博士論文

Discovery of microRNA genes and pseudogenes from genomic sequences

（ゲノム配列からのマイクロ RNA 遺伝子と偽遺伝子の発見）

寺井　悟朗

Abstract

Genomic sequences contain various types of genomic elements such as genes, their regulatory elements, transposable elements, and so on. For better understanding of the genome, it is important to discover these genomic elements as thoroughly as possible. Although biological experiments are reliable to identify these elements, they are too costly and labor intensive to be applied to the entire genome. To narrow down genomic regions to be investigated, a number of computational methods have been developed. Although sometimes not entirely accurate, computational methods have an advantage over biological experiments in that they are generally fast enough to be applied to the whole genome and are therefore indispensable for the investigation of genomic sequences.

In this thesis, we propose two computational methods for predicting specific genomic elements. The first method, miRRim2, predicts microRNA (miRNA) genes, which are important genomic elements belonging to non-protein-coding genes. They are suggested to regulate the expression of several thousand protein-coding genes and have been shown to be involved in several human diseases including cancer. Therefore, the discovery of novel miRNA genes has both biological and clinical importance. As transcribed miRNAs have a unique secondary structure that is in most cases evolutionarily conserved, their accurate detection requires the integration of both structural and evolutionary features. In our method, these features are expressed by multi-dimensional vectors. The known miRNAs, represented by sequences of the multi-dimensional feature vectors, were used to train a probabilistic model. To evaluate the accuracy of miRRim2, we designed a cross-validation test in which the whole genome was evaluated. miRRim2 could detect miRNA hairpins more accurately than the other computational prediction methods used previously on the human genome. Furthermore, miRRim2 can infer the location of a mature miRNA sequence, which is useful for deducing the function of predicted miRNAs. In our cross-validation test, miRRim2 has successfully detected the position of the 5′-end of mature miRNAs with a sensitivity and positive predictive value greater than 0.4.

The second method, TSDscan, predicts pseudogenes, which are non-functional copies of functional genes and are one of the prominent genomic elements of mammalian genomes. The detection of pseudogenes not only contributes to the correct annotation of genomic sequences but also provides valuable insights into the expression patterns of the parent genes. For example, genes that have their pseudogenes should be expressed in germ line cells (or cells that develop into germ line cells), because only the pseudogenes occurring in these cells are heritable to the descendant genomes. TSDscan detects pseudogenes based on the surrounding sequence signature. TSDscan clearly differs from existing methods in that it does not rely on commonly used features such as the number of exons and the presence of in-frame stop codons, and hence has the potential to detect novel pseudogenes. Indeed, by applying TSDscan to the human genome, we found 654 short (≤300 bp), previously unknown pseudogenes derived from protein-coding genes. In addition, a comprehensive analysis of the identified pseudogenes and their parent genes revealed an interesting tendency, that short pseudogenes are generated more frequently from long parent genes. On the basis of this observation, we propose a new hypothesis for the generation of pseudogenes.

In summary, we developed two computational methods for predicting specific types of genomic elements. The first method, miRRim2, not only predicts miRNA genes accurately but also infers the location of their mature forms. The second method, TSDscan, predicts pseudogenes based on their surrounding sequence signature. By using TSDscan, we identified a novel type of pseudogenes in the human genome. These prediction results provide a more accurate and comprehensive view of these two genomic elements and contribute to a better understanding of the genome.

Acknowledgement

I am greatly thankful to Professor Kiyoshi Asai and Toshihisa Takagi for their direction. This thesis could never be written without their help and support. I also thank the member of the bioinformatics group of the Functional RNA Project for fruitful discussions. Finally, I am indebted to INTEC Inc. for giving me an opportunity to study at top level research institutes.

Contents

Chapter 1

Introduction

The genome of an organism contains essential information for the development and maintenance of cells, tissues, and organs. It contains various types of genomic elements, such as genes, promoters, enhancer elements, pseudogenes, repetitive elements, and transposable elements. Among them, genes are especially important genomic elements because they encode the information needed to produce proteins and functional RNA molecules, which play a central role in various biological processes, such as the cell cycle, metabolism, and immune response. Promoters and enhancer elements control where and when genes are activated; therefore, they can be considered as a switch that turns genes on or off. Transposable elements, such as long interspersed nuclear elements (LINEs) and *Alu* elements, occupy a significant portion of the genomes of higher eukaryotes. They are considered to have important evolutionary roles by giving diversity to the genome (Deininger et al. 2003; Batzer and Deininger 2002). Pseudogenes are one of the prominent genomic elements of mammalian genomes. Although they were considered to be the non-functional copies of genes, several studies suggested that they are involved in the birth of novel functional genes (see below). To understand the genome, it is important to identify these genomic elements as thoroughly as possible and elucidate their biological roles.

Protein coding and non-protein-coding genes

Genes are transcribed into RNA molecules, which are in many cases translated into proteins. Proteins consist of chains of any of the 20 amino acids coded for by the genome and have a great variety of biochemical activities, such as enzymatic reactions, nucleotide polymerization, and chemical modifications. At present, approximately 30,000 protein-coding genes have been found in the human genome (Claverie 2001; Lander et al. 2001; Venter et al. 2001). The RNA molecules that are translated into proteins are called messenger RNAs (mRNAs).

The other transcribed RNA molecules exert their function without undergoing translation. Such RNA molecules are commonly called non-coding RNAs (ncRNAs). Transfer RNAs and ribosomal RNAs are classic and well-known examples of ncRNAs. At present, various types of ncRNAs have been found in the human genome, some of which play essential biological roles. For example, the Xist RNA coats one of the two X chromosomes in female mammalian cells and causes transcriptional inactivation, resulting in dosage compensation of genes on the X chromosome between the sexes (Lyon 1961). Another important example is microRNAs (miRNAs). They are suggested to regulate more than 50% of the protein-coding genes in humans (Lewis et al. 2005; Friedman et al. 2009) and are expected to be involved in virtually all biological processes (Ambros 2001). It is now increasingly recognized that ncRNAs are key players in a broad range of biological processes (Mercer et al. 2009; Ponting et al. 2009; Prasanth and Spector 2007).

Reverse transcription of gene transcripts and its potential roles

In some cases, transcribed RNA molecules are reverse-transcribed and integrated into genomic DNA, giving rise to copies of functional genes. The copies of genes thus generated are commonly called pseudogenes, which were named after the common belief that they are non-functional genomic elements. However, several studies have suggested that the generation of pseudogenes has resulted in the birth of novel functional genes. For example, the Xist RNA described above appears to have originated from *Alu* elements and a pseudogene of the Lnx3 gene (Elisaphenko et al. 2008). The TRIMCyp gene in the owl monkey, which is responsible for HIV resistance, was generated by a fusion of the TRIM5 gene and the CypA pseudogene (Sayah et al. 2004). Some endogenous small interfering RNAs (siRNAs), which negatively regulate functional genes, are expressed from pseudogenes (Watanabe et al. 2008). Therefore, the generation of pseudogenes may be important for the evolution of functional genes.

1.1 Problem statement and our contribution

Among the various types of ncRNAs, miRNAs are especially important molecules whose discovery can have a significant biological and clinical impact. Therefore, much effort has been devoted to their detection by both computational methods (e.g., Berezikov et al. 2005; Hertel et al. 2006; Lim et al. 2003a) and biological experiments (e.g., Lagos-Quintana et al. 2002; Landgraf et al. 2007; Cummins et al. 2006). In this thesis, we present:

- *A new method for predicting miRNA genes and their mature forms*

Pseudogenes are one of the prominent elements of the human genome. Their detection contributes to the correct annotation of the genome. Many studies have been conducted for the detection of pseudogenes (e.g., Torrents et al. 2003; Zhang et al. 2003; Ohshima et al. 2003). In many cases, pseudogenes lack introns because reverse transcription of transcribed RNA molecules occurs after their introns are spliced out. Existing methods for detecting pseudogenes rely heavily on this lack of introns; however, this strategy is not applicable to pseudogenes derived from genes without introns. In this thesis we present:

- *A method for predicting pseudogenes that does not rely on the lack of introns*

We also performed comprehensive analysis between the pseudogenes we identified and their parent genes. From this analysis, we found a tendency that short pseudogenes are generated more frequently by long genes. On the basis of this observation we present:

- *A hypothetical model for the generation of pseudogenes*

Overview of this thesis

In this introductory chapter, we start with an overview of the biogenesis of miRNAs and the molecular mechanism for generating pseudogenes. In the second chapter, we present miRRim2, a method for predicting miRNA genes and their mature forms. The third chapter presents TSDscan, a method for predicting pseudogenes. A hypothetical model for the generation of pseudogenes, which was deduced from our comprehensive analysis of pseudogenes and their parent genes, is also presented in this chapter. The final chapter summarizes this thesis and gives an outlook for future work.

1.2 Biogenesis of miRNAs and generation mechanism of pseudogenes

Biogenesis of miRNAs

A miRNA gene is transcribed initially as a long RNA molecule, called pri-miRNA, in the nucleus (Fig. 1.1a). It contains one or more hairpin structures that are processed by the enzyme Drosha (Lee et al. 2003). In this study, we refer to the hairpin structure as an “miRNA hairpin.” A miRNA hairpin is processed into a shorter (c.a. 60-bp) hairpin, called pre-miRNA, by Drosha (Fig. 1.1b). Then, it is exported to the cytoplasm by an enzyme called Exportin-5 (Fig. 1.1c; Lund et al. 2004). It is further processed into a double-stranded RNA molecule of approximately 22-nucleotides (nt), which is called a miRNA duplex by the enzyme Dicer (Fig 1.1d; Hutvagner et al. 2001). In general, either strand of the miRNA duplex is loaded into the RISC protein complex and functions as a mature miRNA (Fig. 1.1e; Gregory et al. 2005). The other strand of the miRNA duplex, which we refer to as the “passenger strand,” is degraded rapidly (Fig. 1.1f; Gregory et al.

