Method

Predicting transfer RNA gene activity from sequence and genome context

Bryan P. Thornlow,[1] Joel Armstrong,[1,2] Andrew D. Holmes,[1] Jonathan M. Howard,[1] Russell B. Corbett-Detig,[1,2] and Todd M. Lowe[1,2]

[1]*Department of Biomolecular Engineering, University of California, Santa Cruz, California 95064, USA;* [2]*Genomics Institute, University of California, Santa Cruz, California 95064, USA*

Transfer RNA (tRNA) genes are among the most highly transcribed genes in the genome owing to their central role in protein synthesis. However, there is evidence for a broad range of gene expression across tRNA loci. This complexity, combined with difficulty in measuring transcript abundance and high sequence identity across transcripts, has severely limited our collective understanding of tRNA gene expression regulation and evolution. We establish sequence-based correlates to tRNA gene expression and develop a tRNA gene classification method that does not require, but benefits from, comparative genomic information and achieves accuracy comparable to molecular assays. We observe that guanine + cytosine (G + C) content and CpG density surrounding tRNA loci is exceptionally well correlated with tRNA gene activity, supporting a prominent regulatory role of the local genomic context in combination with internal sequence features. We use our tRNA gene activity predictions in conjunction with a comprehensive tRNA gene ortholog set spanning 29 placental mammals to estimate the evolutionary rate of functional changes among orthologs. Our method adds a new dimension to large-scale tRNA functional prediction and will help prioritize characterization of functional tRNA variants. Its simplicity and robustness should enable development of similar approaches for other clades, as well as exploration of functional diversification of members of large gene families.

[Supplemental material is available for this article.]

Transfer RNAs (tRNAs) are essential for the translation of messenger RNA (mRNA) into proteins for all life. At the gene level in eukaryotes, they are of special interest for their high copy number, strong nucleotide sequence conservation, and variation in expression (Kutter et al. 2011; Schmitt et al. 2014; Pan 2018). tRNA molecules are required in large abundance to meet the dynamic metabolic needs of cells, and tRNA genes are believed to be among the most highly transcribed genes in the genome (Palazzo and Lee 2015; Boivin et al. 2018).

Despite high cellular demands, numerous individual tRNA genes have no direct evidence for expression (Kutter et al. 2011; Palazzo and Lee 2015; Gogakos et al. 2017; Hummel et al. 2019). High duplication rates and consequent weakened purifying selection may lead to an abundance of pseudogenes. Additionally, many of these genes may be tRNA-derived short interspersed nuclear elements (SINEs), which often retain strong promoter elements. However, even after removal of apparent pseudogenes and SINEs, more than 60 human tRNA genes and more than 100 mouse tRNA genes are in constitutively silenced regions of the genome for all tissues and cell lines, suggesting they are never or rarely transcribed (Roadmap Epigenomics Consortium et al. 2015; Bogu et al. 2016; Holmes 2018; Thornlow et al. 2018). Chromatin immunoprecipitation sequencing (ChIP-seq) data support this conclusion, as one multispecies study detected occupancy by RNA Polymerase III (Pol III) for only 224 of 417 high-confidence tRNA genes in human liver, with other mammals showing similar patterns (Kutter et al. 2011).

tRNA gene expression may coevolve with phenotypic differences between species. Data from previous studies suggest that the rate of evolution of protein-coding gene expression levels differs by clade (Li et al. 1996; Brawand et al. 2011; Necsulea and Kaessmann 2014). The rate of evolution of gene expression also varies among noncoding RNA gene families (Meunier et al. 2013; Necsulea and Kaessmann 2014; Necsulea et al. 2014). Because of difficulties in high-throughput, accurate quantification of tRNA abundance, the complexity of tRNA gene expression across mammals is not well understood. The expanding functional repertoire of tRNA transcripts and tRNA-derived small RNAs (Mleczko et al. 2014; Goodarzi et al. 2015; Kirchner and Ignatova 2015; Sun et al. 2018) indicates that changes in tRNA gene expression between species could have profound cellular effects.

Expression of tRNA genes has clear importance for organismal development and contribution to disease, but our understanding of its regulation and evolution is severely lacking for several reasons (Hanada et al. 2013; Schaffer et al. 2014; Yoo et al. 2016). Measuring expression of unique mRNA transcripts has become relatively straightforward. However, tRNA sequencing by the methods originally developed for unmodified small RNAs (e.g., microRNAs) is frequently impeded by numerous RNA modifications at the reverse transcription phase. Only very recently have specialized sequencing library preparation methods been developed to remove or overcome these modifications, enabling effective sequencing (Cozen et al. 2015; Zheng et al. 2015). Furthermore, because fully processed tRNA gene transcripts from

Corresponding authors: lowe@soe.ucsc.edu; rucorbet@ucsc.edu
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.256164.119.

different loci are often identical, simple tRNA-seq abundance measurements are often insufficient to determine the true transcriptional activity at each gene locus. Therefore, to determine which tRNA genes are potentially constitutively active, highly regulated, or silenced, other methods are needed. Several genome-wide methods examine tRNA loci in their generally unique genomic contexts, bypassing the problem of identical mature tRNA transcripts. Such assays include chromatin immunoprecipitation (ChIP) (Roadmap Epigenomics Consortium et al. 2015; Bogu et al. 2016; Thornlow et al. 2018), RNA Polymerase III (Pol III) ChIP-seq (Kutter et al. 2011), and ATAC-seq (Foissac et al. 2019), among others. These high-throughput assays remain cost and resource intensive, so currently available data are often limited to few species and tissues. Nonetheless, these data show that identical tRNA genes do vary in expression profiles (Kutter et al. 2011; Schmitt et al. 2014; Pan 2018), supporting the need to incorporate extrinsic factors into the prediction of when or if tRNA genes are active. The study of the local genomic context is therefore essential and has not been tackled comprehensively by any tRNA gene prediction method.

Here, we begin to resolve these concerns by developing a model to predict whether individual tRNA genes are actively transcribed in at least one tissue, or transcriptionally silent. Previous work has shown that tRNA gene transcription may be inferred based on DNA variation driven by transcription-associated mutagenesis (Thornlow et al. 2018). We leverage this correlation, further enhanced by other genomic features, to infer expression of tRNA genes with high accuracy. This novel advance in tRNA research uses, but does not require, comparative genomic information, enabling its broad application. We show our method using 29 placental mammalian genomes, most of which have no tRNA expression data. We also developed a robust mapping of syntenic tRNA genes across all 29 species. By combining our new method with this comprehensive ortholog set, we have analyzed and compared expression classifications of more than 10,000 tRNA genes, yielding a first look at the rate of tRNA gene regulation evolution in placental mammals, as well as bringing attention to the high frequency of silenced "high-scoring" canonical tRNA genes.

Results

Our goal was to develop a tRNA activity–predictive model that could be applied to as many species as possible. To date, the most facile method for inferring tRNA gene function has been the use of tRNAscan-SE covariance model bit scores, which quantify similarity to primary sequence and secondary structure profiles derived from an alignment of reference tRNAs (Lowe and Eddy 1997; Chan et al. 2019). However, comparison to RNA Polymerase III ChIP-seq data from multiple mouse tissues (Kutter et al. 2011) suggests that high covariance model bit scores do not always correspond to occupancy by RNA Polymerase III (Pol III) (Supplemental Fig. S1). More generally, this is consistent with the idea that tRNAscan-SE bit scores alone are not strongly predictive of gene expression.

To improve prediction of tRNA functional roles and better understand the basis of tRNA gene regulation in mammals, we evaluated many additional sequence features easily obtained from a single reference genome (Fig. 1). We explored genomic features correlated with activity based on comprehensive epigenomic data across 127 human tissues and cell lines (Roadmap Epigenomics Consortium et al. 2015) and then reduced this set to just those yielding the best predictions for our training data (Table 1; Supplemental Table S1).

