignment that simplifies statistical significance estimation. *PLoS computational biology*, **4**, e1000069.

24. Eddy, S.R. (2011) Accelerated Profile HMM Searches. *PLoS computational biology*, **7**, e1002195.

25. Brown, M.P. (2000) Small subunit ribosomal RNA modeling using stochastic context-free grammars. *Proc Int Conf Intell Syst Mol Biol*, **8**, 57-66.

26. de Bruijn, M.H., Schreier, P.H., Eperon, I.C., Barrell, B.G., Chen, E.Y., Armstrong, P.W., Wong, J.F. and Roe, B.A. (1980) A mammalian mitochondrial serine transfer RNA lacking the "dihydrouridine" loop and stem. *Nucleic Acids Res*, **8**, 5213-5222.

27. Helm, M., Brule, H., Friede, D., Giege, R., Putz, D. and Florentz, C. (2000) Search for characteristic structural features of mammalian mitochondrial tRNAs. *RNA*, **6**, 1356-1379.

28. Richly, E. and Leister, D. (2004) NUMTs in sequenced eukaryotic genomes. *Mol Biol Evol*, **21**, 1081-1084.

29. Hazkani-Covo, E., Zeller, R.M. and Martin, W. (2010) Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. *PLoS Genet*, **6**, e1000834.

30. Daniels, G.R. and Deininger, P.L. (1985) Repeat sequence families derived from mammalian tRNA genes. *Nature*, **317**, 819-822.

31. Okada, N. (1991) SINEs: Short interspersed repeated elements of the eukaryotic genome. *Trends in ecology & evolution*, **6**, 358-361.

32. Okada, N., Hamada, M., Ogiwara, I. and Ohshima, K. (1997) SINEs and LINEs share common 3' sequences: a review. *Gene*, **205**, 229-243.

33. Fujishima, K., Sugahara, J., Tomita, M. and Kanai, A. (2010) Large-scale tRNA intron transposition in the archaeal order Thermoproteales represents a novel mechanism of intron gain. *Mol Biol Evol*, **27**, 2233-2243.

34. Randau, L., Pearson, M. and Soll, D. (2005) The complete set of tRNA species in Nanoarchaeum equitans. *FEBS Lett*, **579**, 2945-2947.

35. Fujishima, K., Sugahara, J., Kikuta, K., Hirano, R., Sato, A., Tomita, M. and Kanai, A. (2009) Tri-split tRNA is a transfer RNA made from 3 transcripts that provides insight into the evolution of fragmented tRNAs in archaea. *Proc Natl Acad Sci U S A*, **106**, 2683-2687.

36. Chan, P.P., Cozen, A.E. and Lowe, T.M. (2011) Discovery of permuted and recently split transfer RNAs in Archaea. *Genome Biol*, **12**, R38.

37. Silva, F.J., Belda, E. and Talens, S.E. (2006) Differential annotation of tRNA genes with anticodon CAT in bacterial genomes. *Nucleic Acids Res*, **34**, 6015-6022.

38. Marck, C. and Grosjean, H. (2002) tRNomics: analysis of tRNA genes from 50 genomes of Eukarya, Archaea, and Bacteria reveals anticodon-sparing strategies and domain-specific features. *RNA*, **8**, 1189-1232.

39. Hubert, N., Sturchler, C., Westhof, E., Carbon, P. and Krol, A. (1998) The 9/4 secondary structure of eukaryotic selenocysteine tRNA: more pieces of evidence. *RNA*, **4**, 1029-1033.

40. Mizutani, T. and Goto, C. (2000) Eukaryotic selenocysteine tRNA has the 9/4 secondary structure. *FEBS Lett*, **466**, 359-362.

41. Ioudovitch, A. and Steinberg, S.V. (1999) Structural compensation in an archaeal selenocysteine transfer RNA. *J Mol Biol*, **290**, 365-371.

42. Sherrer, R.L., Ho, J.M. and Soll, D. (2008) Divergence of selenocysteine tRNA recognition by archaeal and eukaryotic O-phosphoseryl-tRNASec kinase. *Nucleic Acids Res*, **36**, 1871-1880.

43. Baron, C., Westhof, E., Bock, A. and Giege, R. (1993) Solution structure of selenocysteine-inserting tRNA(Sec) from Escherichia coli. Comparison with canonical tRNA(Ser). *J Mol Biol*, **231**, 274-292.

44. Itoh, Y., Sekine, S., Suetsugu, S. and Yokoyama, S. (2013) Tertiary structure of bacterial selenocysteine tRNA. *Nucleic Acids Res*, **41**, 6729-6738.