2005). Although a novel type of miRNA gene that bypasses Drosha processing has been reported (Berezikov et al. 2007), to date, most of the identified miRNAs are subject to Drosha processing.

In various eukaryotes, miRNA hairpins can be located in the introns of protein-coding genes. In human, approximately 40% of miRNA hairpins are found in introns. It is not clear whether such intronic miRNAs are processed after or before introns are spliced out. In either way, miRNA hairpins in intronic miRNAs are processed into pre-miRNA and enter into the maturation pathway illustrated in Figure 1.1.

Generation mechanism of pseudogenes

Most pseudogenes are generated by a process called retrotransposition, which is a series of *in vivo* processes involving the reverse transcription of RNA molecules and the integration of the transcripts into the genome. Retrotransposition in eukaryotes can be divided into two types: the long terminal repeat (LTR) type and the non-LTR type. The latter accounts for the majority of retrotransposition events in human (Ostertag and Kazazian 2001). Various types of RNA molecules, including *Alu* RNAs, LINE RNAs, mRNAs, and small ncRNAs, are copied via non-LTR retrotransposition (Dewannieux et al. 2003; Moran et al. 1996; Esnaul et al. 2000; Buzdin et al. 2003; Perreault et al. 2005).

Non-LTR retrotransposition is mediated by a protein encoded by the second open reading frame of LINE-1 (hereafter L1-ORF2p), which has both reverse transcriptase and endonuclease activity (Feng et al. 1996). L1-ORF2p can bind to the 3′-end of RNA molecules and promotes their retrotransposition (Fig. 1.2a). The endonuclease activity of L1-ORF2p creates a cleavage site in genomic DNA (Fig. 1.2 b); this cleavage site is used as a primer, and reverse transcription of the template RNA begins (Fig. 1.2c). The resultant cDNAs are integrated into genomic DNA (Fig. 1.2d). This integration process is called target-site-primed reverse transcription (Luan et al. 1993; Cost et al. 2002).

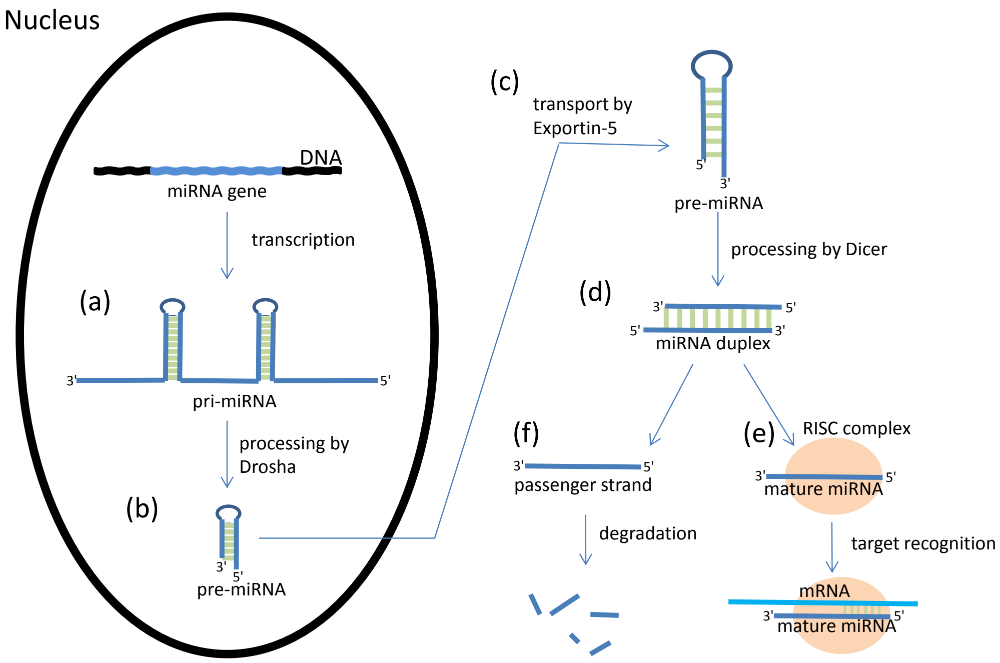


Figure 1.1. Processing of miRNA genes. (a) A miRNA gene is transcribed as a long RNA molecule, called pri-miRNA. (b) Some of the hairpin structures in pri-miRNA are processed into pre-miRNA by Drosha. (c) Pre-miRNAs are exported to the cytoplasm by Exportin-5. (d) Pre-miRNAs are processed into miRNA duplexes by Dicer. (e) One of the two strands is loaded into the RISC complex and functions as a mature miRNA. (f) The other strand, which we refer to as the passenger strand, is degraded rapidly.

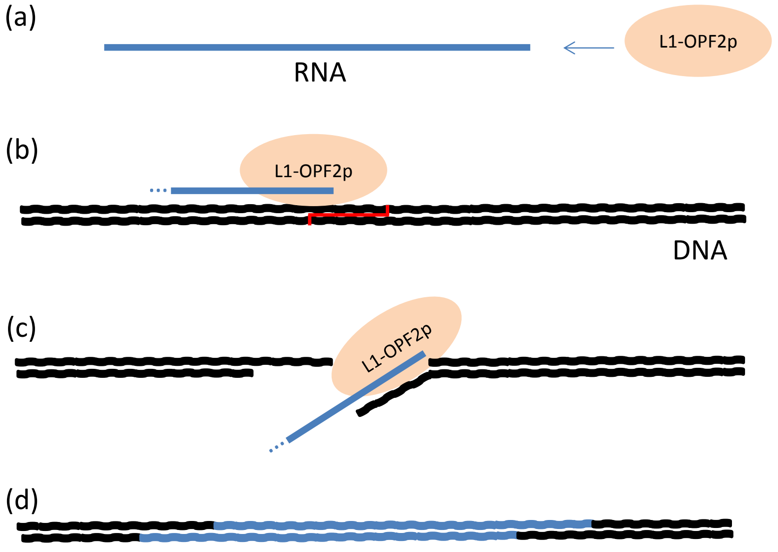


Figure 1.2. Schematic view of retrotransposition. (a) RNA molecules are recognized by L1-ORF2p. (b) The endonuclease activity of L1-ORF2p creates a cleavage site in genomic DNA. (c) Reverse transcription of the template RNA begins using the cleavage site as a primer. (d) The resultant cDNAs are integrated into genomic DNA.

1.3 List of publications

The contents of the Chapter 2 and 3 are the modified version of the article originally published in [1] and [2], respectively.

[1] Terai G, Okida H, Asai K, Mituyama T. (2012) Prediction of conserved precursors of miRNAs and their mature forms by integrating position-specific structural features. *PLoS One* **7**: e44314.

[2] Terai G, Yoshizawa A, Okida H, Asai K, Mituyama T. (2010) Discovery of short pseudogenes derived from messenger RNAs. *Nucleic Acids Res.* **38**:1163-1171.

Chapter 2

Prediction of miRNA based on position-specific structural features

2.1 Introduction

MicroRNA (miRNA) is a well characterized non-coding RNA family that has important roles in various biological processes such as development (Wienholds et al. 2005), cancer (Esquela-Kerscher et al. 2006), and immune response (Lindsay 2008). Therefore, miRNA identification and functional analysis are necessary for the understanding of many biological phenomena. A miRNA is initially transcribed as a long RNA molecule called pri-miRNA which contains one or more hairpin structures that are processed by the enzyme Drosha (Fig. 1.1; Lee et al. 2003). In this study, we refer to the hairpin structure as a "miRNA hairpin". After a miRNA hairpin is processed into a shorter hairpin, called pre-miRNA, by Drosha, it is further processed into a ~22-nucleotide (nt) double-stranded RNA molecule called a miRNA duplex by the enzyme Dicer (Hutvagner et al. 2001). Figure 2.1 illustrates the location of a miRNA duplex as well as the Drosha and Dicer cleavage sites in a putative miRNA hairpin. In general, either strand of the miRNA duplex is loaded into the RISC protein complex and functions as a mature miRNA (Gregory et al. 2005). Another strand of the miRNA duplex, which we refer to as "passenger strand", is rapidly degraded (Gregory et al. 2005).

Previous biochemical and computational studies have revealed several important features that are specific or necessary for miRNA hairpins. For example, miRNA duplex regions within miRNA hairpins generally form stable base pairs (Ambros et al. 2003) and often have an internal small bulge in the middle (Krol et al. 2004; Han et al. 2006). The 5′-end position of mature miRNA is predominantly composed of uracil and tends to be energetically unstable (Krol et al. 2004; Khvorova et al. 2003). Drosha was shown to recognize the outermost base pair in miRNA hairpins (Han et al. 2006) and cleave the molecule at ~13 nt and ~11 nt from the Drosha recognition base pair (DRB; Fig. 2.1). Therefore, the positions around the DRB have unique secondary structural (Han et al. 2006; Saetrom et al. 2006) and evolutionary features, as shown in Results and Discussion. The length between the Drosha cleavage sites is ~60 nt with a small variation (SD = 4.9) for human miRNA hairpins (Saetrom et al. 2006), although it can be longer (~80 nt) in *Drosophila* (Ruby et al. 2007).

