Chapter 3. Distinguishing functional ncRNAs from pseudogenes in mammalian genomes

The results presented in the previous chapter (chapter 2) suggest that many Rfam human ncRNAs appear to be synteny-non-conserved in the mammalian genome after the primate-rodent split. When considering using comparative methods for genome-wide ncRNA finding, one important question is whether synteny-non-conserved ncRNAs tend to be functional genes or pseudogenes. If a considerable proportion of synteny-non-conserved ncRNAs in the genomes under investigation are functional, the strategies that predict ncRNAs only in the alignments of syntenic regions will fail to predict those functional ncRNAs. Conversely, if most synteny-non-conserved ncRNAs are pseudogenes, methods that depend on alignments derived from synteny may be sufficient for genome-wide ncRNA finding.

Before exploring the likelihood of synteny-non-conserved ncRNAs to be pseudogenes, it is necessary to briefly introduce how pseudogenes might be generated, and how they can be computationally identified. Pseudogenes are believed to be generated by either genome duplication or retrotransposition, followed by non-functionalization of a subset of the duplicated copies (for review see Lynch and Conery 2000). The mechanisms that may lead to genome duplications include unequal crossing-over (for review see Graur and Li 2000), and duplication of a segmental (Gu et al. 2002) or entire chromosome (Van de Peer 2004; Dehal and Boore 2005). In so-called retrotransposition, which is a RNA-mediated process, the RNA transcript of a gene is reverse transcribed into DNA, which is then inserted back into the genome at a new location (Maestre et al. 1995). The pseudogenes that are generated through retrotransposition have usually lost the original gene's intron-exon architecture and thus are often referred to as processed pseudogenes, while the pseudogenes generated through duplications of genomic DNA are referred to as non-processed pseudogenes.

Currently, pseudogenes can be computationally identified by searching protein coding genes for indicators of non-functionality. For instance, a duplicated protein pseudogene can be evolutionarily unconstrained, and hence have accumulated random mutations that may destroy its protein gene-like features; a retrotransposed protein pseudogene can completely lose introns (Figure 3-1 A). Several surveys already performed for exploring pseudogenes in the human genome were based on indicators of functionality derived from features of multi-exon protein coding genes (Ohshima et al. 2003; Torrents et al. 2003; Zhang et al. 2003). In particular, by using the ratio of silent to replacement nucleotide substitutions (K_A/K_S), Torrents *et al*. discovered ~20,000 protein pseudogenes in the human genome, where as many as 70% of them were retrotransposed (Torrents et al. 2003). These results, together with the estimate that ~96% of the human protein genes are mouse-synteny-conserved (Mouse Genome Sequencing Consortium 2002), suggest that a protein coding gene sequence that is synteny-non-conserved in mammalian genomes is very likely to be a pseudogene.

However, since the surveys mentioned above were limited to investigating protein pseudogenes, the tendency of synteny-non-conserved ncRNAs to be pseudogenes is unknown. To date, the functionality of the synteny-non-conserved ncRNAs in mammalian genomes has not been systematically investigated. One reason for this is that in mammalian genomes there are abundant ncRNA-derived short interspersed repetitive elements (SINEs) (International Human Genome Sequencing Consortium 2001; Mouse Genome Sequencing Consortium 2002) which make the determination of ncRNA pseudogenes difficult. SINEs are repetitive elements that are amplified in the genomes through retrotransposition (for review see Smit 1999). Most eukaryotic SINEs have evolved from the ncRNAs that are transcribed by RNA polymerase III. Known evolutionary sources of eukaryotic SINEs include tRNA genes, 7SL genes, 5S rRNA genes (for review see Kramerov and Vassetzky 2005). With respect to ncRNA pseudogene identification some of the SINEs in mammalian genomes are so similar, at both the

primary-sequence and structural levels, to functional ncRNAs that even well tuned ncRNA finding algorithms may falsely predict them as real ones. For instance, about 2,700 tRNA genes, which is more than five times of the tRNA genes annotated in the human genome, were initially predicted in the mouse genome (Mouse Genome Sequencing Consortium 2002). In order to generate a smaller, but more confident, set of functional mouse tRNA genes, the Mouse Genome Consortium has used an additional criterion, non-overlapping with the SINEs identified by RepeatMasker (Smit and Green unpublished), to filter the initial prediction. However, there are at least two considerations with such a criterion. First, it may be too arbitrary to hypothesize that all SINEs are pseudogenes. Second, ncRNA pseudogenes that are unrelated to SINEs can not be filtered out. The above case about filtering out tRNA pseudogenes illustrates the difficulty of distinguishing functional ncRNAs from pseudogenes.

It is possible that some synteny-non-conserved ncRNAs are functional genes. Firstly, a synteny-non-conserved ncRNA might be functional and originally synteny-conserved, but has been deleted in the other lineage. Secondly, a synteny-non-conserved ncRNA may be a functional gene as a result of mechanisms creating a functional copy. Perhaps, due to unique features of certain types of ncRNAs, there is a high tendency for these genes to be synteny-non-conserved in mammalian genomes. One argument is that the mechanisms that generate protein pseudogenes may generate synteny-non-conserved but functional ncRNAs, in addition to ncRNA pseudogenes. While a mechanism of pseudogenisation may effectively cause a newly amplified protein gene to lose the association with its upstream regulatory regions, the same mechanism may not necessarily cause the nonfunctionality of a recently amplified ncRNA locus in the genome.

Retrotransposition appears to be one possible mechanism that can lead to the generation of protein pseudogenes, but new and functional ncRNA loci. Since the transcription regulatory elements in the 5’ flanking regions of the protein genes are not contained in mRNA transcripts,

a retrotransposed protein gene, even if it has retained part of the intron-exon structure, should generally be untranscribable. Therefore, a retrotransposed protein gene may become a pseudogene as soon as the redundant sequence is generated (Figure 3-1 A). Conversely, a retrotransposed ncRNA that is not truncated may remain transcribable, if its intragenic promoters are still intact during the process of generating this redundant copy (Figure 3-1 B).

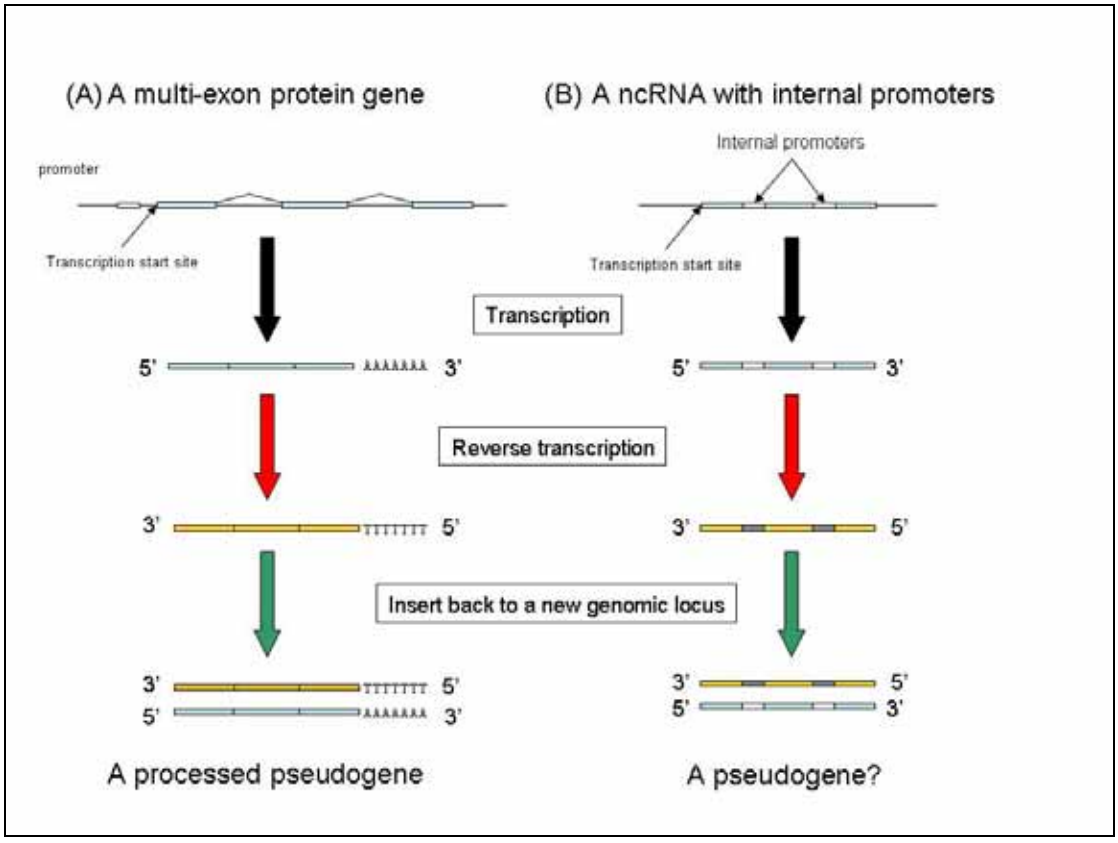

Figure 3-1. Comparison of the gene structures of a retrotransposed protein gene and a hypothetical retrotransposed ncRNA that contain internal promoters.

