Chapter 2. Constraints from comparative genomics for ncRNA finding

Among various approaches for *ab initio* ncRNA finding, comparative algorithms have been claimed to have good performance in identifying structural ncRNAs in test data sets (Rivas and Eddy 2001; di Bernardo et al. 2003; Coventry et al. 2004; Washietl et al. 2005; Pedersen et al. 2006) and simple genomes, such as bacteria and yeasts (Rivas et al. 2001). One algorithm, RNAz, was also claimed to perform well in identifying structural ncRNAs in mammalian genomes (Washietl et al. 2005). One requirement for using these comparative algorithms is that the input data must be sequence alignments.

Recently, some of these comparative algorithms have been applied to finding ncRNAs in vertebrate genomes (Washietl et al. 2005; Pedersen et al. 2006), where the alignments used for prediction were mainly derived from syntenic regions of multiple vertebrate genomes. In this thesis, such type of alignments is referred to as synteny alignments. However, the properties of synteny alignments that may contain ncRNAs are not necessarily comparable to the test data sets used to assess these comparative algorithms. This makes it uncertain whether these algorithms will have the same performance in finding ncRNAs, when synteny alignments are used.

For convenience, some terms are defined here. “Synteny-conserved ncRNAs” is used to indicate ncRNAs, in one organism, that are conserved in the corresponding syntenic regions of other genomes; if an ncRNA is not synteny-conserved, it is referred to as “synteny-non-conserved”; “synteny-conservation ratio” of ncRNAs refers to the ratio of one organism’s ncRNAs that are “synteny-conserved ncRNAs” to the total number.

There are several considerations when using synteny alignments as the target for

genome-wide ncRNA finding. Firstly, if many functional ncRNAs are synteny-non-conserved in the genomes under investigation, finding ncRNAs using only synteny alignments would risk missing a significant number of ncRNAs. To date, the synteny-conservation ratio of different classes of ncRNAs in vertebrate genomes has not been comprehensively surveyed. One obstacle in carrying out such a survey is that classic ncRNAs, which are frequently related to repetitive elements in vertebrate genomes, have generally been removed before building synteny data sets (Schwartz et al. 2003; Frazer et al. 2004; Siepel et al. 2005).

Secondly, if orthologous ncRNAs in the genomes under investigation are so conserved that only a few covariations are found, it may be difficult to determine whether the sequence conservation means the existence of RNA high-order structures or simply of primary-sequence motifs. The number of covariations in alignments of the orthologous ncRNAs may be expected to be greater for more distantly related organisms. This is why the sequence identity of a primary-sequence alignment is usually required to be within certain ranges for comparative ncRNA finding algorithms. For instance, the desired ranges of sequence identity for running QRNA and ddbRNA are 65%-85% (Rivas and Eddy 2001) and 60%-80% (di Bernardo et al. 2003), respectively. Likewise, RNAz implicitly requires that the sequences of orthologous ncRNAs are divergent to a certain extent, because the false positive rate of RNAz was reported to increase when alignments of high identities were used (Washietl et al. 2005). However, so far, no systematic survey has been performed to estimate the abundance of covariations in the orthologous ncRNAs in vertebrate genomes.

This chapter is therefore dedicated to investigating the conservation patterns of ncRNAs in vertebrate genomes, especially in mammalian genomes. A detailed survey of the conservation patterns of both classic (such as tRNAs, rRNAs, and snRNAs) and non-classic (such as miRNAs, snoRNAs, *etc*) ncRNAs in mammalian genomes was performed, in order to provide a solid basis for using the mammalian synteny alignments in genome-wide ncRNA

finding. The conservation patterns explored in this chapter include:

- The synteny-conservation ratios of ncRNAs.
- The abundance of covariations between orthologous ncRNAs.

In the first section of this chapter (section 2.1), a protein-coding gene based strategy for locating the respective syntenic regions of individual human ncRNAs was used. The conservation patterns of multiple classes of human ncRNAs in these human-mouse syntenic regions were then investigated. The synteny-conservation ratios, as well as the abundance of covariations, of the ncRNAs in the human genome with respect to the mouse genome were then calculated. A survey of the abundance of covariations was also performed on the human-mouse synteny-conserved ncRNAs with respect to their best homologues in the zebrafish genome. Based on this data, the possible effects of using real genomic alignments of ncRNAs on the performance of several comparative ncRNA finding algorithms was explored.

One caveat with respect to the syntenic-region locating strategy used in the first section of this chapter is the ignorance of gene-order conservation of ncRNAs. This means that, if there are local changes of the ncRNA copy numbers and/or of the ncRNA gene order within syntenic regions, these will be missed. Since the changes caused by evolutionary events may help explain the observed synteny-conservation ratios of ncRNAs, gene-order conservation is of interest.

In section 2.2, I examined the conservation/change of the physical arrangements of tRNA gene loci in mammalian genomes. This study is intended to explore if the pattern of gene-order conservation may give any insight into the origin of the substantial number of synteny-non-conserved ncRNAs observed in mammalian genomes. In particular, the gene-order conservation of clustered tRNA gene loci in mammalian genome is of interest. This idea was motivated from the observations of many clustered ncRNAs in diverse genomes,

from virus (Wilson et al. 1972), bacteria (Fournier et al. 1974), yeast (Beckmann et al. 1977), to primates (Chang et al. 1986). For instance, a tRNA gene cluster consisting of ~150 tRNA gene loci were found on human chromosome 6 (Mungall et al. 2003). The specific issues I intend to address in section 2.2 are as follows:

- Are there synteny-conserved clusters of tRNA gene loci?
- Are there many gene-order changes in the syntenic tRNA gene clusters?

This study is useful to genome-wide ncRNA finding in several ways. First, it may provide a high-resolution view on how tRNA genes have evolved in mammalian genomes, and may therefore give insights on how alignments should be generated for the purpose of genome-wide ncRNA finding. Second, this study may potentially be useful for distinguishing the tRNA gene loci that are functional, from those that have become pseudogenes. Although the rules derived from the case of mammalian tRNA genes may not necessarily be valid for the cases of other classes of ncRNA genes, this study may provide an independent piece of evidence, which is not biased toward protein genes, to the evolution of mammalian genomes.

2.1. The conservation patterns of vertebrate ncRNAs

2.1.1. Materials and Methods

2.1.1.1. Recruiting human ncRNAs

The genomic loci of human tRNAs were retrieved from Ensembl release 29. Ensembl is a software system that aims to provide a comprehensive annotation of selective eukaryotic genomes (Birney et al. 2006). Different releases of Ensembl may use different versions of genome assemblies. The human genome assembly that is used in Ensembl release 29 is NCBI 35, which was released by NCBI in April 2004. (http://www.ncbi.nlm.nih.gov/genome/guide/human/release_notes.html)

The genomic loci of human tRNAs in Ensembl are annotated using tRNAscanSE, which is a tRNA finding pipeline that integrates several tRNA finding algorithms (Lowe and Eddy 1997). The algorithms used by tRNAscanSE include tRNAscan (Fichant and Burks 1991), eufindtRNA (Pavesi et al. 1994), covels (Eddy and Durbin 1994), and coves (Eddy and Durbin 1994). tRNAscan is a hierarchical and rule-based system to identify intragenic promoters and consensus secondary structures of tRNAs. eufindtRNA was designed to find intragenic promoters of tRNAs. Covels is a search algorithm that uses a covariance model (CM) (see subsection 1.3.3.2.) to detect both primary-sequence and secondary-structure motifs with high specificity in genomes, although it is very slow. In the tRNAscanSE pipeline, both the outputs of tRNAscan and eufindtRNA are combined into one set of candidate tRNA genes, which are further assessed by covels in order to remove false positives. The criterion for deciding true positives is the degree of conservation at both primary-sequence and secondary-structure levels (Lowe and Eddy 1997). The final structural alignments are generated by coves. In Ensembl release 29, there are 498 tRNA genes in the human genome, after excluding pseudogenes and the tRNAs with undetermined codon types.

Other human ncRNAs were retrieved from Rfam 6.1 (Griffiths-Jones et al. 2005). Rfam is a database of curated sequence alignments and CMs of different classes of ncRNAs. The CMs created by Rfam are also used to search for novel ncRNAs in the EMBL nucleotide sequence database (Kanz et al. 2005), which includes sequences of the human genome and the mouse genome. The sequences and the ncRNAs so predicted are also deposited in Rfam. Infernal (a system for “INFERence of RNA ALignment”, http://infernal.janelia.org/) is the software package used by Rfam to build CMs and to find ncRNA-like sequences in the sequence database (Griffiths-Jones et al. 2005).

The coordinates of Rfam ncRNAs in the human genomic contigs were retrieved from Rfam.full, which was downloaded from the Rfam ftp site (ftp://ftp.sanger.ac.uk/pub/databases/

Rfam/). The coordinates were converted to human chromosomal coordinates using software libraries provided by the Ensembl Project written in the Perl programming language referred to as Application Programming Interfaces (APIs). Although there have been newer releases of Ensembl since the analyses in this thesis were performed, NCBI 35 has continued to be used by a number of later releases of Ensembl (releases 30 ~ 36). This procedure of mapping ncRNAs to the human genome is exactly the same as that used for generating the ncRNA annotation of Ensembl releases 30 ~ 36.

2.1.1.2. Searching for human-mouse synteny-conserved ncRNAs

The alignments of human-mouse syntenic regions were retrieved from Ensembl Compara release 29 (Clamp et al. 2003) using the Ensembl Compara Perl APIs. The Ensembl Compara database is the component of Ensembl that contains comparative genomic information, including predictions of orthology relationships between protein-coding genes and synteny alignments among different genomes. The genome assemblies used by Ensembl Compara release 29 include human NCBI 35 and mouse NCBI M33 (http://www.ncbi.nlm.nih.gov/genome/seq/NCBIContigInfo.html).

The existence of synteny-conserved ncRNAs in candidate alignments was searched using cmsearch and Rfam CMs. cmsearch is a program of the Infernal package that can use a Rfam CM trained using a particular type of ncRNAs to search for new occurrences of ncRNAs of the same type. Given a sequence, cmsearch can align it to a Rfam CM and return high scoring matches. cmsearch reports matches with bit scores (for more details about bit scores see subsection 1.3.3.1). The regions with bit scores higher than corresponding family-specific thresholds pre-determined by Rfam (Griffiths-Jones et al. 2003) were considered to be ncRNA loci.

In order to correctly include classic ncRNAs in genomic regions that are missing from available resources of genome-wide alignments, an approach was adopted which takes

advantage of the syntenic regions defined by human-mouse orthologous protein-coding genes. This approach allows the identification of missing synteny-conserved ncRNAs in initially unaligned syntenic regions. The basic idea is that, if the relation of a particular ncRNA to its 5' and 3' flanking protein-coding genes has been preserved in evolution, a synteny-conserved ncRNA may also be found in the corresponding syntenic region defined by synteny-conserved protein-coding genes in the other genome (Figure 2-1, a).

One issue when using this strategy to find the synteny-conserved ncRNAs is the ambiguity in assigning orthology to protein-coding genes retrieved from different genomes. For instance, ambiguity can occur whenever multiple protein-coding genes, which are paralogous to each other in one organism, appear orthologous to a particular gene in the other organism. Such many-to-one or even many-to-many relationships between protein-coding genes may cause difficulties in determining unique human-mouse syntenic regions for individual human ncRNAs. In order to control the complexity of finding the appropriate syntenic regions, best reciprocal protein homologs (UBRHs), where there is only one uniquely best hit in both directions between two genomes, were used in the following analyses. Each pair of UBRHs (UBRHP) consists of two homologous members from the human and mouse genomes, respectively. All UBRHPs between these two genomes were retrieved from Ensembl Compara release 29. The 5' and 3' flanking protein-coding genes nearest to a particular human ncRNA, which are also the members of two consecutive UBRHPs, were used to define the boundaries of the corresponding mouse syntenic region (Figure 2-1, a).

Synteny-conserved counterparts of human ncRNAs in the mouse (UBRHPs-bound) syntenic regions were obtained by using WU-BLAST alignment algorithm to scan the UBRHP-bound mouse genome sequence with the human ncRNA sequence. The threshold used for filtering alignment hits was set to be at least 40% identity. Certainly, the cost of this heuristic is an inevitable decrease in sensitivity; however hits with low percent identities (<

50%) are also unsuitable for using existing algorithms for *ab initio* ncRNA finding. The existence of synteny-conserved ncRNAs was further verified using Infernal and Rfam CMs. Human ncRNAs that were found to be conserved in the syntenic regions were labelled as "synteny-conserved ncRNAs"; otherwise they were labelled as "synteny-non-conserved ncRNAs". It should be noted that the set of UBRHPs, and accordingly, UBRHPs-bound syntenic regions, can change between releases of Ensembl, even if exactly the same genome assemblies were used. Such changes result from improvements in the annotations of protein-coding genes in Ensembl. However, the annotation of genes in the mouse genome (NCBI M33) was constant through Ensembl releases 29 ~ 31, so there were essentially no major changes in the set of UBRHPs-bound syntenic regions in the Ensembl Compara database of these Ensembl releases.

Several complicated situations could be encountered when using the UBRHPs based approach to find synteny-conserved ncRNAs: 1) ncRNAs at either end of chromosomes may not be flanked by members of UBRHPs (Figure 2-1, b); 2) the members of two consecutive UBRHPs may be partitioned into two different chromosomes (Figure 2-1, c); 3) the relationships of UBRHPs-bound blocks between two genomes may be inconsistent due to some unknown evolutionary events (Figure 2-1, d). Each of these three situations makes the search process more difficult, and might thus cause false negatives in determining synteny-conserved ncRNAs.

In order to reduce the false negatives caused by the first and the second situations, either the 5' or the 3' member of the flanking UBRHP of a particular ncRNA was used as the anchoring point to extend the candidate sequence blocks for searching for a synteny-conserved ncRNA in the second genome. The cases of the second situation are marked as "inter-chromosomal translocation" (Figure 2-1, c).