45. Hatfield, D.L., Lee, B.J., Price, N.M. and Stadtman, T.C. (1991) Selenocysteyl-tRNA occurs in the diatom Thalassiosira and in the ciliate Tetrahymena. *Mol Microbiol*, **5**, 1183-1186.

46. Novoselov, S.V., Rao, M., Onoshko, N.V., Zhi, H., Kryukov, G.V., Xiang, Y., Weeks, D.P., Hatfield, D.L. and Gladyshev, V.N. (2002) Selenoproteins and selenocysteine insertion system in the model plant cell system, Chlamydomonas reinhardtii. *EMBO J*, **21**, 3681-3693.

47. Obata, T. and Shiraiwa, Y. (2005) A novel eukaryotic selenoprotein in the haptophyte alga Emiliania huxleyi. *J Biol Chem*, **280**, 18462-18468.

48. Cassago, A., Rodrigues, E.M., Prieto, E.L., Gaston, K.W., Alfonzo, J.D., Iribar, M.P., Berry, M.J., Cruz, A.K. and Thiemann, O.H. (2006) Identification of Leishmania selenoproteins and SECIS element. *Mol Biochem Parasitol*, **149**, 128-134.

49. Eisen, J.A., Coyne, R.S., Wu, M., Wu, D., Thiagarajan, M., Wortman, J.R., Badger, J.H., Ren, Q., Amedeo, P., Jones, K.M. *et al.* (2006) Macronuclear genome sequence of the ciliate Tetrahymena thermophila, a model eukaryote. *PLoS Biol*, **4**, e286.

50. Lobanov, A.V., Delgado, C., Rahlfs, S., Novoselov, S.V., Kryukov, G.V., Gromer, S., Hatfield, D.L., Becker, K. and Gladyshev, V.N. (2006) The Plasmodium selenoproteome. *Nucleic Acids Res*, **34**, 496-505.

51. Lobanov, A.V., Fomenko, D.E., Zhang, Y., Sengupta, A., Hatfield, D.L. and Gladyshev, V.N. (2007) Evolutionary dynamics of eukaryotic selenoproteomes: large selenoproteomes may associate with aquatic life and small with terrestrial life. *Genome Biol*, **8**, R198.

52. Jiang, L., Ni, J. and Liu, Q. (2012) Evolution of selenoproteins in the metazoan. *BMC Genomics*, **13**, 446.

53. Mariotti, M., Ridge, P.G., Zhang, Y., Lobanov, A.V., Pringle, T.H., Guigo, R., Hatfield, D.L. and Gladyshev, V.N. (2012) Composition and evolution of the vertebrate and mammalian selenoproteomes. *PLoS One*, **7**, e33066.

54. Gobler, C.J., Lobanov, A.V., Tang, Y.Z., Turanov, A.A., Zhang, Y., Doblin, M., Taylor, G.T., Sanudo-Wilhelmy, S.A., Grigoriev, I.V. and Gladyshev, V.N. (2013) The central role of selenium in the biochemistry and ecology of the harmful pelagophyte, Aureococcus anophagefferens. *ISME J*, **7**, 1333-1343.

55. da Silva, M.T., Caldas, V.E., Costa, F.C., Silvestre, D.A. and Thiemann, O.H. (2013) Selenocysteine biosynthesis and insertion machinery in Naegleria gruberi. *Mol Biochem Parasitol*, **188**, 87-90.

56. Eichinger, L., Pachebat, J.A., Glockner, G., Rajandream, M.A., Sucgang, R., Berriman, M., Song, J., Olsen, R., Szafranski, K., Xu, Q. *et al.* (2005) The genome of the social amoeba Dictyostelium discoideum. *Nature*, **435**, 43-57.

57. Zhang, Y., Romero, H., Salinas, G. and Gladyshev, V.N. (2006) Dynamic evolution of selenocysteine utilization in bacteria: a balance between selenoprotein loss and evolution of selenocysteine from redox active cysteine residues. *Genome Biol*, **7**, R94.

58. Bult, C.J., White, O., Olsen, G.J., Zhou, L., Fleischmann, R.D., Sutton, G.G., Blake, J.A., FitzGerald, L.M., Clayton, R.A., Gocayne, J.D. *et al.* (1996) Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii. *Science*, **273**, 1058-1073.

59. Slesarev, A.I., Mezhevaya, K.V., Makarova, K.S., Polushin, N.N., Shcherbinina, O.V., Shakhova, V.V., Belova, G.I., Aravind, L., Natale, D.A., Rogozin, I.B. *et al.* (2002) The complete genome of hyperthermophile Methanopyrus kandleri AV19 and monophyly of archaeal methanogens. *Proc Natl Acad Sci U S A*, **99**, 4644-4649.

