Predicting transfer RNA gene activity from genome variation data

Authors and Affiliation:

Bryan Thornlow[1], Joel Armstrong[1,2], Andrew Holmes[1,2], Russell Corbett-Detig[1,2], Todd Lowe[1,2]

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

ABSTRACT:

Transfer RNA (tRNA) genes are among the most highly transcribed genes in the genome due 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 post-transcriptional modifications and high sequence identity across transcripts, has severely limited our collective understanding of tRNA gene expression regulation and evolution. We establish correlates to tRNA gene expression and develop a tRNA gene classification method that minimally requires a single reference genome per species of interest and achieves accuracy comparable to epigenomic and ChIP-Seq assays. We observe that 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 classifications in conjunction with a novel tRNA gene ortholog set spanning 29 placental mammals to infer that tRNA gene expression regulation evolves slowly. The success of our method has fundamental implications to tRNA annotation and variant prioritization. Its simplicity and robustness suggests facile application to other clades and timescales, as well as to explore functional diversification of members of large gene families.

INTRODUCTION:

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

Despite these high cellular demands, many individual tRNA genes currently have little or no evidence for expression. High duplication rates and consequent weakened purifying selection might lead to an abundance of pseudogenes. Additionally, many of these genes may actually be tRNA-derived short interspersed nuclear elements (SINEs), which often retain their promoter elements. However, even after removal of apparent pseudogenes and tRNA-derived SINEs, more than 60 human tRNA genes and over 100 mouse tRNA genes have no evidence for expression in any of more than 100 tissue and cell lines, based on epigenomic data (Thornlow et al. 2018). Chromatin-immunoprecipitation sequencing (ChIP-Seq) data supports this conclusion, as a recent study detected expression of only 224 of 417 high-confidence tRNA genes in human liver, with other mammals showing similar patterns (Kutter et al. 2011).

The evolution of tRNA gene expression may underlie significant phenotypic differences between species as well. Data from prior studies has suggested that the rate of evolution of protein coding gene expression levels differs significantly across clades (Necsulea and Kaessmann 2014; Brawand et al. 2011; Li et al. 1996). The rate of evolution of gene expression also appears to vary among non-coding RNA gene families (Meunier et al. 2013; Necsulea and Kaessmann 2014; Necsulea et al. 2014). Due to difficulties in quantifying tRNA gene expression, the conservation of tRNA expression across mammals at the gene level is not well characterized. The expanding functional repertoire of tRNA transcripts and fragments (Kirchner and Ignatova 2015; Goodarzi et al. 2015; Mleczko, Celichowski, and Bąkowska-Żywicka 2014; Sun et al. 2018) indicates that changes in tRNA expression between species could have profound phenotypic effects.

Expression of tRNA genes has clear importance for organismal development and disease treatment, but our understanding of its regulation and evolution is severely lacking for several reasons. Measurement of unique mRNA and long non-coding RNA (lncRNA) transcripts is straightforward as these molecules can be directly sequenced. However, sequencing of tRNA molecules by traditional methods is impeded by their abundance of post-transcriptional modifications. tRNAs are by far the most modified RNAs, and only very recently have specialized library preparation methods been developed to remove these modifications, enabling reasonably quantitative sequencing (Cozen et al. 2015; Zheng et al. 2015). Furthermore, because many tRNA transcripts are identical at the sequence level, even perfect transcript abundance measurements may not be sufficient to annotate tRNA expression at the gene level. In order to understand how individual tRNA genes are regulated, the study of the local genomic context is therefore essential.

Several alternative genome-wide methods examine tRNA loci in their genomic context to bypass the problem of identical mature tRNA transcripts. Such assays include chromatin immunoprecipitation (ChIP; Bogu et al. 2015; Roadmap Epigenomics Consortium et al. 2015; Thornlow et al. 2018), RNA Polymerase III (Pol III) ChIP-Seq (Kutter et al. 2011), and ATAC-Seq (Foissac et al. 2018), among others. These high-throughput assays require specialized reagents and are often cost- and resource-intensive, so resulting data are generally limited to a few species and a handful of tissues or cell lines. Nonetheless, these limited data indicate that tRNA genes that are identical in sequence may vary in expression (Pan 2018; Kutter et al. 2011; Schmitt et al. 2014), suggesting that tRNA gene expression is partially regulated by factors extrinsic to tRNA genes.

Here, we resolve many of these concerns by developing a simple method to predict which tRNA genes are actively transcribed and which are inactive. Our previous work has shown that relative tRNA expression may be inferred based on DNA variation driven by transcription-associated mutagenesis (Thornlow et al. 2018). We leverage this correlation and introduce novel and previously published genomic correlates to gene expression to create a model for inference of expression of tRNA genes. Our method requires only one target genome in order to predict its constituent tRNA gene activity, so that it may be applied broadly. We demonstrate our method using 29 placental mammal genomes, many of which have no tRNA expression data. We also present a full set of tRNA gene orthologs encompassing the tRNA gene sets of all 29 species. We use this in conjunction with our expression classifications to infer that tRNA gene function appears to evolve slowly and that changes between predicted activity states are rare across mammalian evolution.

RESULTS:

Our classifier uses only data obtainable from reference genomes and alignments. Our goal was to develop a model that could be applied to as many species as possible, and therefore used only features obtainable from a single reference genome. To build our classifier, we explored features that might be correlated with gene activity, and then reduced this superset to just those yielding the best predictions for our training and test data.

Towards this, we evaluated and incorporated three types of function-predictive statistics. First, mutation and selection should leave genomic signatures in genome variation data. In particular, our previous study found that tRNA gene conservation is highest for expressed tRNAs presumably due to stronger purifying selection. Conversely, the putatively neutral flanking regions are more divergent in highly expressed tRNA genes due to consequences of transcription-associated mutagenesis (Thornlow et al. 2018). Thus, we designed our classifier to incorporate features of sequence divergence and polymorphism in and around tRNA genes.

Second, we reasoned that highly expressed tRNAs should encode strong internal promoter sequences and fold correctly to encode a functional mature tRNA. Both primary sequence and secondary structure are included in tRNAscan-SE bit scores, quantitating these core qualities (Chan et al., 2019). Orioli et al. (2011) also demonstrated that tRNA genes with longer and more immediate transcription termination sequences may

	Feature	Importance
Extrinsic	Average PhyloP Score in tRNA Gene Sequence	0.068
	Average PhyloP Score in 5' Flanking Region	0.087
	CpG Density Across tRNA Locus	0.322
	Exons Within 75 Kilobases	0.017
	Observed/Expected CpG Islands Score Upstream of tRNA Gene	0.232
	tRNA Genes Within 10 Kilobases	0.020
Intrinsic	Canonical tRNA Secondary Structure	0.069
	Distance To Nearest TTTT Transcription Termination Sequence	0.053
	tRNA Decoding Redundancy	0.077
	tRNAscan-SE General Bit Score	0.055

Table 1: Both intrinsic (tRNA-specific) and extrinsic (genome context) features are integral to the model. The relative importance of each feature as measured by decrease in node impurity in scikit-learn (Pedregosa et al. 2011) is shown for each feature included in the model (see Methods). Higher importance scores indicate greater contribution to discrimination between active and inactive tRNA genes by the model. tRNA decoding redundancy refers to the number of other tRNA genes sharing the same anticodon, and canonical tRNA secondary structure is based on the constrained minimum free energy (Lorenz et al. 2011).

be more highly expressed as well. Indeed, we found that tRNAscan-SE general bit scores (Chan et al., 2019) and distance to transcription termination sites are significantly correlated with tRNA gene activity based on epigenomic data (Spearman rank correlation, $p < 1e-5$ for all comparisons).

Third, because protein coding gene expression depends heavily on the local chromatin context, we explored features of the surrounding genomic environment. In particular, protein coding genes in regions rich in G/C content, and specifically CpG (or CG) dinucleotides, are known to be more frequently expressed (Gardiner-Garden and Frommer 1987; Krinner et al. 2014). Gardiner-Garden and Frommer (1987) defined CpG Islands scores, which measure the observed frequency of CpG dinucleotides compared to their expected frequency, given the G/C content of the region. We found that these scores are significantly correlated with tRNA gene expression as well (Spearman rank correlation, $p < 8e-30$). We also found that tRNA genes are more likely to be expressed if adjacent to protein-coding genes ($p < 4.7e-5$) or other tRNA genes ($p < 1.36e-6$). Based on these observations, we expected that a combination of both intrinsic and extrinsic features could enable inference of tRNA gene expression in lieu of direct measurement.

To choose a model, we tested several common frameworks, including random forest (RF), logistic regression, and support vector machines, to classify tRNA genes. We found that the RF classifier was most effective, as it achieves the greatest area under the Receiver Operating Characteristic (ROC) curve (AUC, Figure 1) based on ten-fold cross-validation of the human tRNA gene data and application to mouse tRNA gene data (see Methods). We developed our initial model with 30 features (Supplementary Table 1) and gradually reduced it to 10 (Table 1). We used the correlation between each feature and the activity labels, as well as intercorrelation between features, and selected the most informative subset of features (M. A. Hall 1998; M. Hall et al. 2009). The final model performed equally well compared to the initial model, and RF remained the best option.

Features derived from CpG islands are most informative. To better understand our classifier, we determined the relative importance of each feature, in terms of the normalized mean decrease in node impurity across all decision trees in the model (Table 1; Pedregosa et al. 2011). We also used the Spearman rank correlation to determine the significance of the correlation between each feature and the activity measurements, to assess the strength of each feature individually. While all features contribute to model accuracy and are significantly correlated with the activity labels (Spearman rank correlation, $p < 1e-5$ for all features), the most informative features are derived from CpG islands at each tRNA locus, extolling the

importance of chromosomal context and the ability to regulate expressed tRNA genes by altering methylation state (Gardiner-Garden and Frommer 1987).

In development of our method, we assessed the correlation of the length of each pre-tRNA's 3' tail, measured by the distance from the end of the tRNA gene sequence to the beginning of its transcription termination sequence (Allison and Hall 1985; Koski et al. 1980), with tRNA gene activity. Multiple studies on tRNA transcription termination (Maraia, Kenan, and Keene 1994; Hamada et al. 2000; Orioli et al. 2011; Arimbasseri, Rijal, and Maraia 2013) observed that transcription sites vary in sequence, length, and proximity to their associated gene, and that distantly located or weak termination sequences allow for significant readthrough, which might lead to coupling of tRNA maturation and production of downstream-encoded RNA transcripts. We found that active tRNA genes in placental mammals are more likely to have more proximal transcription termination sequences (Spearman rank, $p < 6.5e-6$), perhaps indicating stronger selection to preserve these termination sequences for more frequently transcribed genes, as extended 3' trailer sequences may impede normal post-transcriptional processing.

While tRNAscan-SE bit scores reflect elements of primary sequence and secondary structure known to be associated with functional tRNA genes, other intrinsic features are comparably informative of tRNA activity (Table 1). However, tRNAscan-SE bit scores are also used in the tRNAscan-SE high-confidence filter, which was applied to each tRNA gene set prior to classification. This may lead to underestimation of its relative importance to the classifier. Nonetheless, the importance of CpG density suggests that it is the surrounding chromatin environment more than the tRNA gene sequence that regulates tRNA gene expression, although tRNA sequence features and the surrounding genomic context both likely play significant roles.

Our classifier is at least 90% accurate in classifying mouse tRNA genes. We next tested the accuracy of our classifier using previously collected epigenomic and tRNA expression data. We obtained activity classifications based on chromatin-IP data for all human and mouse tRNAs (Roadmap Epigenomics Consortium et al. 2015; Bogu et al. 2015). For both species, we excluded tRNA genes for which epigenomic data was not available, and tRNA genes that are contained within large segmental duplications (see Methods). Our training set includes 371 human tRNA genes, 304 active and 67 inactive. The activity labels were determined using chromatin-IP (ChIP) data from (Roadmap Epigenomics Consortium et al. 2015) for human tissues and analogous data from (Bogu et al. 2015) for mouse tissues. All tRNA genes located near active transcription start sites in at least one tissue were considered active. All others were considered inactive, as although many of these may actually be highly regulated to the extent that their expression is not detected by our ChIP data, we have no evidence that these tRNA genes are being transcribed. Using ten-fold cross validation, we achieved 93.0% accuracy, correctly classifying 345 of 371 human tRNA genes. Upon applying this same model, without re-training, to the test set of 387 mouse tRNA genes (262 active, 125 inactive), we achieved 92.0% accuracy, correctly classifying 356 tRNA genes (Figure 2A-C).

We did not identify a pattern connecting the 26 incorrectly classified human tRNA genes. These tRNA genes represent 21 different anticodons, indicating that there is not one specific group that the model struggles to classify. For mouse, 15 different anticodons are represented by the 31 misclassified genes. The misclassified genes are borderline cases that are perhaps highly regulated or weakly expressed, and no individual feature appears to be most responsible for their misclassification (Supplementary Table 2). Several of these may also reflect errors in our epigenomic data, which captures over 100 human tissues and 9 mouse tissues but may not detect variation in expression across individuals. Misclassified tRNA genes may also act primarily as insulators or serve other functions not fully captured by our feature set.