Therefore, this chapter is dedicated to distinguishing functional ncRNAs from ncRNA pseudogenes in the context of genomic sequences. There are two purposes in this chapter:

- To explore whether human synteny-non-conserved ncRNAs tend to be pseudogenes
- To evaluate novel rules that may be useful for distinguishing functional ncRNAs from ncRNA pseudogenes

Mammalian tRNA genes were chosen for further investigation. One reason for this decision is that many features of functional tRNA genes have been well studied. For example, a tRNA molecule can fold into a cloverleaf-like secondary structure; tRNA genes have internal promoters, which consist of A and B boxes (DeFranco et al. 1980); mammalian tRNA genes tend to cluster in the genomes (Lasser-Weiss et al. 1981). It was therefore hoped that, by integrating the information of sequence similarity, anticodon types, clustering, *etc*., evidence might possibly be found to determine if synteny-non-conserved tRNA genes in the mammalian genomes tend to be pseudogenes.

In the first part of this chapter (section 3.1), I investigate whether the human synteny-non-conserved tRNA genes that were retrieved from Rfam tend to be pseudogenes. The conservation of secondary structures and conservation of promoters, as well as conservation of primary sequences, were used to infer the functionality of the human synteny-non-conserved tRNA genes. The idea is that, if certain tRNA genes are pseudogenes, their sequences may have accumulated mutations which may change the features important for the functionality of tRNAs. The specific questions I address here include:

- Is there a clear-cut difference between the bit-score distributions of synteny-non-conserved tRNA genes and synteny-conserved tRNA genes?
- Do synteny-non-conserved tRNA genes tend to have more unstable structural features than synteny-conserved tRNA genes do?
- Do synteny-non-conserved tRNA genes tend to have degraded internal promoters?

A particular property of tRNA genes is that they frequently exist in synteny conserved clusters, as examined in chapter 2. In the second part of this chapter, I explore whether properties of copies of tRNA genes that are clustered and copies that are un-clustering are different and whether there is any evidence that can relate this to the likelihood of being

pseudogenes. Clustering seems to be an effective strategy to ensure each transcription unit can be accessed with generally equal probability by transcription machinery. Evidence suggests that clustering is important for regulating expression of ncRNAs. It has been demonstrated that clustered miRNA genes tend to be co-expressed (Baskerville and Bartel 2005). Besides, a cluster of 40 miRNA genes has been found in the human imprinted 14q32 domain and only the maternally inherited genes are expressed (Seitz et al. 2004).

I therefore hypothesized that non-clustered tRNA genes tend to be pseudogenes. Two tests were therefore designed to evaluate this hypothesis:

- Is there an enrichment of non-clustered tRNA genes in the low-scoring group which are more likely to be pseudogenes?
- Are clustered tRNA genes sufficient for covering 46 types of anticodons that are necessary for protein translation? If so, this would be evidence that non-clustered tRNA genes are not absolutely required for protein translation, supporting hypothesis that they could be pseudogenes.

3.1. Are Rfam synteny-non-conserved tRNA genes functional?

3.1.1. Materials and methods

The coordinates of human and mouse tRNA genes were retrieved from RFAMSEQ of Rfam 4.1 (Griffiths-Jones et al. 2003) and then converted to chromosomal coordinates in the human and mouse genomes respectively. The reference genome assemblies are human NCBI 33 and mouse NCBI M30. The bit scores of the Rfam tRNA genes were calculated using Infernal and the tRNA covariance model (CM) of Rfam 4.1 (Griffiths-Jones et al. 2003). The

human tRNA genes predicted using tRNAscanSE were retrieved from Ensembl release 19 by using the Ensembl Perl APIs (Birney et al. 2004).

In order to compare the bit-score distributions of the Rfam tRNA genes and the tRNAscanSE-predicted tRNA genes with that of *bona fide* tRNA genes, a trusted set of functional tRNA genes from the human genome is required. However, only a few experimentally verified human tRNA genes are available (Sprinzl and Vassilenko 2005). One consideration is that the bit-score distribution of a small number of tRNA genes may be biased and thus unsuitable for use as the reference distribution. Therefore, I decided to recruit Rfam tRNA genes that are human-mouse synteny-conserved as a trusted set of functional tRNA genes. Since synteny conservation has been widely accepted as a strong indication for the existence of functional elements, the human-mouse synteny-conserved tRNA genes are very likely to be functional tRNA genes. The sequences of these tRNA genes were prepared using the Ensembl Compara Perl APIs to search syntenic regions identified by Ensembl Compara release 19 (Clamp et al. 2003).

The preservation of structural features of tRNA genes was evaluated by using Infernal to align these sequences to Rfam tRNA CM. For the purpose of checking the conservation of the internal promoters in these tRNA genes, eufindtRNA (Pavesi et al. 1994) was used (for a brief introduction of Infernal and eufindtRNA, see materials and methods, section 2.1, chapter 2).

3.1.2. Results

3.1.2.1. Distribution of the Rfam bit scores of tRNA genes

808 human and 452 mouse tRNA genes were retrieved from Rfam (release 4.1). At first glimpse, it seems that there are more tRNA genes in the human genome than in the mouse genome; however, a substantial portion of the mouse genome assembly NCBI M30 is composed of sequences from whole genome shotgun sequencing, which has not been scanned

by Rfam 4.1. The number of the tRNA genes in the mouse genome is therefore an underestimate.

Interestingly, both the bit-score distributions of the human and the mouse tRNA genes sequences are bimodal (Figure 3-2, see "Rfam-human" and "Rfam-mouse" respectively). The bimodal bit-score distribution of the human tRNA genes seems to consist of two well-shaped distributions, which have modes at 65 and at 30 respectively.

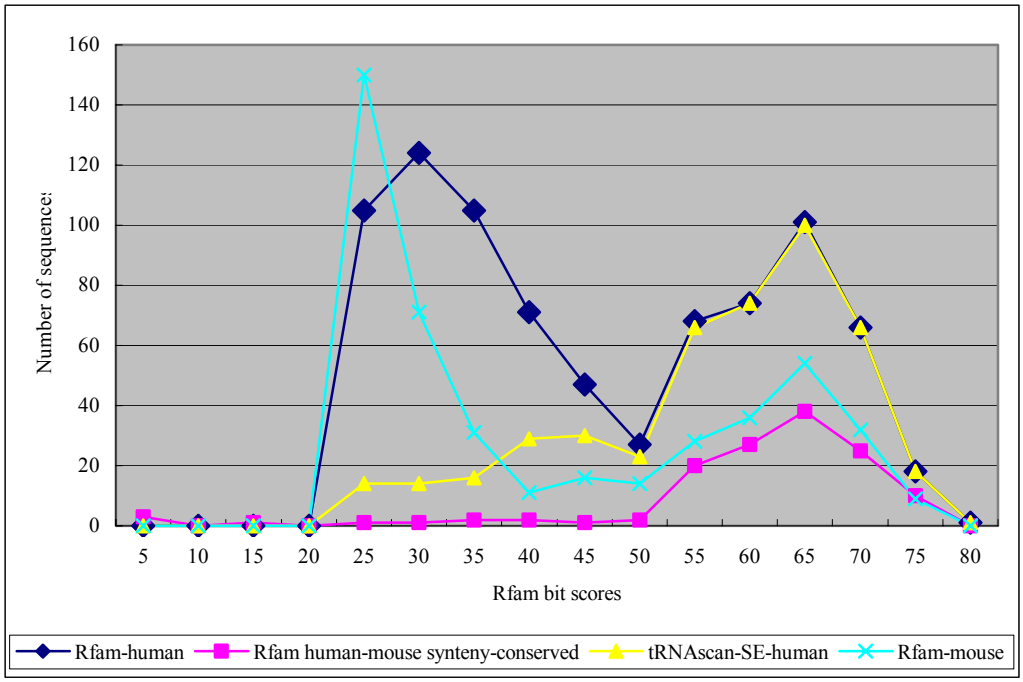

Figure 3-2. Distributions of Rfam bit scores of tRNA genes of different categories

The bin size of Rfam bit scores is 5. Almost no tRNA genes (except the human numt-tRNAs) have bit scores less than 25 because Rfam has used 25 bits as the gathering threshold for tRNA genes.