60. Kendall, M.M., Liu, Y., Sieprawska-Lupa, M., Stetter, K.O., Whitman, W.B. and Boone, D.R. (2006) Methanococcus aeolicus sp. nov., a mesophilic, methanogenic archaeon from shallow and deep marine sediments. *Int J Syst Evol Microbiol*, **56**, 1525-1529.

61. Hendrickson, E.L., Kaul, R., Zhou, Y., Bovee, D., Chapman, P., Chung, J., Conway de Macario, E., Dodsworth, J.A., Gillett, W., Graham, D.E. *et al.* (2004) Complete genome sequence of the genetically tractable hydrogenotrophic methanogen Methanococcus maripaludis. *J Bacteriol*, **186**, 6956-6969.

62. Salinas-Giege, T., Giege, R. and Giege, P. (2015) tRNA biology in mitochondria. *Int J Mol Sci*, **16**, 4518-4559.

63. Grosjean, H., de Crecy-Lagard, V. and Marck, C. (2010) Deciphering synonymous codons in the three domains of life: co-evolution with specific tRNA modification enzymes. *FEBS Lett*, **584**, 252-264.

64. Cozen, A.E., Quartley, E., Holmes, A.D., Hrabeta-Robinson, E., Phizicky, E.M. and Lowe, T.M. (2015) ARM-seq: AlkB-facilitated RNA methylation sequencing reveals a complex landscape of modified tRNA fragments. *Nat Methods*, **12**, 879-884.

65. Leung, Y.Y., Kuksa, P.P., Amlie-Wolf, A., Valladares, O., Ungar, L.H., Kannan, S., Gregory, B.D. and Wang, L.S. (2016) DASHR: database of small human noncoding RNAs. *Nucleic Acids Res*, **44**, D216-222.

66. Giege, R., Sissler, M. and Florentz, C. (1998) Universal rules and idiosyncratic features in tRNA identity. *Nucleic Acids Res*, **26**, 5017-5035.

67. Perry, J., Dai, X. and Zhao, Y. (2005) A mutation in the anticodon of a single tRNAala is sufficient to confer auxin resistance in Arabidopsis. *Plant Physiol*, **139**, 1284-1290.

68. Kimata, Y. and Yanagida, M. (2004) Suppression of a mitotic mutant by tRNA-Ala anticodon mutations that produce a dominant defect in late mitosis. *J Cell Sci*, **117**, 2283-2293.

69. The 1000 Genomes Project Consortium. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56-65.

70. The 1000 Genomes Project Consortium. (2015) A global reference for human genetic variation. *Nature*, **526**, 68-74.

71. NCBI Resource Coordinators. (2015) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, **43**, D6-17.

72. Leahy, S.C., Kelly, W.J., Altermann, E., Ronimus, R.S., Yeoman, C.J., Pacheco, D.M., Li, D., Kong, Z., McTavish, S., Sang, C. *et al.* (2010) The genome sequence of the rumen methanogen Methanobrevibacter ruminantium reveals new possibilities for controlling ruminant methane emissions. *PLoS One*, **5**, e8926.

73. Koonin, E.V., Makarova, K.S. and Aravind, L. (2001) Horizontal gene transfer in prokaryotes: quantification and classification. *Annu Rev Microbiol*, **55**, 709-742.

74. Polz, M.F., Alm, E.J. and Hanage, W.P. (2013) Horizontal gene transfer and the evolution of bacterial and archaeal population structure. *Trends Genet*, **29**, 170-175.

75. Marck, C. and Grosjean, H. (2003) Identification of BHB splicing motifs in intron-containing tRNAs from 18 archaea: evolutionary implications. *RNA*, **9**, 1516-1531.

76. Sugahara, J., Fujishima, K., Morita, K., Tomita, M. and Kanai, A. (2009) Disrupted tRNA Gene Diversity and Possible Evolutionary Scenarios. *J Mol Evol*, **69**, 497-504.

77. Gibbs, R.A., Weinstock, G.M., Metzker, M.L., Muzny, D.M., Sodergren, E.J., Scherer, S., Scott, G., Steffen, D., Worley, K.C., Burch, P.E. *et al.* (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*, **428**, 493-521.

78. Borodulina, O.R. and Kramerov, D.A. (1999) Wide distribution of short interspersed elements among eukaryotic genomes. *FEBS Lett*, **457**, 409-413.