Classification without alignment or annotation is similarly accurate. We have ensured that our data is applicable to any species with a sequenced genome. Here, we used Cactus graphs to generate alignments, but we recognize that Cactus graphs are not yet ubiquitous among genomic studies. There may be interest in applying our methods to species for which no alignments or annotations have been developed. Therefore, we have included in our package an option to remove alignment-based and annotation-based features. Use of this model led to slight decreases in accuracy in both human (AUC = 0.932, 92.5% accuracy compared to AUC = 0.942 and 93.0% accuracy in the full model) and mouse (AUC = .935, 90.1% accuracy

compared to AUC = 0.964 and 92.0% accuracy in the full model), which may be exacerbated upon application to more phylogenetically distant species.

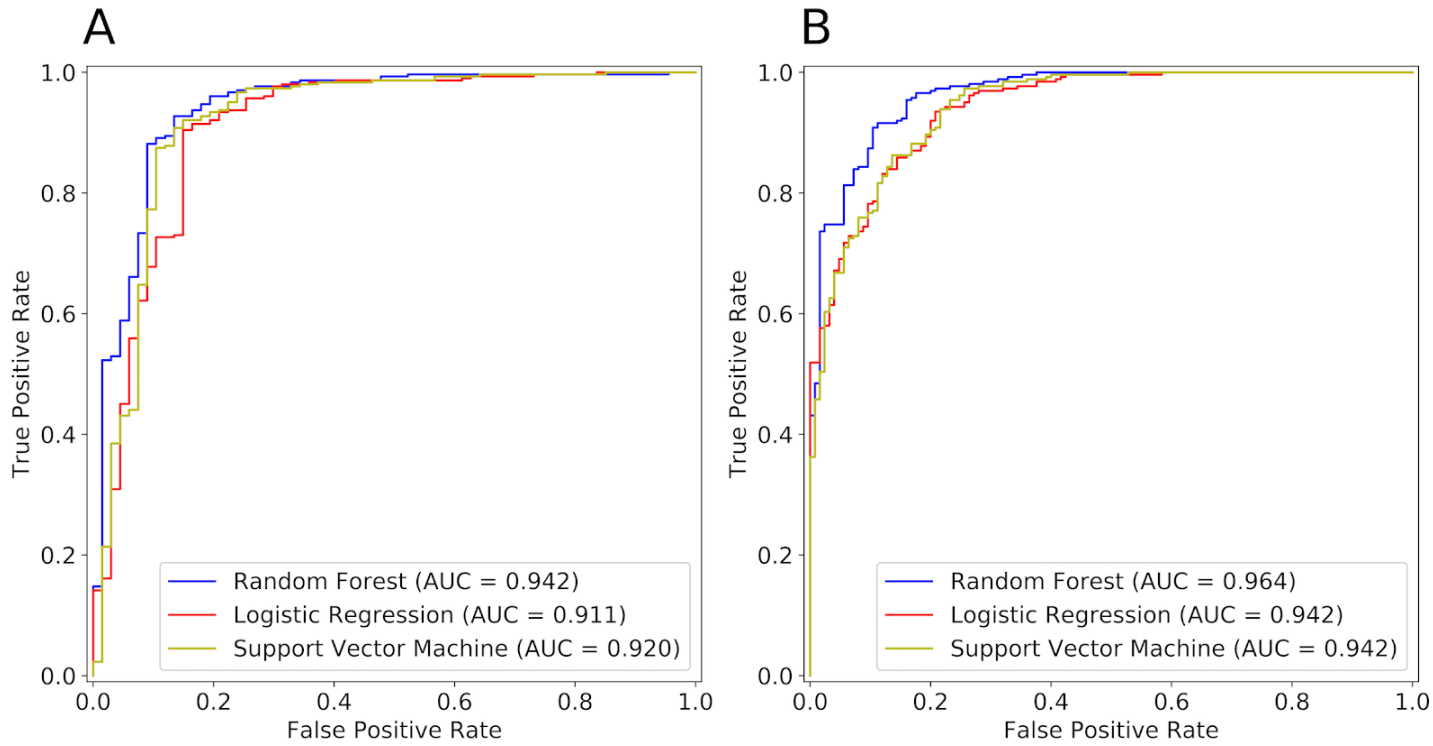


Figure 1: Random forest outperforms other standard binary classification methods. We tried several different classification methods, including random forest (bue), logistic regression (red) and support vector machine (yellow). We applied each method to both the human training data with ten-fold cross-validation (A) as well as after training each model on the human data and testing on the mouse data (B).

Previously collected ChIP-Seq and ATAC-Seq data validate our classifications. We used ChIP-Seq data from (Kutter et al. 2011) to validate our model. In this study, ChIP-Seq was conducted on the livers of six species using RNA Polymerase III (Pol3). We compared our own tRNA gene classifications for four of these species (*Mus musculus*, *Macaca mulatta*, *Rattus norvegicus* and *Canis lupus familiaris*) to their Pol3 ChIP-Seq data. We found comparable agreement between our classifications, the activity states derived from epigenomic ChIP data, and the Pol3 ChIP-Seq read counts (Figure 2). We also found that the relationship between the probability scores in our classifications and the Pol3 ChIP-Seq read counts is consistent across all four species, suggesting similar accuracy in classification across all four species (Figure 2C-F). Our predictions are consistent and similarly accurate when compared to mouse muscle and testes ChIP-Seq data (Supplementary Figure 1).

We predicted many tRNA genes to be active despite a lack of Pol3-based expression in the liver. This may be because our model does not account for differential expression across tissues and instead is trained to predict tRNA genes as active if epigenomic data suggests expression in at least one tissue. For example, in the mouse genome, 262 total tRNA genes are active in at least one tissue based on the epigenomic data, but 89 of these are not expected to be active specifically in the liver based on the same data. Likewise, our model predicted that 122 macaque tRNA genes, 149 dog tRNA genes and 68 rat tRNA genes are active despite no recorded Pol3 activity in the liver (Kutter et al. 2011). Although chromatin-IP has not been performed on macaque, rat and dog tRNA loci, we find that virtually all tRNA genes with measured Pol3-based expression are predicted to be active by our classifier. Among tRNA genes with Pol3 ChIP-Seq read-counts greater than zero, we predicted that 95.3% are active in mouse, 97.7% in macaque, 98.9% 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.

To further validate our predictions, we used genome-wide ATAC-seq data (Foissac et al. 2018), captured in liver, CD4 and CD8 cells for the cow, pig and goat genomes. We measured the ATAC-Seq peak across these tissues for the regions up to 250 bases up- and down-stream of each tRNA gene (Supplementary Figure 2). As before, this data spans only a small subset of tissues, and many tRNA loci that do not show activity in these tissues but predicted active by our model may be active in other tissues. Among tRNA genes with ATAC-Seq read-counts greater than zero, we predicted 90.4% as active in cow, 95.1% in goat and 90.8% in pig. These results are comparable to measurements obtained from ChIP-Seq data in mouse, macaque, rat and dog. This slight decrease in accuracy may be due to differences in resolution between ATAC-Seq and ChIP-Seq, or the increased phylogenetic distance from human to the ungulate species, relative to rodent species.

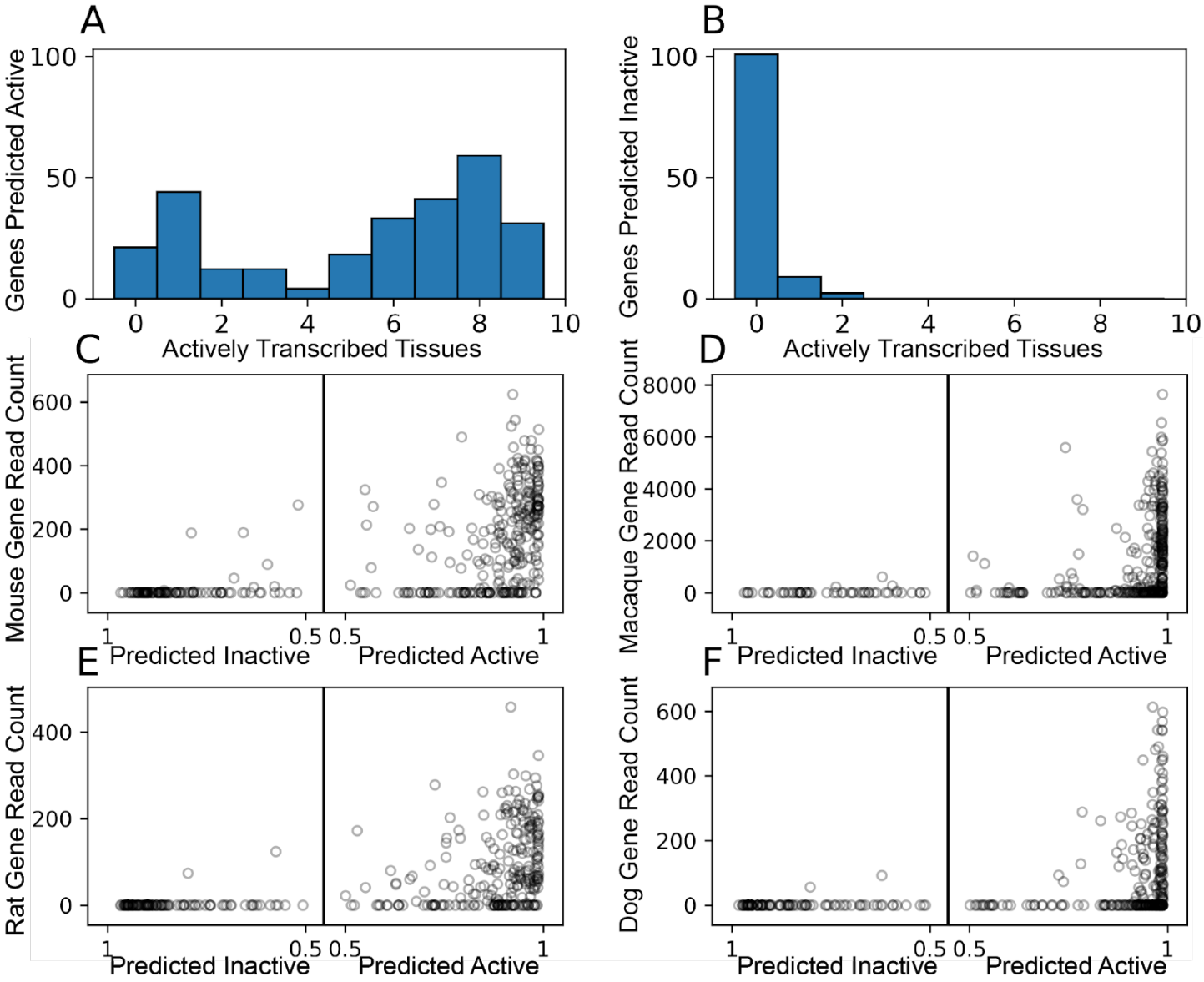


Figure 2: Classification of expression based on genomic data achieves similar results to ChIP-Seq expression analysis in four species. The number of mouse tRNA genes predicted active (A) and inactive (B) are compared to the number of tissues in which they are actively transcribed according to Bogu et al. (2015). We consider any mouse tRNA gene active if it is actively transcribed in at least one tissue. Probability scores output by the classifier are shown on the horizontal axis where tRNA genes further left are predicted inactive with greater probability and tRNA genes further right are predicted active with greater probability. The vertical axis indicates Pol III ChIP-Seq read counts from Kutter et al. (2011) for each tRNA gene in mouse (C), macaque (D), rat (E) and dog (F).