One interpretation of these results is that the bimodal distribution represents two groups of evolutionarily distinct tRNA genes. This idea is supported by the similarity between the high-scoring part of this bimodal distribution and the bit-score distributions of other sets of tRNA sequences. For example, the contour of the bit-score distribution of the tRNAscanSE-predicted human tRNA genes (Figure 3-2, "tRNAscanSE-human") is very similar to the high-scoring part of the bimodal bit-score distribution. In addition, the bit-score

distribution of the trusted set of *bona fide* tRNA genes (Figure 3-2, "Rfam human-mouse synteny-conserved") is also very similar. Only 9% (12/133) of the trusted *bona fide* tRNA genes have bit scores lower than 50. This comparison suggests that the high-scoring mode represents the bit-score distribution of human *bona fide* tRNA genes.

At this stage, this evidence is not convincing enough to conclude that the low-scoring tRNA genes are more likely to be pseudogenes. For example, the small bump in the distribution for "human-tRNAscanSE" within the range of 35 to 50 suggests that some *bona fide* tRNA genes may have bit scores indistinguishable from what are presumed to be tRNA pseudogenes (Figure 3-2, "tRNAscanSE-human"). In addition, the existence of a prominent low-scoring peak in the bit-score distribution of the tRNA genes predicted by Rfam does not really favour the hypothesis that "the low-scoring tRNA genes are pseudogenes". If the low-scoring tRNA genes are pseudogenes and the descendants of ancient functional tRNA genes, the random drifts caused by neutral mutations would be expected to result in a tail at the left side of the bit-score distribution, rather than generating an obviously bimodal distribution.

Consequently, I evaluated additional information, such as loss of primary-sequence and secondary-structure features, to look for additional evidence that low-scoring tRNA genes might be pseudogenes. Such information cannot be directly inferred from the bit scores of individual tRNA genes. An Rfam bit score for a particular ncRNA is actually a statistical evaluation of its degree of conservation at both primary-sequence and secondary-structure levels. It turns out that two factors can contribute to low bit scores for a tRNA gene: 1) the loss of the capability to fold into cloverleaf-like secondary structure; 2) the loss of the internal promoter which is required for being recognized by RNA polymerase III in order to generate functional tRNAs. These factors are further explored in subsections 3.1.2.2. and in 3.1.2.3. respectively.

3.1.2.2. Moderate preservation of secondary structures in the low-scoring and synteny-non-conserved tRNA genes

The number of non-canonical base pairs in Rfam tRNA predictions, as compared to a reference tRNA structure, is plotted. For the synteny-non-conserved tRNA genes with bit scores lower than 50, the mode of the number of non-canonical base pairs that may make the secondary structures unstable is 3 and the average is ~5 (Figure 3-3). In other words, for a low-scoring tRNA gene, there is on average slightly more than 1 non-canonical base pair per stem region (*i.e.* 4 stems in a tRNA molecule in its functional form).

However, even for the synteny-conserved tRNA genes which are more likely to be *bona fide* tRNA genes, the mode is 2 non-canonical base pairs in their stem regions (Figure 3-3) and the average is 2.6. This suggests that for one stem region of a tRNA, one non-canonical base pair can still be tolerated and its secondary structure can still be preserved. The evidence suggests that there is moderate preservation of structural features in the low-scoring tRNA genes and a moderate level of non-canonical base pairs may be tolerated. The degree of loss of structural features provides only limited support for the view that these low-scoring tRNA genes tend to be pseudogenes.

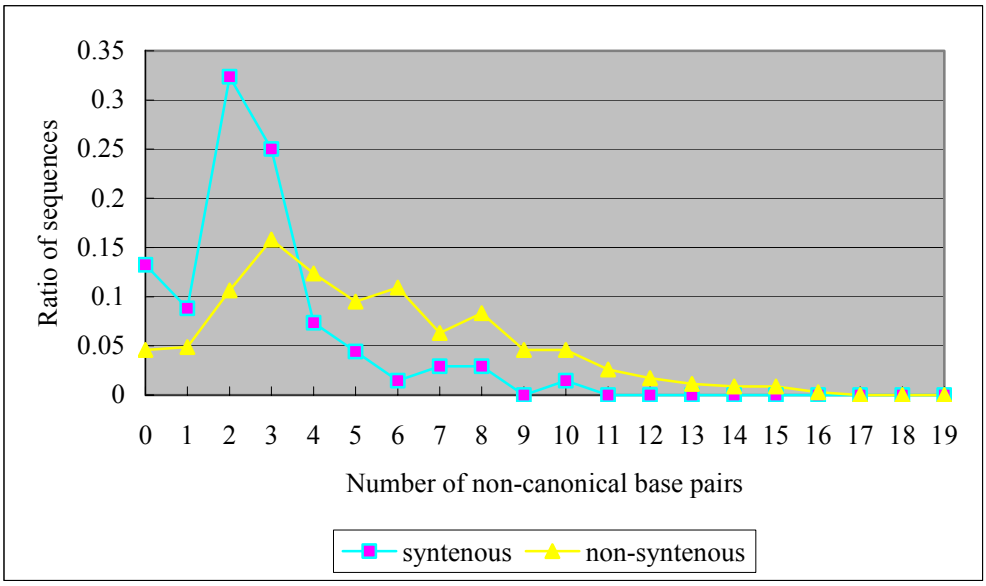

Figure 3-3. Distributions of numbers of the non-canonical base pairs in human tRNA genes

The synteny-conserved and the synteny-non-conserved tRNA genes are aligned to the tRNA consensus structures by using Infernal and the Rfam tRNA CM. Non-canonical base pairs that may destabilize the secondary structures of these tRNA genes are counted, except that G-U base pairs are tolerated.

3.1.2.3. Degradation of the internal promoters in the low-scoring tRNA genes

The genomic loci containing tRNA genes need to be transcribed into tRNA molecules in order to function in cells. If these low-scoring tRNA genes are not transcribable, they are pseudogenes. In order to be transcribable a functional promoter is required. The internal promoters of the tRNA predictions were evaluated using the eufindtRNA algorithm (see methods in subsection 2.2.1.6 of the materials and methods of section 2.2). Previously in subsection 2.2.3.4 in chapter 2, two cases of promoter-specific degradation of synteny-non-conserved tRNAscanSE-predicted tRNA gene loci were found. Here, pseudogenization through promoter-specific degradation is investigated more generally in synteny-non-conserved and low-scoring Rfam tRNA genes.

The results reveal that, about three-quarters (339/441) of the low-scoring tRNA genes do not have intact promoters in their intragenic regions. According to current knowledge, these low-scoring tRNA genes in the human genome cannot be transcribed into tRNAs by

eukaryotic RNA polymerase III. This is good evidence which indicates that the set of low-scoring tRNA genes is enriched with pseudogenes. This result suggests that in the human genome there is a group of tRNA-related pseudogenes, where their internal promoters are degraded, while their secondary structures are moderately conserved.

3.1.2.4. Tracing the evolutionary origins of low-scoring tRNA genes

The finding that the majority of low-scoring tRNA genes appear to have more significantly degraded internal promoters than secondary structures and may be pseudogenes, suggests the hypothesis that mutations that degrade internal promoters have a selective advantage in mammalian evolution. It seems possible that degradation of internal promoters might be the most effective mechanism for disabling tRNA genes, since aberrant tRNA genes with mutations that make RNA secondary structures unstable would be still transcribable and lead to abnormal protein translation and damage the cell.

If selective degradation of synteny-non-conserved tRNA genes were an important mechanism in the human evolution, it would be reasonable that the human genome would contain numerous tRNA genes which have lost functional promoters, but not yet lost their secondary structures. In order to test this hypothesis, it was proposed to demonstrate that random mutations are unlikely to generate tRNA genes, where their internal promoters have degraded and structural features are still moderately preserved.

Consequently, a simulation, where a random mutation model is applied to the ancestors of these low-scoring tRNA genes, was planned. The initial step for preparing this simulation was to find an appropriate ancestral sequence for each low-scoring tRNA gene. The considerations for finding the ancestral sequences of these low-scoring tRNA genes are discussed in the following two subsections (3.1.2.4.1. and 3.1.2.4.2.).