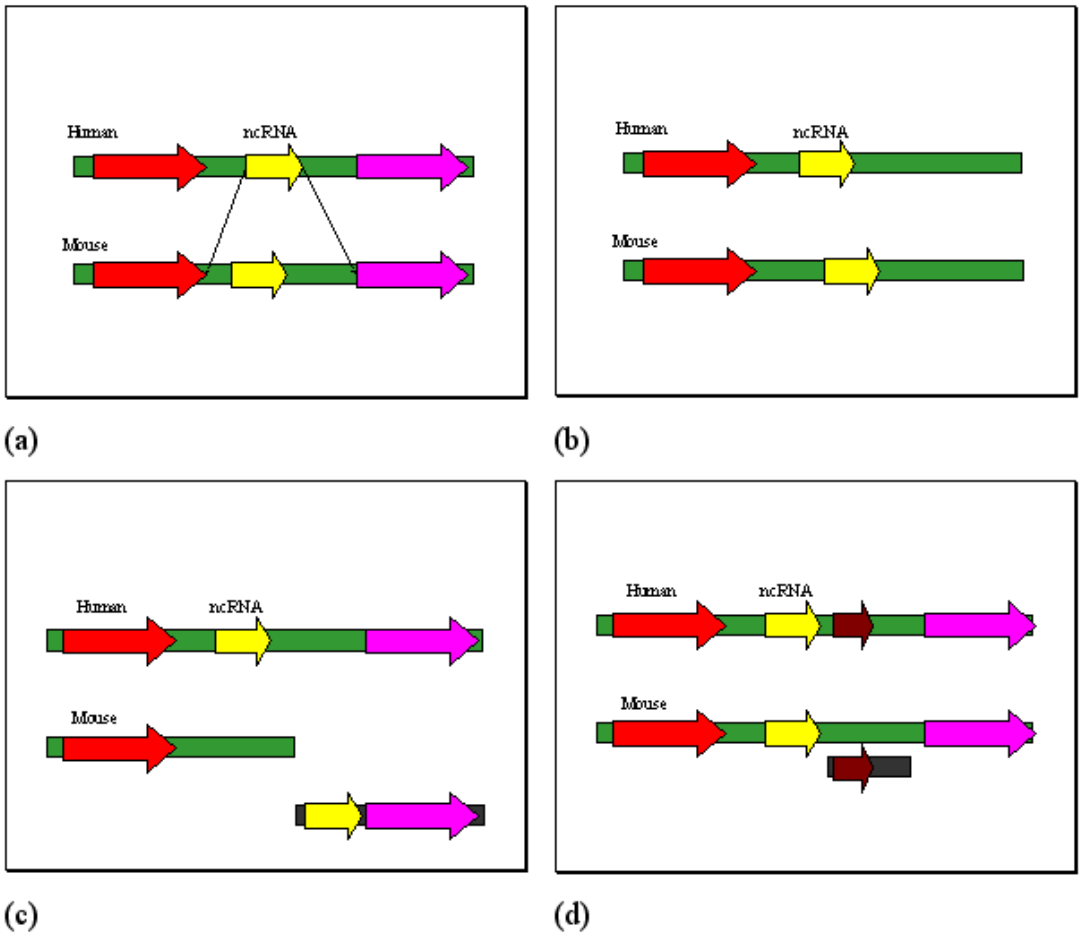

Figure 2-1. Physical relations of human and mouse synteny-conserved ncRNAs to UBRHPs-bound syntenic regions

Red arrows: one pair of unique best reciprocal protein homologues (UBRHP) in the 5' flanking region of one ncRNA. Magenta arrows: one UBRHP in the 3' flanking region of one ncRNA. Yellow arrows: synteny-conserved ncRNAs. (a) The mouse members of two consecutive UBRHPs are on the same chromosome. (b) ncRNAs that are near the ends of chromosomes are flanked by only one UBRHPs (either in the 5' or in the 3' flanking region). (c) The mouse members of two consecutive UBRHPs are separated into two chromosomes. (d) The relationship of UBRHPs-bound blocks becomes incompatible between two genomes due to unknown evolutionary events.

For the third situation, however, it is unknown how to determine the real evolutionary event leading to the finding of pairs of protein-coding genes that are out of order (Figure 2-1, d, the brown arrows). It is possible that, in these regions, there might have been inter-chromosomal rearrangements, pseudogenisations of duplicated genes, *etc*. Consequently, it is difficult to define a clear rule to avoid possible false negatives in such complicated cases. To partially address this problem, one additional measure was adopted. In recruiting two consecutive UBRHPs to define a suitable syntenic block for one ncRNA, next adjacent

UBRHPs was tried (Figure 2-1, d, the magenta arrows) if the initial UBRHPs was not on the same chromosome as their 5' and 3' flanking UBRHPs (Figure 2-1, d, compare the red and brown arrows). These regions were marked as "complicated regions".

In addition, I also considered cases where there might be segmental inversions in the UBRHPs-bound syntenic regions. I took the incompatibility of the strand combinations of the UBRHPs in different genomes as an indicator of segmental inversions. The argument is that, when there are no segmental inversions, the strand combination of the respective members from 5' and 3' UBRHPs in the first genome should be consistent with the strand combination in the second genome.

2.1.1.3. Determination of covariations between orthologous ncRNAs

To determine covariations between orthologous ncRNAs, cmalign was used. cmalign is a program of the Infernal package that can simultaneously align multiple sequences to a Rfam CM corresponding to a particular type of ncRNAs. Given a set of ncRNAs of the same type, cmalign returns an alignment augmented with secondary-structure annotation, as shown in Figure 2-2. Such an output was then processed to determine types of mismatches, which can either be covariations or just unpaired changes, between stem regions of orthologous ncRNAs.

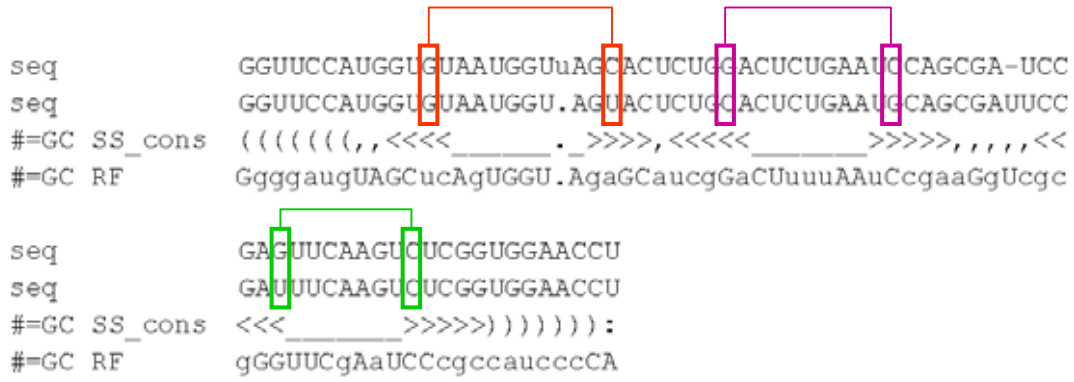

Figure 2-2. A multi-sequence secondary-structure alignment generated by cmalign

Mismatches in double-stranded regions were further categorized into three subtypes. An incomplete covariation is a case where only one base was changed at a base-paired position, such that a conversion occurs between a non-canonical pairing (G-U) and a canonical pairing (G-C or A-U) (*e.g.* red boxes in Figure 2-2). A complete covariation is a case where paired bases were simultaneously mutated to other types of valid pairing, such as G-C to C-G (*e.g.* magenta boxes in Figure 2-2), A-U, U-G, or U-A. A base change that results in a non-canonical and non G-U pairing is referred to as an unpaired change (*e.g.* green boxes in Figure 2-2).

The reason for separating incomplete covariations from complete covariations is that the former type of covariation is a weaker signal for indicating the existence of secondary structures than the latter type. For instance, when the information of covariations is calculated using the standard mutual information (MI) measure (Chiu and Kolodziejczak 1991; Gutell et al. 1992), covariations consisting only of GC and GU pairings do not contribute. However, incomplete covariations still provide useful information for RNA secondary structure prediction (Hofacker et al. 2002; Lindgreen et al. 2006), and should be included in covariation analysis. Thus, in this thesis, the numbers of incomplete covariations and complete covariations were counted separately.

2.1.2. Evaluating different approaches for finding human-mouse synteny-conserved ncRNAs

2.1.2.1. Using the synteny alignments retrieved from public-domain resources

By using the human-mouse syntenic regions that were retrieved from Ensembl Compara release 19, only 26.7% (133/498) of human tRNA genes predicted by tRNAscanSE were found to have synteny-conserved counterparts in the mouse genome (NCBI M30). By using the later releases of the Ensembl Compara database (19-31) where different assemblies of

human (NCBI 35) and mouse (NCBI M32 and NCBI M33) genomes were used, even fewer synteny-conserved tRNA genes could be found. The differences caused by using different Ensembl Compara database releases were due to the changes of strategies for building synteny used by Ensembl. One reason for these changes was to avoid Ensembl Compara containing alignment artefacts caused by repetitive elements. These results show that using existing resources for comparative genomics cannot be relied upon to give a correct estimate of the synteny-conservation ratios of classic ncRNAs between mammalian genomes.

Fortunately, a useful insight was gained from the investigation of tRNA gene clusters in mammalian genomes. A relevant finding is the identification of multiple human-mouse synteny-conserved tRNA gene clusters (for details see section 2.2). As many as ~68% (338/498) of human tRNA genes predicted by tRNAscanSE were found to be in the human-mouse synteny-conserved tRNA gene clusters, although some of their respective synteny-conserved counterparts in the mouse genome might have been lost in evolution.

These results suggest that the real synteny-conservation ratio of human and mouse tRNA genes is much higher than the highest number (26.7%) derived from syntenic alignments retrieved from the Ensembl Compara database alone. Using other public-domain resources of comparative genomics would be unlikely to make much difference, because the algorithms used for creating syntenic alignments in the different releases of the Ensembl Compara database have also been used by these other resources (Schwartz et al. 2003; Frazer et al. 2004). I concluded that the synteny alignments provided by public-domain databases were inadequate for the purpose of generating a comprehensive set of human-mouse synteny-conserved ncRNAs.

2.1.2.2. Using the UBRHPs-bound syntenic regions

Using the UBRHPs-based approach, 74.5% (371/498) of the human tRNA genes that are predicted by tRNAscanSE were found to be conserved in the mouse syntenic regions. These

results suggest that, for finding the human-mouse syntenic regions of classic ncRNAs, the UBRHPs-based approach is likely to be much more effective than using the syntenic regions retrieved from public-domain resources (such as the Ensembl Compara release 29) of comparative genomics.

2.1.3. Results

2.1.3.1. The synteny-conservation ratios of human ncRNAs from Rfam

Since the UBRHPs-bound syntenic regions strategy for finding human-mouse synteny-conserved tRNA genes proved successful, it was further used to identify other human-mouse synteny-conserved ncRNAs. 4,201 unique human ncRNA genomic loci were recruited from Rfam 6.1 for analysing their patterns of conservation in human-mouse syntenic regions. These ncRNAs correspond to 157 classes of ncRNAs (41% of 379 classes of ncRNAs in Rfam 6.1).

Analysing the patterns of conservation of these ncRNAs in human-mouse syntenic regions revealed that the synteny-conservation ratios vary greatly among the different classes. For example, 73.6% of human miRNAs were found to be synteny-conserved; however, only 1.1% of miscellaneous ncRNAs were synteny-conserved (Table 2-1). Overall, 78.1% of the human ncRNAs identified by Rfam6.1 were not found to be conserved in the corresponding mouse syntenic regions. The overall initial estimated synteny-conservation ratio for human ncRNAs is only 21.9%.

In order to evaluate whether the calculated synteny-conservation ratios of human and mouse ncRNAs might be affected by the quality of the mouse genome assembly, the assembly status for the UBRHPs-bound syntenic region corresponding to each human ncRNA was determined. 63.8% of the mouse UBRHPs-bound syntenic regions, where the synteny-non-conserved ncRNAs are supposed to reside, were found to contain genome

sequence fragments labelled either unfinished regions (UR) or whole genome shotgun (WGS) (Table 2-2). It was found that in these UR- or WGS-containing regions there were more synteny-non-conserved ncRNAs than synteny-conserved ncRNAs (compare Table 2-3 with Table 2-2). On average, 63.8% of the synteny-non-conserved ncRNAs and 59.8% of the synteny-conserved ncRNAs are in mouse UR-WGS-containing syntenic regions. The P-value (*Chi*-square test) is far less than 0.001. This result suggests that there is an association between the inability to detect synteny-conserved ncRNAs and the quality of the mouse genome assembly. Consequently, the synteny-conservation ratio for the human ncRNAs that were retrieved from Rfam should be higher than ~22%, because some synteny-conserved ncRNAs will have been missed in mouse UR-WGA regions.

class	mapped to NCBI 35	synteny-conserved	synteny-non-conserved
IRES	8	3 (37.5%)	5 (62.5%)
ribozyme	3	2 (66.7%)	1 (33.3%)
miRNA	87	64 (73.6%)	23 (26.4%)
snoRNA	390	199 (51.0%)	191 (49.0%)
cis-reg	194	96 (49.5%)	98 (50.5%)
tRNA	842	370 (43.9%)	472 (56.1%)
rRNA	350	13 (3.7%)	337 (96.3%)
misc ncRNA	924	10 (1.1%)	914 (98.9%)
snRNA	1403	163 (11.6%)	1240 (88.4%)
Total	4201	920 (21.9%)	3281 (78.1%)

Table 2-1. Conservation of different classes of Rfam human ncRNAs in human-mouse syntenic regions

"IRES" consists of IRES_Bag1, IRES_Bip, IRES_c-myc, IRES_FGF, IRES_L-myc, and IRES_n-myc. "ribozyme" consists of RNaseP_nuc and RNase_MRP. "rRNA" includes 5S_rRNA, 5_8S_rRNA, and SSU_rRNA_5. "cis-reg" consists of Antizyme_FSE, CAESAR, G-CSF_SLDE, GAIT, Histone3, IFN_gamma, IRE, REN-SRE, RRE, SECIS, Spi-1, TAR, and Vimentin3. snRNA consists of U1, U2, U4, U5, U6, U7, U12, and U14. Other ncRNAs, including 7SK, S15, SRP_euk_arch, Telomerase-vert, Vault, and Y., are grouped into "misc ncRNA" (miscellaneous ncRNA).

class	synteny-non-conserved in mouse finished contigs	synteny-non-conserved in mouse UR or WGS
IRES	3 (60.0%)	2 (40%)
ribozyme	0 (0.0%)	1 (100%)
miRNA	6 (26.1%)	17 (73.9%)
snoRNA	61 (31.9%)	130 (68.1%)
cis-reg	37 (37.8%)	61 (62.2%)
tRNA	167 (35.4%)	305 (64.6%)
rRNA	104 (30.9%)	233 (69.1%)
misc ncRNA	346 (37.9%)	568 (62.1%)
snRNA	464 (37.4%)	776 (62.6%)
Total	1188 (36.2%)	2093 (63.8%)

Table 2-2. Distribution of the human synteny-non-conserved ncRNAs in the regions corresponding to mouse finished contigs or UR-WGS-containing regions (regions with unfinished gaps in contig-base sequencing and regions from whole genome shotgun sequencing)

class	synteny-conserved in mouse finished contigs	synteny-conserved in mouse UR or WGS
IRES	2 (66.7%)	1 (33.3%)
ribozyme	1 (50%)	1 (50%)
miRNA	29 (45.3%)	35 (54.7%)
snoRNA	70 (35.2%)	129 (64.8%)
cis-reg	66 (68.8%)	30 (31.3%)
rRNA	6 (46.2%)	7 (53.8%)
tRNA	165 (44.6%)	205 (55.4%)
misc ncRNA	0 (0%)	10 (100%)
snRNA	31 (19%)	132 (81%)
Total	370 (40.2%)	550 (59.8%)

Table 2-3. Distribution of human synteny-conserved ncRNAs in the regions corresponding to mouse finished contigs or UR-WGS-containing regions (regions with unfinished gaps in contig-base sequencing and regions from whole genome shotgun sequencing)

These results show that human ncRNAs are more likely to be synteny conserved in mouse syntenic regions containing only mouse finished contig based sequence (FCS) than in regions that are unfinished (UR) or whole genome shotgun (WGS), but that the effect is small. The average synteny-conservation ratio only increases from ~22% (920/4201) to ~24% (370/1558) when only FCS is considered (see the statistics in the context of mouse finished

contigs in Table 2-2 and Table 2-3). There is a much bigger variation of synteny-conservation ratio between categories. When ncRNAs are considered by category, an inverse correlation was found between the average copy numbers and the synteny-conservation ratios (Figure 2-3).