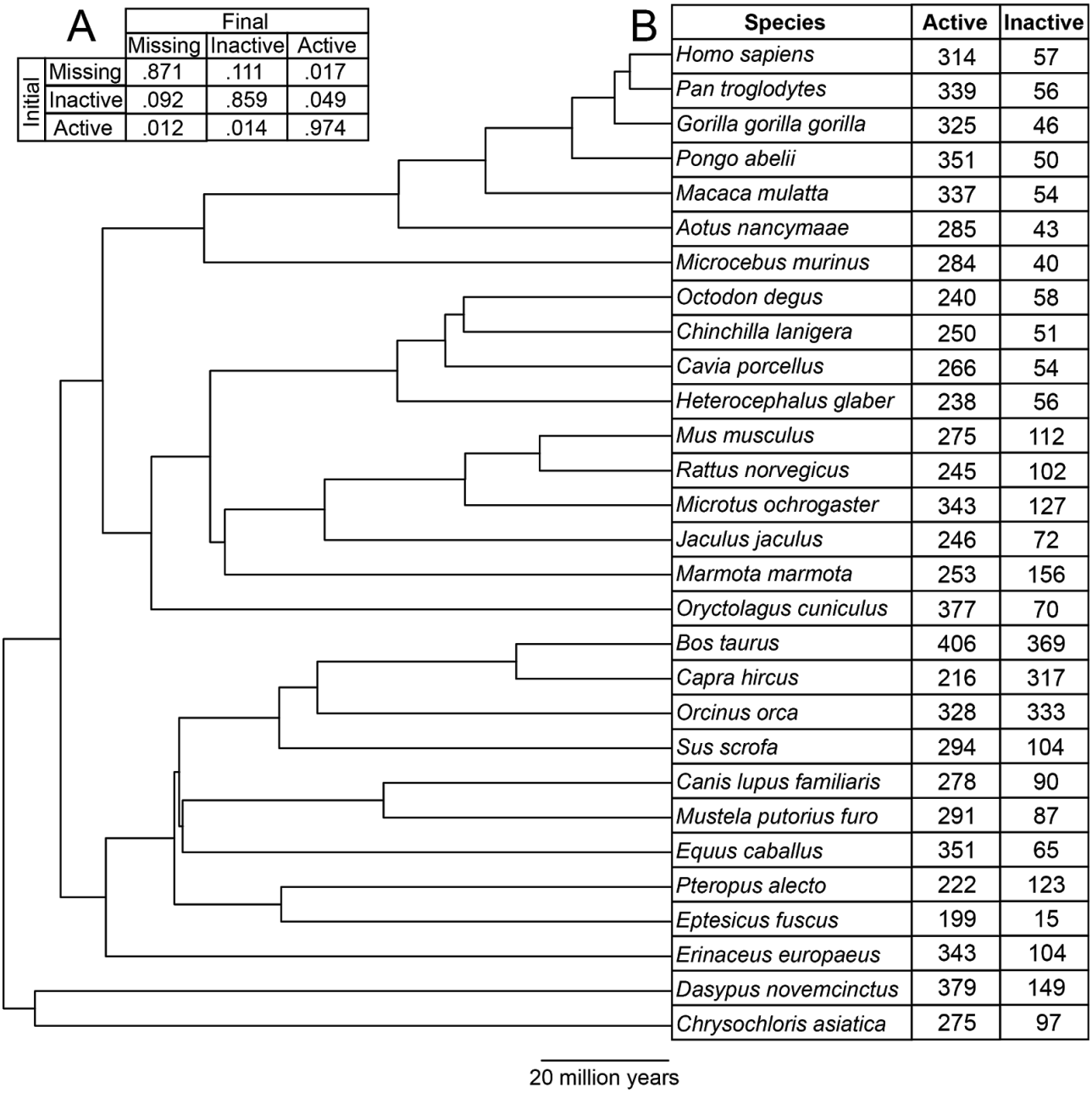


Figure 3: Changes between activity states are rare among placental mammal tRNA genes. Transition rate probabilities between each predicted activity state over a branch length of 10 million years are shown when the model is fit to tRNA data from all species in the phylogeny (A). The number of tRNAs in each class are shown alongside the phylogenetic tree of species involved in this study (B) (Hedges et al. 2006). To generate transition probabilities, the classifications for each tRNA gene ortholog set were modeled as character states and were fit to a Markov model using RevBayes (Hohna et al. 2016).

tRNA gene classifications follow similar distributions across the eutherian phylogeny. Following validation upon comparison to previously collected data, we applied our model to 29 mammalian species (Figure 3). We also determined the distributions of active and inactive tRNAs by anticodon across all species (Supplementary Table 3). Most species have approximately 250-350 predicted active genes, comprising roughly 75% of their tRNA gene sets. We observe generally similar distributions by clade, with few exceptions. Notably, *Bos taurus*, *Capra hircus* and *Orcinus orca* (cow, goat and orca, respectively) have a greater number of predicted inactive tRNA genes. All three of these species have more than 300 tRNA genes predicted inactive while no other species has more than 156. This may be due expansions of tRNA genes in a common ancestor, but is more likely the result of a decreased ability of tRNAscan-SE to discriminate tRNA genes from tRNA-derived SINEs (short interspersed nuclear elements) in these species (Chan et al., 2019). Consistent

with Tang et al. (2009), we identified 406 active cow tRNA genes, and 437 when including those found in segmental duplications (Supplementary Table 4), the most of any species in our phylogeny. Of the 439 confident tRNA genes identified by Tang et al. (2009) in the cow genome, we find 396 genes with exactly matching sequences in our tRNA gene set, and classify 346 of these as active. tRNA gene sequences found by Tang et al. (2009) that were not found in our tRNA set are most likely due to differences in genome assembly.

As an additional check, we ensured that all species had at least one tRNA gene predicted to be active for each expected anticodon (Supplementary Table 3). This excludes anticodons with no known active tRNA genes in human or mouse. We observed three exceptions to this rule. *C. hircus* is the only species that does not have a tRNA-Ser-TGA gene predicted to be active. *C. hircus* is also one of only three species that has any tRNA-Ser-GGA genes, and it has 27 such genes, one of which was predicted to be active. It is possible that *C. hircus* may use one or several of these genes to translate TCA codons through wobble pairing. We also predict that the lone tRNA-Leu-TAA gene in *Erinaceus europaeus* (hedgehog) is inactive. The probability score for this prediction is 0.51, nearly the minimum possible probability score. This uncertainty, combined with the lack of copies of this gene, indicates that it may be active. Additionally, the *Chinchilla lanigera* (chinchilla) genome does not contain any high-confidence selenocysteine (SeC) tRNA genes, which is most likely due to an incomplete genome assembly, as selenocysteine tRNAs are required in mammals (Johansson, Gafvelin, and Arnér 2005).

Outlier tRNA genes may become foci of future experimentation. We identified several anomalous lineage-specific expansions. Most notably, *E. europaeus* has 123 tRNA-Lys-CTT genes. All of these have high tRNAscan-SE general bit scores (Supplementary Figure 3) and are unlikely to be pseudogenes, and 98 of these are classified as active. tRNA-derived SINEs can proliferate rapidly via retrotransposition, and it is possible that a rapid expansion occurred in this species through this mechanism, primarily targeting areas within or near regions of high transcription. These loci have minimal changes to their sequences, possibly indicating a relatively recent expansion. While few of these tRNA genes have completely identical sequences, 43 have identical 3' flanking regions, extending 10 bases after the tRNA gene (CCATTTGTTG). An additional 24 genes share a different 3' flanking motif (CCAGATGTTG). The immediate flanking regions of actively transcribed tRNA genes are expected to be highly divergent due to transcription-associated mutagenesis (Thornlow et al. 2018). One possible explanation for this discrepancy is that these motifs are part of a fused SINE element responsible for the proliferation of the tRNA-derived SINE element. A similar expansion has occurred in *Microtus ochrogaster* (prairie vole), whose genome contains 131 tRNA-Lys-CTT genes, 66 of which are classified as active. 25 of these genes are exactly identical in tRNA gene sequence, suggesting a similarly recent expansion through retrotransposition.

To our knowledge, no actively transcribed suppressor tRNA genes have been demonstrated in primate genomes. However, we have classified a tRNA-Sup-CTA in *Pan troglodytes* (chimpanzee) as active, although experimentation is necessary to support that this tRNA is actively transcribed. While nonsense suppression via readthrough by near-cognate tRNAs has been demonstrated several times in metazoa (Jungreis et al. 2011; Yacoubi et al. 2012; Loughran et al. 2018; Roy et al. 2015) and suppressor tRNAs have been found in other species (Beier and Grimm 2001; Valle, Morch, and Haenni 1987), evidence for an actively transcribed suppressor tRNA in a primate species is of great interest for future experimentation. This tRNA gene is a well-conserved ortholog of the human tRNA-Gln-CTG-6-1 gene, and exhibits several chimpanzee-specific nucleotide substitutions (Supplementary Figure 4) that are not expected to significantly impair the function of its transcript. The similarity of the sequence of this gene to other tRNA-Gln-TTG genes suggests an additional anticodon shift, prior to the divergence of placental mammals. Although our classifier was created in order to better understand tRNA gene expression, it may also be used for rapid screening for outliers that may exhibit unique biology. This tRNA-Sup-CTA gene will be a focus of future experimentation and represents one of many examples of outliers brought to light by our method.

tRNA genes predicted to be active are more deeply conserved. In addition to application of our classifier to placental mammal genome data, we have developed a complete set of placental mammal tRNA

gene orthologs using a Cactus graph (Paten, Diekhans, et al. 2011; Paten, Earl, et al. 2011; Nguyen et al. 2015; see Methods). Of the 11,752 tRNA genes in our 29-species alignment, 3,582 genes in total, or about 122 per species, appear to be species-specific. The remaining 8,170 genes were condensed into 1,097 ortholog sets. On average, each of the tRNA genes is present in 7.4 species, indicating that tRNA genes are generally either fairly deeply conserved or recently evolved (Supplementary Figure 5). This is consistent with prior studies in *Drosophila* that showed that tRNA genes can be "core" or "peripheral" (Rogers, Bergman, and Griffiths-Jones 2010).

To further investigate the relationship between activity and conservation, we identified an ancient "core" set of tRNA genes found in all primates (Supplementary Figure 6). We found 97 tRNA genes for which all seven primate species (human, chimpanzee, gorilla, orangutan, macaque, *Microcebus murinus* and *Aotus nancymaae*) have a syntenic ortholog, with no upper limit on other species in the phylogeny. This most likely represents the set of tRNA genes 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, nearly a full set (Figure 4). The only standard amino acid not represented is cysteine. tRNA-Cys genes are often present in high numbers, and every species in the phylogeny has at least 17 such genes. However, these genes are prone to accumulating substitutions in their sequences. Therefore, the lack of a "core" eutherian tRNA-Cys gene may be due to relatively rapid evolution of this gene family in particular, or difficulty in alignment due to their high level of variation in sequence.

In 15 of these 97 ortholog sets, we predicted at least one gene predicted as active and at least one gene as inactive. In all cases, we predict more genes to be active than inactive, further suggesting that tRNA genes predicted to be active are also more deeply conserved. Additionally, out of all 1,097 ortholog sets, 750 contain only tRNA genes predicted to be active, approximately mirroring the distribution of active to inactive tRNA genes predicted at the species level and further indicating that active tRNA genes tend to be more conserved than inactive tRNA genes.

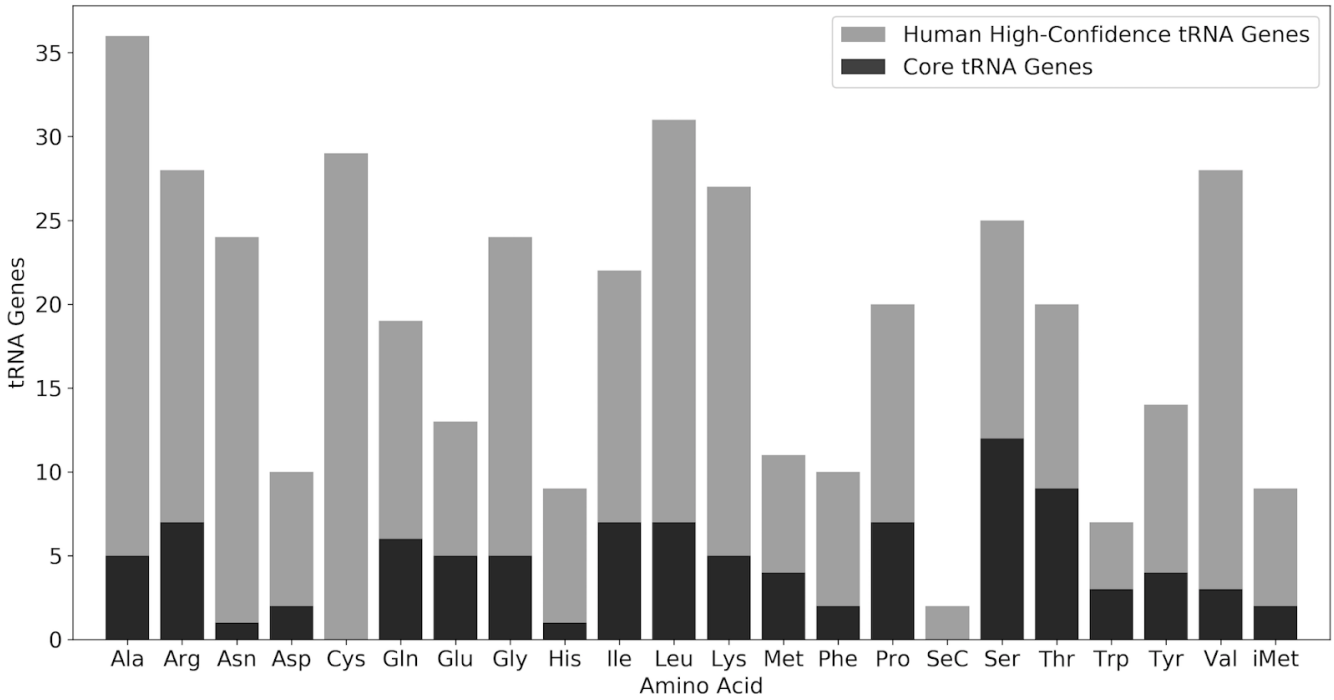


Figure 4: Cysteine is the only standard amino acid without a "core" tRNA. All human active and inactive high-confidence tRNA genes are shown in gray for each amino acid, and those found in the core set of 97 tRNA genes are shown in black. Cysteine is one of the highest copy-number tRNA gene families, but is the only standard amino acid with zero core tRNA genes.

Substitutions in tRNA anticodons accompany changes in predicted activity state. Among our 1,097 ortholog sets, we identified 111 instances in which an ortholog set contains genes corresponding to

multiple different anticodons. 52 of these 111 anticodon shifts are non-synonymous, resulting in a change in the corresponding amino acid. Ortholog sets containing anticodon shifts are significantly enriched for changes in predicted activity, as 36 of 111 such ortholog sets contain at least one active and at least one inactive tRNA gene (Hypergeometric test, $p < 2.3e-4$). Likewise, of the 52 ortholog sets containing non-synonymous anticodon substitutions, 18 contain at least one active and at least one inactive tRNA gene, indicating significant enrichment for predicted activity state transitions (Hypergeometric test, $p < 2.8e-3$).