3.1.2.4.1. Weak evolutionary relation of low-scoring tRNA genes with bona fide *human tRNA genes*

A sensible conjecture is that the ancestral sequences of the low-scoring tRNA genes are *bona fide* human tRNA genes. According to the discussions above (for details see subsections 3.1.2.1. , 3.1.2.2. , and 3.1.2.3.), it is conceivable that *bona fide* tRNA genes are enriched in the sets of human-mouse synteny-conserved tRNA genes, the tRNAscanSE-predicted high-scoring tRNA genes, and the tRNA genes in manually-curated tRNA repositories. However, the search for the evolutionary origins of the low-scoring tRNA genes proved difficult. Using WU-BLAST a possible ancestor could be found for less than one-quarter (101/441) of the low-scoring tRNA genes. In addition, less than half of the low-scoring tRNA genes were found to have homologous sequences in the sets of tRNAscanSE-predicted human tRNA genes and of tRNA compilation (Sprinzl et al. 1998).

3.1.2.4.2. Strong evolutionary relation of low-scoring tRNA genes with mitochondrial tRNAs

Because of the failure to find the ancestral sequences for the majority of the low-scoring tRNA genes from the set of *bona fide* human tRNA genes, it was necessary to consider other sources of tRNA genes that might be the evolutionary ancestors of the low-scoring tRNA genes. In eukaryotic cells, the nuclear genome is not the only sequence that contains tRNA genes. Some intracellular organelles, such as mitochondria and chloroplasts, have their own tRNA genes in their organelle genomes. The tRNA genes of these organelles are divergent, at the primary-sequence level, from the vertebrate nuclear tRNA genes. They are another possible origin of the low-scoring tRNA genes.

The sequences of the low-scoring tRNA genes were searched against the genomic sequence of the human mitochondrion (GenBank accession number: NC_001807.4), and better matches were found to human mitochondrial tRNA genes than to trusted tRNA genes in many cases (human-mouse synteny-conserved tRNA genes) (Table 3-1). In addition, 239 of the sequences that did not appear to have any homologous sequence in the set of human tRNA

genes matched human mitochondrial tRNA genes. The average identity of the 280 tentative nuclear mitochondrial tRNA sequences (numt-tRNAs) to human mitochondrial tRNA genes is 84.8%. The average coverage of these alignments to the full length of the mitochondrial tRNA genes is 85.3%. The evidence strongly suggests that many low-scoring tRNA genes in the human nuclear genome are derived from the human mitochondrial tRNA genes, and not from the tRNA genes in the human nuclear genome.

More similar to the human nuclear tRNA genes	128 (29%)
More similar to the human mitochondrial tRNA genes	280* (64%)
None	33 (7%)
All the human low-scoring tRNA genes	441 (100%)

Table 3-1. Numbers of the human low-scoring tRNA genes which are more similar to either the human nuclear tRNA genes or the human mitochondrial tRNA genes.

"None" is used to indicate the low-scoring tRNA genes which are not significantly similar to either human nuclear tRNA genes or mitochondrial tRNA genes. "*" indicates that 239 out of the 280 low-scoring tRNA genes do not have homologous sequences in the set of human tRNA genes.

For the 128 human tRNA genes that are more similar to human nuclear tRNA genes than to mitochondrial tRNA genes, 71.9% (92/128) of them were recognised using eufindtRNA. This means that the majority of human-nuclear-tRNA-derived low-scoring tRNA sequences still preserve their internal promoters to a certain extent. Consequently, the hypothesis which asserts that there might be selection for mutations that degrade the promoters of the tRNA genes in mammals does not appear to apply to tRNA genes derived from other human tRNA genes.

3.1.2.5. Searching for nuclear mitochondrial tRNAs in mammalian genomes

3.1.2.5.1. Finding nuclear mitochondrial tRNA sequences in the human genome

Since the Rfam tRNA CM (covariance model) is not specifically trained for finding nuclear mitochondrial tRNA sequences (numt-tRNAs) in the human genome, there may be

other human numt-tRNAs which were not identified by Rfam. In order to discover as many numt-tRNAs as possible, blastz and the human mitochondrial genome were used to search for nuclear mitochondrial sequences (numt-seqs) in the whole human genome (NCBI 33). Blastz was used since it is well tuned for aligning genomic sequences (Schwartz et al. 2003).

177 human genomic loci were found to be similar to mitochondrial sequences. Many loci contain more than one nuclear mitochondrial genes (numt-genes). The arrangements of mitochondrial genes in these loci are mostly consistent with those of the real mitochondrial genes encoded in the human mitochondrial genome. It is therefore reasonable to infer that the numt-genes of each locus have been co-transferred into the nuclear genome. There are 627 numt-tRNAs in the 177 human loci of numt-seqs. The average identity between these numt-tRNAs and the human mitochondrial tRNA genes is 84.5%. The average coverage of these alignments to the full-length mitochondrial tRNA genes is 85.3%. None of the 627 tRNA genes overlap with known repetitive elements except tRNAs. Only 30 out of the 627 sequences were found to have homologous sequences in the set of trusted *bona fide* human tRNA genes (human-mouse synteny-conserved tRNA genes). By using eufindtRNA, only 33 out of the 627 sequences were found to have RNA Pol III promoters. The discovery of human numt-tRNAs could explain the low-scoring mode in the bimodal score distribution of the human tRNA genes identified by Rfam well (Figure 3-4, human numt-tRNAs). Although the curve for numt-tRNAs does not fit exactly with the low-scoring group of the bimodal distribution, it is almost parallel.

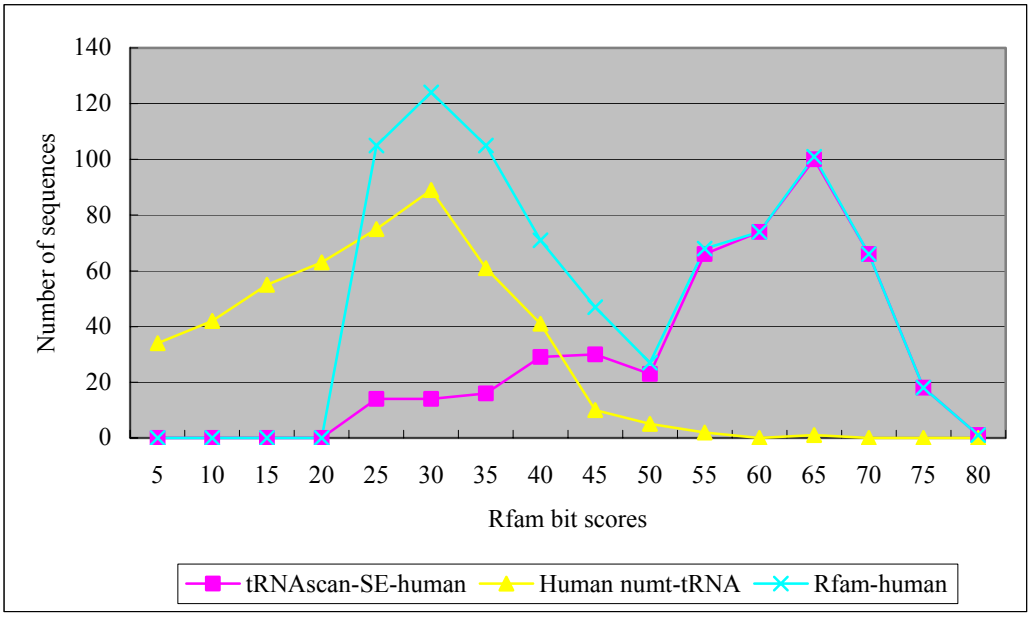

Figure 3-4. Distributions of Rfam bit scores of tRNA genes of human-numt, Rfam-human, and tRNAscanSE tRNA genes.

3.1.2.5.2. Few numt-tRNAs in the mouse genome

Following the discovery of numt-seqs related sequences in the human genome the same analysis was repeated for the mouse genome. In contrast to the discovery of numerous numt-tRNAs in the human genome, far fewer numt-tRNAs could be found in the mouse genome. The bit-score distribution of the mouse low-scoring tRNA genes is obviously different from that of the human low-scoring tRNA genes (Figure 3-2, Rfam-mouse). 86% (217/252) of the mouse low-scoring tRNA genes from Rfam 4.1 are SINEs. Surprisingly, only 64 numt-tRNAs were found in the mouse genome assembly NCBI M30. Not only is the number of numt-seqs smaller than that in the human nuclear genome, but also the average length for each locus of integration is shorter. There are on average 1.7 numt-tRNAs per locus of mouse numt-seq (64 numt-tRNAs / 38 loci), while there are on average 3.5 numt-tRNAs per locus of human numt-seq (627 numt-tRNAs / 177 loci).