To create our predictive model, we evaluated and incorporated two types of function-predictive statistics: intrinsic features related to tRNA gene sequence, and extrinsic features derived entirely from the genomic context. First, we reasoned that highly expressed tRNA genes should generally encode strong internal promoter sequences, and their transcripts must fold stably into the canonical tRNA structure. Both types of information are incorporated into tRNAscan-SE bit scores (Chan et al. 2019). Furthermore, our previous study found that tRNA gene conservation is highest for actively transcribed tRNA genes, presumably because of stronger purifying selection on required sequence features (Thornlow et al. 2018). Thus, we included tRNA gene conservation in the form of the phyloP score, a nucleotide-level quantitative measure of conservation using multiple alignments (Pollard et al. 2010). We also assessed the correlation of gene activity with the length of each pre-tRNA's 3′ tail, measured by the nucleotide distance from the end of the mature tRNA gene to the beginning of the poly(T) transcription termination sequence (Koski et al. 1980; Allison and Hall 1985). Multiple studies on tRNA transcription termination (Maraia et al. 1994; Hamada et al. 2000; Orioli et al. 2011; Arimbasseri et al. 2013) observed that the RNase Z-trimmed 3′ sequences vary in overall length, composition, and terminator strength [poly(T) length], each potentially affecting tRNA maturation and processing.

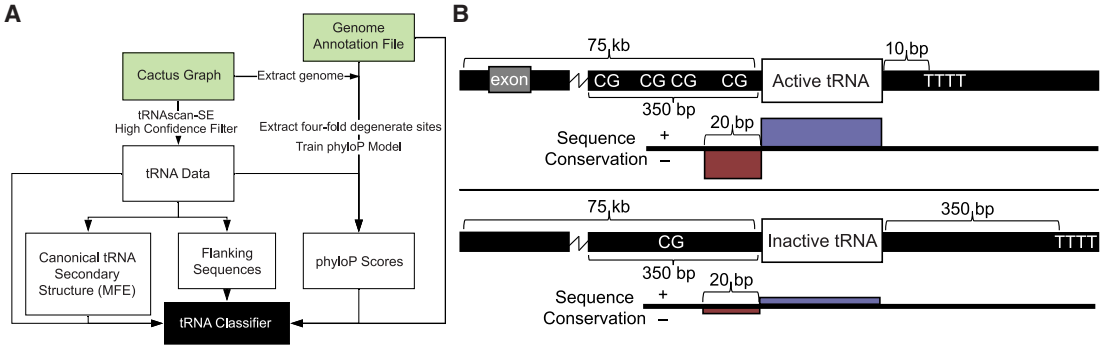

Figure 1. Schematic of tRNA activity classifier and key features used in prediction. (*A*) Flowchart of analysis pipeline, which extracts tRNA information solely from genomic data and classifies tRNA genes as active or inactive. Green blocks indicate files not created by the pipeline. By default, the method uses a Cactus graph (Armstrong et al. 2019), which is a reference-free whole-genome alignment, and a genome annotation file as input. (*B*) Active tRNA genes generally have more CpG dinucleotides in their 350-bp upstream flanking regions, more proximal transcription termination sequences ("TTTT"), are within 75 kb of more exons, and have more highly conserved gene sequences and more evolutionarily divergent 20-nt 5′ flanking regions.

Table 1. Both intrinsic (tRNA-specific) and extrinsic (genome context) features are integral to the model

Feature group	Feature name Activity Probability	Feature importance — —	Active mean (95% CI) — —	Inactive mean (95% CI) — —	*TRE-CTC1-1* Active +0.994	*TRK-CTT11-1* Inactive −0.975	*TRQ-TTG3-1* Active +0.984	*TRQ-TTG4-1* Inactive −0.895
Intrinsic	Total number of tRNA genes with identical anticodon	0.089	11.1 (10.4, 11.8)	17.9 (15.3, 20.2)	14	15	6	6
	Minimum free energy of canonical tRNA secondary structure	0.074	−27.4 (−27.8, −27.0)	−23.5 (−24.6, −22.5)	−26.4	−16.0	−22.0	−24.1
	tRNAscan-SE general bit score	0.070	76.2 (75.3, 77.3)	69.9 (67.3, 72.6)	73.2	56.6	66.9	58.1
	Average phyloP score in tRNA sequence	0.063	0.86 (0.79, 0.92)	0.35 (0.25, 0.46)	1.37	−0.081	0.51	−0.047
	Distance to nearest TTTT transcription termination sequence	0.040	15.9 (13.1, 19.2)	66.9 (43.5, 94.7)	8	351	7	14
Extrinsic	CpG density across tRNA locus	0.303	0.043 (0.041, 0.045)	0.018 (0.014, 0.023)	0.045	0.014	0.024	0.014
	Observed/expected CpG islands score upstream of tRNA gene	0.225	0.67 (0.64, 0.69)	0.27 (0.21, 0.33)	0.68	0.092	0.62	0.11
	Average phyloP score in 5′ flanking region	0.092	−2.61 (−2.74, −2.45)	−1.21 (−1.44, −0.99)	−3.88	0.17	−2.50	0.009
	tRNA genes within 10 kb	0.023	1.97 (1.75, 2.18)	0.84 (0.54, 1.19)	3	0	5	0
	Exons within 75 kb	0.019	35.3 (32.2, 38.7)	21.6 (17.2, 26.7)	88	43	37	0

All features included in the model with their relative importance values as measured by decrease in node impurity by *scikit-learn* (Pedregosa et al. 2011; Methods). Greater feature importance scores indicate greater contribution to discrimination between active and inactive tRNA genes by the model. The active mean and inactive mean columns refer to the mean value across all human tRNA genes in our training set that are known to be active and inactive, respectively, with 95% confidence intervals (CI) in parentheses, calculated for each mean using bootstrapping. Minimum free energy of canonical tRNA secondary structure refers to the minimum free energy when constrained to folding into the canonical cloverleaf secondary structure (Lorenz et al. 2011). For calculating CpG-related statistics, we consider the tRNA locus to begin 350 bp upstream and end 350 bp downstream from the gene. To calculate the phyloP score in the 5′ flanking region, we considered only the 20 bp immediately upstream of each tRNA gene. As examples, the human tRNA genes predicted most likely active (*TRE-CTC1-1*; GtRNAdb ID: Glu-CTC-1-1) and most likely inactive (*TRK-CTT11-1*; GtRNAdb ID: Lys-CTT-11-1), across all human tRNAs, are shown, as well as two examples from the same anticodon family (tRNA-Gln-TTG), one active and one inactive. For the distance to the nearest transcription termination sequence, if the motif "TTTT" was not found within 350 nt of a tRNA gene, 351 was used as its value, as is the case for *TRK-CTT11-1*.

We found that tRNAscan-SE bit scores and average phyloP scores across tRNA gene sequences are significantly correlated with tRNA gene activity based on epigenomic data. We also found that the total number of tRNA genes with identical anticodons and the distance to transcription termination sites are significantly anti-correlated with activity, as higher anticodon redundancy in tRNA genes and tRNA genes with more distal transcription termination sites are more frequently inactive (Spearman's rank correlation, $P < 1 \times 10^{-4}$ for all comparisons) (Roadmap Epigenomics Consortium et al. 2015; Thornlow et al. 2018).

Second, because mRNA expression depends heavily on local chromatin context, we explored features of the genomic environment. Protein-coding genes in regions rich in CG, or CpG, dinucleotides are known to be more frequently expressed (Gardiner-Garden and Frommer 1987; Krinner et al. 2014). Gardiner-Garden and Frommer define CpG islands scores as the observed frequency of CpG dinucleotides compared to their expected frequency given the G + C content of a region. We found that these scores, when calculated for the 350 bases upstream of each gene, are significantly correlated with active tRNA genes (Spearman's rank correlation, $P < 2.1 \times 10^{-24}$). Similarly, the frequency of CpG dinucleotides spanning from 350 bases upstream to 350 bases downstream from each tRNA gene is even more significantly correlated with expression ($P < 1.9 \times 10^{-27}$) (Roadmap Epigenomics Consortium et al. 2015; Thornlow et al. 2018).