Transitions between active and inactive are rare. We have fit our ortholog sets and their predicted activity states to a Markov model of evolution of multiple discrete characters using RevBayes (Hohna et al. 2016; Figure 3; see Methods). By fitting our data to the model, we estimated the transition probabilities to and from three states - active, inactive, and missing (no detected ortholog). We solved only for the transition rate parameters, holding the phylogeny constant. Consistent with the premise that local gene duplication is the dominant mode of evolution of tRNA genes, our model suggests that active tRNA genes, which are generally born from prior active tRNA genes, overwhelmingly tend to remain active, and the same is true of inactive tRNA genes (Figure 3).

Although such transition events are rare, our classifier does well to detect them. There are 188 human-mouse ortholog pairs for which we have epigenomic data, and in 173 of them, human and mouse have the same activity state. However, we correctly classify 183 human and 181 mouse tRNA genes within this set, indicating that most activity state changes between these species are detectable. If inputs to the model are unobtainable and true activity states for only one species are available, assuming that the activity state of orthologous tRNA genes remains constant across closely related species would yield mostly accurate results. However, our classifier represents an improvement over this assumption.

Generally, tRNA genes in a given activity state are most likely to remain in that state, over branch lengths of 10 million years. Overall, transitions of any kind are rare, and the active state appears to be most stable (Figure 3A), but surprisingly, inactive tRNA genes overwhelmingly remain inactive as well. However, we observed some variance in the relative transition probabilities within clades. 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). 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 in functional and evolutionary genomics. There are many obstacles preventing direct measurement of expression at the gene level. Nonetheless, here we show that, in order to determine relative levels of expression of tRNA genes, direct measurement can be effectively replaced by inference through data derived from genome sequences and alignments. We have shown that binary expression levels can be reliably predicted through features both intrinsic to tRNA genes, which relate directly to tRNA function and processing, and extrinsic, which relate to the surrounding genomic environment.

Our classifier demonstrates remarkable accuracy in predicting tRNA gene activity. However, there are important caveats to our method. Most importantly, the epigenomic ChIP data used to create the “active” and “inactive” labels for both human and mouse data is imperfect. It is possible that many inactive tRNAs are in fact highly regulated and do exhibit expression, but only at such specific times in development or specific tissues that their activity is not captured by previously conducted assays. ChIP captures the chromatin environment across many tissues and is by far the most comprehensive data available for tRNA gene activity, yet is still an indirect approximation of transcript abundance because it does not take into account tRNA processing rate or stability. However, that comprehensive tRNA gene expression data is unavailable motivates the creation of our classifier. Available data indicate that our estimates are comparable to experimental results, but with much greater ease of use and cost-effectiveness.

Our method focuses on tRNA genes specifically, but the basic principles underlying our model might be applied to any duplicated gene family. For example, ribosomal RNA (rRNA) genes are present in many identical copies and are transcribed at similar levels to tRNA genes. Indeed, many rRNA genes have no detected expression in plants and yeast, and rRNA gene copy number and relative transcription appear to have fundamental implications for organismal fitness (Ide et al. 2010). We have also previously shown that histone protein coding genes qualitatively similar genetic variation to tRNA genes (Thornlow et al. 2018). If one was interested in detecting relative levels of expression of rRNA or histone protein coding gene loci, a model similar to ours could be developed. This core idea therefore has the potential to revolutionize functional and evolutionary genomic inference for large multi-copy gene families.

Similarly, our methods may be applied to other species. We have previously demonstrated that transcription-associated mutagenesis creates signatures at plant and insect tRNA gene loci that are qualitatively similar to those observed in human and mouse (Thornlow et al. 2018). While some features may not extend quite so far phylogenetically, and others may require recalibration or processing, a similar model could be developed for other clades of eukaryotes outside of placental mammals. For the purposes of the present study, we have chosen to work with placental mammals as the corresponding data is of relatively high quality.

Our approach may expand in several directions as more tRNA data is collected. For example, variation within populations, and potentially within highly proliferative tissues, may be useful for predicting relative transcript expression among gene families. Our previous work showed that actively transcribed tRNA genes tend to accumulate more rare single nucleotide polymorphisms (SNPs) in both their flanking regions and within the gene sequences (Thornlow et al. 2018). Therefore, we expect that, when population variation data is available for more species, our approach might be adapted to infer expression differences at narrower timescales. With more data, our model may also be expanded to accommodate non-binary classification of expression levels, as tRNA gene expression regulation in reality is more nuanced than that genes are either active and inactive.

Importantly, our classifier handles strongly intercorrelated features well. When attempting to model a single biological process, all parts of the process are inevitably correlated. This is visible in the current model, as we use two features related to CpG dinucleotides and two features derived from PhyloP scores (Table 1). We also include tRNAscan-SE general bit scores and canonical secondary structure, although the general bit scores account for both primary sequence and secondary structure (see Methods). Using highly correlated features leads to underestimation of their importance to the model (Strobl et al. 2008; Tolosi and Lengauer 2011). Nevertheless, we have included intercorrelated features because while understanding contributions of each feature to the model is of interest, achieving the greatest possible performance of the model is paramount.

While tRNA genes predicted to be active are most well conserved, it is surprising that inactive tRNA genes are also strongly conserved. This is true not only in their sequence but in their expression, as tRNA genes predicted to be inactive are far more likely to remain inactive than to become pseudogenes or active genes. While our epigenomic data captures over 100 tissues, it is temporally limited, and some of these genes may be transcribed only at very specific stages in development not captured by these assays. However, other tRNA genes may be biologically important for functions other than production of mature transcripts. A subset of apparently silent human tRNA genes are present in the introns of protein coding genes, and may regulate their transcription. Prior studies have also demonstrated that tRNA genes act as powerful insulators bounding chromatin regulatory programs (Raab et al. 2012).

Here we demonstrate reliable classification of tRNA genes using open-source software that requires little input data and can easily be expanded in the future. Annotations created by our method will be useful in aiding design of tRNA characterization experiments, as well as for variant effect interpretation, as polymorphic sites within or near more highly active tRNA genes may have more significant fitness consequences. Our approach focuses on species up to 100 million years since divergence, but we anticipate that similar approaches at different time-scales, using variation data within populations, or even within highly proliferative

tissues, may be possible in the near future. Moreover, our method demonstrates the importance of the local genomic context in gene-level expression, and its success suggests that the same strategy may be used to reliably predict expression of other multi-paralog gene families as well. This work informs the broader question of tRNA gene function evolution, illustrating that tRNA gene expression regulation is likely dependent on many factors, including those intrinsic to the tRNA gene sequence as well as those derived from the varied genomic environment.

METHODS

Developing the classifier. For the training data, we used coordinates from the high-confidence tRNA set in the hg38 genome from GtRNAdb (Chan and Lowe 2016). For all test species, we extracted the genomes from our Cactus alignment, ran tRNAScan-SE 2.0 using the default parameters and applied the EukHighConfidenceFilter to exclude tRNA pseudogenes and tRNA-derived SINEs (Chan et al. 2019). We then used custom Python scripts to find tRNA genes identical in gene sequence and 20-nucleotide 5' and 3' flanking regions. If any of these non-unique loci did not align to any tRNA loci in any other species, they were removed from our alignment, as these loci are most likely assembly errors. For genomes in which at least 85% of nucleotides were found on chromosomes, we also excluded from our analysis all tRNA genes that were not found on chromosomes.

We used the PHAST (Hubisz, Pollard, and Siepel 2011) and HAL (Hickey et al. 2013) toolkits, as well as custom Python scripts to find all other feature sets, using only each reference genome, files output from tRNAScan-SE (Chan and Lowe 2019), and annotation files. We used scikit-learn to test all classification methods and used Random Forest in our final model, as it outperformed other standard frameworks (Pedregosa et al. 2011).

Eliminating uninformative features. To reduce our feature set to only features contributing to classification, we first used the Spearman rank correlation test for each feature in the human training set, in order to determine whether any perfectly correlated variables were present (Guyon and Elisseeff 2003). When no features were perfectly correlated, we used the CfsSubsetEval from Weka (M. Hall et al. 2009) to remove uninformative features. To test and apply our model, we eventually moved to the scikit-learn package, so that obtaining feature data and generating predictions could be performed in the same program (Pedregosa et al. 2011). From here, we identified feature importance values through a function built in to scikit-learn.

Two of our features rely on PhyloP scores, and therefore alignment across species. This led to missing data, as some tRNA genes did not align to regions in any other species. These features are significantly correlated with tRNA gene expression (Thornlow et al. 2018) and could not be excluded from the model. When PhyloP data were unobtainable, values for each tRNA gene were replaced with the mean value for that feature across all tRNA genes. This enabled classifications to be made for all tRNA genes in all species. Many of these unalignable genes are likely species-specific. Indeed, we observe a significant correlation between the number of unalignable tRNA genes and the phylogenetic distance to the species' nearest neighbor ($p < 8.8e-6$, Spearman rank, Supplementary Figure 8A). However, many of these tRNA genes are also found on relatively small scaffolds, and the percentage of unalignable tRNA genes is also significantly inversely correlated with contig N50 ($p < 3.6e-5$, Spearman rank, Supplementary Figure 8B). Importantly, the model is developed to discriminate among tRNAs that are already relatively high-quality as judged by tRNAscan-SE 2.0 (Chan et al., n.d.). Both genome and alignment quality are active areas of genomics research and are very quickly improving. As such, we believe that the effectiveness of our method will steadily improve with better genome assembly.

We used output from RNAfold from ViennaRNA (Lorenz et al. 2011) to estimate constrained minimum free energy (Table 1). tRNAscan-SE also outputs a secondary structure score (Chan et al. 2019). We find that these scores are highly correlated to the RNAfold output (Spearman rank correlation, $p < 2e-4$), indicating that they may be interchangeable. To determine whether to use one instead of the other, we created a separate model that included both scores, and found that RNAfold output had greater importance to the model. We also tested a model that includes all bit scores output by tRNAscan-SE, including secondary structure, HMM and

isotype-specific bit scores. We find that upon their inclusion, the importance of the tRNAscan-SE general bit score decreases slightly, but it remains more important to the model than any of the other tRNAscan-SE scores, supporting the use of the tRNAscan-SE general bit score over the other scores.

Refining the classifier. We also experimented with Random Forest hyperparameters. Our final model uses 250 estimators, a minimum of 2 samples required to split an internal node, a maximum depth of 4 nodes, and is trained through 10-fold cross-validation. For all species, we used the pipeline on GitHub (github.com/bpt26/tRNA_classifier/) to generate feature data (Supplementary Figure 9). For the human and mouse data, we used epigenomic classifications from Roadmap Epigenomics Consortium et al. (2015) and Bogu et al. (2015), which used chromatin data to determine which regions of the genome were actively transcribed in the hg19 and mm10 genomes, respectively. Because our alignment uses the hg38 genome, we excluded annotated human tRNA genes found in hg38 that did not liftover (Casper et al. 2018) to hg19, as no epigenomic data is available for these genes. Additionally, we considered any tRNA loci that were identical from 80 nucleotides upstream to 40 nucleotides downstream of the gene start and end to be within segmental duplications. These were excluded from our training and testing sets, because their epigenomic activity states were likely not unique to each locus and were therefore not suitable for training and testing our model. For the human tRNA gene set, because our epigenomic data is derived from hg19 annotations, we also removed any tRNAs in segmental duplications in the hg19 genome, even if this was resolved in hg38, as the epigenomic data was unreliable. tRNA genes in segmental duplications in all species are excluded from our tRNA gene counts (Figure 3) as several of these may be sequencing errors.

Creating an ortholog set. We used the hal2maf tool to generate multiple alignments of all tRNA loci (Hickey et al. 2013). Because of the inclusion of alt contigs in the Cactus graph, we did not use the --noDupes flag. 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 also used previously compiled human/mouse, human/dog and human/macaque tRNA gene ortholog pairs (Holmes et al, in prep). These ortholog pairs are syntenic and were generated through study of surrounding protein-coding genes. To combine these datasets, we found all instances in which each tRNA gene in an ortholog pair had aligned to mutually exclusive sets of species in the Cactus graph. For each of these instances, we collapsed these mutually exclusive ortholog sets into one ortholog set. This created a more complete ortholog set and prevented overestimation of tRNA gene birth and death rates. The ortholog pairs from Holmes et al were determined through in-depth studying of each tRNA locus, and are therefore likely more accurate than those obtained from the Cactus graph. Overall, we found that 29 Holmes et al human-mouse ortholog pairs align to each other, 152 align to mutually exclusive sets of species, and 17 align to overlapping sets of species. Therefore, we combined each of these 152 human-mouse tRNA gene pairs into a larger ortholog set. We also removed from our ortholog set all tRNA loci that were in segmental duplications, as described above, and also did not align to tRNA genes in any other species, as these were most likely assembly errors.

Our method differs from Tang et al. (2009), who found 22 unique sequences across 150 tRNA genes to be conserved beyond placental mammals, in that we use synteny and not only sequence similarity to produce our alignments. Consequently, we consider fewer tRNA genes to be deeply conserved.