There are various hypotheses that might explain the difference between the numbers of numt-tRNAs in the human genome and in the mouse genome. However, before designing

strategies to test these hypotheses, the effect of the quality of the mouse genome assembly on identifying numt-seqs needs to be addressed. Unlike the high coverage of clone-based sequences used in the current human genome assembly, the mouse genome assembly NCBI M30 consists of sequences from both whole genome shotgun (WGS) and high throughput genome sequencing (HTGS). One limitation of WGS sequence assembly is its inability of resolving duplicated regions. If there were numerous recent integrations of the mitochondrial genomic sequence into the mouse nuclear genome, it is possible that the numt-seqs could still be quite similar to one another and thus inappropriately collapsed by WGS sequence assembly. In order to confirm that there is a significant difference between the numbers of the numt-seq loci in the human and mouse genomes respectively, the latter value should be reassessed in the future when more clone-based sequences are used in the mouse genome assembly.

3.1.2.5.3. Effects of numt-tRNAs on finding mammalian tRNAs

The presence of numt-seqs in the human genome has not been considered in the annotation of the human genome. For example, at least five tRNAscanSE-predicted tRNA genes were found within regions of numt-seqs in the human genome. It is unknown whether human numt-tRNAs can be transcribed into functional tRNAs in human cells (for further discussion see subsection 3.1.2.6.). Numt-seqs are also frequently ignored in annotations provided by public-domain genome databases. Unlike the annotation of repetitive elements, consideration of numt-seqs is not part of the procedure in pipelines of genome annotation. In addition, most of the mitochondrial genes are not included in the current release of RepBase (released on 10/09/2004) and there are only two mitochondrial tRNA genes from *G. gallus* in RepBase.

3.1.2.6. Are numt-tRNAs functional?

The existence of numt-seqs in the nuclear genome has been known for some time (Tsuzuki et al. 1983), and their evolutionary dynamics have been discussed in a number of

papers (Mourier et al. 2001; Tourmen et al. 2002; Woischnik and Moraes 2002; Hazkani-Covo et al. 2003; Ricchetti et al. 2004). Most related research suggests that nuclear mitochondrial protein-coding genes (numt protein-coding genes) are pseudogenes. One important factor is that the genetic code of the genes encoded in mitochondrial genomes is different from that of the genes encoded in nuclear genomes. Presumably numt protein-coding genes cannot be translated into functional proteins.

In contrast, the functions of numt-tRNAs have never been explicitly discussed. The arguments, which have been used to infer that numt protein-coding genes should be pseudogenes, may not be applicable to the case of numt-tRNAs. The functions of numt-tRNAs do not depend on being translated into proteins. Numt-tRNA genes could be functional if they were transcribed into tRNA molecules. The following two subsections (3.1.2.6.1. and 3.1.2.6.2.) are therefore dedicated to finding evidence to support the hypothesis that human numt-tRNAs were initially functional while other nuclear mitochondrial sequences (non-tRNA numt-seqs) lost functions upon integration of numt-seqs into nuclear genomes.

3.1.2.6.1. Comparing patterns of mutations of numt-tRNAs and non-tRNA numt-seqs

In order to investigate the possibility that numt-tRNAs were once functional, the patterns of mutation in numt-tRNAs and other non-tRNA numt-seqs were compared. The hypothesis is that, in order to protect the organism from the deleterious effects of transcripts of numt-tRNAs, mutations that disable these genes would accumulate more rapidly than in non-tRNA numt-genes which might be expected to be inactive upon initial insertion. In other words, differences between the patterns of mutations in numt-tRNAs and in non-tRNA numt-seqs might be considered as evidence that either numt-tRNAs or non-tRNA numt-seqs were once functional.

By aligning various human numt-seqs to the human mitochondrial genome, numbers of mutations in human numt-tRNAs and in human non-tRNA numt-seqs were counted separately.

Unexpectedly, on average numt-tRNAs were found to be slightly more conserved than other non-tRNA numt-seqs (Figure 3-5). This result suggests that while evolutionary pressures on human numt-tRNAs and human non-tRNA numt-seqs may be different; overall human numt-tRNAs are not degraded faster than human non-tRNA numt-seqs. In addition, there is no obvious difference between the substitution patterns of the numt-tRNAs and the non-tRNA numt-seqs (Figure 3-6).

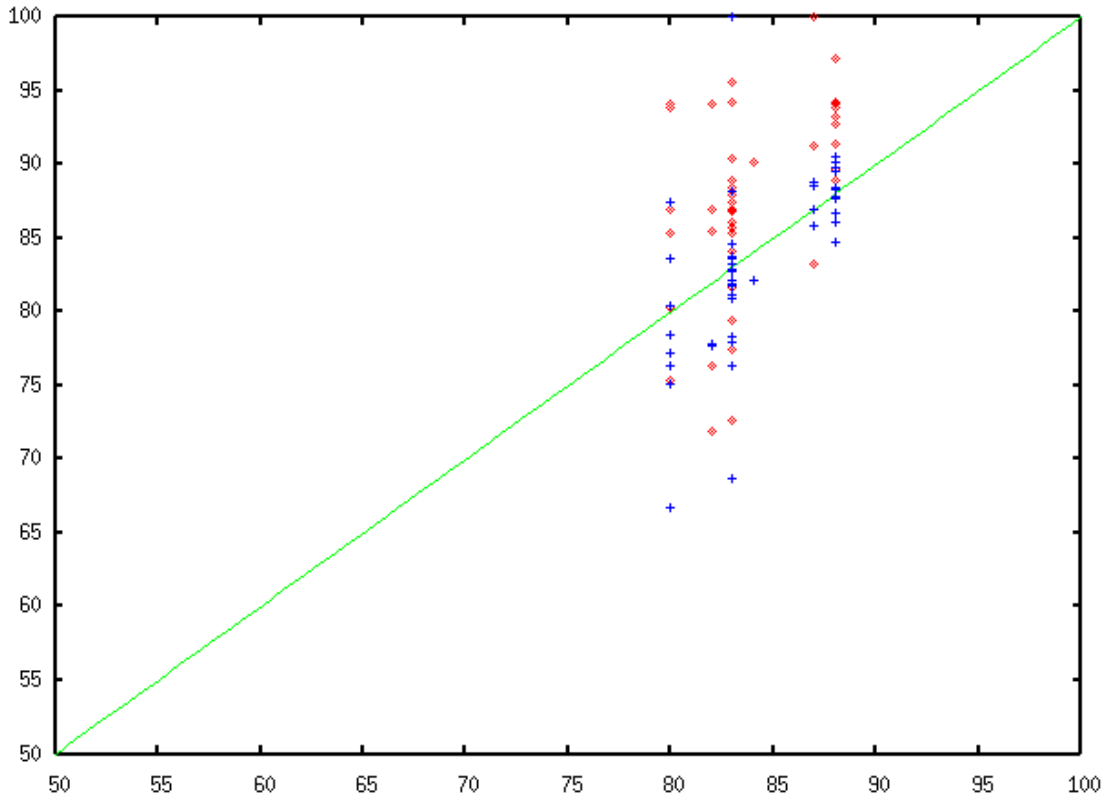

Figure 3-5. Distribution of identities of human numt-tRNAs and human non-tRNA numt-seqs in 80-90 percent identity regions to the human mitochondrial genome

The red points indicate numt-tRNAs and the blue crosses indicate non-tRNA numt-seqs. The green line is the diagonal line (x=y). Numt-tRNAs and non-tRNA numt-seqs were separated from all numt-seqs (found by using blastz) with 80-90 percent identities to the human mitochondrial genome. There are 43 numt-tRNAs and 43 non-tRNA numt-seqs in this plot. The y-axis is the identities of numt-tRNAs or non-tRNA numt-seqs to their corresponding human mitochondrial genes. The x-axis is the identities to the human mitochondrial genome for respective numt-seqs, in which the numt-tRNAs or non-tRNA numt-seqs are embedded.