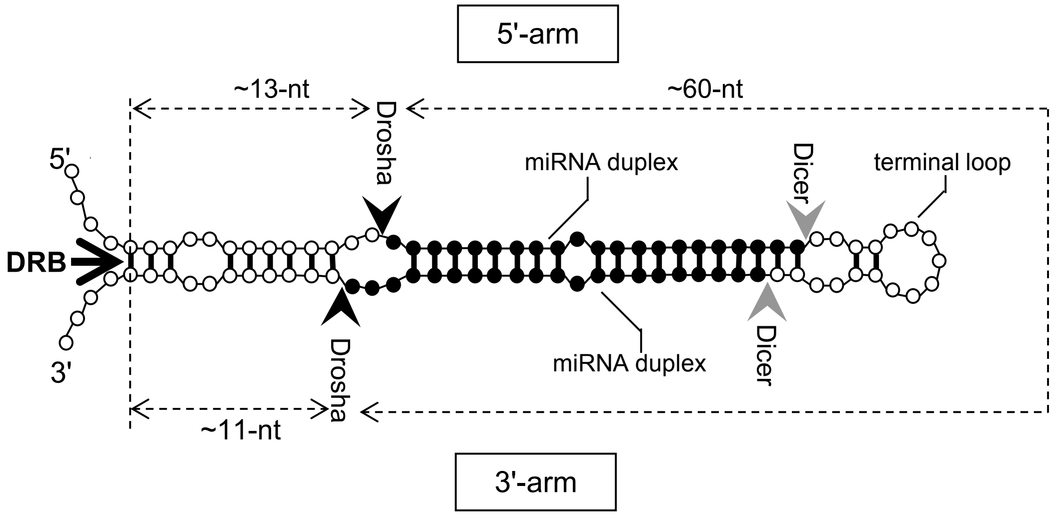


Figure 2.1 Schematic view of a putative miRNA hairpin. Cleavage sites by Drosha and Dicer are indicated by black and gray arrowheads, respectively. A miRNA duplex is represented by black circles. An arrow at the left side indicates the Drosha recognition base pair (DRB) (see text). The 5'-arm and 3'-arm indicate the 5'- and 3'-sides of a stem region in a miRNA hairpin, respectively.

The features described above can be useful for identifying conserved miRNA hairpins, and several methods have been proposed that take these features into account. In miRScan (Lim et al. 2003a) and Berezikov et al. (2005), evolutionary and structural features in each part of miRNA hairpins were used for detecting miRNA hairpins. RNAmicro (Hertel et al. 2006) used the support vector machine (SVM) for discriminating miRNA hairpins and other ncRNAs. SSCprofiler (Oulas et al. 2009) and miRRim (Terai et al. 2007) used a hidden Markov model (HMM) to model evolutionary and secondary structural features of miRNA hairpins. The above methods focus on the detection of miRNA hairpins, and cannot infer the location of mature miRNAs.

More recently, several groups report new methods for detecting miRNA hairpins based on detailed structural and nucleotide features. MiRPred (Brameier and Wiuf 2007) detects miRNA hairpins based on ensemble of secondary structural motifs. Agrawal et al. (Agarwal et al. 2010) used context-sensitive HMM to model sequence and structure of miRNA hairpins. MiPred (Jiang et al. 2007) used the Random Forest algorithm to integrate local structural characteristics and global structural stability of miRNA hairpins. Liu et al. (2012) extracted sequence-structure motifs from miRNA hairpins and used them to distinguish true and non-miRNA hairpins. These methods, however, do not take evolutionary features into account. To use these methods to detect conserved miRNA hairpins, an additional screening procedure based on the evolutionary features has to be developed, which is not a simple task because, as shown below, miRNA hairpins have a unique and complex pattern of evolutionary conservation.

In this study, we developed a new method, miRRim2, which can not only detect conserved miRNA hairpins, but also infer their mature forms. In miRRim2, each position of a miRNA hairpin is expressed as a multidimensional feature vector to detect position-specific features; therefore, a miRNA hairpin is expressed as a sequence of the feature vectors. miRNA hairpins, expressed by sequences of feature vectors, are modeled using conditional random fields (CRFs) (Lafferty et al. 2001), which optimize feature weights so that a trained model can most probably discriminate between miRNA hairpins and background data. The probabilistic model used in miRRim2 has several sub-components, each of which corresponds to a specific component of miRNA hairpins, such as mature miRNA, passenger strand, and terminal loop regions; therefore, the position-specific features of each component are appropriately modeled.

Recently, many miRNA hairpins have been identified that are not evolutionarily conserved. A recent study shows that the expression level of these non-conserved miRNA hairpins are very low, and that they are almost free of selective pressure (Liang et al. 2009). Another recent study has suggested that non-conserved miRNA hairpins may disappear quickly during the course of evolution (Lu et al. 2008). Because the biological relevance of non-conserved miRNA hairpins remains elusive, we focus on the detection of conserved miRNA hairpins, of which the biological importance is evolutionarily supported.

2.2 Methods

2.2.1 Evolutionary and secondary structural features of miRNA hairpins

Before entering technical details, we show the position-specific features of miRNA hairpins which we would like to model with probabilistic framework. Figure 2.2 shows the evolutionary and structural features of 306 conserved miRNA hairpins in human, which we refer to “core miRNA hairpins” (see Materials and Details). In this figure, both of the PhastCons and PhyloP score represent the degree of evolutionary conservation in each position, which are calculated based on multiple alignment between species (Siepel 2005; Siepel 2006). The base pair potential represents the likelihood of forming a base pair in each position, which is calculated from the predicted secondary structure (see below). The position 0 in the x-axis indicates to 5'-ends of miRNA duplexes in the 5'-arm.

Overall, the PhastCons, PhyloP, and base pair potential are highly correlated, indicating that highly conserved regions tend to form base pairs in miRNA hairpins.

Especially, the miRNA duplex regions are more strongly conserved and form more stable base pairs than their surrounding regions. The outside regions of the miRNA hairpin (position -20 or less, and 80 or more) are generally less conserved than the internal regions, as has been already reported (Berezikov et al. 2005; Terai et al. 2007).

The Drosha recognition base pair (DRB; Fig. 2.1) in the 5'-arm of the miRNA hairpin is located around position -13, where the base pair potential drops sharply, as previously reported (Han et al. 2006; Saetrom et al. 2006). Interestingly, PhastCons and the PhyloP scores also drop at the same position. The same propensity was observed around the DRB in the 3'-arm (position +11), when we adjusted position 0 to the 3′-ends of the miRNA duplex regions in the 3'-arm (Fig. S2.1).

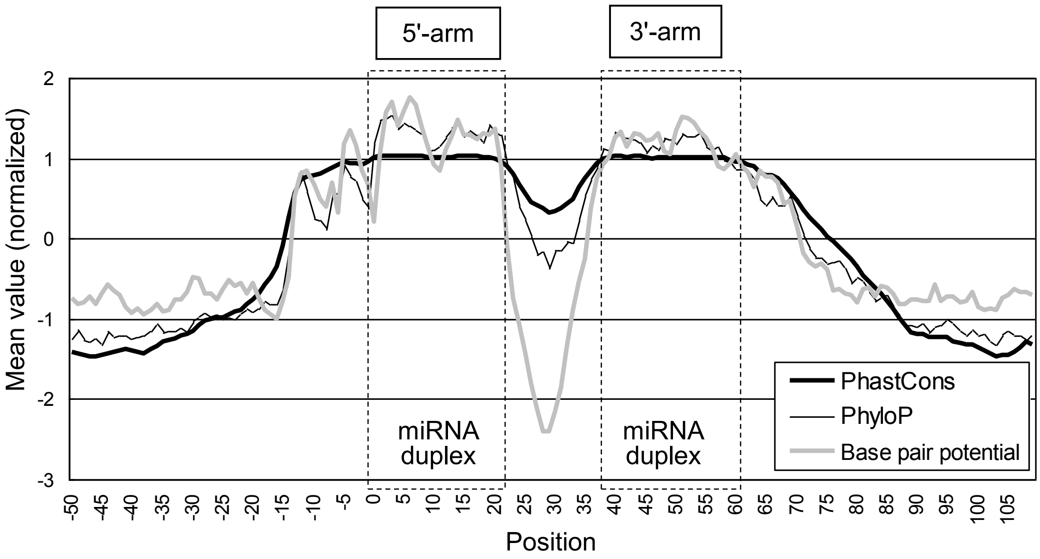


Figure 2.2. PhastCons scores, PhyloP scores, and base-pair potential averaged in each position. Position 0 indicates to 5' ends of miRNA duplexes in the upper strand of miRNA hairpins. Dotted rectangles indicate the approximate location of the miRNA duplex.

Next we focused on the difference between the mature miRNA and passenger strand. Mature miRNAs are more strongly conserved than passenger strands (Fig. 2.3a), and the 5′-ends of mature miRNAs are less likely to form a base pair than passenger strands (Fig. 2.3b). These differences were only found for mature miRNAs in the 5'-arm. For mature miRNAs in the 3'-arm, these propensities were very weak (Fig. S2.2).