79. Bao, W., Kojima, K.K. and Kohany, O. (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA*, **6**, 11.

80. Nishihara, H., Plazzi, F., Passamonti, M. and Okada, N. (2016) MetaSINEs: Broad Distribution of a Novel SINE Superfamily in Animals. *Genome Biol Evol*, **8**, 528-539.

81. Paquin, B., Kathe, S.D., Nierzwicki-Bauer, S.A. and Shub, D.A. (1997) Origin and evolution of group I introns in cyanobacterial tRNA genes. *J Bacteriol*, **179**, 6798-6806.

82. Biniszkiewicz, D., Cesnaviciene, E. and Shub, D.A. (1994) Self-splicing group I intron in cyanobacterial initiator methionine tRNA: evidence for lateral transfer of introns in bacteria. *EMBO J*, **13**, 4629-4635.

83. Soma, A., Onodera, A., Sugahara, J., Kanai, A., Yachie, N., Tomita, M., Kawamura, F. and Sekine, Y. (2007) Permuted tRNA genes expressed via a circular RNA intermediate in Cyanidioschyzon merolae. *Science*, **318**, 450-453.

84. Maruyama, S., Sugahara, J., Kanai, A. and Nozaki, H. (2010) Permuted tRNA genes in the nuclear and nucleomorph genomes of photosynthetic eukaryotes. *Mol Biol Evol*, **27**, 1070-1076.

85. Nawrocki, E.P., Jones, T.A. and Eddy, S.R. (2018) Group I introns are widespread in archaea. *Nucleic Acids Res*, **46**, 7970-7976.

86. Yarham, J.W., Elson, J.L., Blakely, E.L., McFarland, R. and Taylor, R.W. (2010) Mitochondrial tRNA mutations and disease. *Wiley interdisciplinary reviews. RNA*, **1**, 304-324.

87. Abbott, J.A., Francklyn, C.S. and Robey-Bond, S.M. (2014) Transfer RNA and human disease. *Front Genet*, **5**, 158.

88. Simone, D., Calabrese, F.M., Lang, M., Gasparre, G. and Attimonelli, M. (2011) The reference human nuclear mitochondrial sequences compilation validated and implemented on the UCSC genome browser. *BMC Genomics*, **12**, 517.