Fitting a Markov model. Using our classifications for all tRNAs in all species in conjunction with our ortholog sets, we compiled a table of all activity classifications for all tRNA genes. 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 use TimeTree instead of the phylogeny built into our Cactus alignment, as we wanted to estimate transition rates over millions of years. We allowed transitions to and from the missing state because the missing state includes pseudogenes, absent tRNA genes and alignment errors. Therefore, transitions to and from the missing state would include duplication and deletion events, which certainly occur, as well as transitions between filtered pseudogenes and active or inactive genes, which are conceivable by substitutions to the tRNA gene sequence or flanking regions. We held the phylogeny constant and allowed RevBayes to optimize only the Q matrix using our tRNA data. We then determined transition probabilities over 10 million year branch lengths using the built-

in function getTransitionProbabilities() in RevBayes. We applied this method to all tRNAs in our ortholog set for all species as well as those in specific clades (Supplementary Figure 7).

One might expect that predicted transitions in activity are be caused by substitutions in the internal promoter regions of the genes. Such substitutions strongly affect tRNAScan-SE bit scores, as these are especially strongly conserved elements among tRNA genes, and bit scores are measures of alignment of sequences to prototypical tRNA genes. To test whether activity transition predictions were driven by bit score effects of substitutions, we removed the tRNAScan-SE bit score feature from our model and compared the predicted number of transitions in activity among orthologous gene sets. Our standard model predicts 216 activity transition events, while the model with bit score removed predicts 208 such events. Transitions in 191 ortholog sets are predicted by both models, indicating that while substitutions within tRNA genes certainly affect predicted activity state, the majority predicted transition events do include changes to extrinsic features among orthologous genes.

DATA ACCESS

The tRNA classifier pipeline is available on GitHub (https://github.com/bpt26/tRNA_classifier/) along with the training data necessary to run the classifier. Classification distribution data for all species are available in Supplemental Materials. Multiple alignment files generated from the Cactus graph are available upon request. The guide tree for the Cactus graph was:

(((((((((Homo_sapiens:0.00655,Pan_troglodytes:0.00684)Anc32:0.00422,Gorilla_gorilla_gorilla:0.008964)Anc29:0.009693,Pongo_abelii:0.01894)Anc25:0.015511,Macaca_mulatta:0.043601)Anc20:0.08444,Aotus_nancymaae:0.08)Anc15:0.08,Microcebus_murinus:0.10612)Anc11:0.083494,((((Jaculus_jaculus:0.1,(Microtus_ochrogaster:0.14,(Mus_musculus:0.084509,Rattus_norvegicus:0.091589)Anc30:0.047773)Anc26:0.06015)Anc21:0.122992,(Heterocephalus_glaber:0.1,(Cavia_porcellus:0.065629,(Chinchilla_lanigera:0.06,Octodon_degus:0.1)Anc31:0.06)Anc27:0.05)Anc22:0.06015)Anc16:0.05,Marmota_marmota:0.1)Anc12:0.05,Oryctolagus_cuniculus:0.21569)Anc08:0.04)Anc04:0.040593,(((Sus_scrofa:0.12,(Orcinus_orca:0.069688,(Bos_taurus:0.04,Capra_hircus:0.04)Anc23:0.09)Anc17:0.045488)Anc13:0.02,((Equus_caballus:0.109397,(Felis_catus:0.098612,(Canis_lupus_familiaris:0.052458,Mustela_putorius_furo:0.08)Anc28:0.02)Anc24:0.049845)Anc18:0.02,(Pteropus_alecto:0.1,Eptesicus_fuscus:0.08)Anc19:0.033706)Anc14:0.03)Anc09:0.025,Erinaceus_europaeus:0.278178)Anc05:0.021227)Anc02:0.023664,((Loxodonta_africana:0.098842,Chrysochloris_asiatica:0.04)Anc06:0.05,Dasypus_novemcinctus:0.169809)Anc03:0.02)backbone_root;

ACKNOWLEDGMENTS

We thank the R.B.C.-D. and T.M.L. laboratories for suggestions and feedback. This work was supported by National Human Genome Research Institute/NIH Grant 2R01HG006753-04A1 (to T.M.L.) and National Institute of General Medical Sciences/NIH Grant R35GM128932-01 (to R.B.C.-D.). B.P.T. was funded by NIH Grant 5F31HG010584-02.

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

DISCLOSURE DECLARATION

We declare no competing interests.

REFERENCES

Allison, D. S., and B. D. Hall. 1985. “Effects of Alterations in the 3’ Flanking Sequence on in Vivo and in Vitro Expression of the Yeast SUP4-O tRNATyr Gene.” *The EMBO Journal* 4 (10): 2657–64.

Arimbasseri, Aneeshkumar G., Keshab Rijal, and Richard J. Maraia. 2013. “Transcription Termination by the Eukaryotic RNA Polymerase III.” *Biochimica et Biophysica Acta* 1829 (3-4): 318–30.

Beier, H., and M. Grimm. 2001. “Misreading of Termination Codons in Eukaryotes by Natural Nonsense Suppressor tRNAs.” *Nucleic Acids Research* 29 (23): 4767–82.

Bogu, Gireesh K., Pedro Vizán, Lawrence W. Stanton, Miguel Beato, Luciano Di Croce, and Marc A. Marti-Renom. 2015. “Chromatin and RNA Maps Reveal Regulatory Long Noncoding RNAs in Mouse.” *Molecular and Cellular Biology* 36 (5): 809–19.

Boivin, Vincent, Gabrielle Deschamps-Francoeur, Sonia Couture, Ryan M. Nottingham, Philia Bouchard-Bourelle, Alan M. Lambowitz, Michelle S. Scott, and Sherif Abou-Elela. 2018. “Simultaneous Sequencing of Coding and Noncoding RNA Reveals a Human Transcriptome Dominated by a Small Number of Highly Expressed Noncoding Genes.” *RNA* 24 (7): 950–65.

Brawand, David, Magali Soumillon, Anamaria Necsulea, Philippe Julien, Gábor Csárdi, Patrick Harrigan, Manuela Weier, et al. 2011. “The Evolution of Gene Expression Levels in Mammalian Organs.” *Nature* 478 (7369): 343–48.

Casper, Jonathan, Ann S. Zweig, Chris Villarreal, Cath Tyner, Matthew L. Speir, Kate R. Rosenbloom, Brian J. Raney, et al. 2018. “The UCSC Genome Browser Database: 2018 Update.” *Nucleic Acids Research* 46 (D1): D762–69.

Chan, Patricia P., Brian Y. Lin, Allysia J. Mak, and Todd M. Lowe. 2019. “tRNAscan-SE 2.0: Improved Detection and Functional Classification of Transfer RNA Genes.” https://doi.org/10.1101/614032.

Chan, Patricia P., and Todd M. Lowe. 2016. “GtRNAdb 2.0: An Expanded Database of Transfer RNA Genes Identified in Complete and Draft Genomes.” *Nucleic Acids Research* 44 (D1): D184–89.

Cozen, Aaron E., Erin Quartley, Andrew D. Holmes, Eva Hrabeta-Robinson, Eric M. Phizicky, and Todd M. Lowe. 2015. “ARM-Seq: AlkB-Facilitated RNA Methylation Sequencing Reveals a Complex Landscape of Modified tRNA Fragments.” *Nature Methods* 12 (9): 879–84.

Foissac, Sylvain, Sarah Djebali, Kylie Munyard, Nathalie Villa-Vialaneix, Andrea Rau, Kevin Muret, Diane Esquerre, et al. 2018. “Livestock Genome Annotation: Transcriptome and Chromatin Structure Profiling in Cattle, Goat, Chicken and Pig.” *bioRxiv*. https://doi.org/10.1101/316091.

Gardiner-Garden, M., and M. Frommer. 1987. “CpG Islands in Vertebrate Genomes.” *Journal of Molecular Biology* 196 (2): 261–82.

Goodarzi, Hani, Xuhang Liu, Hoang C. B. Nguyen, Steven Zhang, Lisa Fish, and Sohail F. Tavazoie. 2015. “Endogenous tRNA-Derived Fragments Suppress Breast Cancer Progression via YBX1 Displacement.” *Cell* 161 (4): 790–802.

Guyon, Isabelle, and André Elisseeff. 2003. “An Introduction to Variable and Feature Selection.” *Journal of Machine Learning Research: JMLR* 3 (Mar): 1157–82.

Hall, Mark A. 1998. “Correlation-Based Feature Subset Selection for Machine Learning.” *Thesis Submitted in Partial Fulfillment of the Requirements of the Degree of Doctor of Philosophy at the University of Waikato*. https://ci.nii.ac.jp/naid/10018668219/.

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. “The WEKA Data Mining Software: An Update.” *ACM SIGKDD Explorations Newsletter* 11 (1): 10–18.

Hamada, M., A. L. Sakulich, S. B. Koduru, and R. J. Maraia. 2000. “Transcription Termination by RNA Polymerase III in Fission Yeast. A Genetic and Biochemically Tractable Model System.” *The Journal of Biological Chemistry* 275 (37): 29076–81.

Hickey, Glenn, Benedict Paten, Dent Earl, Daniel Zerbino, and David Haussler. 2013. “HAL: A Hierarchical Format for Storing and Analyzing Multiple Genome Alignments.” *Bioinformatics* 29 (10): 1341–42.

Höhna, Sebastian, Michael J. Landis, Tracy A. Heath, Bastien Boussau, Nicolas Lartillot, Brian R. Moore, John P. Huelsenbeck, and Fredrik Ronquist. 2016. “RevBayes: Bayesian Phylogenetic Inference Using Graphical Models and an Interactive Model-Specification Language.” *Systematic Biology* 65 (4): 726–36.

Hubisz, Melissa J., Katherine S. Pollard, and Adam Siepel. 2011. “PHAST and RPHAST: Phylogenetic Analysis with Space/time Models.” *Briefings in Bioinformatics* 12 (1): 41–51.

Ide, Satoru, Takaaki Miyazaki, Hisaji Maki, and Takehiko Kobayashi. 2010. “Abundance of Ribosomal RNA Gene Copies Maintains Genome Integrity.” *Science* 327 (5966): 693–96.

Johansson, Linda, Guro Gafvelin, and Elias S. J. Arnér. 2005. “Selenocysteine in Proteins—properties and Biotechnological Use.” *Biochimica et Biophysica Acta (BBA) - General Subjects* 1726 (1): 1–13.

Jungreis, Irwin, Michael F. Lin, Rebecca Spokony, Clara S. Chan, Nicolas Negre, Alec Victorsen, Kevin P. White, and Manolis Kellis. 2011. “Evidence of Abundant Stop Codon Readthrough in Drosophila and Other Metazoa.” *Genome Research* 21 (12): 2096–2113.

Kirchner, Sebastian, and Zoya Ignatova. 2015. “Emerging Roles of tRNA in Adaptive Translation, Signalling Dynamics and Disease.” *Nature Reviews. Genetics* 16 (2): 98–112.

Koski, R. A., S. G. Clarkson, J. Kurjan, B. D. Hall, and M. Smith. 1980. “Mutations of the Yeast SUP4 tRNATyr Locus: Transcription of the Mutant Genes in Vitro.” *Cell* 22 (2 Pt 2): 415–25.

Krinner, Simone, Asli P. Heitzer, Sarah D. Diermeier, Ingrid Obermeier, Gernot Längst, and Ralf Wagner. 2014. “CpG Domains Downstream of TSSs Promote High Levels of Gene Expression.” *Nucleic Acids Research* 42 (6): 3551–64.

Kumar, Sudhir, Glen Stecher, Michael Suleski, and S. Blair Hedges. 2017. “TimeTree: A Resource for Timelines, Timetrees, and Divergence Times.” *Molecular Biology and Evolution* 34 (7): 1812–19.

Kutter, Claudia, Gordon D. Brown, Angela Gonçalves, Michael D. Wilson, Stephen Watt, Alvis Brazma, Robert J. White, and Duncan T. Odom. 2011. “Pol III Binding in Six Mammals Shows Conservation among Amino Acid Isotypes despite Divergence among tRNA Genes.” *Nature Genetics* 43 (10): 948–55.

Li, W. H., D. L. Ellsworth, J. Krushkal, B. H. Chang, and D. Hewett-Emmett. 1996. “Rates of Nucleotide Substitution in Primates and Rodents and the Generation-Time Effect Hypothesis.” *Molecular Phylogenetics and Evolution* 5 (1): 182–87.

Lorenz, Ronny, Stephan H. Bernhart, Christian Höner Zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F. Stadler, and Ivo L. Hofacker. 2011. “ViennaRNA Package 2.0.” *Algorithms for Molecular Biology: AMB* 6 (November): 26.

Loughran, Gary, Irwin Jungreis, Ioanna Tzani, Michael Power, Ruslan I. Dmitriev, Ivaylo P. Ivanov, Manolis Kellis, and John F. Atkins. 2018. “Stop Codon Readthrough Generates a C-Terminally Extended Variant of the Human Vitamin D Receptor with Reduced Calcitriol Response.” *The Journal of Biological Chemistry* 293 (12): 4434–44.

Maraia, R. J., D. J. Kenan, and J. D. Keene. 1994. “Eukaryotic Transcription Termination Factor La Mediates Transcript Release and Facilitates Reinitiation by RNA Polymerase III.” *Molecular and Cellular Biology* 14 (3): 2147–58.