The previous comparison considers the effect of sequence quality on the apparent ncRNA synteny-conservation ratio. Another factor is assembly completeness. Among the ncRNAs that were investigated, surprisingly low synteny-conservation ratios were found between human and mouse 5S rRNA genes (5S rDNAs). One concern is that the mouse genome assembly (NCBI M33) may have missed *bona fide* 5S rDNAs. Prior to the large-scale sequencing of the human and the mouse genomes, 5S rDNAs were known to be exist as tandem repeats in both genomes (Little and Braaten 1989; Suzuki et al. 1994). It is possible that the strategy of whole genome shotgun sequencing may lead to the omission of tandem repeats, such as 5S rDNAs.

In order to clarify if there are tandemly arranged 5S rDNAs in the mouse genome assembly used in this chapter, a reliable mouse 5S rDNA (GenBank accession number: X71804) was used to search for all 5S rDNAs in NCBI M33. This mouse 5S rDNA sequence, which was published before any large-scale genome sequencing projects were finished, is one unit of the 5S rDNA tandem repeats in the mouse genome (Hallenberg et al. 1994). The result indicates that no such tandem repeats can be found in NCBI M33, while the 5S rDNA tandem repeats can be found in the human genome assembly NCBI 35. In addition, this mouse 5S rDNA is perfectly identical (100%) to the human 5S rDNA. Consequently, the evidence does not suggest that functional 5S rDNAs become synteny-non-conserved after the primate-rodent split. The apparent low synteny-conservation ratio of human and mouse 5S rDNAs is most likely an artefact caused by the missing of *bona fide* 5S rDNAs in NCBI M33.

During the preparation of this thesis, a new mouse genome assembly NCBI M36 is available and the 5S rDNA tandem repeats can be found in this genome assembly. This result

suggests that the quality of the mouse genome assembly has been improved since the release of NCBI M33. NCBI M36 may be a suitable genome assembly for re-estimating the synteny-conservation ratios of human and mouse ncRNAs.

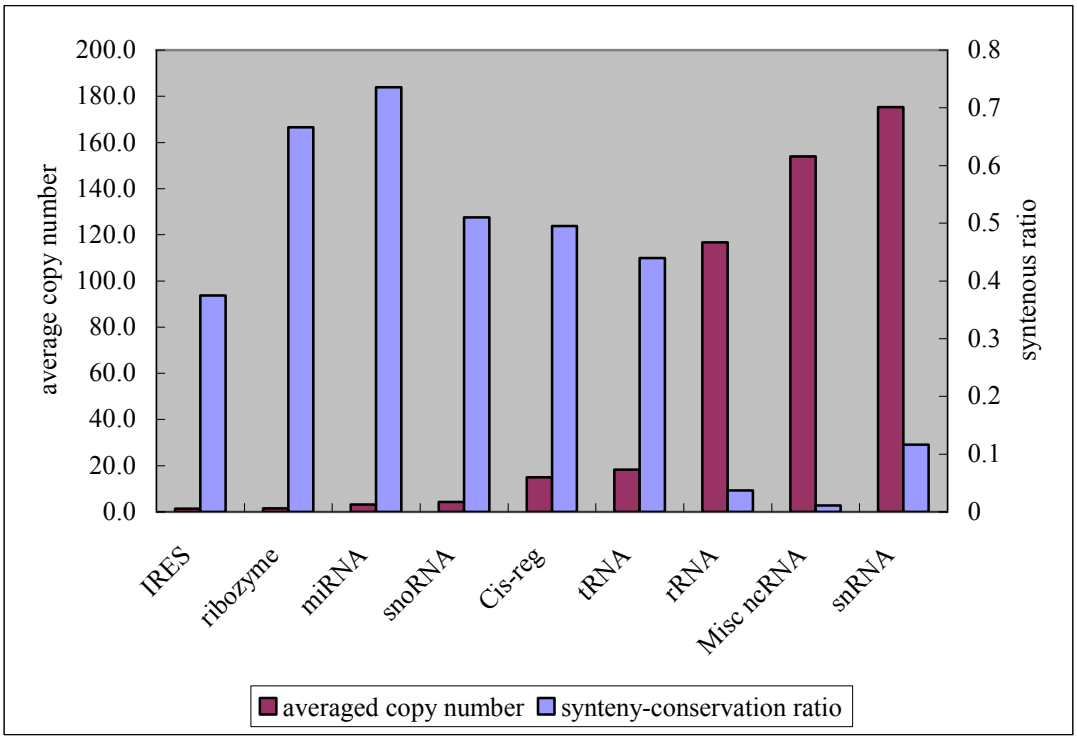

Figure 2-3. Synteny-conservation ratios and average copy numbers for different categories of human ncRNAs (mapped by Rfam)

2.1.3.2. Effect of genome rearrangements on synteny conservation

In order to assess any relationship between genome rearrangements and the estimated synteny-conservation ratios of ncRNAs, chromosome-compatibility and strand-compatibility were taken as the indicators of inter-chromosomal rearrangement and intra-chromosomal rearrangement (for the method see subsection 2.1.1.2.). In the cases where the gene orders and the strand-relationship of the ncRNAs and their flanking genes have been conserved, the syntenic regions were assigned as evolutionary-intact regions.

From this analysis, the syntenic blocks between human and mouse were categorised as

intact regions, segmental inversions, inter-chromosomal translocations, and 'complicated' regions, *i.e.* where evolutionary processes are unclear (Figure 2-1, d). The number of synteny-conserved and synteny-non-conserved ncRNAs in each of these regions is listed in Table 2-4. The synteny-conservation ratio of ncRNAs in the syntenic blocks with inter-chromosomal translocations is not significantly different from that in the intact syntenic blocks (*Chi*-square test, P-value >> 0.1). The synteny-conservation ratio of ncRNAs in the syntenic blocks of the complicated type appears significantly lower than that in the intact syntenic blocks (*Chi*-square test, P-value << 0.001), however this could be an artefact where some synteny-conserved ncRNAs were missed in these regions due to difficulties with the UBRHPs-based method in such regions. It is possible that the method used in this chapter to find synteny-conserved ncRNAs was vulnerable to certain types of genome rearrangements. For instance, if an event of genome rearrangement has changed the linear order of a ncRNA with respect to its flanking synteny landmarks (*i.e.* the protein-coding genes that can be used to define syntenic blocks), this ncRNA may be mistakenly classified as a synteny-non-conserved one. It can be inferred that the calculated synteny-conservation ratios of ncRNAs might be underestimated due to genome rearrangements in "complicated" regions.

The synteny-conservation ratio of ncRNAs in the syntenic blocks with segmental inversions, which are a type of intra-chromosomal rearrangements, is much higher than that in the intact syntenic blocks (*Chi*-square test, P-value << 0.001). No obvious explanation could be found to explain this surprising observation, however such an affect has been reported before. Inversions were found to reduce recombination dramatically (for review see Hoffmann et al. 2004).

synteny conditions	synteny-conserved	synteny-non-conserved	subtotal
evolutionary-intact	579 (24%)	1800 (76%)	2379
segmental inversion	131 (48%)	141 (52%)	272
inter-chromosomal translocation	51 (24%)	163 (76%)	214
complicated	153 (11%)	1183 (89%)	1336

Table 2-4. Numbers of the human-mouse synteny-conserved and the synteny-non-conserved ncRNAs in regions which have undergone different evolutionary events

2.1.3.3. Few covariations in human-mouse synteny-conserved ncRNAs

The aligned sequences of the set of human-mouse synteny-conserved ncRNAs were assessed for covariations as previously defined (see subsection 2.1.1.3.). 64% of human-mouse synteny-conserved tRNAs and 54% of human-mouse orthologous snRNAs were found to not contain any covariations. In addition, no covariations could be found in 70% of human-mouse synteny-conserved miRNAs and in 51% of human-mouse synteny-conserved snoRNAs. Since incomplete covariations are weaker signals than complete ones (see subsection 2.1.1.3.), the cases with only one incomplete covariation were combined with exactly conserved ones (*i.e.* these with no mutations in stem regions), as shown in columns "0-1" base involved in covariations in the following tables (see Table 2-5 and Table 2-7).

On average, 73% of human-mouse synteny-conserved ncRNAs do not provide useful number of covariations (Table 2-5). These results suggest that the alignments of human-mouse synteny-conserved ncRNAs do not contain sufficient covariations for ncRNA finding. Even though the average identity of human-mouse synteny-conserved ncRNAs is 86%, which is only slightly greater than the upper limit of identities requested by some algorithms (*i.e.* ddbRNA and QRNA), covariations are not enriched in the mismatches between the members of each orthologous ncRNA pair. Much of the primary-sequence difference between human-mouse synteny-conserved ncRNAs is attributed to mutations that were found in the

single-stranded regions, and to mutations that may destabilize the stem regions.

Bases in covariations	0-1	2-10	11-23	Subtotal
cis-reg	83 (86%)	13 (14%)	0 (0%)	96 (100%)
misc ncRNA	5 (50%)	4 (40%)	1 (10%)	10 (100%)
IRES	0 (0%)	2 (67%)	1 (33%)	3 (100%)
miRNA	54 (84%)	10 (16%)	0 (0%)	64 (100%)
ribozymes	0 (0%)	2 (100%)	0 (0%)	2 (100%)
rRNA	0 (0%)	12 (92%)	1 (8%)	13 (100%)
snoRNA	139 (70%)	60 (30%)	0 (0%)	199 (100%)
snRNA	110 (67%)	41 (25%)	12 (7%)	163 (100%)
tRNA	282 (76%)	74 (20%)	14 (4%)	370 (100%)
Subtotal	673 (73%)	218 (24%)	29 (3%)	920 (100%)

Table 2-5. Numbers of the human-mouse synteny-conserved ncRNAs that contain various numbers of covariations

	Human-mouse	Human-zebrafish
cis-reg	0.6 (96)	0 (1)
misc ncRNA	3.6 (10)	33.0 (2)
IRES	7.7 (3)	N/A
miRNA	0.7 (64)	3.2 (20)
ribozyme	5.5 (2)	N/A
rRNA	6.2 (13)	9.0 (4)
snoRNA	1.2 (199)	3.5 (2)
snRNA	2.2 (163)	2.1 (79)
tRNA	1.4 (370)	1.1 (185)

Table 2-6. Average numbers of bases involved in covariations per sequence of the human-mouse synteny-conserved ncRNAs and of the human-zebrafish orthologous ncRNAs

N/A: no synteny-conserved ncRNAs found. Each parenthesized value is the number of sequences for respective category of ncRNAs.

Bases in covariations	0-1	2-10	11-33	Subtotal
cis-reg	1 (100%)	0 (0%)	0 (0%)	1 (100%)
Misc ncRNA	0 (0%)	0 (0%)	2 (100%)	2 (100%)
miRNA	7 (35%)	12 (60%)	1 (5%)	20 (100%)
rRNA	0 (0%)	3 (75%)	1 (25%)	4 (100%)
snoRNA	1 (50%)	1 (50%)	0(0%)	2 (100%)
snRNA	51 (64.6%)	25 (31.6%)	3 (3.8%)	79 (100%)
tRNA	133 (71.9%)	52 (28.1%)	0 (0%)	185 (100%)
Subtotal	193 (65.9%)	93 (31.7%)	7 (2.4%)	293 (100%)

Table 2-7. Numbers of the human-mouse-zebrafish orthologous ncRNAs that contain various numbers of covariations

2.1.3.4. Only a few covariations in the human-zebrafish best-fit ncRNAs

From the conclusion that there are insufficient covariations between human and mouse synteny-conserved ncRNAs (for details see subsection 2.1.3.3.), it is reasonable to infer that successful detection of ncRNAs through using comparative ncRNA finding approaches may require more distantly related species than human and mouse. Zebrafish was therefore used in order to investigate if comparing the human genome with other vertebrate genomes can provide significantly more covariations for the purpose of ncRNA finding.

Initially, the zebrafish ncRNAs that are synteny-conserved to human-mouse synteny-conserved ncRNAs were searched in the human-zebrafish UBRPHs-bound syntenic regions; however, only 110 out of 920 human-mouse synteny-conserved ncRNAs could be matched to 58 non-redundant zebrafish ncRNAs. This is most likely due to the lost of synteny between these distantly related species.

In order to recruit more human-zebrafish orthologous ncRNAs, WU-BLAST (Gish 1996-2004) was used to perform a whole genome search for homologues for individual human-mouse synteny-conserved ncRNAs. The best hit for each ncRNA was used for further analysis. 31.8% (293/920) of 920 human-mouse synteny-conserved ncRNAs matched to 112 non-redundant zebrafish ncRNAs. Taking the number of covariations from human-mouse

synteny-conserved ncRNAs as the reference, the number of covariations was found to increase in the human-zebrafish orthologous miRNAs and snoRNAs. However, there were not significantly more covariations in the human-zebrafish orthologous tRNAs and snRNAs than in the human-mouse synteny-conserved ones (Table 2-6). In fact, there were no useful covariations in 65.9% (193/293) of the human-zebrafish orthologous ncRNAs (Table 2-7).

2.1.3.5. Using real genomic alignments to assess the performances of ncRNA finding algorithms

The credibility of existing comparative ncRNA finding algorithms generally comes from benchmarks against adopted test data sets created by aligning well-curated ncRNAs, and not the alignments of ncRNA-containing genomic sequences. For example, one of the popular data sets is the alignments of ncRNAs retrieved from Rfam. These Rfam ncRNAs are different from real genomic sequences in that their 5' and 3' flanking sequences have been carefully trimmed. It is possible that additional noise may be introduced to complicate the detection of consensus RNA motif, if alignments of real genomic sequences, instead of Rfam seed sequences, are used.

In the following test, pairwise and three-way genomic alignments of human tRNA genes were generated to assess the performances of RNAz, QRNA, and ddbRNA. In particular, an additional 20 bases from both the 5' and 3' flanking regions of human tRNA genes were included when generating the alignments. The reason for including (2 x 20) bases is that, including longer flanking sequences to generate alignments may result in a significant drop of identities and only a few of the generated alignments may have identities within the identity range preferred by the three algorithms under test. On the other hand, including flanking sequences shorter than 20 bases may not introduce noise into alignments and the property of the generated alignments is still similar to that of the alignments of curated tRNAs.

One thousand pairwise alignments and one thousand three-way alignments were generated by using ClustalW 1.83. Three algorithms, RNAz, QRNA, and ddbRNA, were

tested on these alignments using their default parameters. These algorithms are ncRNA classifiers. Given a sequence alignment, they will determine whether the sequences as a whole are ncRNAs or not. The result reveals that the performances of none of these algorithms are as good as claimed in their respective papers (Table 2-8). For example, in the original paper of RNAz, the sensitivity was as high as ~95% for detecting tRNA genes by using alignments of identities within 60% ~ 100%; however, using the genomic alignments of human tRNA genes, the sensitivity is only ~49%, when pairwise alignments of identities no less than 60% are used (Table 2-8). In addition, changing the threshold of alignment identity does not improve the sensitivity of any of the algorithms.

In order to rule out the possibility that the bias of using only human tRNA genes could cause the drop in sensitivities, a positive control was performed by using the alignments of human tRNA genes without the 5' and 3' flanking regions. The sensitivity of RNAz on this positive control data set is 94% (data not shown), which is close to the published value (95%) (Washietl et al. 2005). Consequently, the incorporation of flanking regions of human tRNA genes in the test alignments is the only obvious explanation that contributes to the drop in sensitivity of these ncRNA-finding algorithms. These results clearly indicate that it is much harder to identify ncRNAs from the alignments of real genomic sequences than from the alignments of curated ncRNAs.