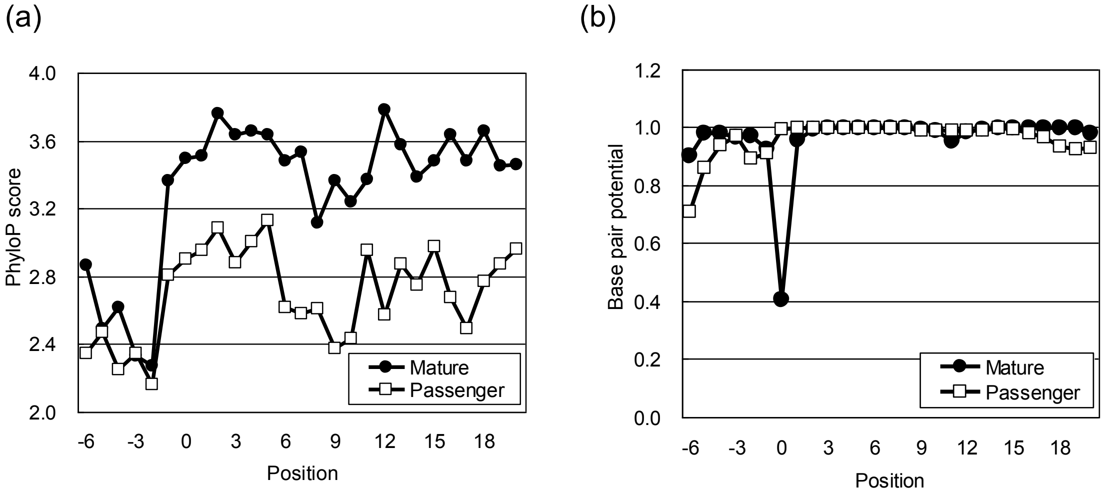


Figure 2.3. Difference between mature miRNA and passenger strand. Median values of (a) PhyloP scores and (b) base pair potentials are plotted in each position. Position 0 indicates the 5'-ends of mature miRNA or passenger strand.

2.2.2 Features used in our method

To utilize the evolutionary and structural features described above, we expressed each genomic position *i* as a 7-dimensional vector $\mathbf{o}^{(i)}$. Dimensions 1–4 represent evolutionary conservation, which are calculated from multiple alignments between species. Dimensions 5 and 6 represent secondary structural features, which are calculated from the predicted secondary structure. Dimension 7 represents a nucleotide in each position. Below, we summarize the content of a feature vector $\mathbf{o}^{(i)}$. Details of the calculation of $\mathbf{o}^{(i)}$ are described in Materials and Details.

Dimension 1 of $\mathbf{o}^{(i)}$ is the PhastCons score (Siepel et al. 2005) of position *i*, which is calculated from multiple alignment between species. Dimension 2 and 3 is the PhastCons score in position i-20 and i+20, respectively. Dimension 4 is the PhyloP score (Siepel et al. 2006) which is another measure of evolutionary conservation. The important difference between the PhyloP and PhastCons scores is that the PhyloP score is calculated independent of neighboring positions; therefore, the PhyloP score is more appropriate for evaluating the degree of evolutionary conservation at each position. In contrast, the PhastCons score is more sensitive for detecting continuous conserved regions (Fujita et al. 2011).

Dimension 5 is the base pair potential which represents the likelihood of forming a base pair in each position. The base pair potential is calculated as the maximum value of base pair probabilities assigned to each position. The base pair probabilities can be calculated by McCaskill's algorithm (McCaskill 1990). Dimension 6 is the base pair distance which represents the distance between a predicted base pair. For example, if position *i* is predicted to form a base pair with position *j*, the base pair distance of position *i* is $j - i$.

Dimension 7 simply represents the nucleotide (A, U, G, or C) in each position.

2.2.3 The probabilistic model used in this study

Because each genomic position is expressed by a 7-dimensional vector, a long genomic region is represented by a sequence of 7-dimensional vectors, and each miRNA hairpin is a sequence segment hidden in it. To detect miRNA hairpins from a long genomic region, we used a probabilistic model called a conditional random field (CRF) (Lafferty et al. 2001), which is recently beginning to be used in biological sequence analyses and achieves better performance than existing methods (Do et al. 2006a; Do et al. 2006b; Sato et al. 2005; DeCaprio et al. 2007).

The probabilistic model employed here consists of 12 sub-models (Fig. 2.4). The left and right sides of the Flanking sub-model represent the upstream and downstream regions of the miRNA duplex, respectively. The Loop sub-models represent the regions between miRNA duplexes. The Mature and Passenger sub-models represent the mature miRNA and passenger strand, respectively. The Non-miRNA sub-model represents regions that are not miRNA hairpins. As described above, either strand of the miRNA duplex can become mature miRNA; however, in some cases, both strands become mature miRNA. Therefore, there are 3 types of mature miRNA location in miRNA hairpins. Our architecture has 3 paths, each of which corresponds to one of the 3 types of mature miRNA location. A given sequence segment is considered a miRNA hairpin if it is derived from 1 of these 3 paths with a high probability. Similarly, a sequence segment that is expected to be derived from the Mature sub-model is considered a mature miRNA sequence.

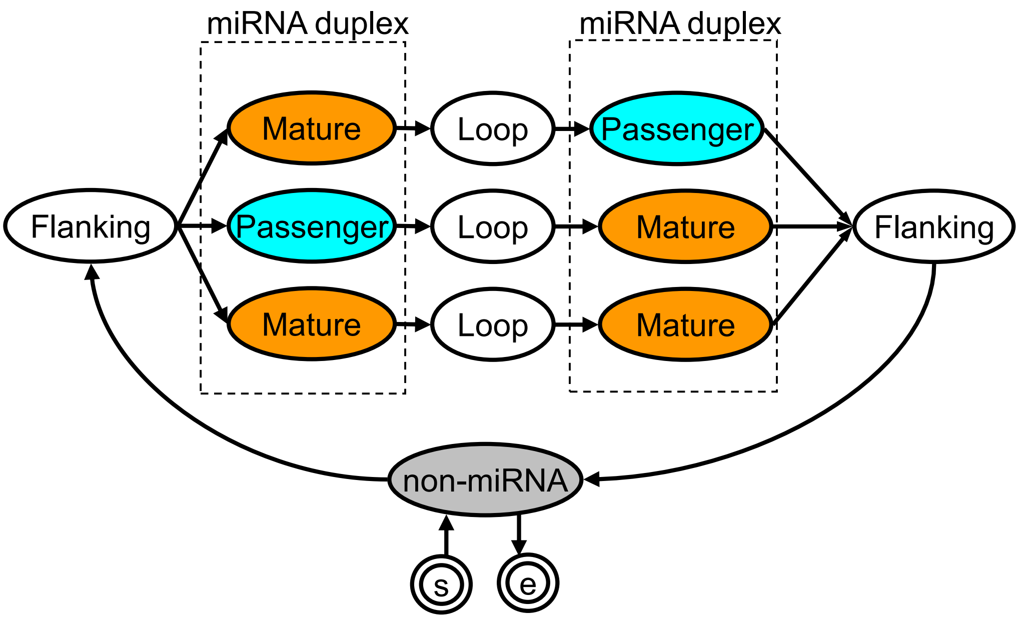


Figure 2.4. Architecture of the Model. Each sub-model is represented by an oval. The circled "s" and "e" represent a start and end state, respectively. Dotted rectangles indicate sub-models corresponding to a miRNA duplex.

Scores for detecting miRNA hairpins and mature miRNAs

Using the above CRF model, we calculated the probability that each genomic position *i* is a miRNA hairpin, which we denoted as $\mathrm{P}^{\mathrm{mi}}{}_{i}$, using the Forward-Backward algorithm (see Materials and Details). We considered a continuous sequence segment of 80 base pairs (bp) or more with a $\mathrm{P}^{\mathrm{mi}}{}_{i} > T$ as a predicted miRNA hairpin, where T is a probabilistic threshold from 0 to 1.

We also calculated the probability of position *i* being the 5′-end position of a mature miRNA region, which we denoted as $\mathrm{P}^{\mathrm{5end}}{}_{i}$. The position with $\mathrm{P}^{\mathrm{5end}}{}_{i} > T$ is considered to be the 5′-end of a mature miRNA.

2.3 Results and discussion

2.3.1 Accuracy for predicting miRNA hairpins

To evaluate the accuracy of miRRim2, we designed a genome-wide cross-validation, in which the whole human genome was used for training and test data. Briefly, we selected a particular human chromosome and scanned it using miRRim2 that was trained using the core miRNA hairpins on the remaining chromosomes. To mimic a realistic situation, miRNA hairpins were excluded from training data if they were homologous to miRNA hairpins on the chromosome to be scanned. This procedure was repeated for all chromosomes. So the whole human genome was used for evaluation. Further details were described in Materials and Details.

The accuracy of miRRim2 is shown in Figure 2.5a. When the number of predicted miRNA hairpins was 216, miRRim2 could detect 180 core miRNA hairpins, indicating that miRRim2 was highly accurate at this threshold. For comparison, we obtained four publically available prediction results, and evaluated them using the same core miRNA hairpins (Figure 2.5a). miRRim2 could detect more core miRNA hairpins when the number of predicted miRNA hairpins was adjusted to be the same. The genomic coordinates of miRNA hairpins predicted by berezikov (Berezikov et al. 2005), miRscan (Lim et al. 2003b), and miRRim (Terai et al. 2007) were obtained from supplemental data of these articles. Predicted miRNA hairpins of RNAmicro were obtained from the "Predicted miRNA track" of our fRNA database (Mituyama et al. 2009).