Meunier, Julien, Frédéric Lemoine, Magali Soumillon, Angélica Liechti, Manuela Weier, Katerina Guschanski, Haiyang Hu, Philipp Khaitovich, and Henrik Kaessmann. 2013. “Birth and Expression Evolution of Mammalian microRNA Genes.” *Genome Research* 23 (1): 34–45.

Mleczko, Anna M., Piotr Celichowski, and Kamilla Bąkowska-Żywicka. 2014. “Ex-Translational Function of tRNAs and Their Fragments in Cancer.” *Acta Biochimica Polonica* 61 (2): 211–16.

Necsulea, Anamaria, and Henrik Kaessmann. 2014. “Evolutionary Dynamics of Coding and Non-Coding Transcriptomes.” *Nature Reviews. Genetics* 15 (11): 734–48.

Necsulea, Anamaria, Magali Soumillon, Maria Warnefors, Angélica Liechti, Tasman Daish, Ulrich Zeller, Julie C. Baker, Frank Grützner, and Henrik Kaessmann. 2014. “The Evolution of lncRNA Repertoires and Expression Patterns in Tetrapods.” *Nature* 505 (7485): 635–40.

Nguyen, Ngan, Glenn Hickey, Daniel R. Zerbino, Brian Raney, Dent Earl, Joel Armstrong, W. James Kent, David Haussler, and Benedict Paten. 2015. “Building a Pan-Genome Reference for a Population.” *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 22 (5): 387–401.

Orioli, Andrea, Chiara Pascali, Jade Quartararo, Kevin W. Diebel, Viviane Praz, David Romascano, Riccardo Percudani, et al. 2011. “Widespread Occurrence of Non-Canonical Transcription Termination by Human RNA Polymerase III.” *Nucleic Acids Research* 39 (13): 5499–5512.

Palazzo, Alexander F., and Eliza S. Lee. 2015. “Non-Coding RNA: What Is Functional and What Is Junk?” *Frontiers in Genetics* 6 (January): 2.

Pan, Tao. 2018. “Modifications and Functional Genomics of Human Transfer RNA.” *Cell Research* 28 (4): 395–404.

Paten, Benedict, Mark Diekhans, Dent Earl, John St John, Jian Ma, Bernard Suh, and David Haussler. 2011. “Cactus Graphs for Genome Comparisons.” *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 18 (3): 469–81.

Paten, Benedict, Dent Earl, Ngan Nguyen, Mark Diekhans, Daniel Zerbino, and David Haussler. 2011. “Cactus: Algorithms for Genome Multiple Sequence Alignment.” *Genome Research* 21 (9): 1512–28.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. “Scikit-Learn: Machine Learning in Python.” *Journal of Machine Learning Research: JMLR* 12 (Oct): 2825–30.

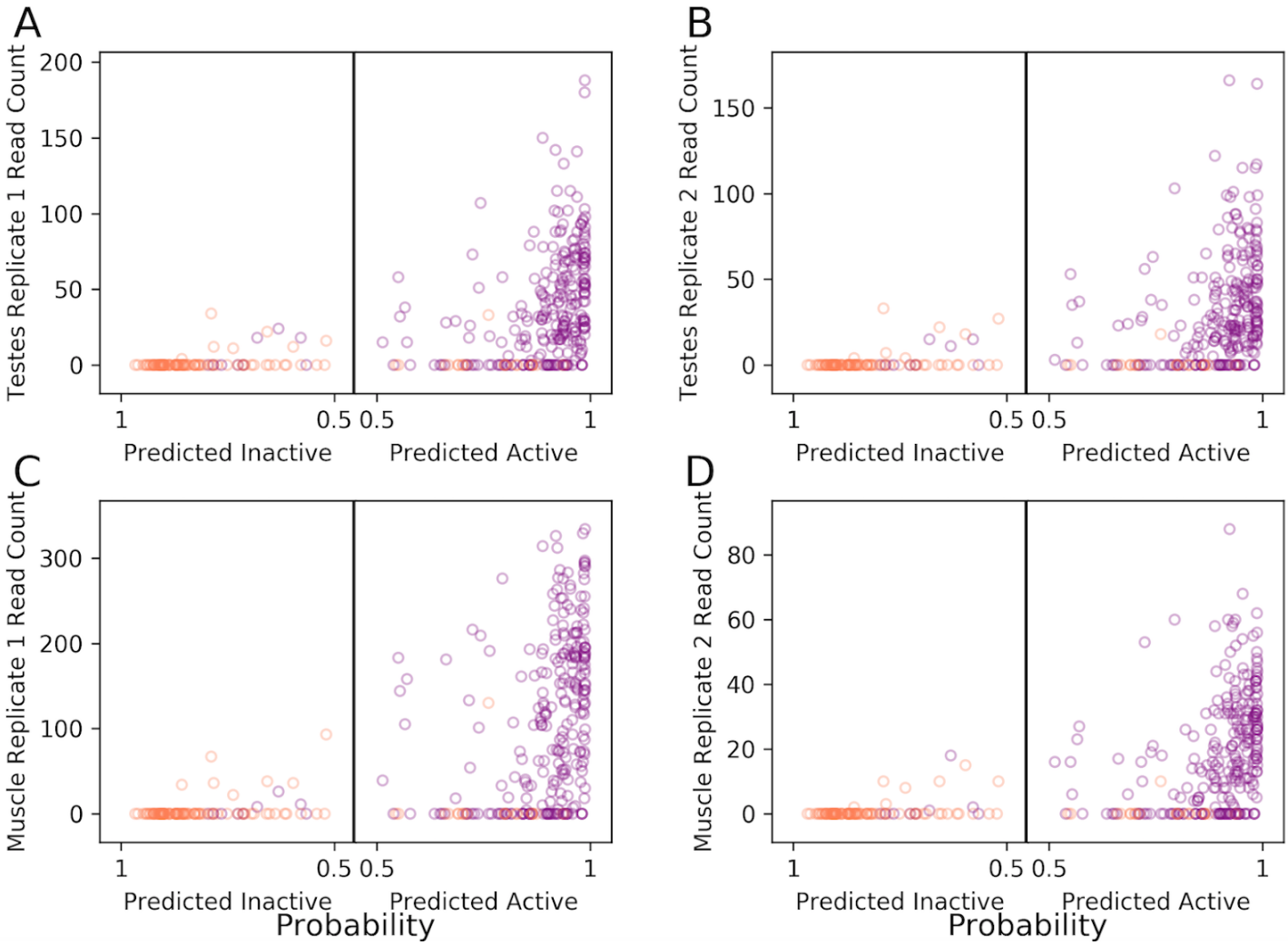
Raab, Jesse R., Jonathan Chiu, Jingchun Zhu, Sol Katzman, Sreenivasulu Kurukuti, Paul A. Wade, David Haussler, and Rohinton T. Kamakaka. 2012. “Human tRNA Genes Function as Chromatin Insulators.” *The EMBO Journal* 31 (2): 330–50.

Roadmap Epigenomics Consortium, Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, et al. 2015. “Integrative Analysis of 111 Reference Human Epigenomes.”

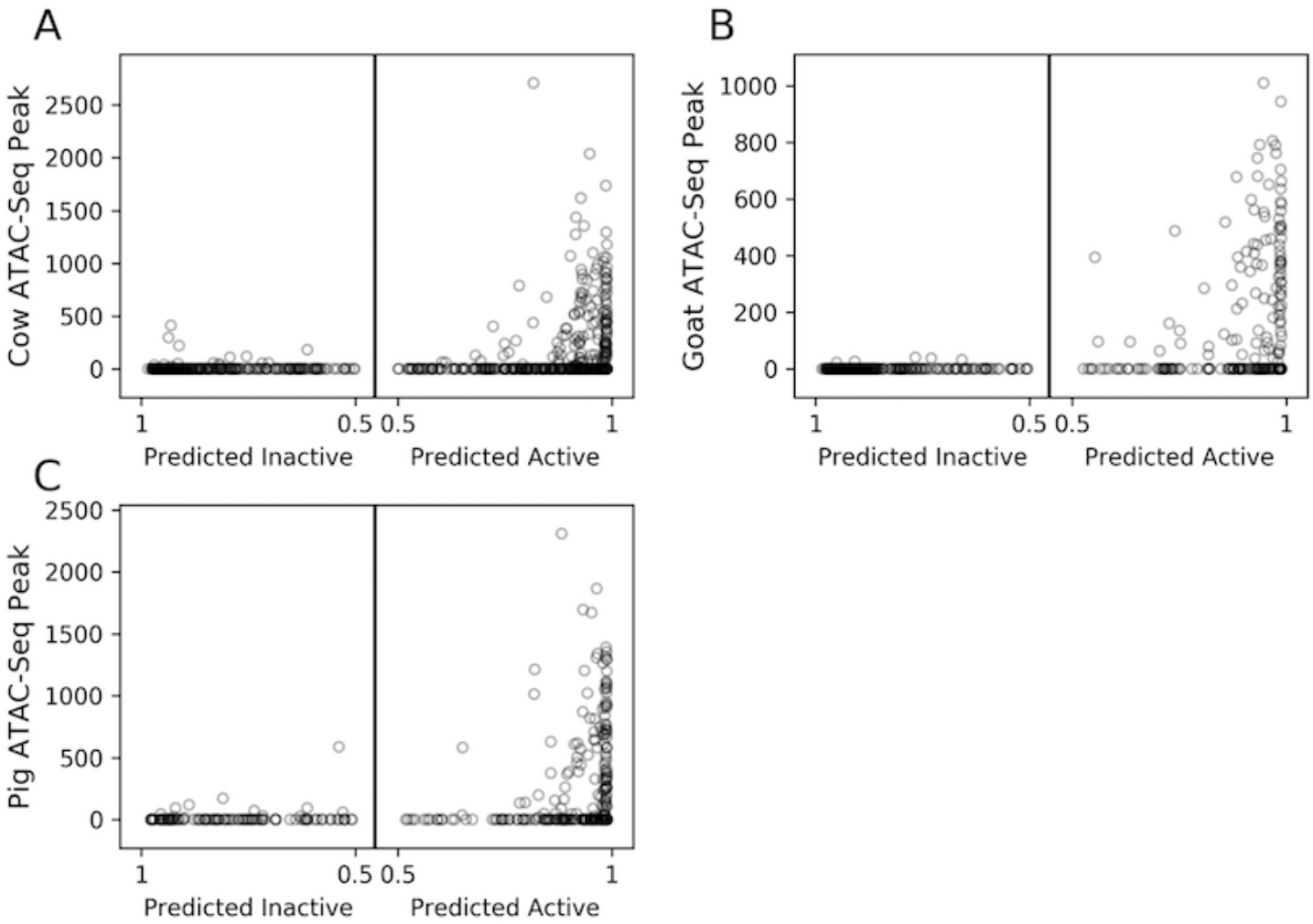
Nature 518 (7539): 317–30.
Rogers, Hubert H., Casey M. Bergman, and Sam Griffiths-Jones. 2010. “The Evolution of tRNA Genes in Drosophila.” *Genome Biology and Evolution* 2 (July): 467–77.
Roy, Bijoyita, John D. Leszyk, David A. Mangus, and Allan Jacobson. 2015. “Nonsense Suppression by near-Cognate tRNAs Employs Alternative Base Pairing at Codon Positions 1 and 3.” *Proceedings of the National Academy of Sciences of the United States of America* 112 (10): 3038–43.
Schmitt, Bianca M., Konrad L. M. Rudolph, Panagiota Karagianni, Nuno A. Fonseca, Robert J. White, Iannis Talianidis, Duncan T. Odom, John C. Marioni, and Claudia Kutter. 2014. “High-Resolution Mapping of Transcriptional Dynamics across Tissue Development Reveals a Stable mRNA–tRNA Interface.” *Genome Research* 24 (11): 1797–1807.
Strobl, Carolin, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. 2008. “Conditional Variable Importance for Random Forests.” *BMC Bioinformatics* 9 (July): 307.
Sun, Chunxiao, Ziyi Fu, Siwei Wang, Jun Li, Yongfei Li, Yanhong Zhang, Fan Yang, et al. 2018. “Roles of tRNA-Derived Fragments in Human Cancers.” *Cancer Letters* 414 (February): 16–25.
Tang, Dave T. P., Evgeny A. Glazov, Sean M. McWilliam, Wesley C. Barris, and Brian P. Dalrymple. 2009. “Analysis of the Complement and Molecular Evolution of tRNA Genes in Cow.” *BMC Genomics* 10 (April): 188.
Thornlow, Bryan, Josh Hough, Jackie Roger, Henry Gong, Todd Lowe, and Russell Corbett-Detig. 2018. “Transfer RNA Genes Experience Exceptionally Elevated Mutation Rates.” *Proceedings of the National Academy of Sciences*. https://doi.org/10.1101/229906.
Tolosi, Laura, and Thomas Lengauer. 2011. “Classification with Correlated Features: Unreliability of Feature Ranking and Solutions.” *Bioinformatics* 27 (14): 1986–94.
Valle, R. P., M. D. Morch, and A. L. Haenni. 1987. “Novel Amber Suppressor tRNAs of Mammalian Origin.” *The EMBO Journal* 6 (10): 3049–55.
Yacoubi, Basma El, Basma El Yacoubi, Marc Bailly, and Valérie de Crécy-Lagard. 2012. “Biosynthesis and Function of Posttranscriptional Modifications of Transfer RNAs.” *Annual Review of Genetics*. https://doi.org/10.1146/annurev-genet-110711-155641.
Zheng, Guanqun, Yidan Qin, Wesley C. Clark, Qing Dai, Chengqi Yi, Chuan He, Alan M. Lambowitz, and Tao Pan. 2015. “Efficient and Quantitative High-Throughput tRNA Sequencing.” *Nature Methods* 12 (9): 835–37.