	RNAz (three-way)	ddbRNA (three-way)	RNAz (pairwise)	ddbRNA (pairwise)	QRNA (pairwise)
All	64.2% (642/1000)	36.2% (362/1000)	61.1% (611/1000)	36.2% (362/1000)	36.6% (366/1000)
Identities >=50%	75.7% (115/152)	57.9% (88/152)	53.8% (148/275)	42.2% (116/275)	46.5% (128/275)
Identities >=60%	75% (6/8)	37.5% (3/8)	48.8% (20/41)	31.7% (13/41)	36.6% (15/41)
Identities >=70%	NA	NA	44.4% (8/18)	5% (1/18)	27.8% (5/18)

Table 2-8. Estimating sensitivities of ncRNA-finding algorithms by using the alignments of genomic sequences of human tRNA genes

Additional 20 bases from both the 5' and 3' flanking regions of human tRNA genes are included when generating alignments of human paralogous tRNA genes. NA means in 1000 alignments, none of them have identities greater than certain thresholds as indicated in the first column of this table. In parentheses, numerators are the numbers of alignments that are correctly classified as ncRNAs. Denominators are the numbers of alignments with identities within a certain range indicated in the first column of this table.

2.1.4. Discussions

2.1.4.1. Practicality of ncRNA prediction based on comparative genomics

With the results already presented in this section (section 2.1), pairwise and three-way alignments of vertebrate genomes do not appear to be ideal data sets for ncRNA finding algorithms. Firstly, there are limited numbers of covariations between orthologous ncRNAs and high primary sequence conservation (see subsections 2.1.3.3. and 2.1.3.4.). Secondly, algorithms that take alignments as input data may be unable to properly score RNA motifs from genome alignments (see subsection 2.1.3.5.).

The difference between the performance of ncRNA finding algorithms on these data sets and their published performance is due to the different data sets used. Many comparative ncRNA finding algorithms have been trained and tested using alignments of ncRNAs, such as seed sequences used to build the Rfam CMs. These alignments are referred to as synthetic alignments in this thesis, because they are not generated directly by aligning genomic sequences. ncRNA finding algorithms perform better on synthetic alignments than genomic alignments. Also, while few, if any, covariations could be found in human-mouse syntenic ncRNAs, there were larger numbers of covariations in these synthetic alignments. One reason

for the difference is that they were generated from more distantly related organisms. A second reason is that the alignments also contained paralogous ncRNAs. Comparison reveals that paralogous ncRNAs can provide more covariations than comparison of orthologous ncRNAs. Synthetic alignments of ncRNAs from ncRNA databases (such as Rfam) may include paralogous ncRNAs. By contrast, synteny alignments should contain few, if any, paralogous ncRNAs.

Under the situation of few covariations in vertebrate ncRNA alignments, the use of multi-way alignments of more than three genomes is an alternative choice that should be considered. In a recent report, eight-way genome alignments were used for genome-wide ncRNA finding (Pedersen et al. 2006). However, several cases presented by Pedersen *et al*. demonstrated that candidate regions of ncRNAs are very well conserved and only a few putative compensatory mutations could be found. In other words, the evidence presented in Pedersen *et al*.'s report actually indicates good conservation at the primary-sequence level. These cases should therefore be considered only as good candidates for functional elements, but not necessarily good candidates for RNA structural motifs.

I therefore conclude that, although comparative ncRNA finding algorithms have been used to find ncRNA in multiple vertebrate genomes, there are still concerns with the results presented in relevant papers. Further examining the ncRNA conservation patterns in multiple vertebrate genomes may be required, in order to determine the potential of using multi-way alignments of vertebrate genomes for ncRNA finding.

It is possible that multi-way ncRNA alignments from sufficient vertebrate genomes will contain enough variations and covariations for ncRNA finding algorithms to work effectively. However, a serious issue for practical genome-wide ncRNA finding is the quality of genome alignments that must be scanned by these algorithms. Up to now, a significant proportion of existing vertebrate genome assemblies are composed of sequences generated from whole

genome shotgun sequencing (WGS). Compared to genome assembly composed of mainly clone based sequencing, genome assemblies consisting of much WGS may contain more sequence misassignment errors and unfinished regions (Cheung et al. 2003). It can be inferred that WGS may result in missing synteny-conserved ncRNAs (false negatives). Even when finished contig sequences are used, multi-way genome alignments provided by public-domain resources may still miss synteny-conserved ncRNAs. For instance, in the 10-way vertebrate genome alignments generated using the Pecan algorithm, a new comparative-genomics resource provided by Ensembl, only 114 human tRNA gene loci were found to be aligned to their synteny-conserved counterparts in other species (data not shown). This number is much smaller than that found using the UBRHPs-based approach (371 loci, see subsection 2.1.2.2.), even though the mouse genome assembly used to generate Pecan alignments consists mainly of finished contig sequences. The UBRHPs-based approach is useful for evaluating ncRNA conservation, as it has been used here, but cannot be used in *de novo* ncRNA prediction as it relies on the location of ncRNAs in one species already being known. An additional source of false negatives, when using ncRNA finding algorithms that depend on genome alignments, will be ncRNAs which are genuinely synteny-non-conserved. In genomes that are distantly related, numerous ncRNAs may be synteny-non-conserved. Such a situation has been demonstrated by the low synteny-conservation ratio of human and zebrafish ncRNAs (see subsection 2.1.3.4.). A similar situation was also encountered when comparing the human and chicken genomes (Hillier et al. 2004).

When evaluating ncRNA finding algorithm performance on genome alignments, it is also necessary to consider the number of false positives. Recently ncRNA finding algorithms were applied to a high-quality set of 28-way vertebrate genome alignments consisting mainly of finished contig sequences and corresponding to 1% of the human genome sequence (Washietl et al. 2007). This is part of the ENCODE project (The ENCODE Project Consortium 2007).

The ncRNA finding algorithms were found to have successfully detected the small number of known ncRNAs. However with an evaluation using shuffled alignments that preserved the dinucleotide frequency to that of the 28-way genome alignments, Washietl *et al*. estimated that these comparative algorithms for genome-wide ncRNA finding may suffer from a high false positive rate, 50% ~ 70%.

All in all, in the context of using existing vertebrate genome assemblies and their alignments, I conclude that the effectiveness of ncRNA finding algorithms that are based on comparative genomics is limited.

2.1.4.2. Proportion of human ncRNAs which are human-mouse synteny-non-conserved

In the process of collecting synteny-conserved ncRNAs to assess comparative algorithms for genome-wide ncRNA finding, the occurrence of synteny-non-conserved ncRNAs was also established. Synteny-conservation ratios of ncRNAs were calculated from this and were found to vary substantially for ncRNAs in different categories (see subsection 2.1.3.1.). At first sight the ratios for all categories appear substantially lower than published estimates of for protein coding genes (Mouse Genome Sequencing Consortium 2002), which were estimated as high as 96%. However there are substantial differences in the protein and ncRNA data sets from which the synteny-conservation ratios have been calculated which should be considered before any conclusions are drawn. For ncRNA genes in vertebrate genomes it is very difficult to determine which predictions are *bona fide* ncRNAs and which are ncRNA pseudogenes. Estimating synteny-conservation ratios for *bona fide* ncRNAs of various classes in vertebrate genomes is therefore difficult. For protein genes it is much easier to determine which ones are pseudogenes and the figures quoted were calculated after pseudogenes have been excluded, unlike figures for ncRNAs.

If many synteny-non-conserved ncRNAs are pseudogenes, the synteny-conservation ratio of human and mouse ncRNAs may be significantly higher than estimated previously in this

section (2.1). Apart from the effect of pseudogenes, there are several other factors that will contribute to an underestimate of the synteny-conservation ratios of ncRNAs, though only to a small extent. Firstly, some uncertain type(s) of genome rearrangements may potentially cause artefacts in finding synteny-conserved ncRNAs (for details see subsection 2.1.3.2.). However, even if the real synteny-conservation ratio of ncRNAs in "complicated" regions is comparable to that under other evolutionary conditions, the overall synteny-conservation ratio of ncRNAs would only be ~2% higher than previously estimated. Secondly, ~40% of the mouse genome assembly (NCBI M33) used in this section was composed of whole genome shotgun sequencing (WGS). However here too, the effect is small, and estimated to have lowered the synteny-conservation ratio by only ~2% (for details see subsection 2.1.3.1.). The major uncertainty relates to the functionality of synteny-non-conserved ncRNA. This issue is further explored in the next chapter (chapter 3).

2.2. Gene-order conservation of mammalian tRNA genes

2.2.1. Materials and methods

2.2.1.1. Recruiting mammalian tRNA gene loci

The genomic loci of the human and mouse tRNA genes were retrieved from Ensembl release 40. These tRNA gene loci were predicted by tRNAscanSE (Lowe and Eddy 1997). The human and mouse genome assemblies used in the following analysis are NCBI 36 and NCBI M36, respectively. They are the most updated assemblies that have been annotated by Ensembl (April 2007, http://www.ensembl.org/index.html). Unlike the previous mouse genome assemblies that consist of many sequences generated from whole genome shotgun sequencing (WGS), NCBI M36 is a highly polished genome assembly, where most of the sequence is composed of finished contig sequences (http://www.ncbi.nlm.nih.gov/projects/genome/seq/NCBIContigInfo.html). Investigating the gene-order conservation of mammalian

tRNA genes using this higher quality mouse genome assembly should therefore be far less affected by genome assembly artefacts.

One issue when trying to understand the evolution of tRNA genes is that, by comparing two genomes, it is difficult to determine whether a difference (*i.e.* an unaligned tRNA gene symbol, referred to subsequently as a 'gap') in an alignment between them is caused by the deletion and/or degradation of tRNA genes in one genome or the insertion of tRNA genes in the other. One way to try and distinguish between these possibilities is to recruit a set of tRNA gene loci, as an external reference, from a third genome that is an outgroup of the first two. An organism that has split from a common ancestor of placental mammals (including human and mouse) before the primate-rodent split can suffice for this purpose. In the following analysis, opossum was used which is a species of marsupials. Marsupials diverged from placental mammals about 180 millions years ago (Lawn et al. 1997). By using such an external reference, the evolutionary event that led to a gene order difference in human and mouse may possibly be inferred. For instance, when considering alignments of tRNA gene clusters if a symbol insertion found in a human-mouse tRNA symbol alignment remains an insertion in a human-opossum tRNA symbol alignment, this insertion is likely to be the result of a duplication or transposition event that occurred in the genome of the human ancestors. Likewise, a deletion and/or degradation of a tRNA gene locus after the primate-rodent split may also be inferred. The tRNA gene loci of the opossum genome were retrieved from Ensembl release 40 and the opossum genome assembly used in the following analysis is MonDom4.

There is one concern about using the tRNA gene arrangements in the opossum genome. The sequence assembly of the opossum genome consists mainly of the sequences from whole genome shotgun sequencing (http://www.ensembl.org/Monodelphis_domestica/index.html). For this reason, the opossum tRNA gene loci are used only for inferring the evolutionary

history after the primate-rodent split, but not that before the primate-rodent split, *i.e.* apparent differences in gene order unique to opossum were ignored.

2.2.1.2. Identifying the syntenic tRNA gene clusters

The steps for identifying synteny-conserved tRNA gene clusters are presented in the flowchart in Figure 2-4. In comparing the tRNA gene order in the human and mouse genomes, the first genome is the human genome and the second genome is the mouse genome. The tRNA gene loci were sub-grouped into clustered and non-clustered ones (singlets), respectively. A threshold of the maximal distance allowed between the nearest neighbouring tRNA genes in a cluster was defined to be 1 mega bases. This threshold was set as the minimum distance required to ensure the super cluster (*e.g.* 150 tRNA gene loci) that spans several mega bases on human chromosome 6 remained a single unit.

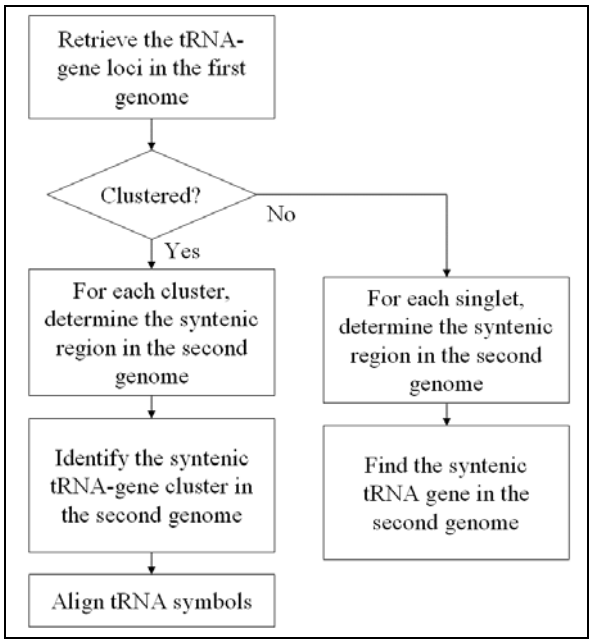

Figure 2-4. The procedure of identifying the syntenic tRNA gene clusters in mammalian genomes

For each human tRNA gene cluster, the syntenic region in the mouse genome was determined using the UBRHPs-based approach (for details see subsection 2.1.2). Each human tRNA gene cluster, together with the corresponding tRNA gene cluster in the syntenic region

in the mouse genome, becomes a pair of synteny-conserved tRNA gene clusters. The conservation of tRNA gene order was investigated by comparing the arrangements of tRNA gene loci in each pair of human-mouse syntenic clusters.

2.2.1.3. Assigning symbols to mammalian tRNA gene loci

A general approach for investigating gene-order rearrangements is to represent genes as symbols and then compare their order (for review see Sankoff and El-Mabrouk 2000). In investigating the tRNA gene-order conservation, I followed a similar strategy. Each tRNA gene locus was thus assigned with a symbol according to its features. These features include the anticodon types and the genomic orientation. For example, there are two different anticodons, GCA and ACA, used by tRNAs for carrying cysteines (tRNA-Cys). Cys1 was used to represent the tRNA-Cys gene loci that have the anticodon GCA. Cys2 was used to represent the tRNA-Cys2 gene loci that have the anticodon ACA. If a Cys1 was on the forward strand of a chromosome, a suffix "F" was added. Conversely, Cys1R was used when a tRNA-Cys1 gene locus was on the reverse strand of a chromosome. A lookup table of the relations between anticodon types and tRNA gene symbols can be found in Table A 1, Appendix A.

There is one consideration in the use of a set of anticodon based tRNA gene symbols. If there are transitions of anticodon types, finding two loci with the same anticodon types does not necessarily mean that both loci should have evolved from a common ancestral locus. Likewise, a mismatch of the anticodon types does not necessarily mean that the two tRNA gene loci should have evolved from two distinct ancestral loci.