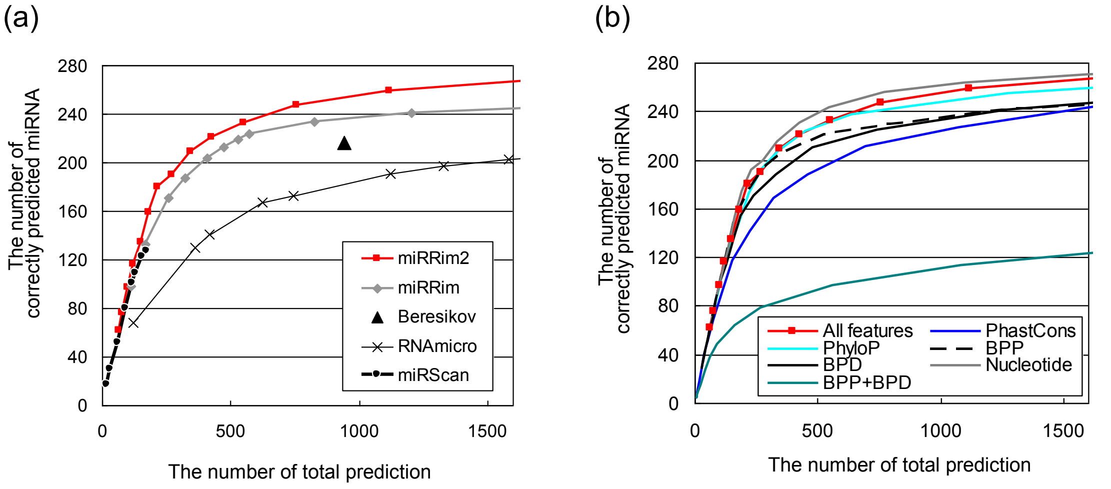


Figure 2.5. Accuracy for detecting the core miRNA hairpins. (a) The accuracy of miRRim2 together with four previously performed computational predictions is shown. (b) The change of the accuracy when one type of features is excluded. BPP: base-pair potential, BPD: base-pair distance.

In order to evaluate the contribution of each type of feature to the prediction accuracy, we excluded 1 or more dimension(s) from the feature vector $\mathbf{o}^{(i)}$ and investigated changes of the prediction accuracy. The result is shown in Figure 2.5b. The exclusion of PhastCons scores (dimensions 1–3) caused significant reduction of the prediction accuracy. The PhyloP score (dimension 4), on the other hand, had only a small effect on the prediction accuracy, indicating that only the PhastCons scores are almost sufficient for capturing the conservation pattern of miRNA hairpins.

Two types of secondary structural features (base pair potential and distance: dimensions 5 and 6) individually contribute to the prediction accuracy, although these features were dependent on each other. When the two types of secondary structural features (dimensions 5 and 6) were simultaneously excluded, the prediction accuracy was greatly reduced, indicating that not only conservation but also secondary structural features were discriminative. The nucleotide (dimension 7) had a slightly bad effect on the prediction accuracy.

2.3.2 Accuracy for predicting mature miRNAs

Figure 2.6a shows the accuracy for detecting the 5′-end of a mature miRNA based on the cross-validation described above. Inferring the 5′-end of a mature miRNA is important because the first 8 bp from the 5′-end is so called "seed region" and plays a pivotal role in the recognition of target genes.

The prediction accuracy was measured by sensitivity and positive predictive value (PPV), which were defined as:

$$\mathrm{Sensitivity} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$$

$$\mathrm{PPV} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$$

where TP, FP, and FN are the number of true positives, false positives, and false negatives, respectively. In this evaluation, the 5′-ends of true mature miRNAs within the core miRNA hairpins were defined as positive sites, and the other positions within the core miRNA hairpins were defined as negative sites. If the predicted 5′-ends were (or were not) positive sites, they were considered as true (or false) positives. Positive sites that were not detected were considered as false negatives.

Our method achieved sensitivity and PPV slightly above 0.4, which is better than our null model. In the null model, all the uracils are considered as 5'-end of mature miRNA. Each uracil has a penalty score, which is designed such that uracils in plausible positions have low penalty score (for details, see supplemental Methods S2.1). When we consider the predicted sites that were 1 bp different from positive sites as true positives, sensitivity and PPV increased to about 0.55. Similarly, when we allowed a 2 bp difference, sensitivity and PPV increased to about 0.65.

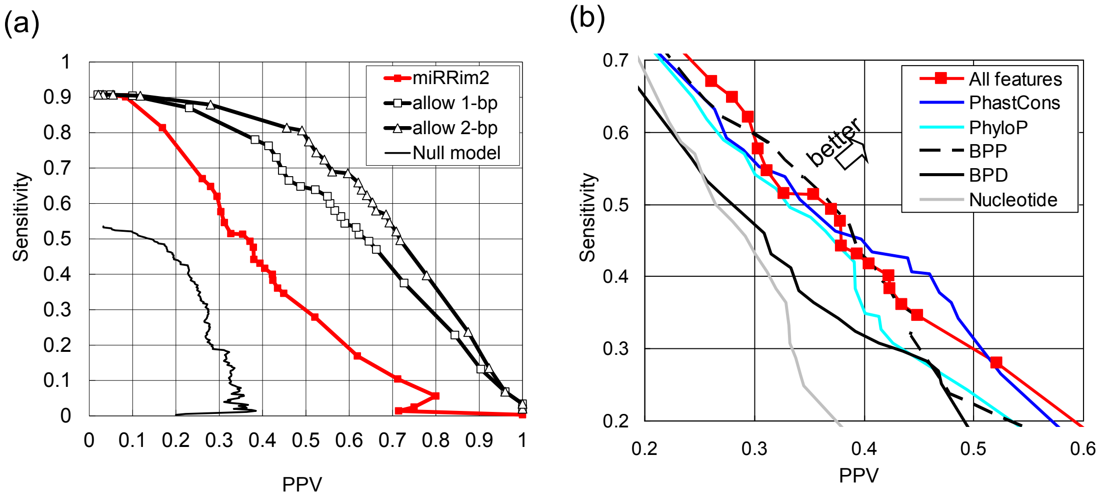


Figure 2.6. Accuracy for detecting the 5'-end of mature RNAs. (a) Sensitivity-PPV plot for mature miRNA prediction. (b) The change of the accuracy when one type of features is excluded. BPP: base-pair potential, BPD: base-pair distance.

Mature miRNAs in the 5'-arm of miRNA hairpins were more accurately predicted than those in the 3'-arm (Fig. S2.3) because the differences between mature miRNA and passenger strand are only found in the 5'-arm (see Fig. 2.3 and Fig. S2.2).

There are several methods that can identify mature miRNAs (Nam et al. 2005; Yousef et al. 2006; Helvik et al. 2007; Sheng et al. 2007; Gkirtzou et al. 2010). Among them, only one tool MatureBayes (Gkirtzou et al. 2010) is specifically designed for predicting 5'-end of mature miRNA. In the other tools, the main purpose is to identify miRNA hairpins (Nam et al. 2005; Yousef et al. 2006; Sheng et al. 2007) or the Drosha cleavage sites (Helvik et al. 2007), not the location of mature miRNAs. Therefore, we compared our results with MatureBayes. The prediction results of MatureBayes were obtained using the web server of MatureBayes. We used the nucleotide sequences of the core miRNA hairpins as input data of MatureBayes. The sensitivity and PPV of MatureBayes was 0.30 and 0.18, respectively. The main reason for the lower accuracy of MatureBayes than that of miRRim2 may be the difference of training data. MatureBayes was trained using both conserved and non-conserved miRNA hairpins. On the other hand, miRRim2 was trained using only conserved miRNA hairpins which were probably more uniform in terms of nucleotide contents and hairpin length than non-conserved ones.

Figure 2.6b shows the feature contribution to the prediction accuracy of mature miRNA. The feature that contributed the most was nucleotide (dimension 7), which may be mainly due to the fact that the 5′-ends of mature miRNAs are predominantly composed of uracil. The second most important feature was the base pair distance (dimension 6). The base pair distance was useful for identifying the approximate positions of the Drosha cleavage sites (Fig. 2.1) because the distance between a pair of Drosha cleavage sites in a miRNA hairpin is ~60 bp on average with a small variation (Helvik et al. 2007). Therefore, it may also help to identify the 5′-end of a mature miRNA. The small contribution of the PhyloP score (dimension 4) may be due to the difference in evolutionary conservation between mature miRNA and passenger strand (Fig. 2.3a).

2.3.3 Accuracy for predicting miRNA hairpins not used in the cross-validation

The latest version of miRBase (v.18) contains 142 conserved miRNA hairpins (mean PhastCons score > 0.5) that are not included in the core miRNA hairpins. About a half of them are not included because they are newly discovered after the release of miRBase version 14.0 from which the core miRNA hairpins were constructed, and another half were excluded because they were supported by only a single literature (see Methods and Details). We investigated whether miRRim2 could detect these miRNA hairpins (Fig 2.7). Several of the 142 miRNA hairpins could be accurately detected. For example, 8 of them were included in the top 11 candidates predicted by miRRim2. Overall, however, miRRim2 as well as the other four methods could not detect many of the 142 miRNA hairpins. We found that the pattern of evolutionary conservation of the 142 hairpins was considerably different from that of the core miRNA hairpins (Fig. S2.4). The latest version of miRBase contains many miRNA hairpins discovered by using the deep-sequencing technology. The innovation of the deep-sequencing technology may have greatly enhanced the discovery of miRNA hairpins. We, however, speculate that the deep-sequencing is too sensitive, so that it can sometimes detect miRNA-like molecules that are accidentally processed by the Drosha and Dicer. At any rate, miRRim2 seems superior to, or at least comparable with, the other computational predictions in the detection of the 142 miRNA hairpins.