	Feature
Extrinsic	**Average PhyloP Score in tRNA Gene Sequence**
	Average PhyloP Score in 3' Flanking Region
	Average PhyloP Score in 5' Flanking Region
	CpG Density Across tRNA Locus
	Distance To Nearest Exon
	Exons Within 2 Megabases
	Exons Within 75 Kilobases
	Minimum PhyloP Score in tRNA Gene Sequence
	Minimum PhyloP Score in 5' Flanking Region
	Minimum PhyloP Score in 3' Flanking Region
	Observed/Expected CpG Islands Score Across tRNA Locus
	Observed/Expected CpG Islands Score Downstream of tRNA Gene
	Observed/Expected CpG Islands Score Upstream of tRNA Gene
	Percentage of CpG Dinucleotides Upstream of tRNA Gene
	Substitutions in tRNA Gene Sequence Relative to Outgroup
	Substitutions in 3' Flanking Sequence Relative to Outgroup
	Substitutions in 5' Flanking Sequence Relative to Outgroup
	tRNA Genes Within 10 Kilobases
	tRNA Genes Within 400 Kilobases
Intrinsic	**Canonical tRNA Secondary Structure**
	Distance To Nearest Transcription Termination Sequence of Any Kind
	Distance To Nearest TTTT Transcription Termination Sequence
	Distance To Nearest TTTTT Transcription Termination Sequence
	tRNA Genes Corresponding to the Same Amino Acid
	tRNA Genes Corresponding to the Same Codon
	tRNAscan-SE General Bit Score
	tRNAscan-SE HMM Bit Score
	tRNAscan-SE Isotype-Specific Bit Score
	tRNAscan-SE Secondary Structure Bit Score
	tRNA Sequence Unique In Genome

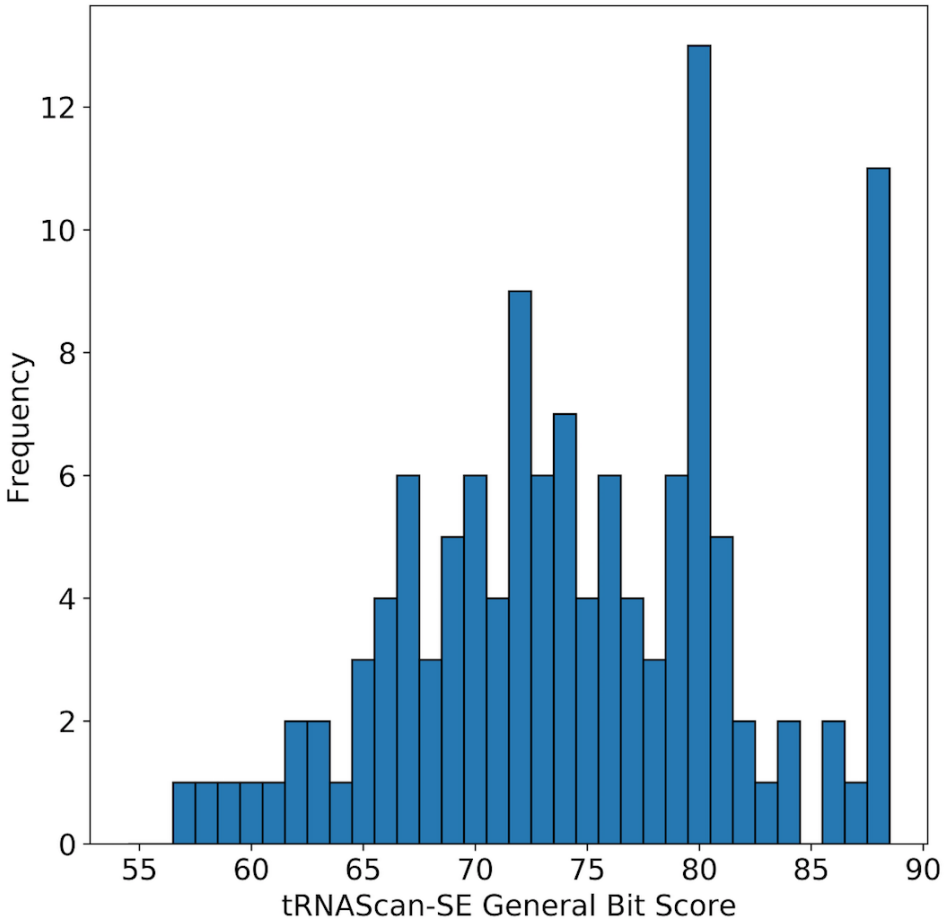
Supplementary Table 1: Our initial model used many features both intrinsic to tRNA structure and function as well as extrinsic and derived solely from the genomic context. All features considered for the model are shown, with those in the final model bolded. Features were removed either for lack of contribution to the model, overlapping contribution with other features, or difficulty in obtaining data due to lack of alignment. "Transcription Termination Sequence of Any Kind" refers to "TTTT", "TTTTT", as well as alternative sequences demonstrated in Orioli et al. (2011).



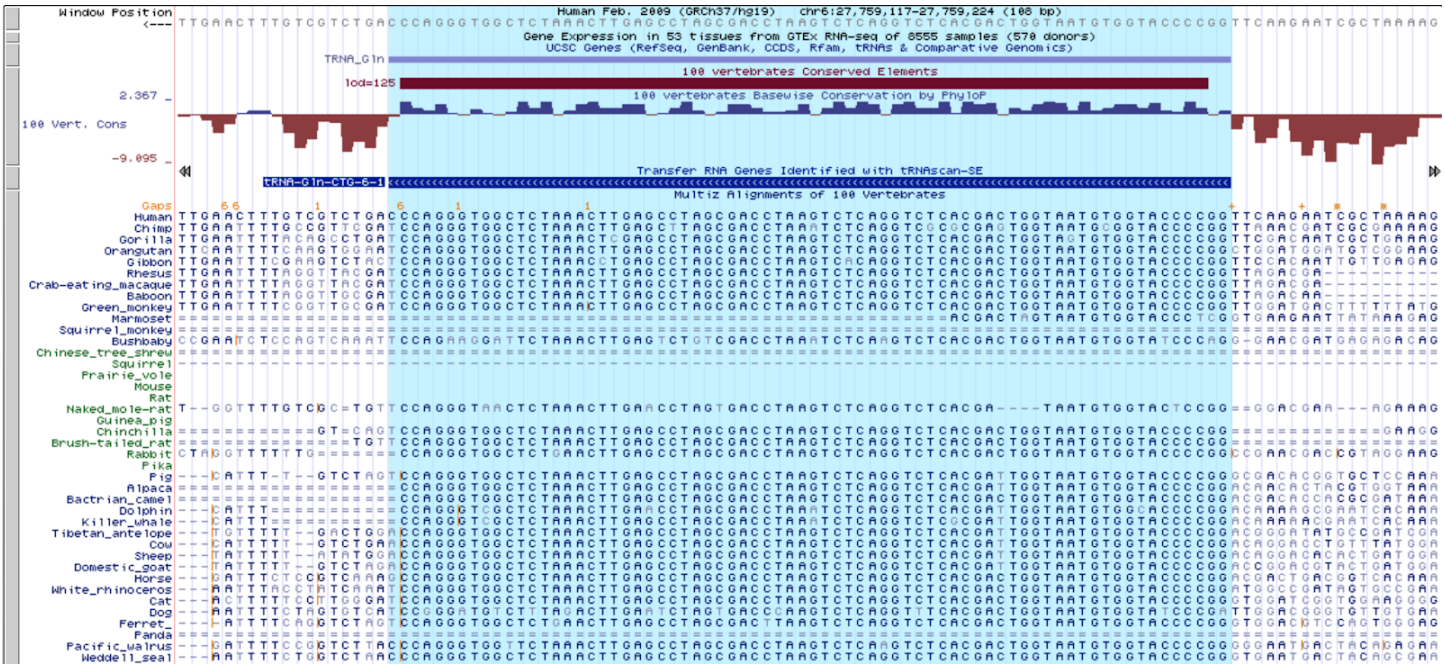
Supplementary Figure 1: Comparison of tRNA gene classifications to ChIP-Seq data is consistent across tissues. Mouse ChIP-Seq data compared to tRNA gene classifications and ChIP data for both testes replicates and both muscle replicates. Probability scores output by the classifier are shown on the horizontal axis where tRNA genes furthest left are predicted inactive with greater probability and tRNA genes furthest right are predicted active with greater probability. The vertical axis indicates read counts in testes (A, B) and muscle (C, D) from Kutter et al. (2011). Purple indicates measured activity in at least one tissue from Bogu et al. (2015) while orange indicates no measured activity.



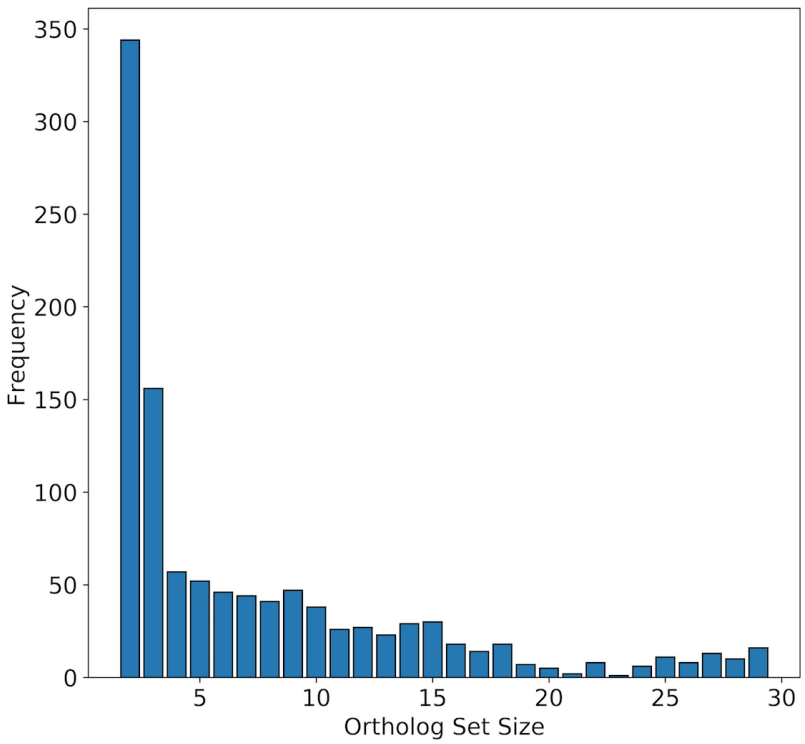
Supplementary Figure 2: Comparison of classifications to ATAC-Seq data suggests similar accuracy across clades. The mean ATAC-Seq peaks across liver, CD4 and CD8 cells in cow (A), goat (B) and pig (C) within 250 base pairs of each tRNA gene are shown against the classification probabilities for each gene as in Supplementary Figure 1.



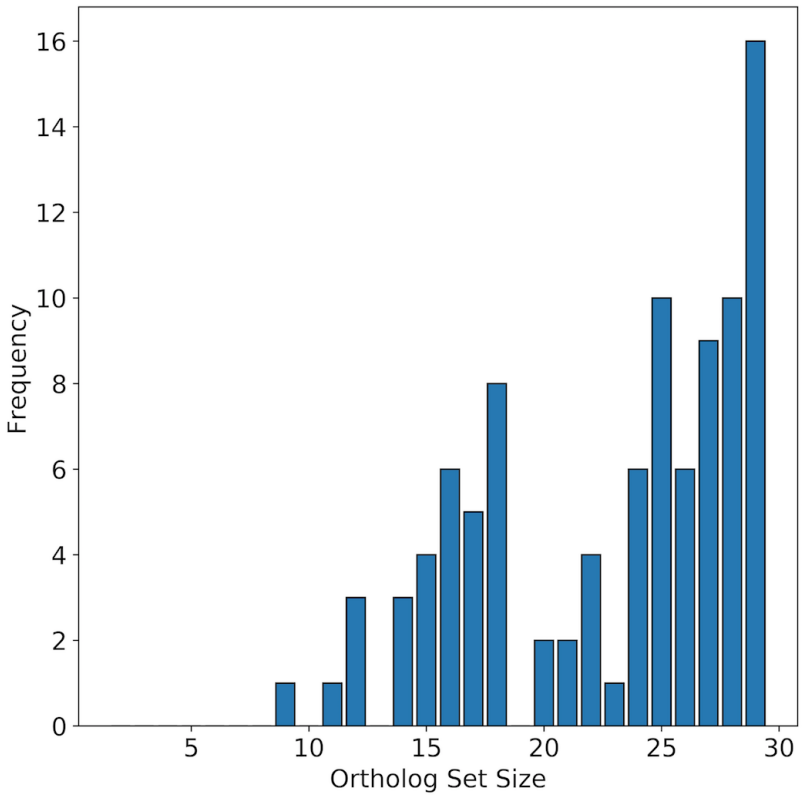
Supplementary Figure 3: Most *Erinaceus europaeus* tRNA-Lys-CTT genes have high tRNAScan-SE general bit scores. Histogram showing the number of tRNA-Lys-CTT genes in the *E. europaeus* genome by tRNAScan-SE general bit score. For reference, the average active human tRNA gene has a bit score of 76.2 and the average inactive human tRNA gene has a bit score of 69.3.