In order to compensate for this limitation of the anticodon-type tRNA gene symbols in the gene-order comparison, another set of tRNA gene symbols based on sequence identities was also created. The steps are as follows. Firstly, all human tRNA gene loci were classified according to their anticodons. For example, there are two anticodons, UUU and CUU, for

tRNAs that carry the amino acid lysine. All lysine-tRNA genes, which carry either one of the two anticodons, were grouped together. Secondly, using the TIGR Gene Indices Clustering Tools (TGICL) (TIGR 2002-2003), each group of tRNA genes was further divided into subgroups according to pairwise sequence identities. The grouping was performed by Cap3 (called by TGICL) (Huang and Madan 1999) using default parameters. Subgroup assignments were performed automatically using TIGR. For example, Thr-tRNAs were divided into S_Thr_1, S_Thr_2, and S_Thr_3 subgroups. Forty subgroups were so created. The pairwise sequence identities within individual subgroups range from 94% to 100%. Sequences in each group are fairly homogeneous at the primary-sequence level. Each subgroup was used as a unique sequence type of tRNA genes. For the purpose of comparing the tRNA gene orders in different genomes, each tRNA gene loci in the human, mouse, and opossum genomes was assigned with the best-hit sequence type according to its sequence identities to all sequence types. The sequence-type symbols of tRNA genes were used to find anticodon transitions that may cause the generation of gaps in the anticodon-type symbol alignments.

2.2.1.4. Filtering out possible tRNA-like SINEs

In this tRNA gene-symbol based comparison one issue is filtering out the large number of tRNA-like SINEs which are present mammalian genomes. If too many are included, many false gaps will be generated when comparing the gene orders of two different genomes. In practice, it is very difficult to prepare a comprehensive list of free of the many tRNA-like SINEs. For instance, there are, in the mouse genome, thousands of species-specific SINEs that are related to tRNA genes (Mouse Genome Sequencing Consortium 2002). This is discussed in more depth in the introduction to chapter 3, however for the purposes here, only mouse tRNA genes with tRNAscanSE bit scores greater than 40 were included. There are two reasons for setting this threshold. First, in the set of 2,345 tRNA genes of low scores (tRNAscanSE bit score < 40), 97.3% (2,282) of them overlap with SINEs. Secondly, the

bit-score distribution of the mouse tRNA genes reveals a bi-modal distribution (data not shown), where bit score 40 seems to be a point that can preserve as many normal mouse tRNA genes as possible, while most of the tRNA-like SINEs can be removed. After this filtering, 504 tRNA gene loci in the mouse genome were recruited for this study, while without any particular filtering, there are by coincidence 504 human tRNA gene loci. Only 11.1% (55 / 504) of the high-scoring tRNA gene loci in the mouse genome overlap with SINEs.

For the opossum tRNA gene loci only the simple pseudogene filter by tRNAscanSE was used to clean the data set of the opossum tRNA gene loci. This is due to there being relative little knowledge about repetitive elements in the opossum genome during the preparation of this manuscript.

2.2.1.5. Types of gene-order conservation

The tRNA gene symbols of the human and mouse tRNA gene clusters were initially aligned using a dynamic programming implementation in Biojava (http://biojava.org). Except in the cases of perfect-type conservation, there were gaps in the tRNA symbol alignments of the human-mouse or human-opossum syntenic tRNA gene clusters. According to the source of the unaligned symbols, these gaps were assigned as either insertions or deletions. The unaligned tRNA symbols that were from the human genome were assigned as insertions. Conversely, when the unaligned symbols were from the other genome, either the mouse or opossum genome, the gaps were assigned as deletions. This convention was used only for indicating the source of gaps in symbol alignments, without implying anything about the evolutionary origin of these gaps.

The gene symbols from the two genomes are aligned on both strands to generate two separate alignments, *i.e.* for human and mouse one is the human-forward-strand *versus* mouse-forward-strand alignment; the other is the human-forward-strand *versus* mouse-reverse-strand alignment. The two symbol alignments automatically generated by using

the Biojava were then examined manually. The purpose of this step was to decide which alignment can best explain the evolutionary relationship between the human-mouse synteny-conserved tRNA gene clusters. In some cases, this decisions was not easy to make, especially when there had been chromosomal inversions in the tRNA gene clusters after the primate-rodent split. In cases where there were also synteny-conserved protein-coding genes intervening in the synteny-conserved tRNA gene clusters, these protein-coding genes were used as landmarks. These intervening protein-coding genes could be used to sub-divide tRNA gene clusters into smaller sub-clusters allowing conservation of tRNA gene orders within these sub-clusters.

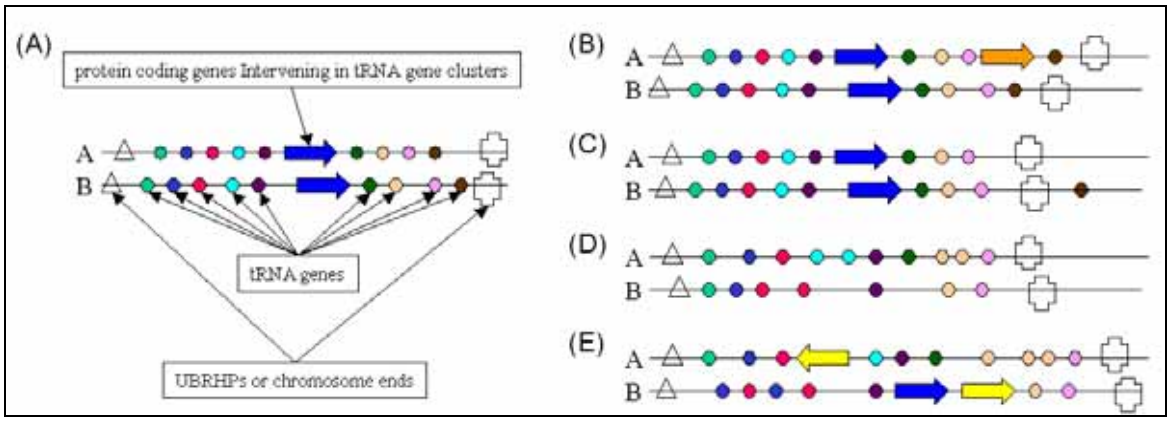

Figure 2-5. Different types of tRNA gene-order conservation

Five types of conservation patterns of the mammalian tRNA genes were defined as follows (see also Figure 2-5):

- "Perfect" conservation (Figure 2-5, A) refers to a pair of syntenic tRNA gene clusters in which the arrangement of all functional elements, including tRNA genes and intervening protein-coding genes, has been completely conserved and all the symbols can be perfectly aligned.

- "Sub-perfect" conservation refers to a pair of synteny-conserved clusters where there are

minor differences between them. "Sub-perfect type-one" conservation (Figure 2-5, B) is used when there is between-syntenic-clusters inconsistency in the physical arrangement of protein-coding genes intervening in the clustered tRNA genes. "Sub-perfect type-two" conservation (Figure 2-5, C) is used when there are non-syntenic tRNA genes at the ends of the syntenic clusters.

- "Gapped" conservation (Figure 2-5, D) refers to a pair of synteny-conserved clusters where a few tRNA gene loci are not aligned.

- "Complicated" conservation (Figure 2-5, E) refers to a pair of synteny-conserved clusters where there may have been multiple genome rearrangements. The existence of a complicated case is inferred when there are multiple gaps in the tRNA symbol alignment. Besides, the linear relations of the protein-coding genes in the neighbourhood of tRNA gene loci may have also changed.

- "Single" conservation refers to the case where, in a tRNA gene cluster, only one synteny-conserved tRNA gene locus was found in the corresponding syntenic region in the second genome.

2.2.1.6. Checking the conservation of the internal promoters of tRNA genes

For the purpose of checking the conservation of the internal promoters in these tRNA genes, eufindtRNA (Pavesi et al. 1994) was used. eufindtRNA is a tRNA-finding algorithm that can recognize the features of important promoting elements, such as A and B boxes, termination signals, and relative spacing between signals, for the transcription of eukaryotic tRNAs. The relaxed mode of eufindtRNA was used here to evaluate only the integrity of intragenic control regions. The stringent mode of eufindtRNA, which can also assess the quality of termination signals, was not used in the following analysis, because evidence suggests that some variations in termination signals are allowed (Gunnery et al. 1999).

2.2.2. Results

2.2.2.1. 32 human-mouse synteny-conserved tRNA gene clusters

Among the 504 tRNA gene loci in the human genome, 92 (18%) loci are not clustered (singlets) (Table 2-9). There are more singlets (27%, 134/504), and also fewer clustered tRNA gene loci in the mouse genome than in the human genome. The significance of this finding is unclear given that we know the data sets used are not entirely clean of loci such as tRNA-like SINEs.

	number of tRNA genes	number of clusters	number of clustered tRNA gene loci	number of non-clustered tRNA gene loci (singlets)
human	504 (100%)	38	412 (82%)	92 (18%)
mouse	504 (100%)	48	370 (73%)	134 (27%)
opossum	991 (100%)	121	597 (60%)	394 (40%)
opossum (bit score >= 40)	546 (100%)	46	408 (75%)	138 (25%)

Table 2-9. The statistics of clustered tRNA gene loci in the human, mouse, and opossum genomes

	human tRNA gene loci in clusters	synteny-conserved clusters	human tRNA gene loci in synteny- non-conserved clusters	human tRNA gene loci in the synteny-conserved clusters
human-mouse	412 (100%)	32	29 (7%)	383 (93%)
human-opossum	412 (100%)	28	181 (44%)	231 (56%)

Table 2-10. The synteny conservation of clustered human tRNA gene loci

Eighty-two percent and seventy-three percent of the tRNA gene loci (Table 2-9) in the human and mouse genomes were grouped into 38 and 48 clusters, respectively (for the detailed lists see Table A 2 and A 3, Appendix A). Thirty-two pairs of human and mouse tRNA gene clusters were found to be synteny-conserved (for a detailed list see Table A 4 in

Appendix A). 93% (383/412) of the tRNA gene loci that are clustered in the human genome are within the human-mouse synteny-conserved tRNA gene clusters (Table 2-10). The conservation of tRNA gene order was then investigated by aligning the symbols of the 32 human-mouse pairs of synteny-conserved tRNA gene clusters. The gene order comparison was performed primarily by using the anticodon-type symbols of tRNA gene loci. The result reveals some unaligned regions in the tRNA symbol alignments (Table 2-12). Among the 383 clustered human tRNA gene loci that reside in the human-mouse synteny-conserved clusters, 230 loci (60%) can be aligned without much uncertainty. A special case is the alignment of human cluster 4.1.36 and mouse cluster 5.1.26. In the initial alignment of this pair of syntenic clusters, only 10 out of the 36 human loci can be aligned. By manual curation, a track of 15 tRNA gene loci that can be aligned in an inverted way was found (Figure 2-6).

Figure 2-6. The conservation pattern of the human tRNA gene clusters 4.1.36 and its syntenic cluster in the mouse genome (see next page)

tRNA gene loci are represented in two ways: (1) the ones in rounded rectangles with symbols indicating the codon type of tRNA genes; (2) the ones that are plotted in red dots, indicating the loci whose evolutionary origins cannot be unambiguously assigned based on sequence identity. Dotted-rounded rectangles are used to indicate the unitary blocks that repeat for multiple times in both the human and mouse genomes. Arrows are used to indicate the orientation of these repetitive blocks, where the red ones are used to indicate the complete unitary blocks, and the cyan and magenta ones are used to indicate the incomplete unitary blocks. Red lines are used to indicate the possible region of a chromosomal inversion.

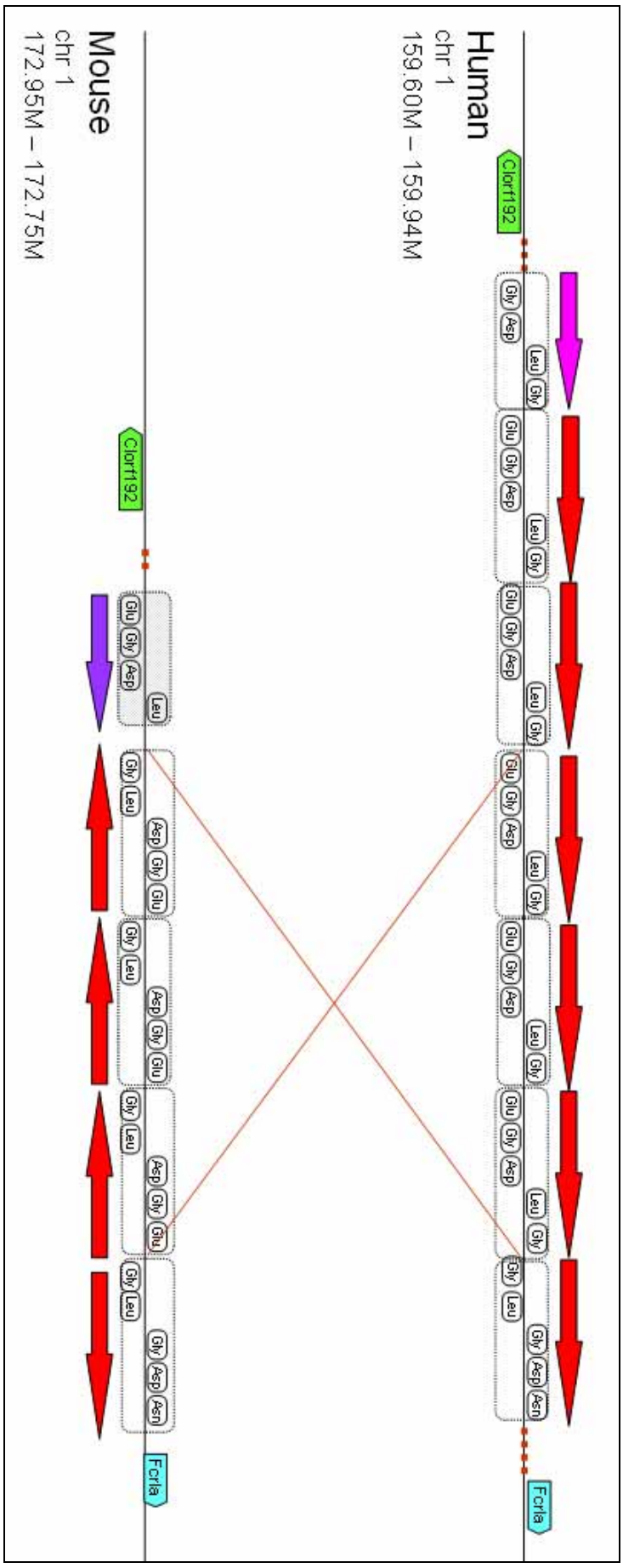

Figure 2-6 (for figure legend see the previous page)

2.2.2.2. Anticodon transitions are rare after the primate-rodent split

The conservation of gene order was also evaluated by comparing the arrangements of the sequence-type symbols. The purpose here was to find if there was any evidence of anticodon transitions that could cause mutated tRNAs to carry different amino acids. The result reveals that, as expected, anticodon transitions in mammalian genomes are very rare. By comparing the human and mouse synteny-conserved tRNA sequence types, only six anticodon transitions were found (Table 2-11). The observed anticodons in these six human tRNA gene loci are not consistent with the expectations inferred from their respective sequence types. The transitions from tRNA-Cys to tRNA-Ser and tRNA-Tyr in human cluster 17.7.20 are also supported by the conserved arrangement of the tRNA gene loci in the corresponding mouse syntenic cluster, in which there are only tRNA gene loci of anticodon type Cys1 and sequence type S_Cys_1.

cluster ID	Coordinate	observed anticodon type	observed sequence type	expected anticodon type	bit score
3.1.42	chromosome:NCBI36:1:147561290:147561360:-1	Val3	S_Gly_1	Gly2/Gly3	60.62
3.1.42	chromosome:NCBI36:1:146185653:146185726:1	Asn2	S_Asn_1	Asn1	52.07
14.6.150	chromosome:NCBI36:6:27379547:27379618:-1	Thr3	S_Met_1	Met1	46.44
14.6.150	chromosome:NCBI36:6:28811185:28811256:-1	Val4	S_Ala_1	Ala3/Ala4	64.08
17.7.20	chromosome:NCBI36:7:148886066:148886138:1	Tyr1	S_Cys_1	Cys1	49.4
17.7.20	chromosome:NCBI36:7:148936400:148936471:1	Ser4	S_Cys_1	Cys1	62.1

Table 2-11. Transitions of the anticodons of tRNA gene loci

2.2.2.3. Numerous gaps between synteny-conserved human and mouse clusters

There were numerous gaps between synteny-conserved human and mouse clusters (Table 2-12). As many as 40% of the human loci in these gene clusters were insertions in symbol alignments. According to the distribution pattern of gaps in the symbol alignments, the synteny-conserved tRNA gene clusters were further grouped into the five conservation types

(Table 2-13) (for the definitions of the five types, see subsection 2.2.1.5. , Materials and Methods). ~65% (267/412) of the human clustered tRNA gene loci are within the human-mouse synteny-conserved clusters where there are multiple gaps in their symbol alignments ("gapped", Table 2-13). Other statistics about the conservation types, aligned loci, *etc*. of the human-mouse synteny-conserved clusters are listed in Table 2-13.