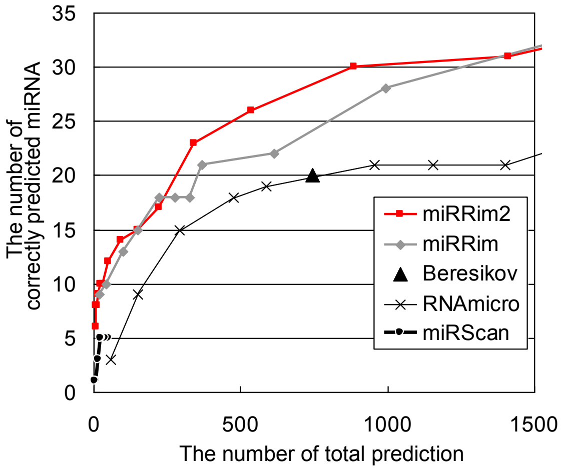


Figure 2.7. Accuracy for detecting the miRNA hairpins other than core miRNA hairpins.

2.3.4 Application to the *Ciona intestines* genome

To assess the effectiveness of miRRim2 on independent data, we applied it to the *Ciona intestinalis* genome, which was recently demonstrated to contain miRNA genes (Norden-Krichmar et al. 2007; Hendrix et al. 2010). For *Ciona intestinalis*, only *Ciona savignyi* is suitable for a comparative study because the other sequenced chordates are evolutionarily too distant (data not shown); therefore, the PhastCons and PhyloP scores were not available. We propose that alternative conservation scores can be calculated from pairwise genomic comparisons. As an alternative to the PhyloP score, we used the following simple score: we assigned a score = 1 to the positions with a matched nucleotide in pairwise genomic alignments, and 0 for the other positions. As an alternative to the PhastCons score, we used the alignment probability, which is a measure of the correctness of a given alignment column. Although the meaning of the alignment probability and PhastCons score is fundamentally different, they are similar in that both of them take higher values in continuous conserved segments, even if nucleotides in a certain alignment column do not coincide. Therefore, we used the alignment probability instead of the PhastCons score. These alternative scores were calculated using the Last program (Kiełbasa et al. 2011), which can not only perform fast genome-wide pairwise alignments, but can also calculate the alignment probability for each alignment column.

We trained miRRim2 using the core miRNA hairpins in human and used it to scan the *Ciona intestinalis* genome. Hendrix et al. (2010) identified 380 miRNA hairpins in *Ciona intestinalis* using deep-sequencing experiments. We compiled miRNA hairpins identified by Hendrix et al. and those included in miRBase v.18, and obtained 419 miRNA hairpins in *Ciona intesinalis*. The detection/prediction performance is shown in supplemental Figure S2.5a. Briefly, miRRim2 detected 47 and 73 miRNA hairpins when the number of predicted miRNA hairpins was 115 and 649, respectively. We found that the low sensitivity (73/419) was derived from the fact that only about 80 of miRNA hairpins in *Ciona intestinalis* were conserved in *Ciona savignyi* (data not shown). Because our method was designed to detect conserved miRNA hairpins, it could not detect non-conserved miRNA hairpins. In the 73 miRNA hairpins detected by miRRim2, it correctly predict the 5'-end of mature miRNA with sensitivity and PPV about 0.4 (Figure S2.5b).

Among the predicted miRNA hairpins, we found 10 candidates in which the locations of predicted mature miRNAs were in good agreement with deep-sequencing results. Table 2.1 shows the genomic coordinates of the 10 candidates. *Cand_1–Cand_4* have the same nucleotide sequence. *Cand_5* and *Cand_6* also have the same nucleotide sequence. Therefore, they may have been generated by very recent genomic duplications, although we cannot exclude the possibility that some of them are artifacts generated by the misassembly of the genomic sequence. Figure 2.8 shows 5’-end positions of predicted mature miRNAs and those identified the deep-sequencing experiments by Hendrix et al. (2010). In many cases, predicted 5’-ends (coloured nucleotides) are located near the 5’-ends identified by the deep-sequencing results (black arrows). *Cand_1–Cand_4* were not reported by Hendrix et al. (2010), because they were located on recently sequenced genomic regions. *Cand_8* and *Cand_9* also were not reported by Hendrix et al. (2010) possibly due to the small number of sequencing reads.

Table 2.1. Promising candidates

Name	Genomic coordinate[a)]	Clustered miRNA
Cand_1	chr13p:43397-43483	cand_2
Cand_2	chr13p:46031-46118	cand_1
Cand_3	scaffold_121:338389-338465	cand_4
Cand_4	scaffold_121:340167-340242	cand_3
Cand_5	scaffold_280:34403-34489	None
Cand_6	chr10q:2147938-2148025	cin-mir-4054
Cand_7	chr10q:2150133-2150203	cin-mir-4091
Cand_8	chr04q:5406744-5406831	None
Cand_9	chr03q:5428973-5429063	cin-mir2235
Cand_10	chr02q:7039737-7039806	None

a) Genomic coordinate of ci2 genome (Mar. 2005 Assembly).

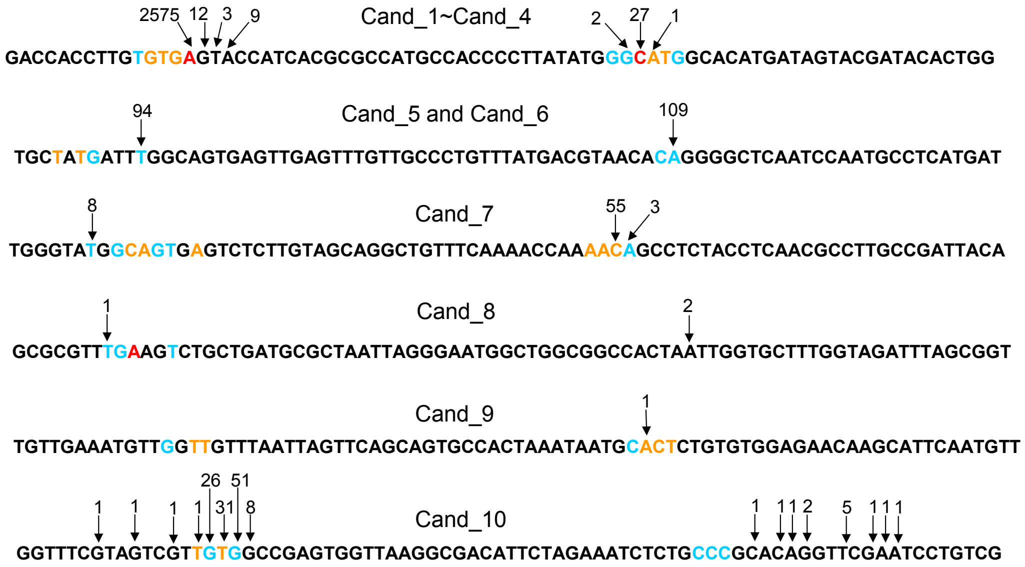

Cand_1~Cand_4
2575
12
3
9
2
27
1
GACCACCTTGTGTGAGTACCATCACGCGCCATGCCACCCCTTATATGGGCATGGCACATGATAGTACGATACACTGG
Cand_5 and Cand_6
94
109
TGCTATGATTTGGCAGTGAGTTGAGTTTGTTGCCCTGTTTATGACGTAACACAGGGGCTCAATCCAATGCCTCATGAT
Cand_7
8
55
3
TGGGTATGGCAGTGAGTCTCTTGTAGCAGGCTGTTTCAAAACCAAAACAGCCTCTACCTCAACGCCTTGCCGATTACA
Cand_8
1
2
GCGCGTTTGAAGTCTGCTGATGCGCTAATTAGGGAATGGCTGGCGGCCACTAATTGGTGCTTTGGTAGATTTAGCGGT
Cand_9
1
TGTTGAAATGTTGGTTGTTTAATTAGTTCAGCAGTGCCACTAAATAATGCACTCTGTGTGGAGAACAAGCATTCAATGTT
Cand_10
1
1
1
1
26
31
51
8
1
1 1 2
5
1 1 1
GGTTTCGTAGTCGTTGTGGCCGAGTGGTTAAGGCGACATTCTAGAAATCTCTGCCCGCACAGGTTCGAATCCTGTCG

Figure 2.8. Comparison of 5'-ends of mature miRNAs predicted by miRRim2 and those identified by deep-sequencing. The probability of a predicted 5'-end ($\mathrm{P}^{5\mathrm{end}}i$) is indicated by colours; Black, blue, orange, and red means $0 \leq \mathrm{P}^{5\mathrm{end}}i < 0.05$, $0.05 \leq \mathrm{P}^{5\mathrm{end}}i < 0.1$, $0.1 \leq \mathrm{P}^{5\mathrm{end}}i < 0.4$, and $0.4 \leq \mathrm{P}^{5\mathrm{end}}i$, respectively. Arrows indicate the 5'-ends identified by deep-sequencing experiments by Hendrix et al. The number associated with an arrow indicates the number of reads.

2.4 Conclusions

In this study, we developed the miRRim2 method for detecting miRNA hairpins and their mature forms by integrating evolutionary, secondary structural, and nucleotide features in each position of miRNA hairpins. Our method achieved better prediction accuracy than genome-wide computational screenings previously performed by other groups. By investigating the contribution of each type of feature to the prediction accuracy, it was shown that evolutionary and secondary structural features, but not nucleotide features, are important for detecting miRNA hairpins. For the prediction accuracy of mature miRNAs, it was shown that nucleotide and secondary structural features were more important than evolutionary features. When miRRim2 was applied to the *Ciona intestinalis* genome, several promising candidates were detected. The prediction results for miRNA hairpins, miRNA duplexes, and 5′-ends of mature miRNAs in humans and *Ciona intestines* are available from http://mirrim2.ncrna.org.