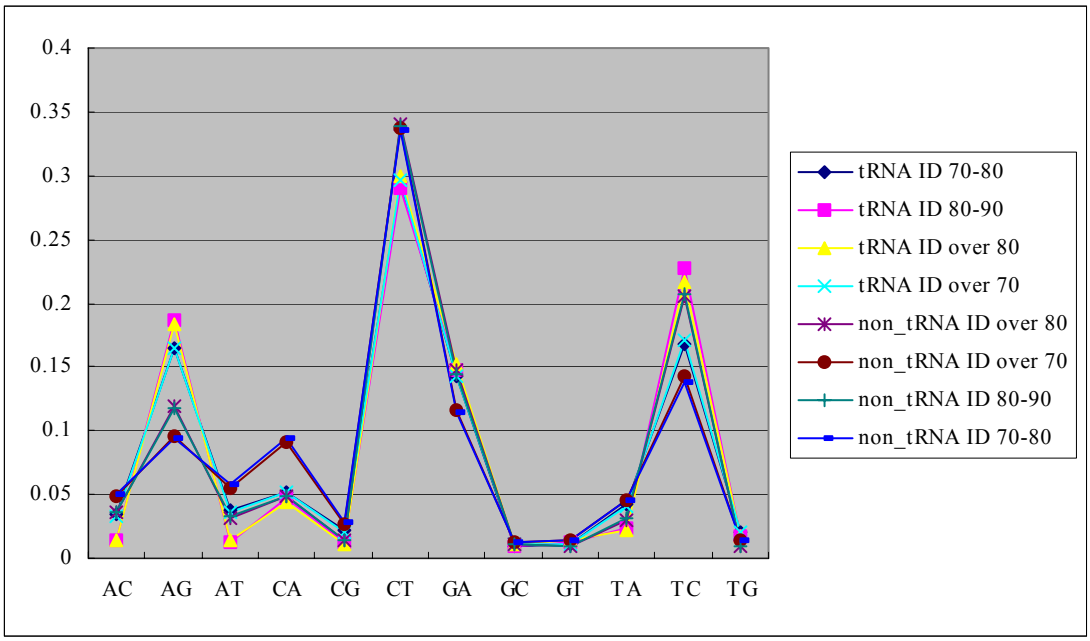

Figure 3-6. Patterns of substitution in the human numt-tRNAs and in the human non-tRNAs embedded in regions with different percent identities to the human mitochondrial genome

"tRNA ID 70-80" indicates the numt-tRNAs embedded in regions with 70-80 percent identities to the human mitochondrial genome and so forth. In the x-axis, "AC" means the base adenosine being substituted with the base cytosine in numt-seqs, and so forth. The y-axis is the normalized ratio of substitutions (*i.e.* number of each type of substitutions normalized by total number of substitutions in each category of numt-tRNAs or non-tRNA numt-seqs).

3.1.2.6.2. Uneven distribution of mutations along human numt-tRNAs

Although the previous results show the overall mutation rate of numt-tRNAs is lower than for non-tRNA numt-seqs, I decided to investigate the distribution of mutations along numt-tRNAs sequences. Given that tRNAs contain internal regulatory elements that promote their transcription, if mutations in numt-tRNAs were found preferentially in positions that could effectively degrade these elements, this would support the hypothesis these numt-tRNAs had initially been active, but subsequently inactivated. Previously counted mutations from alignments between numt-tRNAs and the human mitochondrial genome were therefore counted in bins along the consensus numt-tRNA sequence. The 95% confidence interval for each bin was estimated based on the beta distribution, assuming that the number of mutations was α and that the number of bases in each bin was the sum of α and β.

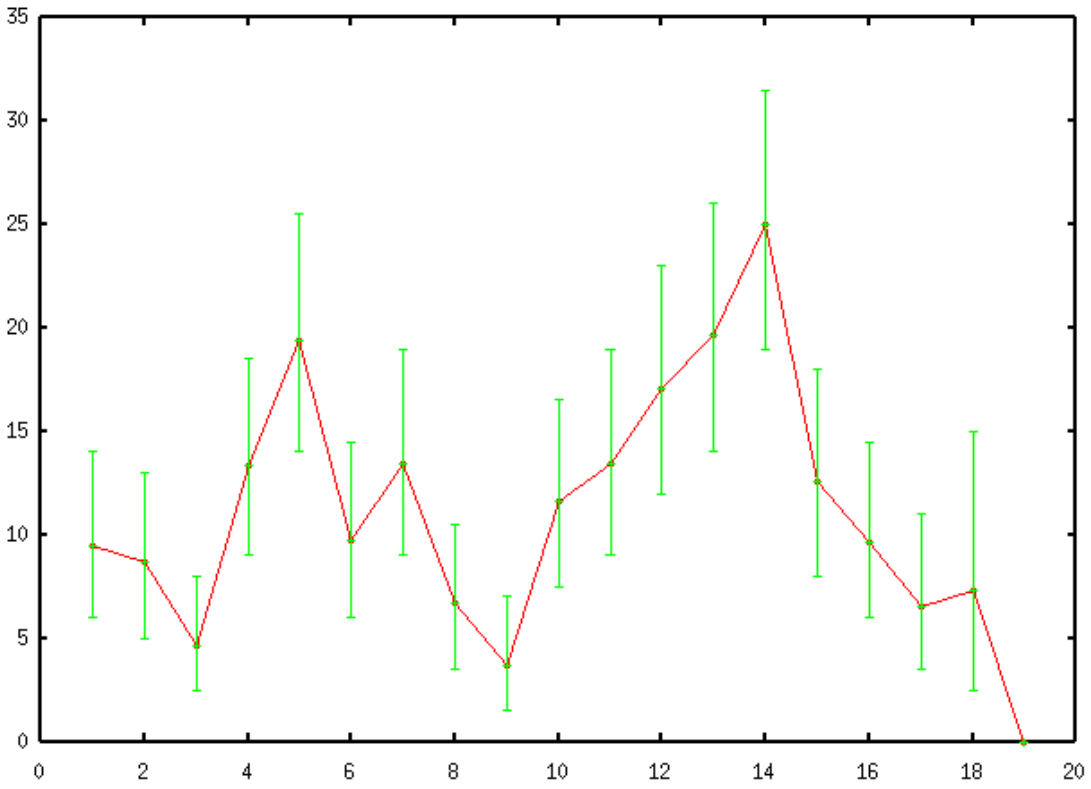

Figure 3-7. Distribution of mutation numbers along human numt-tRNAs

X-axis is the bins along human numt-tRNA sequences. Y-axis is the number of total mutations in each bin. The bin size is 4 bases in length. Forty-three numt-tRNAs are extracted from the numt-seqs with 80 to 90 percent identities to the human mitochondrial genome. The mutations for the first 4 bases for the recruited numt-tRNAs are summed up to give the number of mutations in the first bin and so forth. The green bars are the 95% confidence intervals for bins.

Interestingly, two regions, the 16th to 19th (bin 5) and 52nd to 55th (bin 14) nucleotides, were found to contain significantly more mutations than the 28th to 35th (bin 8 and 9) nucleotide (Figure 3-7). The 95% confidence intervals of mutations for the former two regions do not overlap with those for the 28th to 35th nucleotides. The locations of these two regions are consistent with the positioning of A and B boxes in the nuclear tRNA genes (DeFranco et al. 1980; Galli et al. 1981).

In numt-tRNAs there are significantly more mutations in the positions that correspond to known regulatory regions of human tRNAs and the tRNA promoter finding algorithm eufindtRNA fails to find sequences that score well as promoters. These results might appear consistent with the hypothesis that numt-tRNAs were initially functional when copied into the

mammalian nuclear genomes, but have since become pseudogenes as a result of promoter degradation through selective acceptance of mutations. Unfortunately, proof of this hypothesis needs additional evidence. For example, the mechanism of expression of tRNAs in the mitochondria is different to that of human tRNAs. There is also no evidence to show that expression of mitochondrial tRNAs in the cytoplasm would interfere with the protein synthesis of the genes encoded in nuclear genomes. There are no papers dealing specifically with the fidelity of terminal maturation, aminoacylation, and roles in protein translations if the mitochondrial pre-tRNA transcripts are in the cytoplasm.

3.1.3. Discussion

These results presented in this section (section 3.1) suggest that the 64% of the human synteny-non-conserved tRNA genes retrieved from Rfam are nuclear mitochondrial tRNA genes (numt-tRNAs), whose ancestors are tRNAs in the human mitochondria. With the investigations performed in the previous subsections, these numt-tRNAs should be untranscribable pseudogenes. The pattern of mutation in these numt-tRNAs is interesting and suggestive of pseudogenisation through promoter inactivation. By contrast, the vast majority of the remaining low scoring synteny-non-conserved tRNA genes retrieved from Rfam have sequence similarity to synteny-conserved tRNA genes and ~72% are recognised using eufindtRNA suggesting they have intact promoters and may not be pseudogenes (see subsection 3.1.2.4.2.).

The bit-score distribution appears to be only weakly useful in distinguishing functional tRNA genes from tRNA pseudogenes. The bimodal bit-score distribution observed for low scoring synteny-non-conserved tRNA genes was mainly the result of the special case of numt-tRNAs, however when these were removed any relationships became unclear. This is consistent with the bit-score distributions of other classes of Rfam ncRNAs, where no

particular pattern can be found. With a bit-score distribution that is simply single-modal and heavy-tailed, such as in the case of human U6 snRNA genes identified by Rfam 4.1 (Figure 3-8), it is difficult to choose any clear-cut threshold that might separate functional and non-functional genes. Although ncRNA sequences with higher bit scores are more likely to be synteny-conserved and functional genes, whether ncRNA sequences with lower scores are functional or not cannot be unambiguously determined. Similarly there is little evidence that an ncRNA gene with synteny-non-conserved status is necessarily a pseudogene.