An attempt was made to look for possible relationships between human-mouse non-syntenic tRNA clusters by searching for similarities in the gene order. No significant tRNA gene-order conservation was discovered.

	human tRNA gene loci in the synteny-conserved clusters	insertions in the symbol alignments	aligned human tRNA gene loci in symbol alignment
human-mouse	383 (100%)	153 (40%)	230 (60%)
human-opossum	231 (100%)	104 (45%)	127 (55%)

Table 2-12. The statistics (aligned and inserted regions) of the human-mouse tRNA symbol alignments

conservation type	human tRNA gene clusters	human tRNA-gene loci	aligned loci in the human genome	unaligned loci (insertions)*
perfect	8	17 (4%)	17	0
sub-perfect type one	5	36 (9%)	36	0
sub-perfect type two	4	11 (3%)	9	2
gapped	8	267 (65%)	157	110
complicated	1	42 (10%)	6	36
single	5	10 (2%)	5	5
synteny-non-conserved	7	29 (7%)	0	29
subtotal	38	412 (100%)	230	182

Table 2-13. The statistics of the gene-order conservation of human and mouse tRNA gene clusters

*: There are also 61 deletions in the human-mouse tRNA symbol alignments. Deletions are defined as the additional tRNA symbols in the mouse genome that cannot be aligned to suitable syntenic counterparts in the human genome. Fifty-eight deletions belong to gapped conservation type. Three deletions belong to "single" conservation type.

In addition to clustered tRNA gene loci, some non-clustered tRNA gene loci were also found to be conserved in the corresponding mouse syntenic regions. There are 92 non-clustered tRNA gene loci in the human genome. 37 of them are human-mouse synteny-conserved (see Table A 5, Appendix A).

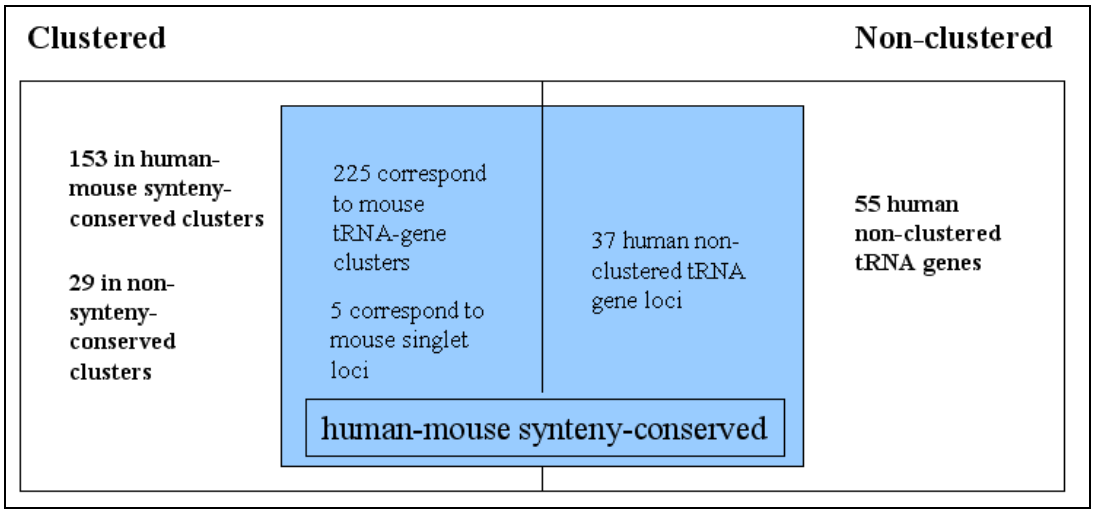

Figure 2-7. Summary of the synteny conservation of human and mouse tRNA gene loci

When the gene order is taken into consideration, only ~53% (267/504) of the human tRNA gene loci are synteny-conserved. This value is much lower, by 21% (74% - 53%), than the previous estimate made under the ignorance of the gene-locus arrangement in each tRNA gene cluster. Obviously, the main source of this big difference is that the arrangements of 153 loci within the synteny-conserved clusters are not conserved (Figure 2-7).

2.2.2.4. The association of the synteny-conservation of tRNA gene clusters with the quality of genome assembly

One factor that may affect the determination of synteny-conservation of tRNA genes is the quality of genome assembly. It is therefore important to explore if the synteny-non-conservation of human tRNA gene loci is associated with unfinished regions or

WGS in the genome assemblies. The investigation reveals that in the synteny-conserved tRNA gene clusters, the gaps in the tRNA symbol alignments are generally not related to the quality of genome assembly (Table 2-14). Within the human-mouse synteny-conserved tRNA gene clusters, all the genomic sequences intervening between each neighbouring tRNA gene loci in the mouse genome are composed of finished contig sequences, but no unfinished contigs nor WGS. Besides, four out of the seven synteny-non-conserved human tRNA gene clusters were found to be in the regions where the genome assembly consists of finished contig sequences.

human tRNA gene clusters	FCS	CSN	WGS*
synteny-conserved clusters	31^{+}	0	0
synteny-non-conserved clusters	4	1	2

Table 2-14. Relation of synteny-conservation of tRNA gene clusters and the quality of the mouse genome assembly
FCS: finished contig sequence; CSN: unfinished contig sequence (with gaps); WGS: whole genome shotgun sequence

*: there are also unfinished gaps in these WGSs.

+: In 3 human-mouse synteny-conserved clusters, the intervening (mouse) genomic sequences between each pair of neighbouring tRNA gene loci are composed of finished contig sequences (FCS), while there are WGSs between the (5' or 3') end tRNA gene loci of a cluster, and the protein-gene boundaries that define the corresponding human-mouse syntenic blocks.

human non-clustered tRNA gene loci	FCS	CSN	WGS
synteny-conserved singlets	36	0	1
synteny-non-conserved singlets	51	1	3

Table 2-15. Relation of synteny-conservation of non-clustered tRNA genes (singlets) and the quality of the mouse genome assembly

The association between the quality of genome assemblies and the synteny conservation of non-clustered tRNA gene loci (singlets) was also evaluated. The inability to find syntenic mouse counterparts to human tRNA gene singlets does not seem to be biased by the quality of genome assembly (Table 2-15). Among the 55 synteny-non-conserved singlets, 51 of the corresponding syntenic regions in the mouse genome are composed of FCS, but no WGS.

These results suggest that the gaps in the synteny-conserved clusters, the synteny non-conservation of at least four human tRNA gene clusters, and the synteny non-conservation of 51 non-clustered human tRNA gene loci, are more likely to be caused by evolutionary events, *i.e.* genome rearrangements, retro-transpositions, degraded genes (pseudogenes), tRNA-related SINEs, *etc*.

2.2.2.5. The information from the tRNA gene loci in the opossum genome

The comparison of the human and opossum tRNA gene loci reveals that there are fewer (28) human-opossum synteny-conserved tRNA gene clusters than human-mouse synteny-conserved clusters (Table 2-10). An example is that no opossum tRNA gene clusters were confirmed to be syntenic counterparts of the super tRNA gene cluster, 14.6.150, which is on human chromosome 6. Besides, more gaps (unaligned human tRNA gene symbols) were found in the human-opossum alignments than in the human-mouse alignments. These findings essentially fit expectations because opossum split from the placental mammals long before the primate-rodent split and the genome assembly quality is much lower.

The arrangement of tRNA genes in the opossum genome provides information that can help us understand tRNA gene evolution in mammalian genomes. The insertions and deletions in the human-mouse tRNA symbol alignments, can be re-categorized by examining the 3-way, human-mouse-opossum, alignments of the tRNA gene symbols and applying the following rules:

- If an inserted tRNA gene symbol is found in opossum in the human-opossum tRNA symbol alignment, this symbol insertion may represent a deletion or degradation of a tRNA gene locus in the mouse genome after the primate-rodent split.
- If an inserted tRNA gene symbol cannot be found in opossum in the human-opossum tRNA symbol alignment, this symbol insertion may represent an insertion of a tRNA gene locus in the human genome after the primate-rodent split.
- If a deleted tRNA gene symbol is also missing from opossum in the human-opossum

tRNA symbol alignment, this symbol deletion may represent an insertion of a tRNA gene locus in the mouse genome after the primate-rodent split.

- If a deleted tRNA gene symbol can be found in opossum in the human-opossum tRNA symbol alignment, this symbol deletion may represent a deletion or degradation of a tRNA gene locus in the human genome after the primate-rodent split.

The re-categorization of gaps in the human-mouse tRNA symbol alignment was performed using the above rules.

human tRNA gene clusters	insertions in the human-mouse alignments	Post primate-rodent-split insertions in the human genome	Post primate-rodent-split deletions/degradations in the mouse genome
6.1.3	1	1	0
13.5.17	10	9	1
16.6.2	1	1	0
17.7.20	2	NA	NA
18.8.4	1	1	0
20.11.2	0	0	0
23.13.2*	2	0	1
24.14.14	9	0	7
26.15.2	1	1	0
30.16.5	2	1	1
33.17.8	2	1	1
37.19.2*	2	0	2
Subtotal	44	15	13

Table 2-16. Evolutionary origin of the insertions in the human-mouse tRNA symbol alignments

NA: not available. The placement of gaps in the alignments is not unique.

*: these tRNA gene clusters are not human-mouse synteny-conserved, but are human-opossum synteny-conserved.

human tRNA gene clusters	deletions in the human-mouse alignments	Post primate-rodent-split insertions in the mouse genome	Post primate-rodent-split deletions/degradations in the human genome
13.5.17	1	1	0
17.7.20	34	NA	NA
18.8.4	1	0	1
20.11.2	1	1	0
Subtotal	38	2	1

Table 2-17. Evolutionary origin of the deletions in the human-mouse tRNA symbol alignments

NA: not available. The placement of many gaps in the alignments is not unique.

Based on the information derived from comparing the human-opossum synteny-conserved tRNA gene clusters, 28 insertions (*i.e.* the unaligned tRNA symbols in the human genome) can be re-classified to 15 post primate-rodent-split insertions of tRNA gene loci in the human genome, and 13 post primate-rodent-split deletions/degradations of tRNA gene loci in the mouse genome (Table 2-16). Two human tRNA gene clusters that are not human-mouse synteny-conserved were found to be human-opossum synteny-conserved (23.13.2 and 37.19.2, Table 2-16). These two clusters may have been deleted/degraded in the mouse genome after the primate-rodent split. Besides, among the deletions in the human-mouse tRNA symbol alignments, there are two post primate-rodent-split insertions of tRNA gene loci in the mouse genome, and one post primate-rodent-split deletion/ degradation of a tRNA gene locus in the human genome (Table 2-17).

2.2.2.6. Duplicated multi-loci blocks in the mammalian tRNA gene clusters

There are several human-mouse synteny-conserved tRNA gene clusters in which gaps in the tRNA symbol alignments cannot be unequivocally placed, due to the existence of so many unaligned regions in the tRNA symbol alignments. Human cluster 3.1.42 is a classic example (Figure 2-8). In the human cluster 3.1.42, not only the arrangement of the tRNA gene loci, but also the relation of the tRNA gene loci to the neighbouring protein-coding genes has changed.

One question that arises from these observations is about the mechanism by which tRNA gene loci in mammalian genomes evolve. Are there any particular rules that govern the changes of tRNA gene orders in these syntenic clusters? Or is the rearrangement of the tRNA gene loci in these synteny-conserved clusters generally random?

Interestingly, the arrangement of the opossum tRNA gene loci provides useful information on this issue. By comparing the arrangements of tRNA gene loci as well as neighbouring protein-coding genes in the human, mouse, and opossum genomes, a vague picture about the evolution of the tRNA gene loci in the human cluster 3.1.42 is revealed (Figure 2-8). My conclusions are summarized as follows:

- The syntenic clusters contain four distinct blocks, A, B, C, and D, of protein-coding genes. The gene order in each block is quite conserved among the human, mouse, and opossum genomes.

- The arrangements of the first three blocks, including A, B, and C, consisting of protein-coding genes, are quite conserved in the mouse and opossum genomes. However, in the human genome, the arrangement of A, B, and C is as C_R-A-B. The subscript "R" indicates that the C block is on the reverse strand. It can be inferred that there might be one segmental inversion in the human genome after the primate-rodent split.

- Between the C and D protein-gene blocks, the arrangements of tRNA gene loci in the human, mouse, and opossum genomes is very different.

- There are multiple species-specific multi-tRNA-loci duplications in each cluster. No common unit blocks of these species-specific duplications were found among the human cluster, 3.1.42, and its syntenic clusters in the mouse and opossum genomes. In the human cluster, 3.1.42, there are two blocks of Gln2-Asn1 tRNA gene loci, two blocks of Gln2-His1 loci, and two duplicated blocks of Asn1-Asn1 loci. In the syntenic tRNA gene cluster in the mouse genome, there are three duplicated blocks of Asn1-His1 tRNA gene loci, two duplicated blocks of Glu1-Gly3 loci. In the syntenic cluster in the opossum genome, there are at least seven types of duplicated blocks, where each distinct type consists of unique combinations of different tRNA gene loci.

- In the human cluster 3.1.42, there are 16 tRNA-Asn1 gene loci which are arranged into

several separated sub-clusters consisting of varied numbers of tRNA-Asn1 gene loci. By contrast, there are 7 tRNA-Asn1 gene loci that are interspersed in the syntenic mouse cluster. 15 out of the human 16 tRNA-Asn1 gene loci were found to have better intra-cluster (other tRNA gene loci in the same cluster, 3.1.42) hits than inter-cluster hits (other tRNA gene loci not in cluster 3.1.42). This means that these tRNA-Asn1 gene loci in the human cluster 3.1.42 are more likely to be generated by intra-cluster duplications than by inter-cluster duplication. In addition to at least three duplicated blocks of two tRNA-Asn1 gene loci, there appear to have been a number of tandem duplications of single tRNA-Asn1 gene loci.