We also previously found that the putatively neutral regions flanking highly expressed tRNA genes are more divergent, consistent with transcription-associated mutagenesis (Thornlow et al. 2018). We observed that the average phyloP score of the 20-nt 5′ flanking regions of tRNAs is significantly anti-correlated with tRNA gene activity, because active tRNA genes more often have highly divergent flanking regions ($P < 8.9 \times 10^{-16}$) (Roadmap Epigenomics Consortium et al. 2015; Thornlow et al. 2018). Finally, based on an expectation for increased chromatin accessibility for tRNA genes near other genes, we found that tRNA genes are indeed more likely to be in an active chromatin state if near protein-coding genes ($P < 8.9 \times 10^{-5}$) or other tRNA genes ($P < 9.7 \times 10^{-7}$).

We hypothesized that some combination of both intrinsic and extrinsic features could enable robust computational inference of potential for tRNA gene activity (Fig. 1). To develop an integrated model, we tested several common frameworks, including random forest (RF), logistic regression, and support vector machines. The RF classifier was most effective, achieving the greatest area under the Receiver Operating Characteristic curve (AUC) (Fig. 2A,B) based on 10-fold cross-validation of human tRNA gene data and subsequent application, without retraining, to mouse tRNA gene data (Methods). For reference, we have included information for all data sets used for testing, training, and validation of our classifier in Supplemental Table S2.

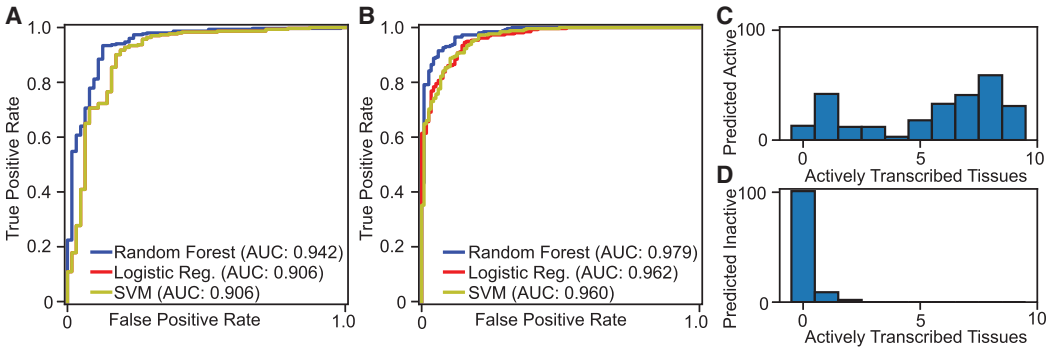

Figure 2. Random forest classifier achieves 94% accuracy on mouse tRNA genes. Receiver operating characteristic curves for random forest (blue), logistic regression (red), and support vector machine (yellow) upon application to human training data with 10-fold cross-validation (*A*) and mouse test data (*B*) are shown. The number of mouse tRNA genes predicted as active (*C*) and inactive (*D*) are compared to the number of tissues in which they are actively transcribed according to Bogu et al. 2016. We considered a mouse tRNA gene active if it is actively transcribed in at least one tissue.

Features derived from CpG islands are most informative

To better understand and improve our classifier, we determined the relative importance of each feature in our random forest model (Table 1; Pedregosa et al. 2011). All features contribute to model accuracy and are significantly correlated with the activity labels (Spearman's rank correlation, $P < 1 \times 10^{-4}$ for all features). Among high-confidence tRNA genes, as determined by tRNAscan-SE (Chan et al. 2019), the most informative features predictive of activity (feature importance) (Table 1; Pedregosa et al. 2011), are derived from CpG content at each tRNA locus. On comparing our CpG data to epigenomic data for each mouse tRNA gene (Bogu et al. 2016), we find that both CpG density and CpG islands scores are exceptionally highly correlated with breadth of activity (Spearman's rank, $P < 2.4 \times 10^{-78}$ for CpG Density, $P < 5.0 \times 10^{-61}$ for CpG Islands Score) (Supplemental Fig. S2). This supports the idea that CpG-derived genomic data are particularly highly informative of tRNA gene activity.

One might expect that the tRNAscan-SE general bit score should be the most informative single feature. However, we start with relatively high-quality tRNAs with likely pseudogenes already removed using the tRNAscan-SE bit score-based high-confidence filter (Chan et al. 2019). Thus, for a starting set of tRNAs already vetted for reasonably strong features, the contribution of the tRNAscan-SE bit score to the model is marginally smaller than other features not previously used to estimate gene function. By incorporating both tRNA gene sequence and genome context (Supplemental Fig. S3), our classifier represents a substantial improvement over using tRNAscan-SE covariance bit scores alone (Supplemental Fig. S1B).

Our classifier is 94% accurate in classifying mouse tRNA genes based on epigenomic data

Because our classifier was trained using comprehensive epigenomic data mined from human tRNA gene loci (Supplemental Table S3; Roadmap Epigenomics Consortium et al. 2015), we required an independent data set to test our model. Therefore, we tested the accuracy of our classifier using epigenomic data evaluating histone marks at mouse tRNA genes across nine tissues from Bogu et al. (2016) (Supplemental Tables S2, S4). Our mouse tRNA gene set contains 376 genes, with 259 observed as active and 117 believed silent based on epigenomic data (Supplemental Table S4; Methods). Our classifier predicted that 264 of these genes are active and 112 are inactive, correctly categorizing 353 tRNA genes and achieving 93.9% accuracy (Fig. 2B–D). Of the 23 misclassified mouse tRNA genes, 14 are misclassified as active and nine are misclassified as inactive. We note that these genes are not biased by isotype, nor by genomic location, and are therefore most likely misclassified for a variety of reasons (Supplemental Table S5). To ascertain the performance of our classifier on nonconserved tRNA loci between human and mouse, we also tested the classifier on only the 184 mouse tRNA genes in our test set without syntenic human orthologs. We correctly classify 167 such genes, achieving 90.8% accuracy in this highly biased subset of tRNA genes.

Classification without alignment or annotation is similarly accurate

We developed our method such that it could potentially be applied to any species with a sequenced genome. For best performance, we used a Cactus graph (Paten et al. 2011a,b; Nguyen et al. 2015; Armstrong et al. 2019), which is a reference-free whole-genome alignment. Usage of a Cactus graph enhances detection of synteny and facilitates extraction of alignments for specific regions in multiple genomes. The Cactus graph used in this study includes 29 mammalian genomes (Supplemental Table S6).

Nonetheless, we recognize that Cactus graphs are not yet available for all species. To accommodate species for which no alignments or protein-coding gene annotations have been developed, we included an option to omit the requirement of this feature information. Use of this simplified classifier led to decreases in accuracy in both human (AUC = 0.927, 91.8% accuracy compared to AUC = 0.942 and 93.2% accuracy in the full model) and mouse (AUC = 0.974, 92.6% accuracy compared to AUC = 0.979 and 93.9% accuracy in the full model), which may be exacerbated upon application to more phylogenetically distant species.