Supplementary Figure 4: Chimpanzee ortholog of human tRNA-Gln-CTG-6-1 may be an actively transcribed suppressor tRNA gene. Gene browser screenshot showing the PhyloP and Multiz Alignment tracks of the human tRNA-Gln-CTG-6-1 gene, for which the chimpanzee genome has a tRNA-Sup-CTA gene. The sequence of the tRNA gene is highlighted in cyan. The chimpanzee sequence (second line from top) exhibits several unique substitutions. Most notably, the CTG anticodon (here from right to left in the center of the highlighted region) has changed to CTA in chimp, accompanied by other changes to the body of the tRNA. The PhyloP track also shows strong divergence upstream and downstream of the tRNA gene sequence, consistent with high levels of expression.



Supplementary Figure 5: tRNA genes are generally either fairly deeply conserved or recently evolved. Size distribution of all ortholog sets, with genes present in at least 2 species. 3,582 of the 11,752 tRNA genes in our alignment are species-specific, excluded from this figure. The remaining 8,170 tRNA genes are condensed into 1,097 ortholog sets.



Supplementary Figure 6: Core tRNA genes are very deeply conserved. Histogram of sizes of the 97 tRNA gene ortholog sets present in at least all 7 primate species.

A

Initial \ Final	Missing	Inactive	Active
Missing	.871	.111	.017
Inactive	.092	.859	.049
Active	.012	.014	.974

B

Initial \ Final	Missing	Inactive	Active
Missing	.403	.341	.256
Inactive	.116	.845	.040
Active	.144	.419	.437

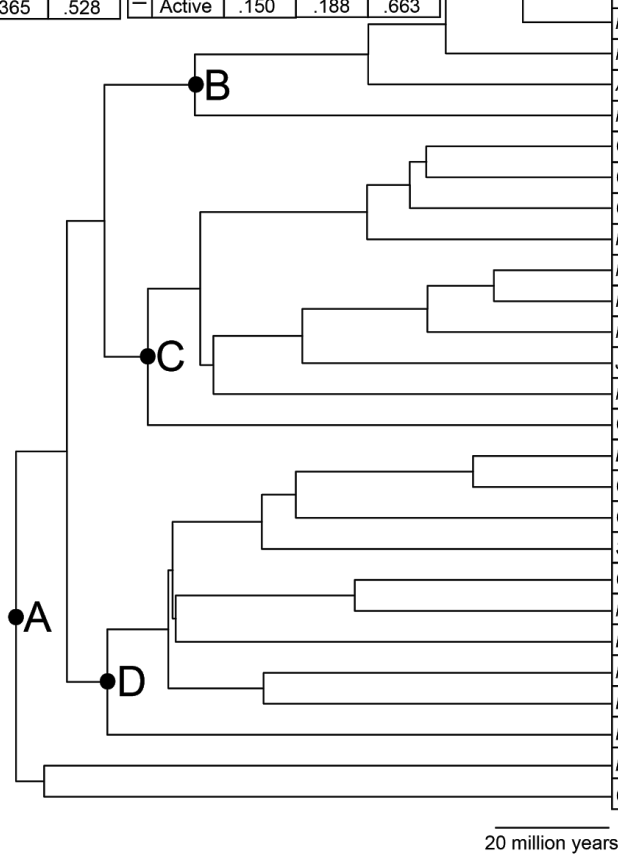
C

Initial \ Final	Missing	Inactive	Active
Missing	.723	.229	.049
Inactive	.163	.656	.182
Active	.106	.365	.528

D

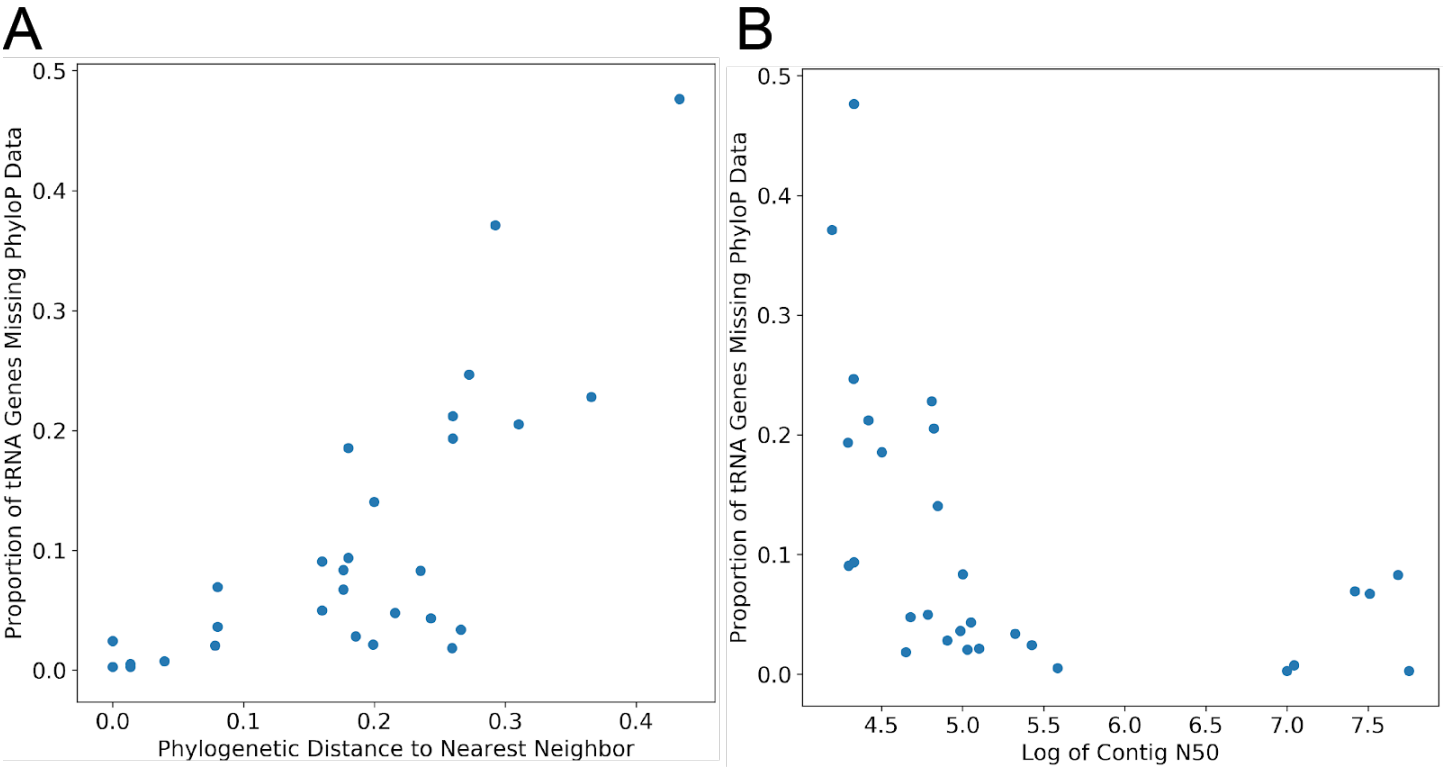
Initial \ Final	Missing	Inactive	Active
Missing	.957	.022	.021
Inactive	.165	.821	.015
Active	.150	.188	.663

E

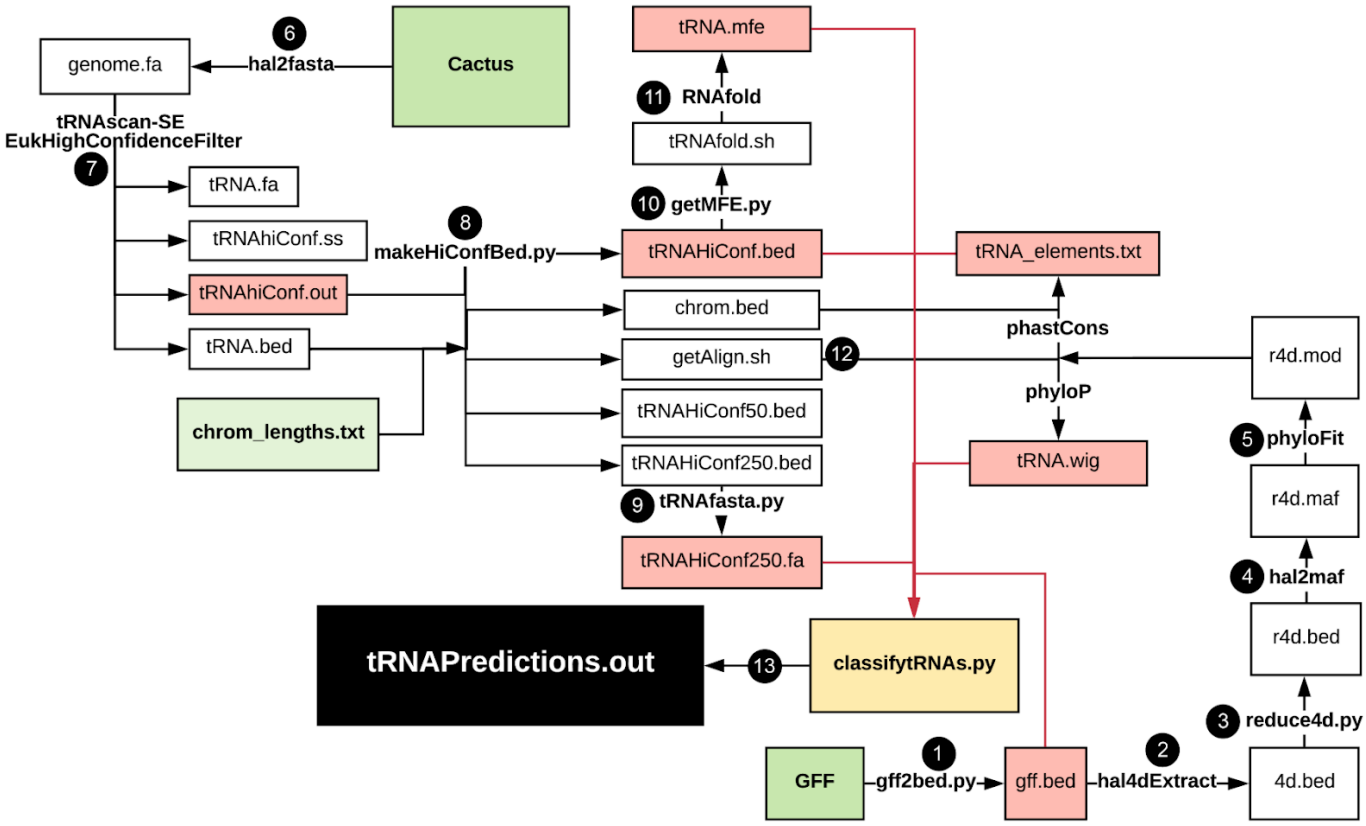


Species	Active	Inactive
Homo sapiens	314	57
Pan troglodytes	339	56
Gorilla gorilla gorilla	325	46
Pongo abelii	351	50
Macaca mulatta	337	54
Aotus nancymaae	285	43
Microcebus murinus	284	40
Octodon degus	240	58
Chinchilla lanigera	250	51
Cavia porcellus	266	54
Heterocephalus glaber	238	56
Mus musculus	275	112
Rattus norvegicus	245	102
Microtus ochrogaster	343	127
Jaculus jaculus	246	72
Marmota marmota	253	156
Oryctolagus cuniculus	377	70
Bos taurus	406	369
Capra hircus	216	317
Orcinus orca	328	333
Sus scrofa	294	104
Canis lupus familiaris	278	90
Mustela putorius furo	291	87
Equus caballus	351	65
Pteropus alecto	222	123
Eptesicus fuscus	199	15
Erinaceus europaeus	343	104
Dasypus novemcinctus	379	149
Chrysochloris asiatica	275	97

Supplementary Figure 7: Clades vary in their activity state transition rates. Transition matrices shown as in Figure 3 for all descendants of labeled nodes (A-D). Phylogeny, species names and predicted active and inactive tRNA gene counts reproduced from Figure 3 (E).



Supplementary Figure 8: Unalignable tRNA genes are likely due to a combination of decreased genome quality and increased phylogenetic distance. The proportion of tRNA genes in a given genome is shown compared to the phylogenetic distance to the nearest neighbor in the phylogeny (A), as well as the log of the Contig N50 for that species (B). Branch length in the phylogeny is measured by expected number of substitutions between species in neutral regions.



Supplementary Figure 9: Classification pipeline uses tRNAscan-SE, HAL toolkit and PHAST. Workflow for classifying tRNAs with instructions for usage found at https://github.com/bpt26/tRNA_classifier. Files outlined in red are the inputs used by classifytRNAs.py, the program responsible for producing the final predictions.