Figure 1

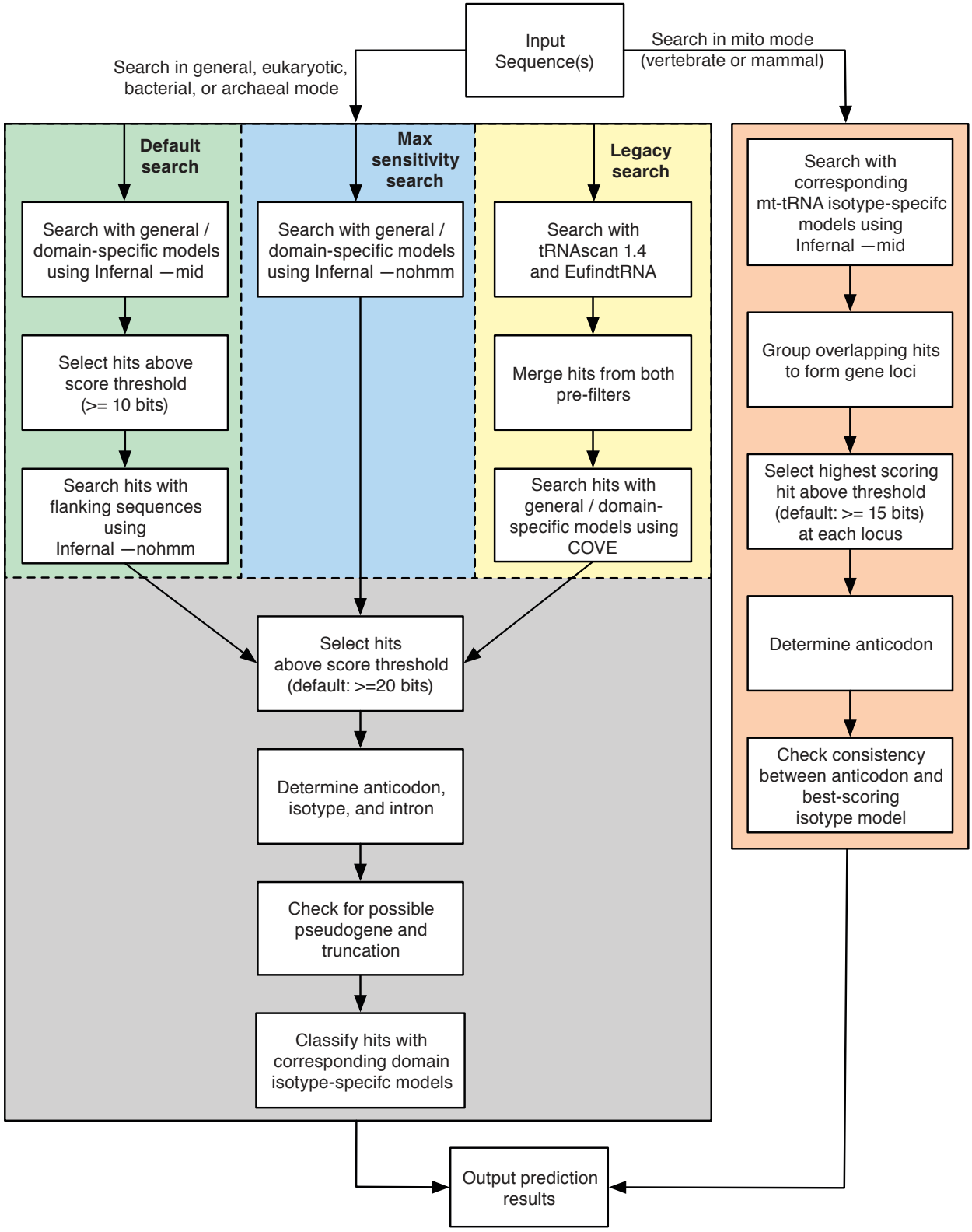
Input Sequence(s)
Search in general, eukaryotic, bacterial, or archaeal mode
Search in mito mode (vertebrate or mammal)
Default search
Max sensitivity search
Legacy search
Search with general / domain-specific models using Infernal —mid
Select hits above score threshold (>= 10 bits)
Search hits with flanking sequences using Infernal —nohmm
Search with general / domain-specific models using Infernal —nohmm
Search with tRNAscan 1.4 and EufindtRNA
Merge hits from both pre-filters
Search hits with general / domain-specific models using COVE
Search with corresponding mt-tRNA isotype-specifc models using Infernal —mid
Group overlapping hits to form gene loci
Select highest scoring hit above threshold (default: >= 15 bits) at each locus
Determine anticodon
Check consistency between anticodon and best-scoring isotype model
Select hits above score threshold (default: >=20 bits)
Determine anticodon, isotype, and intron
Check for possible pseudogene and truncation
Classify hits with corresponding domain isotype-specifc models
Output prediction results

Figure 2

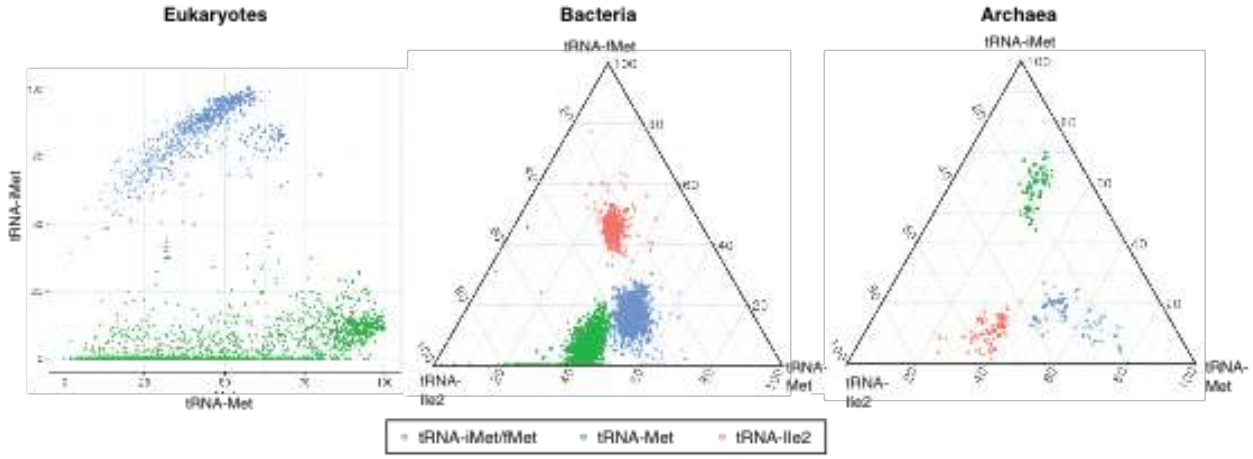

Figure 3