ChIP-seq, DM-tRNA-seq, and ATAC-seq data independently validate our classifications in additional species

To further validate our model, which was trained on human chromatin data, we compared our predictions to RNA Polymerase III (Pol III) ChIP-seq data previously collected from the livers of four species (*Mus musculus*, *Macaca mulatta*, *Rattus norvegicus*, and *Canis lupus familiaris*) (Kutter et al. 2011). Although Pol III ChIP-seq measures Pol III occupancy rather than transcription, it is a requirement for transcription, and our usage of ChIP-seq data instead of transcription data ameliorates the common problem of

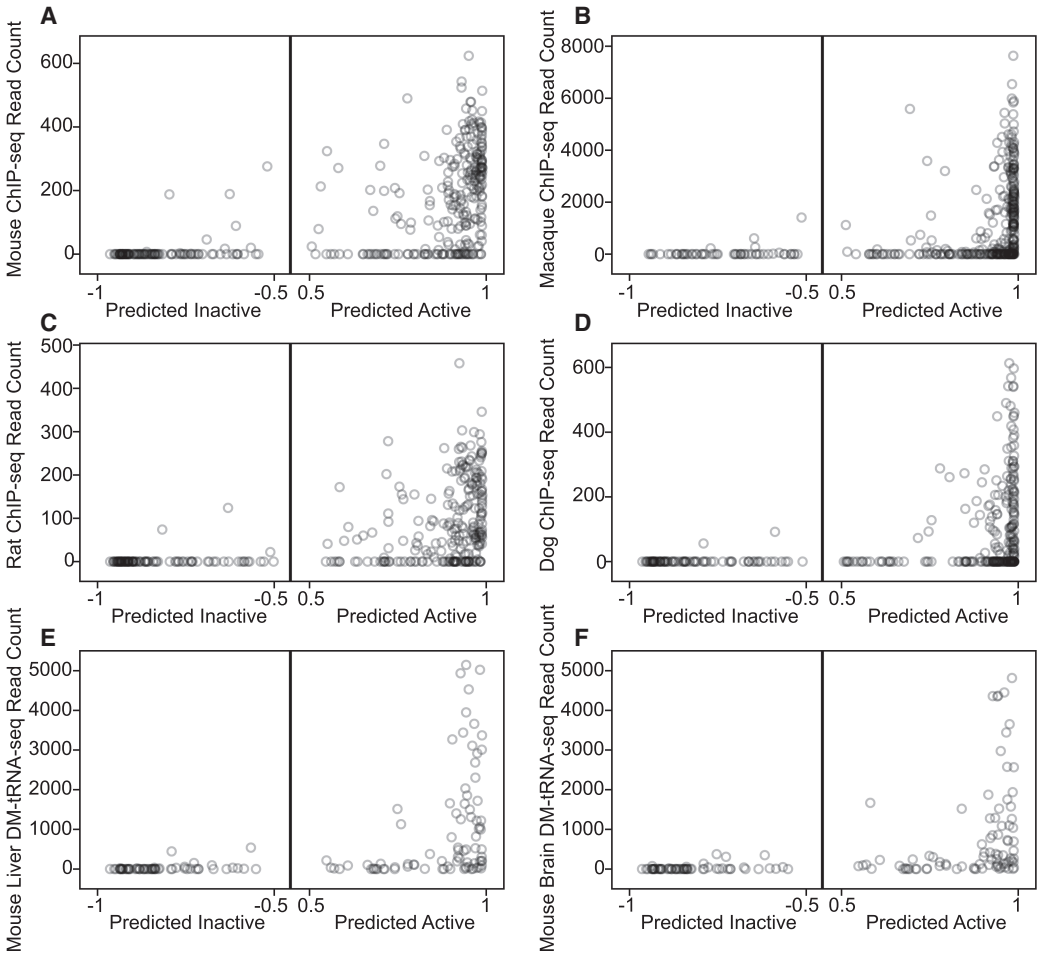

Figure 3. Classification of gene activity based on genomic data achieves similar results to Pol III ChIP-seq analysis in four species and DM-tRNA-seq in two tissues. Probability scores output by the classifier are shown on the *x*-axis, where tRNA genes further *left* are predicted inactive with greater probability, and tRNA genes further *right* are predicted active with greater probability: (*A*) mouse; (*B*) macaque; (*C*) rat; (*D*) dog. The *y*-axis shows Pol III ChIP-seq read counts from the liver of each species for each tRNA gene, from Kutter et al. (2011). Similar patterns are observed for predicted active versus inactive mouse tRNA genes with uniquely mapping DM-tRNA-seq data, comparing to the average normalized read count across three replicates: (*E*) mouse liver; (*F*) mouse brain.

mature tRNA transcripts mapping ambiguously to multiple tRNA loci. We found roughly expected agreement between our classifications and the Pol III ChIP-seq read counts from a single tissue (Fig. 3A–D; Supplemental Fig. S4). Our predictions are similarly accurate when compared to mouse muscle and testes ChIP-seq data (Supplemental Figs. S4A,B, S5).

We predicted many tRNA genes as active despite a lack of Pol III binding at these loci in liver, muscle, and testes. This is a consequence of our methodology, as our model does not predict activity in specific tissues, but is instead trained to predict tRNA genes as active if epigenomic data indicates active transcription in at least one of many tissues (Roadmap Epigenomics Consortium et al. 2015). For example, in mouse, 259 total tRNA genes are active in at least one tissue based on the epigenomic data, but 90 of these (35%) are not expected to be active in the liver based on the same data. Based on human and mouse epigenomic data, a large proportion of tRNA genes are expressed exclusively in stem cells and cell lines (Holmes 2018). This may explain many of the discrepancies we observe in predicting tRNA genes as active that do not have any evidence for Pol III occupancy in one or a small number of differentiated tissues. We predict the brain-specific mouse tRNA gene, *n-TRtct5* (GtRNAdb ID: Arg-TCT-4-1) (Ishimura et al. 2014), which has no ChIP-seq reads in mouse liver, muscle, or testes (Kutter et al. 2011), as active with 0.664 probability (Supplemental Table S4). This is consistent with our goal to predict any tRNA gene with known activity in any tissue as active.

Our model predicted 120 (macaque), 67 (rat), and 142 (dog) tRNA genes as active despite Pol III ChIP-seq read counts of zero in the liver (Kutter et al. 2011). Although ChIP-seq has not been performed on macaque, rat, and dog tRNA loci for any other tissues, we find that virtually all tRNA genes with measured Pol III binding are predicted to be active by our classifier. Among tRNA genes with Pol III ChIP-seq read counts greater than zero, we predicted that 95.3% are active in mouse, 97.3% in macaque, 98.3% in rat, and 98.5% in dog. This consistency in tRNA distributions and classifier behavior across species suggests that the classifier is similarly accurate in mouse, macaque, rat, and dog (Fig. 3; Supplemental Fig. S4).

As additional validation, we compared our predictions to new tRNA transcript abundance data for mouse brain and liver,

collected by our laboratory using DM-tRNA-seq (Supplemental Table S7; Zheng et al. 2015). Compared to other assays, DM-tRNA-seq is a more direct measure of transcript abundance. However, because this sequencing method captures mostly mature tRNA transcripts, we were limited to only the 153 single-copy mouse tRNA loci in our data set, as it is impossible to determine the source loci for transcripts produced by multicopy tRNA genes. We conducted DM-tRNA-seq in mouse liver and brain in three replicates each and compared the average normalized read counts across each tissue to our activity predictions for each single-copy tRNA gene (Fig. 3E,F; Methods). Our DM-tRNA-seq data support the tissue specificity of *n-TRtct5*, as we see moderate expression across all of our brain replicates and detect no expression in any of our liver replicates (Supplemental Table S7). We also found that our DM-tRNA-seq data from mouse liver is significantly correlated with the Pol III ChIP-seq data from mouse liver (Spearman's rank, $P < 3.6 \times 10^{-29}$). When we consider tRNA genes with an average of at least 20 reads in either tissue to be active and all others to be inactive, we achieve 84% accuracy, which may be a low estimate based on the small number of tissues tested.

To validate our predictions in more species, we used ATAC-seq data captured in liver, $CD4^+$, and $CD8^+$ cells for the cow, pig, and goat genomes (Foissac et al. 2019). We compared our predictions to the ATAC-seq peaks across these tissues for the regions spanning 250 bp upstream of and downstream from each tRNA gene (Supplemental Figs. S6, S7). Again, because of the inclusion of only a small subset of tissues in this data, many tRNA loci that do not show activity in these tissues but were predicted as active by our model may be active in other tissues. Among tRNA genes with ATAC-seq peaks, we predicted 90.4%, 95.1%, and 90.8% as active in cow, goat, and pig, respectively. These results are comparable to measurements obtained from ChIP-seq data in mouse, macaque, rat, and dog.

tRNA gene classifications follow similar distributions across the eutherian phylogeny

We applied our model to 29 mammalian species (Fig. 4; Supplemental Tables S6, S8, S9) to glean new insights into the evolution of tRNA complements. We determined the distributions of active and inactive tRNA genes by anticodon across these species (Supplemental Table S10), finding that most species have about 250–350 predicted active genes, comprising ~75% of their tRNA gene sets. We observe similar distributions by clade, with a few exceptions. *Bos taurus*, *Capra hircus*, and *Orcinus orca* (cow, goat, and orca, respectively) have more than 300 tRNA genes predicted to be inactive, but no other species has more than 154. This most likely reflects decreased ability of tRNAscan-SE to discriminate tRNA-derived SINEs from tRNA genes in these species (Chan et al. 2019). Furthermore, we verified that all species had at least one tRNA gene predicted as active for each expected anticodon (Supplemental Tables S10, S11; Grosjean et al. 2010), with only three minor exceptions that may represent genome assembly or classification errors (Supplemental Material).