- Some of the single units of duplicated multi-tRNA-loci blocks in one genome cannot be found in the other genome(s). For instance, the Glu1-Gly3 unit of a pair of duplicated blocks in the mouse genome cannot be found in either the human or opossum syntenic cluster.

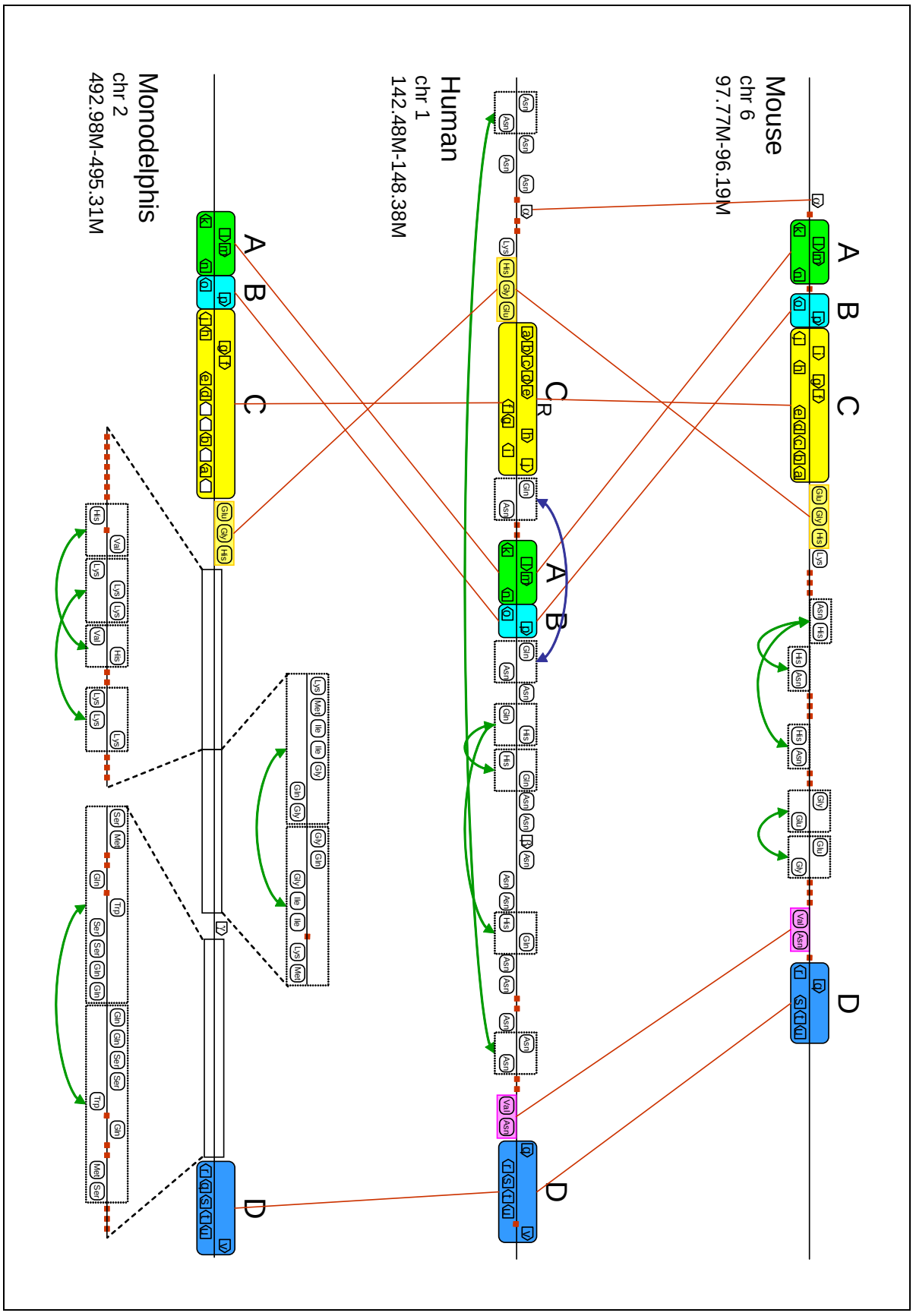

Figure 2-8 (for figure legend see the next page)

Figure 2-8. The conservation pattern of human tRNA gene cluster 3.1.42 and its syntenic clusters in the mouse and opossum genomes

This figure was not prepared to the scale, because it was intended to provide an overview of the putative, both intra-species and inter-species, tRNA gene locus duplications on human chromosome one, 142.48M-148.38M, with respect to the corresponding syntenic regions in the mouse and opossum genomes.

tRNA gene loci are represented in two ways: (1) the ones in rounded rectangles with symbols indicating the codon type of tRNA genes; (2) the ones that are plotted in red dots, indicating the loci whose evolutionary origins cannot be unambiguously assigned based on sequence identity. Color-shaded boxes are used to indicate the inter-species synteny-conserved regions, which are connected by red lines. The dotted boxes around multiple tRNA gene loci are used to indicate the regions that may be involved in intra-species duplications. Curved lines are used to indicate the relation between intra-species duplicated blocks, where the blues ones are used to indicate the blocks of directed duplications, and the green ones are used to indicate the blocks of inverted duplications.

Protein coding genes are represented using arrows. Synteny-non-conserved protein coding genes are represented as open arrows. The symbols for the protein-coding genes used as the landmarks in this figure are as follows:

a	TXNIP	f	NUD17_HUMAN	k	FMO5	p	GJA8	u	ZA20D1
b	LIX1L	g	POLR3C	l	CHD1L	q	BOLA1	v	VPS45A
c	RBM8A	h	ZNF364	m	BCL9	r	HIST2H2AB	α	PDE4DIP
d	ANKRD35	i	CD160	n	ACP6	s	SV2A	β	NP_110423.3
e	PIAS3	j	PDZK1	o	GJA5	t	MTMR11	γ	HIST2H2AA3

2.2.2.7. The synteny conservation of non-clustered tRNA gene loci in mammalian genomes

In addition to the exploration about the evolution of tRNA gene loci in clusters, non-clustered but synteny-conserved tRNA gene loci (singlets) were also investigated in this study. Interestingly, ~78% (29/37) of the human-mouse synteny-conserved tRNA gene singlets were also human-opossum synteny-conserved. All these synteny-conserved tRNA gene singlets were high-scoring (tRNAscanSE bit scores > 64).

2.2.2.8. The association between local duplications and unaligned tRNA gene loci in the human-mouse tRNA symbol alignments

Motivated by the finding of intra-cluster duplicated multi-tRNA gene blocks in the human cluster 3.1.42, and its syntenic clusters in the mouse and opossum genomes, I systematically surveyed the association between local duplications and synteny-non-conserved

tRNA gene loci in mammalian genomes.

The starting point of this survey is to find candidate blocks for local multi-loci duplications. Candidate blocks are defined as repeating multi-loci blocks of 2-6 tRNAs in length that are not necessarily tandemly arranged, e.g. if a 2-locus block re-occurs 4 times, the number of loci involved in the putative duplication is 8, and so forth. If a series of tRNA gene loci of the same anticodon type are tandemly arranged, they are also defined as a type of candidate block. When all human tRNA gene clusters were surveyed, ~20% (108/504) of all human tRNA gene loci were labelled candidate blocks. The existence of local duplications is supported by the observation that, among these 108 loci, ~81% (88/108) have their best (sequence identity) match within the putative regions of human-specific duplications. The remaining ~19% have matches that have only one or two more mismatches than their best hits to the regions outside the putative regions of duplications. The evidence, from the conservation of gene order and the good sequence identities between putative duplicated loci, suggests an association between local duplications and the evolution of tRNA gene loci in mammalian genomes.

Further investigation reveals that local duplications may be implicated in the unaligned tRNA gene loci in synteny-conserved tRNA gene clusters. A substantial proportion of the insertions in the human-mouse tRNA symbol alignments can be explained by species-specific local duplications. ~46% (70) of insertions (153, Table 2-12) overlap with putative human-specific candidate blocks involving multi-tRNA-gene loci; ~16% (25/153) of insertions overlap with human-specific tandem duplications of single tRNA gene locus. In addition, duplications may also associate with the species-specific tRNA gene clusters in mammalian genomes. In the synteny-non-conserved human cluster 1.1.10, there is one pair of candidate blocks, which are arranged in an inverted way. The synteny-non-conserved human cluster, 38.X.3, consists of 3 tRNA-Ile gene loci. The synteny-non-conserved human cluster,

15.6.8, is likely to be the result of a segmental duplication of the human cluster 14.6.150. In summary, 63% of the unaligned tRNA gene loci in the human-mouse tRNA symbol alignments can be explained by local duplications (Table 2-18).

conservation type	unaligned loci (insertions)*	unaligned loci that can be explained by local duplications
sub-perfect type two	2	1 (50%)
gapped	110	66 (60%)
complicated	36	29 (81%)
single	5	0 (0%)
synteny-non-conserved	29	19 (66%)
subtotal	182	115 (63%)

Table 2-18. Local-duplication associated insertions in the human-mouse tRNA symbol alignments

*: The definition of insertion is the same as that in Table 2-13.

2.2.3. Discussions

2.2.3.1. Possible evolutionary events involved in the rearrangements of tRNA gene loci in mammalian genomes

Based on the investigation of gene-order conservation, the human-mouse synteny-conservation ratio of tRNA gene loci is estimated to be only ~53% (see subsection 2.2.2.3. and Figure 2-7). This is lower than the UBRHPs-based estimate of ~74% which did not take into account gene-order and indicates the substantial number of gene-loci whose order is not conserved within tRNA clusters.

One evolutionary event implicated by the low synteny-conservation ratio appears to be local duplication. More than half of the changes between the human-mouse syntenic tRNA gene clusters can be explained as the results of local duplications (see subsection 2.2.2.8. and Table 2-18). In addition to species-specific (post primate-rodent split) duplications, there is

evidence for local duplications before the primate-rodent split. For instance, in the human cluster 4.1.36, three duplicated blocks of five-tRNA-gene loci can be found in both the human and mouse syntenic clusters. Local duplication may be a ubiquitous rule for the evolution of tRNA gene loci in mammalian genomes.

In many cases of putative duplications, the candidate blocks, which may consist of multiple tRNA gene loci, are linked in either a direct or an inverted order. Formally, direct local duplications are called tandem duplications. One mechanism which may generate tandem duplications is unequal crossing-over between sister chromosomes during meiosis (for review see Anderson and Roth 1977). On the other hand, when local duplicated blocks are arranged in an inverted order, the duplications are called inverted duplications. There are at least two possible mechanisms which may generate inverted duplications. First, inverted duplication may be the result of post-tandem-duplication chromosomal inversion. Second, a model with double crossing-overs, which is proposed by Passananti *et al*. (Passananti et al. 1987), can also generate inverted duplications. However, from the investigations already made in this chapter, it is impossible to determine by which mechanism each inverted duplication has been generated. Future work could be to look for evidence to support one of the mechanisms. One possible way to resolve this problem might be to look for existence for replication origins, which is a required feature, proposed by Passananti *et al*., in the generation of inverted duplication.

2.2.3.2. The co-amplification model of the formation of gene clusters

The mechanisms that may lead to gene amplifications through tandem duplications and inverted duplications in one of the daughter strands can also cause the de-amplification of gene loci in the other strand. It has therefore been proposed that local duplications in prokaryotic genomes can act as a dynamic and reversible mechanism that can facilitate adaptation to a variety of environmental conditions (for review see Reams and Neidle 2004).

A co-amplification model has been proposed to explain the generation and maintenance of the clustering of related genes in prokaryotes (Reams and Neidle 2004). One main argument is that clustered genes are more likely to be co-amplified and so equally regulated by gene dosage. Besides, if a gene cluster has been evolutionarily selected by the co-amplification model, the order of genes in this cluster does not need to be strictly conserved.

Interestingly, the differences in tRNA gene order observed between the syntenic counterparts in different mammalian genomes suggest that the co-amplification model may have contributed to the formation and evolution of tRNA gene clusters in mammalian genomes. The findings relevant to the co-amplification model include increases of copy number of tRNA genes through mechanisms leading to local duplications, and the partial conservation tRNA gene orders in mammalian genomes.

One question that remains unanswered is about the advantage to survival conferred by the amplification of tRNA gene loci in mammalian genomes. In prokaryotes, over-expression of gene products caused by gene amplification has been suggested to play a critical role in coping with environmental stresses, such as existence of heavy metals, antibiotics, *etc*. (for review see Romero and Palacios 1997). When a particular selection force disappears, the duplicated loci may be de-amplified through the reversible mechanisms of local duplications. Perhaps, the finding of species-specific duplications of tRNA gene loci in the human, mouse, and opossum genomes, respectively, reflect the differential requirements in the evolution of different mammalian species. Due to local duplications, there is significant difference between the numbers, in the respective genomes, of the tRNA gene loci of particular isoacceptor (anticodon) types. For instance, there are 20 tRNA-Cys1 gene loci in the human cluster, 17.7.20, while there are 52 and 43 loci in the syntenic clusters in the mouse and opossum genomes, respectively.

2.2.3.3. Observations that cannot be explained by the co-amplification model

From the observed synteny-conservation pattern of tRNA gene loci in mammalian genomes, several phenomena were found to be incompatible with the co-amplification model.

Firstly, there are synteny-conserved singlet tRNA gene loci in mammalian genomes. For instance, 29 human non-clustered tRNA gene loci were found to be synteny-conserved in the human-mouse-opossum syntenic regions (Figure 2-9). The synteny conservation of these non-clustered tRNA gene loci strongly suggests they should be functional genes. None of these singlet tRNA gene loci are single copies of respective isoacceptor (anticodon) types. There is also no evidence that these singlets are the degraded remnants of tRNA gene clusters. One question is that, if the co-amplification and clustering is so beneficial to the survival of different mammalian species, why these singlet tRNA gene loci should be still conserved after tens of million years of evolution? During the preparation of this manuscript, no obvious advantages/disadvantages can be proposed to explain this observation.

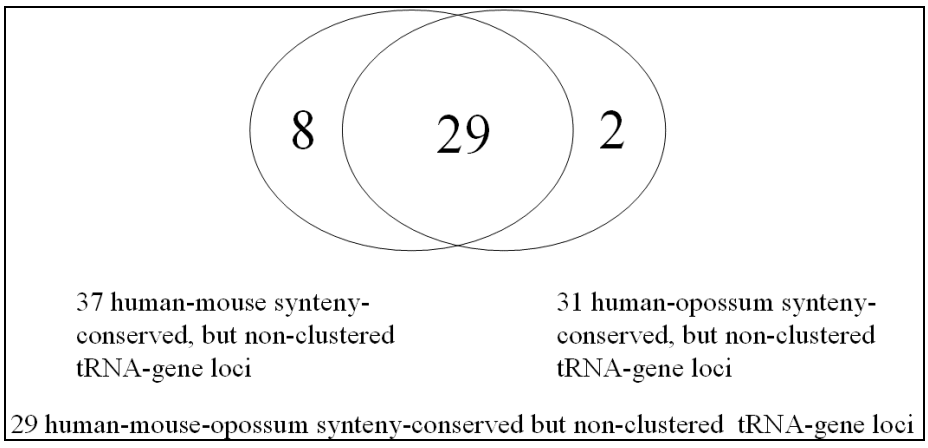

Figure 2-9. the synteny conservation of human non-clustered tRNA gene loci in the syntenic regions of other mammalian genomes

Secondly, there are some synteny-non-conserved human tRNA gene loci, which cannot be explained by local duplication. Possible explanations may include the retro-transpositions,

and the post primate-rodent-split deletions/degradation of tRNA gene loci. These two issues are investigated in the following subsections (2.2.3.4. and 2.2.3.5.).