2.5 Materials and Details

2.5.1 Materials

Construction of core miRNA hairpins and non-miRNA regions

From the 731 human miRNA hairpins in miRBase version 14.0 (Kozomara et al. 2011) that could be mapped on the human genome (version hg18), we selected 398 conserved miRNA hairpins with a mean PhastCons score > 0.5. From these 398 miRNA hairpins, we selected 307 instances that were validated by at least two independent experimental evidences. We checked the presence of experimental evidences for each miRNA hairpin by surveying the literatures listed in miRBase version 14.0. Finally, we excluded the mirtron-type miRNA hairpins (Berezikov et al. 2007). The remaining 306 miRNA were used as training and test data, which we referred to "core miRNA hairpins". We made the length of core miRNA hairpins to be 200-bp by extending upstream and downstream regions.

For some of the core miRNA hairpins, the location of the passenger strand was not annotated. The locations of passenger strands, however, were needed for training the model parameters (see below). In such cases, they were deduced from the location of a mature miRNA assuming the 2-bp 3′-overhang illustrated in Fig. 2.1.

Non-miRNA regions were randomly selected from non-conserved and conserved genomic regions. We selected 10000 instances from non-conserved sequence segments (mean PhastCons score < 0.4) and another 10000 from conserved segments (mean PhastCons score > 0.6). The length of the non-miRNA regions was 200-bp.

2.5.2 Details

Definition of a feature vector

In our method, each genomic position *i* was expressed by a 7-dimensional vector $\mathbf{o}^{(i)}$. Dimension 1 of $\mathbf{o}^{(i)}$ is the PhastCons score (Siepel et al. 2005) of position *i*. Dimension 2 and 3 is the PhastCons score in position *i*-20 and *i*+20, respectively. Dimension 4 is the PhyloP score (Siepel et al. 2006) which is another measure of evolutionary conservation. In this study, we used the PhastCons and PhyloP score calculated based on multiple alignment across 44 vertebrates. The PhastCons and PhyloP score in each position of the human genome were obtained from the USCS genome browser (Fujita et al. 2011).

Dimension 5 is the base pair potential which represents the likelihood of forming a base pair in each position. The base pair potential at position *i*, BPP$_i$, is calculated as:

$$\mathrm{BPP}_i = \max_{-120\leq j\leq +120}(p_{ij})$$

where p$_{ij}$ is a base pair probability between positions *i* and *j* that can be calculated by McCaskill's algorithm (McCaskill 1990). We used the Rfold program (with the option L = 120) (Kiryu et al. 2008) for calculating base pair probabilities in a genome-wide manner.

Dimension 6 is the base pair distance which represents the distance between a predicted base pair. The base pair distance of position *i*, BPD$_i$, is calculated by the following equation:

$$\mathrm{BPD}_i = \left\{\begin{array}{ll} -\infty & \text{if } (\mathrm{BPP}_i < 0.5) \\ J - i & \text{otherwise} \end{array}\right\}$$

where $J = \arg\max_{-120\leq j\leq +120}(p_{ij}) \cdot$

Conversion of continuous values into symbols

Dimensions 1–6 of a 7-dimensional vector **o** are represented by continuous values. We converted them into 5 (or 6) distinct symbols, and this conversion was performed in each dimension. First, continuous values of a given dimension were obtained from the core miRNA hairpins. Continuous values belonging to the 20% or lower percentile were converted into symbol “A.” Similarly, the 20–40%, 40–60%, 60–80%, and 80% or higher percentiles were converted into “B”, “C”, “D”, and “E”, respectively. Dimension 6, which contains negative infinities, was converted into 6 distinct symbols. We converted negative infinities into symbol “F”, and the other continuous values into “A”–“E” using the same procedure as for dimensions 1–5. For dimension 7, we simply assigned “A”, “B”, “C”, and “D” to nucleotides A, U, G, and C, respectively, in order to limit the number of symbols used in our method. Therefore, the feature vector **o** is converted to a symbol vector such as (E, B, E, E, B, C, A).

The architecture of each sub-model

The probabilistic model employed here consists of 12 sub-models (Fig. 2.4). The architecture of the Mature and Passenger sub-models consists of 25 connected states (Fig. S2.6a). Each state has an emission function, e(**o**), that assigns a “weight” to the feature vector **o**. As each state has its own emission function, the sub-models can capture the features in each position of mature miRNAs and passenger strands. Each connection between states has a transition parameter by which the length preference of mature miRNA regions can be modeled. For example, the propensity that mature miRNAs are 22 nt is expressed by assigning a positive large weight to the transition parameter between state 21 and 25 (the broad line in Fig. S2.6a). The Loop sub-model consists of 8 states, each of which is connected to itself (Fig. S2.6b). By using this simple architecture, we can roughly model the length and position-specific features of terminal loop regions without using a large number of states. For the Flanking sub-model, we used 20 linearly connected states (Fig. S2.6c) to capture the features around the DRB, which is located ~11 or 13 nt from the Drosha cleavage sites (Fig. 2.1). The Non-miRNA model consists of a single state with a self transition (Fig. S2.6d).

The emission function

Each state in our CRF model has an emission function e(**o**), which is defined as follows:

$$\mathrm{e}(\mathbf{o}) = \sum_{d=1}^{7} \mathrm{w}_d(\mathrm{o}_d)$$

where **o** is a 7-dimensional feature vector, o_d is a symbol in dimension *d* of **o**, and $\mathrm{w}_d(\mathrm{o}_d)$ is a weight assigned to the symbol o_d in dimension *d*. For example, the vector **o** = (E, B, E, E, B, C, A) has a total weight $= \mathrm{w}_1(\mathrm{E}) + \mathrm{w}_2(\mathrm{B}) + \mathrm{w}_3(\mathrm{E}) + \mathrm{w}_4(\mathrm{E}) + \mathrm{w}_5(\mathrm{B}) + \mathrm{w}_6(\mathrm{C}) + \mathrm{w}_7(\mathrm{A})$. The emission parameter, w_d, was optimized from the training data.

Training emission and transition parameters

In CRFs, the conditional probability distribution P(y| x;**m**) can be directly trained from training data, where, in our case, x is a feature vector sequence, y is a sequence of "labels" assigned to x which reveals the location of certain sequences such as mature miRNA and the terminal loop, and **m** is a vector of emission and transition parameters. Intuitively, the parameters are optimized such that the *predicted* labels agree with the *true* labels as much as possible. This is achieved by iteratively maximizing the conditional log-likelihood of observing *true* labels.

For training the model parameters **m**, we used feature vectors corresponding to miRNA hairpins and non-miRNA regions. According to the annotation of miRBase, we assigned the labels "M", "L", "P", "F", and "N" to each position of the mature miRNA, terminal loop, passenger strand, flanking, and non-miRNA region, respectively. All the parameters **m** = (m_1,m_2...m_J) were initially set to be 0 and were iteratively optimized using the limited-memory quasi-Newton method (L-BFGS) (Liu et al. 1989), which is a general purpose convex optimization algorithm. To prevent over-fitting, we penalized the conditional log-likelihood with the Gaussian prior $\sum_j C_j m_j^2$. In this study, we set C_j = c for j = 1, 2, ..., J. The constant c is determined based on the prediction accuracy for a part of training data (see below).

Determining penalty parameter *c*

In our method, the training data are divided into two groups. The first group was used to optimize transition and emission parameters, and the second one was used to determine an appropriate penalty parameter c. The penalty parameter is chosen from c = 0.1, 1, 10, 50, 100 based on the prediction accuracy for the second group. The prediction accuracy is measured based on F-score, which is a harmonic mean of sensitivity and positive predictive value (PPV). We calculate F-scores at various probabilistic thresholds, P^{mi}_i, and the maximum F-score is used as a measure of the prediction accuracy.

Genome-wide cross validation

We evaluated the accuracy of miRRim2 based on a genome-wide cross validation as follows. First, we selected a particular human chromosome, which we referred to as a "test chromosome". Then, we trained miRRim2 using the core miRNA hairpins and non-miRNA regions on the remaining chromosomes, and used it to scan the test chromosome. To mimic a realistic situation, the core miRNA hairpins were excluded from training data if they were homologous to the core miRNA hairpins in the test chromosome. The information on homologues was obtained from the miFam.dat file in miRBase v.14. The training data are divided into two groups. The first group consists of randomly selected 80% of miRNA hairpins, and the same number of non-miRNA data. The second group consists of the remaining miRNA hairpins and non-miRNA data. The first group was used to optimize transition and emission parameters, and the second one was used to determine an appropriate penalty parameter *c*. This procedure was repeated for all the 24 human chromosomes. So the whole human genome was used for evaluation.

In the genome-wide cross validation, the penalty parameter *c* was determined for each of the 24 human chromosomes. For 16 of the 24 chromosomes, the prediction accuracy for the training data was highest when $c = 10$. Although we can use different *c* for each of the 24 chromosomes, we used $c = 10$ for all the 24 chromosomes.