Establishing mammalian ortholog sets enables further evolutionary analysis of tRNA gene regulation

To investigate the relationship between evolutionary conservation and transcriptional activity, we developed a complete set of placental mammal tRNA gene orthologs using a Cactus graph (Supplemental Tables S6, S12; Armstrong et al. 2019). Cactus graphs are state-of-the-art alignments that allow greater detection of synteny across many species. Of the 11,724 tRNA genes in our 29-species alignment, 3554 genes in total, or about 123 per species on average, appear to be species-specific. The rest were grouped into 1097 ortholog sets. Of these, 750 ortholog groups contain only tRNA genes predicted to be active, approximately mirroring the distribution of active to inactive tRNA genes predicted at the species level (Fig. 4B). On average, each of our 1097 ortholog sets spans 7.4 species, indicating that tRNA genes are generally either fairly deeply conserved or recently evolved (Supplemental Fig. S8). In aggregate, this is consistent with prior studies in *Drosophila* showing that tRNA genes can be "core" or "peripheral" (Rogers et al. 2010).

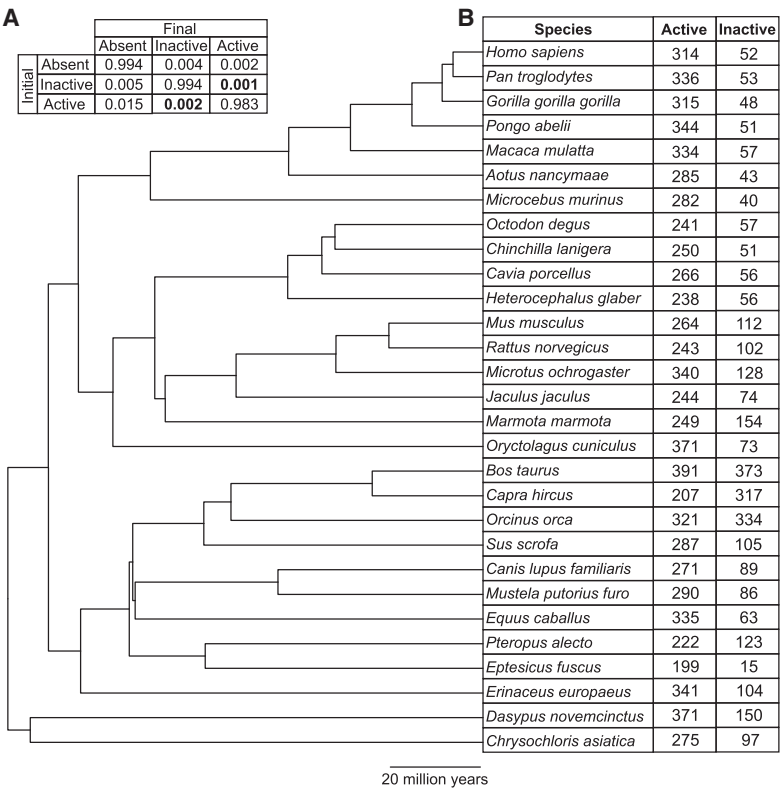

Figure 4. Placental mammals show consistent distributions of predicted active and inactive tRNA genes. (*A*) Estimated transition probabilities between each predicted activity state over a branch length of 1 million years using RevBayes. The probabilities of transition from inactive to active (0.001) and from active to inactive (0.002) are in bold. (*B*) The number of tRNA genes in each predicted activity class are shown for each species in our phylogeny (Hedges et al. 2006), after removal of tRNA genes in segmental duplications. For human and mouse, tRNA genes with no epigenomic data are also excluded from this table (Methods).

We identified a "core" set of 97 primate tRNA genes for which all seven primate species (human, chimpanzee, gorilla, orangutan, macaque, *Microcebus murinus* [gray mouse lemur], and *Aotus nancymaae* [Nancy Ma's night monkey]) have a syntenic ortholog, which are of interest for future experimentation (Supplemental Fig. S9). These represent tRNA genes likely present in the primate common ancestor that have not been lost in any lineage leading to the sampled genomes. These genes encode 19 amino acids (Supplemental Fig. S9A). A single standard amino acid isotype is not represented: cysteine. tRNA-Cys genes are often present in high numbers, and every species in the primate phylogeny has at least 19 of these genes. However, these genes are prone to accumulating nucleotide substitutions, as the human genome contains 23 unique high-confidence tRNA-Cys-GCA gene sequences, the most of any isotype. Therefore, the lack of a "core" eutherian tRNA-Cys gene may be a result of relatively rapid evolution of this gene family, or perhaps difficulty in alignment owing to their high variation in sequence.

In 15 of these 97 "core" ortholog sets, we predicted at least one member of the ortholog group to be active and at least one inactive among the different primate species. Across all 97 "core" ortholog sets, we predict 98% of all member tRNA genes as active, suggesting that deeply conserved tRNA genes are highly likely to be active. Additionally, upon comparison to measurements of Pol III–specific transcription factors (Canella et al. 2010), we find that all 97 "core" human tRNA genes have peaks greater than zero, further demonstrating correlation between conservation and activity.

Transitions between active and inactive are rare

We fit all of our ortholog sets and the predicted activity states of their constituent genes to a Markov model of evolution of discrete characters using RevBayes (Methods; Fig. 4A; Höhna et al. 2016). By fitting our data to the model, we estimated transition probabilities to and from three states: active, inactive, and absent (no detected ortholog). We held the phylogeny constant and solved only for the transition rate parameters. Our model finds that the probability of observing a tRNA gene transition from active to inactive for a given tRNA gene over 1 million years is only 0.002 (Fig. 4A), suggesting that activity state transitions are rare.

Our classifier does well to detect these rare transition events. There are 183 human/mouse ortholog pairs spanning our training and test data sets, and in 171 (93%) of them, human and mouse have the same activity state based on epigenomic data. However, we correctly classified 180 human (98%) and 177 mouse (97%) tRNA genes within this set, indicating that we detected activity state changes between these species, including 11 human tRNA genes and eight mouse tRNA genes whose activity states differ from their orthologous counterparts. Assuming that the activity state of orthologous tRNA genes remains constant across closely related species would yield largely accurate activity state predictions for annotating tRNAs in additional new species. However, our classifier represents an improvement over this assumption and is particularly applicable to species-specific tRNA genes, which are especially common and have no ortholog data.

Inactive tRNA genes that are conserved most often remain inactive (Fig. 4A), hinting at undiscovered biological roles for conserved, apparently silent tRNA genes. We also observed some variation in the relative transition probabilities within clades (Supplemental Fig. S10A–D). Primate tRNA genes are less likely to remain in their initial predicted activity state than rodent tRNA genes. This is consistent with prior studies on the rate of evolutionary change of protein-coding gene expression between clades (Brawand et al. 2011; Necsulea and Kaessmann 2014) but may also reflect differences in sample size between clades. Based on our results, turnover in tRNA gene expression class generally appears to be slow, similar to protein-coding genes (Brawand et al. 2011).