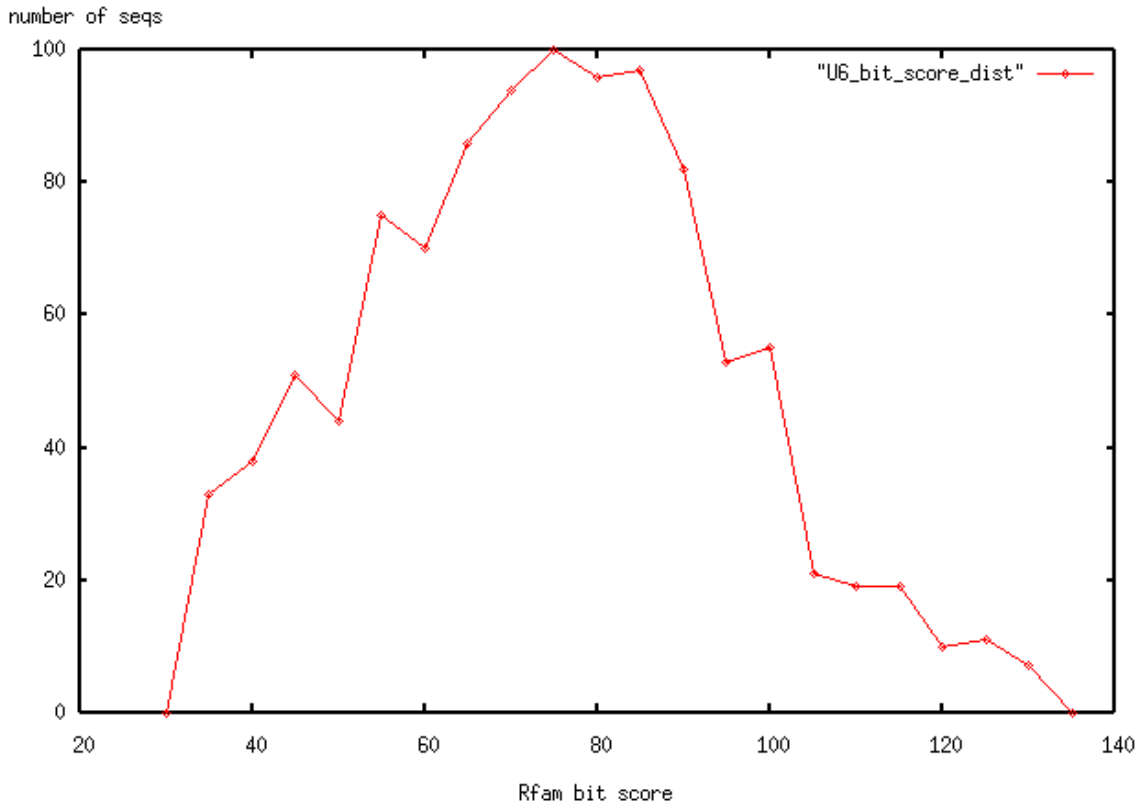

Figure 3-8. Distribution of the Rfam bit scores of the human U6-like sequences identified by Rfam 4.1

The heavy-tailed distributions suggest that, for many classes of ncRNAs in mammalian genomes, the generation of pseudogenes may be a continuous process. It seems that abundant ncRNA pseudogenes in mammalian genomes do not have a strong negative effect on the fitness of organisms. While this is good news for the survival of mammals, it also means that bit score distributions cannot be very helpful in filtering out ncRNA pseudogenes in ncRNA

finding. More specific signals are necessary for distinguishing *bona fide* ncRNAs from ncRNA pseudogenes. One such signal might be whether an ncRNA retains a recognisable internal promoter, however verification of the computational evidence presented here is needed.

3.2. Clustering – a useful criterion for filtering out ncRNA pseudogenes?

3.2.1. Materials and methods

3.2.1.1. Recruiting human and mouse tRNA genes

The human and mouse tRNA genes used in this section were retrieved from Ensembl release 29 by using Ensembl Perl APIs. These genes were predicted by using tRNAscanSE.

3.2.1.2. Defining tRNA-gene clusters

In assessing the features of clustered tRNA genes, one issue concerns deciding a suitable distance criterion, *i.e.* the maximal distance allowed between the nearest neighbouring tRNA genes, for defining tRNA gene clusters. If the selected distance is longer than necessary, more potentially non-clustered tRNAs may be included into clusters. On the other hand, if the selected distance is too short, some clustered *bona fide* tRNA genes may be incorrectly grouped or classified as non-clustered. Several different distances, such as 5-kilo bases and 10-kilo bases, were therefore tried to define tRNA-gene clusters.

3.2.1.3. Comparing the ratios of non-clustered tRNA genes within different bit-score ranges

All human tRNA genes are categorized into five bins according to their bit scores: 20-55, 56-65, 66-75, 76-85, and 86-95. The ratio of tRNA genes that are clustered was calculated separately for each bin. The enrichment of clustered tRNA genes in each bin is determined by comparing the ratios in different bins. The 95% confidence intervals for individual ratios were

estimated based on the beta distribution, assuming that each numerator was α and that each denominator was the sum of α and β.

3.2.1.4. The anticodons required for protein translation

It is known that not all 61 types of anticodons are required for protein translation in eukaryotic cells. Because the interactions between codons and anticodons allow wobble pairs in the third positions (of codons), some codons can share recognition by the same tRNA. Guthrie and Abelson estimated that 46 types of tRNAs that have 45 unique anticodons are sufficient for translation (for review see Guthrie and Abelson 1982). Two types of tRNAs with exactly the same anticodon are used for carrying Met_m and Met_i respectively ("i" indicates translation initiation codon "m" indicates a general non-initiation codon for methionine).

3.2.2. Results

3.2.2.1. Enrichment of mammalian non-clustered tRNA genes in the low-scoring group

A 10-kb distance threshold was initially used to subgroup all human tRNA genes into clustered and non-clustered ones. Among the 608 human tRNA genes predicted by tRNAscanSE, ~65% (125/192) of the tRNA sequences with scores 20-55 were found to be non-clustered. By contrast, ~27% (16/59) with scores 55-65, ~30% (45/152) with scores 65-75, ~25% (42/171) with scores 75-85, and ~26% (9/35) with scores 85-95, are non-clustered (Figure 3-9). These results suggest that non-clustered tRNA genes are enriched in the low-scoring group. There is also a similar finding when the clusters were defined by using the 5-kb distance threshold (data not shown).

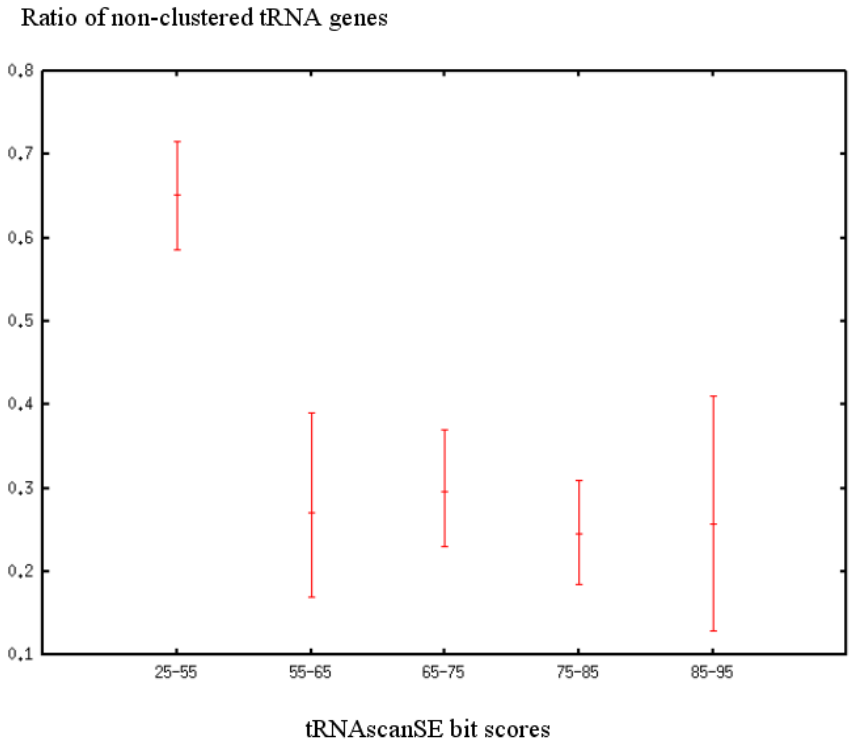

Figure 3-9. The human low-scoring tRNA genes are enriched with non-clustered ones

Each red bar is the 95% confidence interval for each bin. The confidence intervals shown here were estimated as described in subsection 3.2.1.3.