Finally, recent evidence has implied that the co-amplification model may not be the only plausible mechanism for the clustering of tRNA gene loci in the genomes. In the co-amplification model, clustered genes need not to be co-regulated by a cluster-associated enhancer. However, there is evidence that, under different conditions, the relative expression levels of tRNAs of different isoacceptor types may change (Dittmar et al. 2006). One idea is that the internal promoters may provide a basal-level regulation of tRNA transcription, and the non-promoter regulatory regions may be responsible for controlling the differential expression under different situations. Searching for transcription regulatory elements for clustered tRNA gene loci in mammalian genomes is discussed briefly at the end of chapter 5.

2.2.3.4. Degradation or deletion?

Although the co-amplification model is an appealing hypothesis for interpreting the observed conservation patterns of tRNA gene loci in mammalian genomes, not all unaligned tRNA gene loci can be explained by species-specific local duplications or its reversible process (Table 2-18). In order to find other evolutionary events that may also lead to the unaligned regions in the human-mouse tRNA gene symbol alignments, another possibility, the post primate-rodent-split degradation of the sequences of tRNA gene loci, was therefore explored.

For the non-clustered (singlet) and synteny-non-conserved human tRNA gene loci, the search for the evolutionary remnants in their corresponding syntenic regions in the mouse genome proved to be not very informative. For the 54 synteny-non-conserved singlet tRNA gene loci, only short hits could be found by using WU-BLAST. Most of the e-values are much higher than 0.05, except two cases with borderline significance (0.014 and 0.053). Since the evidence is so weak, it is unclear if there has been pseudogenisation through sequence

degradation of singlet tRNA gene loci in the mouse genome.

Interestingly, for the unaligned tRNA gene loci in the human-mouse syntenic clusters, two putative cases of pseudogenisation through sequence degradations were found. None of the two pseudogenes have previously been annotated by Ensembl (using tRNAscanSE). These cases suggest that sequence degradation is implicated in the evolution of clustered tRNA gene loci in mammalian genomes.

The first case is the degraded remnant in the mouse syntenic region of the Gly1-tRNA gene locus in the human cluster 37.19.2, which is a human-mouse synteny-non-conserved cluster. The e-value of the hit is 2.9e-06 (reported by WU-BLAST). The coordinate of the syntenic tRNA gene locus in the mouse genome is chromosome: NCBIM36: 17: 55852840: 55852911: 1.

```
Human            GCGUUGGUGGUAUAGUGGUuAGCAUAGCUGCCUUCCAAGCAGUUGA
Mouse(degraded)  AUAUUGGUAGAAUAGUGGUuAGgAAAGCUGCCUUCCAAA-AGGUGG
SS_cons          (((((((,,<<<<_______._>>>>,<<<<<________>>>>>,,,

Human            -CCCGGGUUCGAUUCCCGGCCAACGCA
Mouse(degraded)  CCCCGGGUUCUAGUCCCAGAUUGCUUA
SS_cons          ,,<<<<<________>>>>>)))))))):
```

Figure 2-10. The structural alignment of a human tRNA gene locus and its syntenic (but degraded) counterpart in the mouse genome

This previously undiscovered mouse tRNA gene locus does not seem to be a functional one. Firstly, the sequence of the promoter, B box, appears to be degraded. Using eufindtRNA, which is a tRNA-finding algorithm based on the promoter conservation of tRNA genes, this sequence was determined to be a worse promoter than the one in the human orthologous tRNA gene. Secondly, even if this mouse tRNA gene could be transcribed, the secondary structure of the generated tRNAs is likely to be unstable. The putative tRNA product of the degraded gene

locus contains 10 non-Watson-Crick (W-C) and non-GU base pairs in the stem regions (red regions on the mouse strand, Figure 2-10). For comparison, there is only one non-canonical base pair potentially de-stabilizing the secondary structure of the tRNA products transcribed from the orthologous human tRNA gene locus (red regions on the human strand, Figure 2-10).

The second case of pseudogenisation is the degraded locus in the human syntenic region of the Arg4-tRNA gene locus in the mouse cluster 10.3.5, which is the syntenic cluster of the human cluster 18.8.4. The e-value of the hit is 7.8e-09 (reported by WU-BLAST). This previously undiscovered human tRNA gene locus, chromosome: NCBI36: 8: 67187730: 67187802: -1, should be a pseudogene, although the secondary structure of the putative tRNA product have largely been preserved (red regions on the human strand, Figure 2-11). Its promoter, B box, has mutated from GGTTCGACT to GGTCCAGCT (corresponding to the RNA sequences in magenta color on the human and mouse strands, resepctively, Figure 2-11). The degradation of the promoter pattern, which cannot be identified by eufindtRNA, suggests that this degraded tRNA gene locus should be untranscribable. This finding is interesting, because it provides an example of pseudogenisation through promoter-specific degradation. Pseudogenization through promoter-specific degradation is investigated and discussed more generally in chapter 3.

```
Mouse            GGGCCAGUGGCGCAAUGGAuAACGCGUCUGACUACGGAUCAGAAGAUUGU
Human(degraded)  AGGCCAGUGGCGCAAGGGAuAACGUGUCUGACCACGCAUCAGAAGAUUGU
SS_cons          (((((((,,<<<<_______._>>>>,<<<<<________>>>>>,,,,,<<

Mouse            AGGUUCGACUCCUACCUGGCUCG
Human(degraded)  AGGUCCAGCUCCUGCCUGGCUCG
SS_cons          <<<________>>>>>)))))))):
```

Figure 2-11. The structural alignment of a mouse tRNA gene locus and its syntenic (degraded) counterpart in the human genome

One advantage of pseudogenisation through promoter-specific degradation is that it is efficient and safe. If pseudogenisation of a tRNA gene locus proceeded through random mutation, accumulated generation by generation until the functions of the tRNA products were fully abolished, it is possible that some intermediate diseased species of tRNAs would be produced and thus decrease the fitness of the affected organism. By contrast, promoter-specific degradation achieves pseudogenisation by mutating only a few residues in the promoter region of a tRNA gene locus. Although only two cases of promoter-specific degradation were found, it is likely that there are other undiscovered degraded tRNA gene loci. Searching for evidence of old pseudogenes can be very difficult, because without functional constraints, pseudogenes may, after millions of years of evolution, have accumulated so many random mutations that sequence similarity search algorithms cannot find the significant remnants. Consequently, determination of the differential contributions made by sequence degradation and deletions, respectively, to the evolution of tRNA gene loci in mammalian genomes is difficult.

2.2.3.5. Finding pseudogenes through the human-mouse tRNA gene symbol alignments

One purpose of investigating the tRNA gene-order conservation is to search for the evidence which can help us to differentiate functional tRNA gene loci from pseudogenes, a topic more broadly discussed in chapter 3. An appealing argument is that synteny-non-conserved tRNA gene loci will tend to be pseudogenes. In addition to this, the human-mouse tRNA gene symbol alignments of synteny conserved tRNAs provide some other insights relevant to the determination of tRNA pseudogenes.

Firstly, several cases of anticodon transitions were found (Table 2-11) and anticodon transitions may potentially be an indicator of tRNA pseudogenes. In order to realize this argument, a brief introduction to tRNA *identity* is necessary. The term, tRNA identity, refers to the amino acid charging specificity of each tRNA molecule by aminoacyl-tRNA synthetases.

For most tRNAs, the determinants of tRNA identity include the anticodon loop as well as the amino acid accepting stem (for review see Giege et al. 1998). It is unknown if these anticodon transitions would change the tRNA identity of the tRNAs produced from the gene loci in Table 2-11. If the tRNA identities of tRNAs with anticodon transitions remained unchanged, there could be incorrect incorporation of amino acids in protein synthesis. Under the consideration related to tRNA identity, the tRNA gene loci with anticodon transitions should be regarded as potential pseudogenes. An alternative possibility may be errors in the human genome sequence. The significance of these tRNA gene loci with anticodon transitions needs further investigation.

Secondly, the human-mouse tRNA gene symbol alignment also reveals at least one synteny-conserved but low-bit-score tRNA gene locus. Such a locus may also represent a candidate pseudogene. The example is the human tRNA-Asp1 gene locus, chromosome: NCBI36: 1: 159768539: 159768610: 1, which is a member of the human cluster 4.1.36. Its bit-score (reported by tRNAscanSE) is 34.08, which is much lower than that (72.92) of its syntenic counterpart, chromosome: NCBIM36: 1: 172873704: 172873775: -1, in the mouse genome. A putative tRNA product from this gene locus may have an unstable amino-acid accepting stem. In addition, this locus may be untranscribable, since its internal promoters might have degraded (data not shown). This finding is consistent with the pseudogenisation mechanism, promoter-specific degradation, which has also been suggested by previous findings in this section (see the examples of Figure 2-10 and Figure 2-11).

2.2.3.6. Other evolutionary events that may be implicated in the evolution of tRNA gene loci in mammalian genomes

The involvement of various evolutionary events, such as local duplications, inversions, and gene degradation, in the evolution of tRNA gene loci in mammalian genomes have been demonstrated in this section. A question is that, what is the involvement of other evolutionary

events, such as retrotranspositions, transpositions, segmental duplications, gene deletions, or even gene transfer from other organisms? In the following discussions, I consider these possibilities under the following conditions, including the species-specific tRNA gene clusters, species-specific singlet tRNA gene loci, and the unaligned tRNA gene loci in synteny-conserved clusters.

For species-specific tRNA gene clusters evolved after the primate-rodent split, an important feature is the pattern of gene arrangement which should have been generated by local duplications. An example is the human cluster 1.1.10, which contains a duplicated block of four tRNA gene loci. There can be two alternative hypotheses to the formation of this cluster. Firstly, it is possible that this human-specific tRNA gene cluster formed before the primate-rodent or even placental-marsupial split. Perhaps, through independent events of genome rearrangements in the mouse and opossum genomes, respectively, the syntenic clusters in either genome have been deleted. Secondly, the human-specific clusters could have evolved after the primate-rodent split. Theoretically, the second hypothesis should be more likely, since the probability of independent segmental deletions in respective genomes should be low. Besides, in the human cluster 1.1.10, interspersed between the duplicated blocks are the primate-specific protein-coding genes (*e.g.* ENSG00000179571, *etc.*) (based on the annotation made by Ensembl). A similar finding was also observed in the human cluster 38.X.3, where two tRNA-Ile2 gene loci are located within the intronic regions of a pair of duplicated genes (*e.g.* ENSG00000205663), which are also primate-specific. In fact, no other tRNA-Ile2 gene loci can be found in the mouse and opossum genomes.

With the evidence collected in this subsection, it can be concluded that segmental deletions in other mammalian genomes are less likely the reason which can explain the existence of species-specific tRNA gene clusters. However, it is still unclear by which mechanism, either retrotranspositions, transpositions, or segmental duplications, the

human-specific clusters have been formed in new genomic loci. Similar situations were also encountered in investigating the evolutionary origin of the synteny-non-conserved singlet tRNA gene loci, and of some of the unaligned loci in the synteny-conserved tRNA gene clusters. A preliminary result indicates that most of the synteny-non-conserved tRNA gene loci in the human genome are not associated with simple repetitive elements, which might be the evidence of retrotranspositions.

2.3. Summary

In the first part of this chapter, the conservation patterns of the human ncRNAs retrieved from Rfam were investigated. The findings and conclusions relevant to comparative ncRNA finding ncRNA finding approaches are summarized as follows:

- Few covariations are found in either human-mouse synteny-conserved ncRNAs or in the human-zebrafish orthologous ncRNAs.
- ncRNA finding algorithms perform worse when applied to genome synteny alignments than on the single ncRNA gene test alignments they were evaluated.
- Multi-vertebrate synteny alignments can contain more co-variations but the performance of ncRNA finding algorithms on them is similarly affected by alignment quality and completeness, resulting in both false positive and false negative predictions.
- The synteny-conservation ratios of categories of Rfam ncRNAs in the human and mouse genomes vary from ~1% to ~74%.
- ncRNAs with more copies in mammalian genomes appear to be less synteny-conserved.
- Genome assembly quality and artefacts resulting from genome rearrangements

(Figure 2-1, d), have only a small effect on calculations of synteny-conservation ratio of Rfam ncRNAs

In the second part of this chapter, the gene-order conservation of mammalian tRNA genes (predicted by tRNAscanSE) was investigated. My findings include that:

- When gene order is considered, only ~53% of the human tRNA gene loci are human-mouse synteny-conserved (see subsection 2.2.2.3. and Figure 2-7). Besides, 6% (29/504) of human tRNA gene loci are in human-specific clusters (see Table 2-10).
- The low gene-order conservation ratio is not biased by the quality of the mouse genome assembly used in this study (see subsection 2.2.2.4.).
- Tandem duplications and inverted duplications may be important reasons for the low gene-order conservation ratio of tRNA gene loci in mammalian genomes (see subsection 2.2.2.8.).
- Promoter-specific degradation may be involved in the pseudogenisation of mammalian tRNA genes (see subsection 2.2.3.4.).

There are a number of hypotheses with respect to the discovery of numerous synteny-non-conserved ncRNAs in mammalian genomes. Finally, I summarize the evidence for or against each of them:

1. Hypothesis: low quality genome assemblies lead to synteny-conserved ncRNAs being misclassified as synteny non-conserved.
   - Evidence for this hypothesis:
     - Synteny-non-conserved ncRNAs (comparing the human genome assembly NCBI 35 and the mouse genome assembly NCBIM 33) were significantly enriched in regions consisting of whole genome shotgun sequencing or

unfinished regions of clone-based sequencing in the mouse genome (see subsection 2.1.3.1. ,Table 2-2 and Table 2-3).

- ◆ Conclusion:
  - Low quality genome assemblies do lead to some ncRNAs being misclassified as synteny non-conserved, but does not explain the majority.

2. Hypothesis: genome duplication and rearrangement can generate synteny-non-conserved ncRNAs.
   - ◆ Evidence for this hypothesis:
     - There are duplicated multi-loci blocks in the mammalian tRNA gene clusters (see subsection 2.2.2.6.).
     - There might be one segmental inversion in the human tRNA gene clusters after the primate-rodent split (see subsection 2.2.2.6. and Figure 2-6).
   - ◆ Conclusion:
     - Analysis of tRNA clusters is highly suggestive that genome duplication and rearrangement is a mechanism for the generation of synteny-non-conserved ncRNAs.

3. Hypothesis: deletion through degradation can generate synteny-non-conserved ncRNAs.
   - ◆ Evidence for this hypothesis:
     - Degraded remnants of tRNAs can be found that correspond to synteny-non-conserved ncRNAs (see subsection 2.2.3.4.)
   - ◆ Conclusion:
     - There is evidence that some synteny-non-conserved ncRNAs are generated through pseudogenisation, degradation and deletion of the corresponding ncRNA in the other species.

4. Hypothesis: retrotransposition can generate synteny-non-conserved ncRNAs.
   - ◆ Evidence for this hypothesis:

- The generation of species-specific tRNA gene clusters (see subsection 2.2.3.6.) could be explained by retrotransposition, but also by other mechanisms.

◆ Conclusion:

- There is no convincing evidence for or against the mechanisms of retrotransposition.