Discussion

Greater understanding of tRNA regulation is a difficult and unmet challenge. There are many obstacles preventing direct measurement of expression at the gene level, including extensive post-transcriptional modifications impeding sequencing, and multiple genomic loci encoding identical transcripts. Nonetheless, we show that accounting for the genomic context allows for improved tRNA gene annotation, and that to determine the transcriptional potential of tRNA genes, direct measurement across many tissues is not necessarily required if the gene sequence and genomic context is known. We leverage features intrinsic to tRNA genes, which relate directly to tRNA function and processing, as well as those extrinsic, which relate to regulation of the chromosomal region.

There are numerous challenges to validating any method for predicting tRNA transcriptional potential. Comprehensive epigenomic data is available for only a few species. Similarly, ChIP- and ATAC-seq data are generally conducted only on a few tissues for a few species of interest. The prevalence of identical tRNA genes in most placental mammal genomes also prevents the identification of source loci for tRNA transcript sequencing data and further limits our ability to support our predictions. However, the relative scarcity of available data motivates the creation of our classifier. Our estimates are comparable to experimental results, but with much greater ease of use and cost-effectiveness. Epigenomic data (Bogu et al. 2016; Holmes 2018) indicate that only 8% of mouse tRNA genes in our test set are active in all nine tissues, and 13% are expressed in only one tissue. This suggests that tissue specificity of tRNA expression is common. This area is of great interest (Ishimura et al. 2014), but few examples have been characterized. Because our classifier infers expression in at least one tissue, our methods will be useful in guiding experiments to find more examples of tightly regulated tRNA genes.

The genome-based nature of this method allows for expansion to incorporate much more data in the future. For example, variation within populations may be useful for predicting relative transcript expression within gene families. We previously determined that actively transcribed tRNA genes accumulate more rare single-nucleotide polymorphisms (SNPs) in both their flanking regions and gene sequences (Thornlow et al. 2018). Therefore, we expect that when population variation data is available for more species, we may infer expression differences at narrower timescales. The model may also be expanded to accommodate nonbinary classification of expression levels in different tissue types and capture the nuance of tRNA gene expression regulation. This approach might also be adapted for the study of other large gene families, because we have previously shown that histone protein-coding genes show similar genetic variation to tRNA genes (Thornlow et al. 2018).

In conclusion, we show reliable classification of tRNA genes using an algorithm that requires little input data and can easily be expanded in the future. Annotations created by our method will be useful in prioritizing tRNA characterization experiments,

as well as interpreting the biological effects of mutations in and surrounding tRNA genes. This work informs the broader question of tRNA gene function evolution, illustrating that tRNA gene expression regulation is dependent on the tRNA gene sequence as well as the varied genomic environment.

Methods

Developing and testing the classifier

For the training data, we used coordinates from human genome assembly GRCh38 for tRNA genes not removed by the tRNAscan-SE high-confidence filter (Chan and Lowe 2016; Chan et al. 2019). For all species, including human and mouse, we extracted the genomes from our Cactus alignment (Supplemental Table S6; Armstrong et al. 2019), ran tRNAscan-SE 2.0, and applied the EukHighConfidenceFilter to exclude tRNA pseudogenes and tRNA-derived SINEs (Chan et al. 2019). We used custom Python scripts to find tRNA loci that were identical from 80 nt upstream to 40 nt downstream from the gene start and end. We considered these segmental duplications and excluded them from classification. If any of these loci also did not align to any tRNA loci in any other species, they were also removed from our ortholog calls, because they most likely represent assembly errors. For genome assemblies in which at least 85% of nucleotides were found on chromosomes, we excluded all tRNA genes not found on chromosomes. For the human tRNA gene set, because our epigenomic data is based on GRCh37 assembly gene annotations, we removed any tRNA genes that were not included in the older assembly, which was determined by performing liftOver (Casper et al. 2018) conversion from GRCh38 to GRCh37, as well as genes in segmental duplications in either assembly.

We used the PHAST (Hubisz et al. 2011) and HAL (Hickey et al. 2013) toolkits to generate phyloP data, and RNAfold (Lorenz et al. 2011) to estimate minimum free energy, using the constraints on secondary structure output by tRNAscan-SE 2.0. We used custom Python scripts in conjunction with tRNAscan-SE 2.0 output and genome annotation files (accession numbers listed in Supplemental Table S6) to obtain data for all other features (Supplemental Code). tRNA alignments generated by the Cactus graph are available in the Supplemental Material. When phyloP data were unobtainable because of lack of alignment, we replaced feature values for each tRNA gene with the mean value for that feature across all tRNA genes in that species, using the SimpleImputer() module in *scikit-learn* (Pedregosa et al. 2011). We used *scikit-learn* to train the model and classify each gene (Pedregosa et al. 2011). Our pipeline and corresponding data are available in the Supplemental Material, as well as at https://github.com/bpt26/tRNA_classifier/. We used Spearman's rank correlation test to ensure that no features were perfectly correlated (Guyon and Elisseeff 2003). We used CfsSubsetEval (Hall et al. 2009) to remove uninformative features and *scikit-learn* to determine feature importance (Pedregosa et al. 2011). To determine the threshold distances for the "Exons Within 75 Kilobases" features, we conducted the Mann–Whitney *U* test for several threshold distances and selected the distance that yielded the smallest *P*-value (Supplemental Table S13). See Supplemental Methods for more details.

To train and test our model, we used epigenomic data from the NIH Roadmap Epigenomics Program (Roadmap Epigenomics Consortium et al. 2015) and the chromatin state-associated gene study in mice (Bogu et al. 2016) for human and mouse tRNA gene activity states, respectively. These studies used histone marks to identify regions of active transcription across 127 human tissues and nine mouse tissues, respectively. In both species, we excluded tRNA genes for which epigenomic data was not available and tRNA genes contained within large segmental duplications. Our training set includes 366 human tRNA genes, 303 active and 63 inactive. For both species, we considered tRNA loci as active if they had an open chromatin state in at least one tissue. We considered all others to be inactive. To determine performance on the human data, we used 10-fold cross-validation, which is commonly used for gene classification studies (McLachlan et al. 2005; Chen et al. 2018; Sethi et al. 2018). We also tested threefold cross-validation but observed very little difference in the model (Supplemental Fig. S11; Supplemental Code). To validate our model, we compared our classifications to ChIP-seq read counts taken directly from Kutter et al. (2011) and ATAC-seq peaks taken directly from Foissac et al. (2019), using liftOver (Casper et al. 2018) conversion to accommodate differences in genome assembly. Sources for training, testing, and validation data are available in Supplemental Table S2. All tRNA data used to generate predictions for each species, as well as their associated predictions and probability scores, are available in Supplemental Tables S8 and S9.

Comparison of predictions to DM-tRNA-seq data

For information on library preparation methods for DM-tRNA-seq assays, see Supplemental Methods. Following sequencing, we used a specialized tRNA sequencing data analysis pipeline, available at https://github.com/UCSC-LoweLab/tRAX, which aligns reads to the tRNA transcripts and reference genome, and computed normalized read counts for each transcript. Because some tRNAs have multiple identical copies in the genome, those sequencing reads were aligned to all corresponding gene loci. To avoid ambiguities, we analyzed only the single-copy tRNA gene loci in this study, as well as only the reads corresponding to whole tRNA molecules.

Creating an ortholog set

We used hal2maf to create 29-way alignments for all tRNA loci of interest for the species in our phylogeny (Hickey et al. 2013). For each tRNA locus, we considered the best aligning tRNA locus from all other species as orthologous, allowing only one ortholog per species per locus. We allowed tRNA genes in segmental duplications to be included, but only if they had an ortholog in at least one other species, because species-specific segmental duplications may be the result of assembly errors. We augmented our ortholog sets with syntenic human/mouse, human/dog, and human/macaque tRNA gene ortholog pairs from Holmes (2018). For all instances in which each tRNA gene in a Holmes (2018) ortholog pair aligned to mutually exclusive sets of species in our Cactus graph, we combined them into one ortholog set. We found that 29 Holmes (2018) human–mouse ortholog pairs align to each other in the Cactus graph, 152 align to mutually exclusive sets of species in our Cactus graph, and 17 align to overlapping sets of species in our Cactus graph. Therefore, we combined the 152 human–mouse tRNA gene pairs with the corresponding ortholog sets defined by our Cactus graph into larger ortholog sets (Supplemental Code).