3.2.2.2. The mammalian clustered tRNA genes can cover 46 necessary anticodons

In this subsection, the functionality of non-clustered tRNA genes is explored indirectly on the basis of the need for their roles in protein translation. If clustered tRNA genes are shown not to include all the anticodons required for protein translation, this will be evidence that non-clustered tRNA genes are necessarily functional. Conversely, if clustered tRNA genes provide all the required anticodons, non-clustered tRNA genes may not be necessarily required for protein translation.

These results indicate that clustered tRNA genes in the human genome can cover all 46 types of tRNAs and exactly satisfy the wobble rules (Table 3-2, compare “yeast” and “clustered” ones). Although, the human non-clustered tRNA genes can also cover 46 types of tRNAs, there are several cases that violate the wobble rules (Table 3-2, compare “yeast” and

"non-clustered" ones). Besides, in the mouse clustered tRNA genes, additional anticodons were found (Table 3-3). These results suggest that the clustered tRNA genes in mammalian genomes may be sufficient to provide the necessary types of tRNAs for translating proteins.

tRNA types	Yeast	Human				
		All	Clustered, dist < 10kb	Clustered, dist < 6kb	Non-clustered, dist < 10kb	Non-clustered, dist < 6kb
Ala	3	3	3	3	3	3
Arg	5	5	5	5	5	5
Asn	1	2	1	1	2	2
Asp	1	1	1	1	1	1
Cys	1	1	1	1	1	1
Gln	2	2	2	2	2	2
Glu	2	2	2	2	2	2
Gly	3	3	3	3	2	2
His	1	1	1	1	1	1
Ile	2	2	2	2	2	2
Leu	5	5	5	5	5	5
Lys	2	2	2	2	2	2
Met*	2	2	2	2	2	2
Phe	1	1	1	1	1	1
Pro	3	3	3	3	1	1
Ser	4	4	4	4	4	4
Thr	3	3	3	3	3	3
Trp	1	1	1	1	1	1
Tyr	1	2	1	1	2	2
Val	3	3	3	3	3	3
Total	45	47	45	45	45	45

Table 3-2. Comparison between types of anticodons of yeast and the human tRNAs

Each number indicates the distinct types of tRNA anticodons corresponding to a particular amino acid. For example, there are 2 distinct types of anticodons found in the yeast tRNA genes corresponding to the tRNAs carrying isoleucine (Ile). Each red box is used to indicate that for a particular amino acid, the number of corresponding anticodon types that can be found in a category (clustered, non-clustered, *etc.*) of human tRNA genes is different from that of the anticodon types found in yeast tRNA genes.

"*" means that there are two types of tRNAs with exactly the same anticodon for Met_i and Met_m respectively.

tRNA types	Yeast	Mouse			
		All	Clustered, dist < 1 mb	Clustered, dist < 10 kb	Clustered, dist < 6 kb
Ala	3	4	4	3	3
Arg	5	6	5	5	5
Asn	1	2	1	1	1
Asp	1	2	1	1	1
Cys	1	2	2	1	1
Gln	2	2	2	2	2
Glu	2	2	2	2	2
Gly	3	4	4	4	4
His	1	2	2	1	1
Ile	2	3	3	2	1
Leu	5	6	5	5	5
Lys	2	2	2	2	2
Met*	2	2	2	2	2
Phe	1	2	2	1	1
Pro	3	4	3	3	3
Ser	4	6	4	4	4
Thr	3	4	3	3	3
Trp	1	1	1	1	1
Tyr	1	2	1	1	1
Val	3	4	4	4	4
total	46	62	53	48	47

Table 3-3. Comparison between types of anticodons of yeast and mouse tRNAs

The color-coding convention used in this table follows that of Table 3-2.

The anticodon types of non-clustered mouse tRNA genes were not listed. The types of anticodons that can be found in non-clustered mouse tRNA genes exceed the essential types of anticodons (the column "yeast"). It is difficult to determine which of them may not be the anticodons of *bona fide* mouse tRNA genes. The purpose of this table is thus to demonstrate that clustered mouse tRNA genes can cover the anticodons essential for protein translation.

"*"means that there are two types of tRNAs with exactly one anticodons for Met_i and Met_m respectively.

3.2.3. Discussion

3.2.3.1. Clustering may be a useful criterion for filtering out tRNA pseudogenes

Three threads of evidence imply that maybe the clustered tRNA genes in the mammalian genomes are functionally more important than the non-clustered tRNA genes are. First, the human low-scoring tRNA genes, which are more likely to be pseudogenes, are significantly enriched with non-clustered tRNA genes. Second, the finding that clustered tRNA genes should be sufficient for protein translation implies that non-clustered tRNA genes may not necessarily be required for protein translation. Third, ~56% of human clustered tRNA genes are human-mouse synteny-conserved, while only ~40% of human non-clustered tRNA genes are human-mouse synteny-conserved (for details see section 2.2 and Figure 2-7).

3.3. Summary

In the first part of this chapter (section 3.1), I explored the tendency of the synteny-non-conserved tRNA genes retrieved from Rfam to be pseudogenes. Results relevant to genome-wide ncRNA finding include that:

- ~65% of human synteny-non-conserved tRNA genes retrieved from Rfam are nuclear mitochondrial tRNA sequences (numt-tRNAs).
- Evidence suggests that these numt-tRNAs are currently non-functional in the human genome. The observed patterns of mutation are weakly suggestive of a mechanism of pseudogenisation that involves promoter inactivation.
- Once numt-tRNAs were disregarded, it was apparent that many of the remaining low-scoring synteny-non-conserved tRNA genes might not necessarily be pseudogenes.

In the second part of this chapter (section 3.2), I explored the functionality of human

non-clustered tRNA genes. The main results are that:

- Low-scoring tRNA genes are enriched with non-clustered tRNA genes.
- Mammalian clustered tRNA genes can provide sufficient types of tRNAs to cover all the anticodons required for protein translation. This is consistent with non-clusters tRNA genes not needing to be functional, but does not demonstrate that they are non-functional.

With respect to the functionality of synteny-non-conserved ncRNAs in mammalian genomes, there are two hypotheses. In the following, I summarize the pieces of evidence for or against each of these:

1. Hypothesis: synteny-non-conserved ncRNA genes are pseudogenes.
   - Evidence against this hypothesis:
     - The majority (71.9%) of human nuclear tRNA derived low-scoring and synteny-non-conserved (Rfam) tRNA sequences still preserve their internal promoters to a certain extent (see subsection 3.1.2.4.2.). They may not be functional tRNA genes but may be transcribable.
     - Some synteny-non-conserved and non-clustered (tRNAscanSE) tRNA gene loci are also high-scoring, suggesting that these loci may not necessarily be pseudogenes (see the high-scoring bins in Figure 3-9).
   - Conclusion:
     - Evidence is weak, but is suggestive that synteny-non-conserved ncRNAs are a mixture of functional ncRNAs and pseudogenes.

2. Hypothesis: non-clustered tRNA genes are pseudogenes.
   - Evidence for this hypothesis:

- The set of low-scoring tRNA genes in the human genome is significantly enriched with non-clustered tRNA genes (see subsection 3.2.2.1. and Figure 3-9).
- Clustered tRNA genes can cover 46 types of anticodons required for protein translation, implying that non-clustered tRNA genes may be functionally less important for translation (see subsection 3.2.2.2.).
- ~56% of human clustered tRNA genes are human-mouse synteny-conserved, while only ~40% of human non-clustered tRNA genes are human-mouse synteny-conserved (see section 2.2 and Figure 2-7).

◆ Evidence against this hypothesis:

- Some non-clustered tRNA genes are high-scoring as well as synteny-conserved in mammalian genomes (see subsection 2.2.2.7.), not suggesting that they are pseudogenes.

◆ Conclusion:

- Evidence is weak, but suggestive that non-clustered tRNAs may be more likely to be pseudogenes.

In conclusion, evidence weakly supports that synteny-non-conserved ncRNAs are a mixture of functional ncRNAs and pseudogenes. Besides, non-clustered tRNA genes may be more likely to be pseudogenes.