Fitting a Markov model

We used a phylogeny from TimeTree (Kumar et al. 2017) and fit our data to a Markov model using RevBayes (Höhna et al. 2016). We held the phylogeny constant and allowed RevBayes to optimize only the Q matrix using our tRNA data (Supplemental Code). We then determined transition probabilities over 1 million years using the RevBayes function getTransitionProbabilities() across all species (Fig. 4A) and by clade (Supplemental Fig. S10A–D).

Data access

All raw and processed sequencing data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under accession numbers GSE140096 and GSE140099. The DM-tRNA-seq data generated in this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession numbers PRJNA588252 and PRJNA588256. All custom scripts generated in this study are available at https://github.com/bpt26/tRNA_classifier and as Supplemental Code.

Acknowledgments

We thank the R.B.C.-D. and T.M.L. laboratories for suggestions and feedback. This work was supported by National Institutes of Health (NIH)/National Human Genome Research Institute Grant Award R01HG006753 (to T.M.L.) and NIH/National Institute of General Medical Sciences Grant Award R35GM128932-01 (to R.B.C.-D.). B.P.T. was funded by NIH/National Human Genome Research Institute Grants T32HG008345 and F31HG010584.

Author contributions: Study design was by B.P.T., R.B.C.-D., and T.M.L. J.A. created and provided the Cactus graph. A.D.H. provided mouse tRNA gene activity labels and ortholog data. J.M.H. collected DM-tRNA-seq data. B.P.T. wrote pipeline for extracting feature data and classifying genes. B.P.T., R.B.C.-D., and T.M.L. wrote the manuscript.

References

Allison DS, Hall BD. 1985. Effects of alterations in the 3′ flanking sequence on *in vivo* and *in vitro* expression of the yeast *SUP4*-o $tRNA^{Tyr}$ gene. *EMBO J* **4:** 2657–2664. doi:10.1002/j.1460-2075.1985.tb03984.x

Arimbasseri AG, Rijal K, Maraia RJ. 2013. Transcription termination by the eukaryotic RNA polymerase III. *Biochim Biophys Acta* **1829:** 318–330. doi:10.1016/j.bbagrm.2012.10.006

Armstrong J, Hickey G, Diekhans M, Deran A, Fang Q, Xie D, Feng S, Stiller J, Genereux D, Johnson J, et al. 2019. Progressive alignment with Cactus: a multiple-genome aligner for the thousand-genome era. bioRxiv doi:10.1101/730531

Bogu GK, Vizán P, Stanton LW, Beato M, Di Croce L, Marti-Renom MA. 2016. Chromatin and RNA maps reveal regulatory long noncoding RNAs in mouse. *Mol Cell Biol* **36:** 809–819. doi:10.1128/MCB.00955-15

Boivin V, Deschamps-Francoeur G, Couture S, Nottingham RM, Bouchard-Bourelle P, Lambowitz AM, Scott MS, Abou-Elela S. 2018. Simultaneous sequencing of coding and noncoding RNA reveals a human transcriptome dominated by a small number of highly expressed noncoding genes. *RNA* **24:** 950–965. doi:10.1261/rna.064493.117

Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M, et al. 2011. The evolution of gene expression levels in mammalian organs. *Nature* **478:** 343–348. doi:10.1038/nature10532

Canella D, Praz V, Reina JH, Cousin P, Hernandez N. 2010. Defining the RNA polymerase III transcriptome: genome-wide localization of the RNA polymerase III transcription machinery in human cells. *Genome Res* **20:** 710–721. doi:10.1101/gr.101337.109

Casper J, Zweig AS, Villarreal C, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, Lee CM, Lee BT, Karolchik D, et al. 2018. The UCSC Genome Browser database: 2018 update. *Nucleic Acids Res* **46:** D762–D769. doi:10.1093/nar/gkx1020

Chan PP, Lowe TM. 2016. GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Res* **44:** D184–D189. doi:10.1093/nar/gkv1309

Chan PP, Lin BY, Mak AJ, Lowe TM. 2019. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. bioRxiv doi:10.1101/614032

Chen L, Li J, Zhang YH, Feng K, Wang S, Zhang Y, Huang T, Kong X, Cai YD. 2018. Identification of gene expression signatures across different types of neural stem cells with the Monte-Carlo feature selection method. *J Cell Biochem* **119:** 3394–3403. doi:10.1002/jcb.26507

Cozen AE, Quartley E, Holmes AD, Hrabeta-Robinson E, Phizicky EM, Lowe TM. 2015. ARM-seq: AlkB-facilitated RNA methylation sequencing reveals a complex landscape of modified tRNA fragments. *Nat Methods* **12:** 879–884. doi:10.1038/nmeth.3508

Foissac S, Djebali S, Munyard K, Vialaneix N, Rau A, Muret K, Esquerré D, Zytnicki M, Derrien T, Bardou P, et al. 2019. Transcriptome and chromatin structure annotation of liver, CD4+ and CD8+ T cells from four livestock species. bioRxiv doi:10.1101/316091

Gardiner-Garden M, Frommer M. 1987. CpG islands in vertebrate genomes. *J Mol Biol* **196:** 261–282. doi:10.1016/0022-2836(87)90689-9

Gogakos T, Brown M, Garzia A, Meyer C, Hafner M, Tuschl T. 2017. Characterizing expression and processing of precursor and mature human tRNAs by hydro-tRNAseq and PAR-CLIP. *Cell Rep* **20:** 1463–1475. doi:10.1016/j.celrep.2017.07.029

Goodarzi H, Liu X, Nguyen HC, Zhang S, Fish L, Tavazoie SF. 2015. Endogenous tRNA-derived fragments suppress breast cancer progression via YBX1 displacement. *Cell* **161:** 790–802. doi:10.1016/j.cell.2015.02.053

Grosjean H, de Crécy-Lagard V, Marck C. 2010. Deciphering synonymous codons in the three domains of life: co-evolution with specific tRNA modification enzymes. *FEBS Lett* **584:** 252–264. doi:10.1016/j.febslet.2009.11.052

Guyon I, Elisseeff A. 2003. An introduction to variable and feature selection. *J Mach Learn Res* **3:** 1157–1182.

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. 2009. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter* **11:** 10–18. doi:10.1145/1656274.1656278

Hamada M, Sakulich AL, Koduru SB, Maraia RJ. 2000. Transcription termination by RNA polymerase III in fission yeast. A genetic and biochemically tractable model system. *J Biol Chem* **275:** 29076–29081. doi:10.1074/jbc.M003980200

Hanada T, Weitzer S, Mair B, Bernreuther C, Wainger BJ, Ichida J, Hanada R, Orthofer M, Cronin SJ, Komnenovic V, et al. 2013. CLP1 links tRNA metabolism to progressive motor-neuron loss. *Nature* **495:** 474–480. doi:10.1038/nature11923

Hedges SB, Dudley J, Kumar S. 2006. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* **22:** 2971–2972. doi:10.1093/bioinformatics/btl505

Hickey G, Paten B, Earl D, Zerbino D, Haussler D. 2013. HAL: a hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics* **29:** 1341–1342. doi:10.1093/bioinformatics/btt128

Höhna S, Landis MJ, Heath TA, Boussau B, Lartillot N, Moore BR, Huelsenbeck JP, Ronquist F. 2016. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Syst Biol* **65:** 726–736. doi:10.1093/sysbio/syw021

Holmes A. 2018. "Analyzing regulation of tRNAs, tRNA fragments, and mRNAs in whole genomes." PhD thesis, University of California, Santa Cruz.

Hubisz MJ, Pollard KS, Siepel A. 2011. PHAST and RPHAST: phylogenetic analysis with space/time models. *Brief Bioinform* **12:** 41–51. doi:10.1093/bib/bbq072

Hummel G, Warren J, Drouard L. 2019. The multi-faceted regulation of nuclear tRNA gene transcription. *IUBMB Life* **71:** 1099–1108. doi:10.1002/iub.2097

Ishimura R, Nagy G, Dotu I, Zhou H, Yang XL, Schimmel P, Senju S, Nishimura Y, Chuang JH, Ackerman SL. 2014. RNA function. Ribosome stalling induced by mutation of a CNS-specific tRNA causes neurodegeneration. *Science* **345:** 455–459. doi:10.1126/science.1249749

Kirchner S, Ignatova Z. 2015. Emerging roles of tRNA in adaptive translation, signalling dynamics and disease. *Nat Rev Genet* **16:** 98–112. doi:10.1038/nrg3861

Koski RA, Clarkson SG, Kurjan J, Hall BD, Smith M. 1980. Mutations of the yeast *SUP4* $tRNA^{Tyr}$ locus: transcription of the mutant genes in vitro. *Cell* **22:** 415–425. doi:10.1016/0092-8674(80)90352-9

Krinner S, Heitzer AP, Diermeier SD, Obermeier I, Längst G, Wagner R. 2014. CpG domains downstream of TSSs promote high levels of gene expression. *Nucleic Acids Res* **42:** 3551–3564. doi:10.1093/nar/gkt1358

Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: a resource for timelines, timetrees, and divergence times. *Mol Biol Evol* **34:** 1812–1819. doi:10.1093/molbev/msx116

Kutter C, Brown GD, Gonçalves A, Wilson MD, Watt S, Brazma A, White RJ, Odom DT. 2011. Pol III binding in six mammals shows conservation among amino acid isotypes despite divergence among tRNA genes. *Nat Genet* **43:** 948–955. doi:10.1038/ng.906

Li WH, Ellsworth DL, Krushkal J, Chang BH, Hewett-Emmett D. 1996. Rates of nucleotide substitution in primates and rodents and the generation–time effect hypothesis. *Mol Phylogenet Evol* **5:** 182–187. doi:10.1006/mpev.1996.0012

Lorenz R, Bernhart SH, Höner Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. 2011. ViennaRNA Package 2.0. *Algorithms Mol Biol* **6:** 26. doi:10.1186/1748-7188-6-26

Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25:** 955–964. doi:10.1093/nar/25.5.955

Maraia RJ, Kenan DJ, Keene JD. 1994. Eukaryotic transcription termination factor La mediates transcript release and facilitates reinitiation by RNA polymerase III. *Mol Cell Biol* **14:** 2147–2158. doi:10.1128/MCB.14.3.2147

McLachlan G, Do KA, Ambroise C. 2005. *Analyzing microarray gene expression data*. Wiley-Interscience, Hoboken, NJ.

Meunier J, Lemoine F, Soumillon M, Liechti A, Weier M, Guschanski K, Hu H, Khaitovich P, Kaessmann H. 2013. Birth and expression evolution of mammalian microRNA genes. *Genome Res* **23:** 34–45. doi:10.1101/gr.140269.112

Mleczko AM, Celichowski P, Bąkowska-Żywicka K. 2014. Ex-translational function of tRNAs and their fragments in cancer. *Acta Biochim Pol* **61:** 211–216. doi:10.18388/abp.2014_1888

Necsulea A, Kaessmann H. 2014. Evolutionary dynamics of coding and non-coding transcriptomes. *Nat Rev Genet* **15:** 734–748. doi:10.1038/nrg3802

Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, Baker JC, Grützner F, Kaessmann H. 2014. The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505:** 635–640. doi:10.1038/nature12943

Nguyen N, Hickey G, Zerbino DR, Raney B, Earl D, Armstrong J, Kent WJ, Haussler D, Paten B. 2015. Building a pan-genome reference for a population. *J Comput Biol* **22:** 387–401. doi:10.1089/cmb.2014.0146

Orioli A, Pascali C, Quartararo J, Diebel KW, Praz V, Romascano D, Percudani R, van Dyk LF, Hernandez N, Teichmann M, et al. 2011. Widespread occurrence of non-canonical transcription termination by human RNA polymerase III. *Nucleic Acids Res* **39:** 5499–5512. doi:10.1093/nar/gkr074

Palazzo AF, Lee ES. 2015. Non-coding RNA: What is functional and what is junk? *Front Genet* **6:** 2. doi:10.3389/fgene.2015.00002

Pan T. 2018. Modifications and functional genomics of human transfer RNA. *Cell Res* **28:** 395–404. doi:10.1038/s41422-018-0013-y

Paten B, Diekhans M, Earl D, John JS, Ma J, Suh B, Haussler D. 2011a. Cactus graphs for genome comparisons. *J Comput Biol* **18:** 469–481. doi:10.1089/cmb.2010.0252

Paten B, Earl D, Nguyen N, Diekhans M, Zerbino D, Haussler D. 2011b. Cactus: algorithms for genome multiple sequence alignment. *Genome Res* **21:** 1512–1528. doi:10.1101/gr.123356.111

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. 2011. Scikit-learn: machine learning in Python. *J Mach Learn Res* **12:** 2825–2830.

Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of non-neutral substitution rates on mammalian phylogenies. *Genome Res* **20:** 110–121. doi:10.1101/gr.097857.109

Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, et al. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* **518:** 317–330. doi:10.1038/nature14248

Rogers HH, Bergman CM, Griffiths-Jones S. 2010. The evolution of tRNA genes in *Drosophila*. *Genome Biol Evol* **2:** 467–477. doi:10.1093/gbe/evq034

Schaffer AE, Eggens VRC, Caglayan AO, Reuter MS, Scott E, Coufal NG, Silhavy JL, Xue Y, Kayserili H, Yasuno K, et al. 2014. *CLP1* founder mutation links tRNA splicing and maturation to cerebellar development and neurodegeneration. *Cell* **157:** 651–663. doi:10.1016/j.cell.2014.03.049

Schmitt BM, Rudolph KLM, Karagianni P, Fonseca NA, White RJ, Talianidis I, Odom DT, Marioni JC, Kutter C. 2014. High-resolution mapping of transcriptional dynamics across tissue development reveals a stable mRNA–tRNA interface. *Genome Res* **24:** 1797–1807. doi:10.1101/gr.176784.114

Sethi A, Gu M, Gumusgoz E, Chan L, Yan KK, Rozowsky J, Barozzi I, Afzal V, Akiyama J, Plajzer-Frick I, et al. 2018. A cross-organism framework for supervised enhancer prediction with epigenetic pattern recognition and targeted validation. bioRxiv doi:10.1101/385237

Sun C, Fu Z, Wang S, Li J, Li Y, Zhang Y, Yang F, Chu J, Wu H, Huang X, et al. 2018. Roles of tRNA-derived fragments in human cancers. *Cancer Lett* **414:** 16–25. doi:10.1016/j.canlet.2017.10.031

Thornlow B, Hough J, Roger J, Gong H, Lowe T, Corbett-Detig R. 2018. Transfer RNA genes experience exceptionally elevated mutation rates. *Proc Natl Acad Sci* **115:** 8996–9001. doi:10.1073/pnas.1801240115

Yoo H, Son D, Jang YJ, Hong K. 2016. Indispensable role for mouse ELP3 in embryonic stem cell maintenance and early development. *Biochem Biophys Res Commun* **478:** 631–636. doi:10.1016/j.bbrc.2016.07.120

Zheng G, Qin Y, Clark WC, Dai Q, Yi C, He C, Lambowitz AM, Pan T. 2015. Efficient and quantitative high-throughput tRNA sequencing. *Nat Methods* **12:** 835–837. doi:10.1038/nmeth.3478

Received August 20, 2019; accepted in revised form December 12, 2019.

Predicting transfer RNA gene activity from sequence and genome context

Bryan P. Thornlow, Joel Armstrong, Andrew D. Holmes, et al.

Genome Res. 2020 30: 85-94 originally published online December 19, 2019
Access the most recent version at doi:10.1101/gr.256164.119