Definition of the miRNA hairpin probability and mature miRNA probability

To detect miRNA hairpins, we defined the probability that each genomic position *i* is a miRNA hairpin, P^{mi_i}, as:

$$P^{mi_i} = \sum_{k \in S_{mirna}} p_{i,k}$$

where S_{mirna} is a set of states belonging to the Flanking, Mature, Passenger, and Loop sub-models, and $p_{i,k}$ is a posterior probability that position *i* is derived from state *k*, which can be calculated by the Forward-Backward algorithm (Baum and Egon 1967). We considered a continuous sequence segment of 80 base pairs (bp) or more with a $P^{mi_i} > T$ as a predicted miRNA hairpin, where *T* is a probabilistic threshold from 0 to 1. Predicted miRNA hairpins of 150 bp or more were discarded.

The probability of position *i* being the 5′-end position of a mature miRNA region, P^{5end_i}, is defined as:

$$P^{5end_i} = \sum_{k \in S_{5end}} p_{i,k} / P^{mi_i}$$

where S_{5end} is a set of the first states in all the Mature sub-models. The position with $P^{5end_i} > T$ is considered to be the 5′-end of a mature miRNA.

Training of miRRim2 for the *Ciona intestinalis* genome

We used the core miRNA hairpins to train miRRim2 and used it to scan the *Ciona intestinalis* genome. The core 306 miRNA hairpins were divided into two groups. The first group consists of randomly selected 80% of miRNA hairpins, and the same number of non-miRNA data. The second group consists of the remaining miRNA hairpins and non-miRNA data. The first group was used to optimize transition and emission parameters, and the second one was used to determine an appropriate penalty parameter c. In this case, $c = 10$ was appropriate.

Chapter 3

Predicting pseudogenes based on their surrounding sequence signature

3.1 Introduction

Pseudogenes are one of the prominent genomic elements of mammalian genomes. Therefore, the identification of pseudogenes can make a significant contribution to the correct annotation of the human genome. In eukaryotes, pseudogenes can be classified into two major types according to their generation mechanism. One is those resulting from genomic duplication; such pseudogenes are therefore called duplicated pseudogenes. The other type comprises those pseudogenes generated by retrotransposition, which is a series of *in vivo* processes involving the reverse transcription of RNA molecules and the integration of the transcripts into the genome. It is known that most pseudogenes in humans are generated by retrotransposition (Pei et al. 2012). Hereafter, we refer to retro-pseudogenes simply as pseudogenes.

Retrotransposition in eukaryotes can be divided into two types; the long terminal repeat (LTR) type and the non-LTR type. The latter accounts for the majority of retrotransposition events in human (Ostertag and Kazazian 2001). Various types of RNA molecules, including *Alu* RNAs, LINE RNAs, mRNAs, and small noncoding RNAs are copied via non-LTR retrotransposition (Dewannieux et al. 2003; Moran et al. 1996; Esnaul et al. 2000; Buzdin et al. 2003; Perreault et al. 2005). An increasing number of versatile roles for retrotransposition have been recognized, such as the generation of novel functional genes and modulation of gene expression. Insertion of LINE-1 and *Alu* in a 3' UTR may reduce gene expression (Faulkner et al. 2009). Retrotransposition may have expanded regulatory elements in the promoter region (Bourque et al. 2008), and some endogenous siRNAs are derived from mRNA pseudogenes (Watanabe et al. 2009). Retrotransposition of LINE-1 may mediate exon shuffling (Moran et al. 1999). Retrotransposition of mRNA is one mechanism for generating functional genes (Babushok et al. 2007).

Non-LTR retrotransposition is mediated by the protein encoded by the second open reading frame of LINE-1 (hereafter L1-ORF2p). This protein has both reverse transcriptase and endonuclease activity (Feng et al. 1996) and promotes retrotransposition of LINE-1 RNAs themselves (Moran et al., 1996). The endonuclease activity of L1-ORF2p creates a cleavage site in genomic DNA (Feng et al. 1996); the cleavage site is used as a primer, and reverse transcription of template RNAs and integration of the resultant cDNAs into the genome occur simultaneously. This integration process is called target-site-primed reverse transcription (TPRT) (Luan et al. 1993; Cost et al. 2002). In addition to LINE-1 RNAs, L1-ORF2p recognizes *Alu* RNAs (Dewannieux et al. 2003) and the mRNAs of protein-coding genes and promotes their retrotransposition, although its recognition efficiency for protein-coding genes is much lower than for LINE-1 and *Alu* (Dewannieux et al. 2003; Esnault et al. 2000; Wei et al. 2001).

In many LINE-1 and pseudogenes observed in the human genome, the 5'-end region of the template transcript has been truncated (Ostertag and Kazazian 2001; Torrents et al. 2003; Zhang et al. 2003). This has long been explained by the inability of L1-ORF2p to copy the entire length of the template RNA during retrotransposition, or degradation of the template RNA before completion of reverse transcription (Ostertag and Kazazian 2001). However, full-length (nontruncated) LINE-1 are also frequently observed (Boissinot et al. 2000; Pavlícek et al. 2002; Myers et al. 2002; Salem et al. 2003). The mechanism for the preferential generation of full-length LINE-1 has not been explained.

In mammals, there are three types of sequence signatures around a sequence element generated by non-LTR retrotransposition (Figure 3.1). The first is a poly-A tract found immediately downstream of the 3'-end of a retrotransposed element. The second is a pair of duplicated sequences surrounding the retrotransposed element, called target site duplications (TSDs). The third is the TTAAAA consensus sequence, which overlaps with the 5'-end of the 5'-TSD. This consensus sequence is recognized by L1-ORF2p endonuclease to create the cleavage site in genomic DNA, but is not always present (Cost et al. 2002; Jurka et al. 1997). The mechanisms generating the poly-A tract and TSD are not fully understood, but the presence of these sequence signatures is an established phenomenon and can be used to detect retrotransposed elements.

In this study, we developed a novel algorithm for detecting pseudogenes based on the presence of the poly-A tract and TSDs and implemented this algorithm as the TSDscan program. Because TSDscan uses general sequence signatures surrounding retrotransposed elements, it is able to detect any type of sequence element generated by non-LTR retrotransposition. TSDscan detected many previously unknown short pseudogenes generated by retrotransposition of mRNA. TSDscan also allows us to analyze detailed characteristics of pseudogenes, such as the length distribution of TSDs

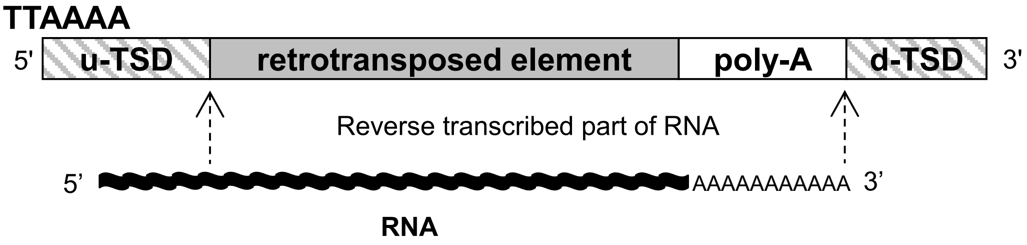


Figure 3.1 Signature sequences of a retroposed element. Several features can be used to identify a retroposed element. A poly-A tract (usually 5–30 bp) is located downstream of the retroposed element and a pair of duplicated sequences, the target site duplications (TSDs), are located at either side of the retroposed element–poly-A structure. The TSD is usually also 5–30 bp. We denote the upstream and downstream TSD sequences as u-TSD and d-TSD, respectively. The AAAA in the TTAAAA consensus sequence overlaps with the u-TSD (Jurka et al. 1997). Reverse transcription of an RNA starts from its 3'-end and does not always copy the entire length (the reverse-transcribed region is indicated with arrows). Therefore, a retroposed element often lacks the 5' region of its parent RNA.

and poly-A tracts, which are useful for further study of the molecular mechanisms of pseudogenes generation. From this analysis, we found that short pseudogenes are more frequently generated from long mRNAs than from short mRNAs. In order to explain this phenomenon in the context of events previously reported to be associated with retrotransposition, we propose that two in vivo processes generate pseudogenes: short parent mRNAs use template-jumping to generate a full-length pseudogene, whereas long parent mRNAs are more likely to be truncated and degraded, after which microhomology generates a short pseudogenes of the mRNA. The findings we presented here provide new insights into the mechanism of retrotransposition.

3.2 Materials and Methods

3.2.1 TSDscan: an algorithm to detect pseudogenes

TSDscan is an algorithm that not only detects pseudogenes but can also predict the length and boundaries of sequence signatures surrounding the pseudogenes. In TSDscan, upstream and downstream sequences of a pseudogene are aligned and scored with a specific scheme. To detect the sequence signatures shown in Figure 3.1, we need to consider that the lengths of poly-A tracts and TSDs are variable and that random mutations accumulate in these sequence signatures. Because the upstream TSD (u-TSD) and downstream TSD (d-TSD) are similar, TSDs can be detected by assigning positive scores to a base match and assigning negative scores to a base mismatch or gaps, analogous to the usual alignment technique for detecting similar regions in genes (Durbin et al. 1998). The poly-A tract is detected by assigning positive scores to the insertion of a poly-A sequence immediately before the d-TSD. Figure 3.2a is an example of the alignment and the scores assigned to each alignment column. For alignment columns corresponding to TSDs (dotted rectangles), a pair of aligned nucleotides has a score defined in the HOXD matrix (Schwartz et al. 2003). The first gap and gaps following it have scores of –400 and –30, respectively. In a poly-A tract located before a d-TSD, insertion of nucleotide A and of other nucleotides (G, T, and C) have scores of +100 and –100, respectively. The total score is the sum of scores assigned to each alignment column. In the case of Figure 3.2a, the total score is 1129.

Figure 3.2b is an example of an alignment containing the TTAAAA consensus sequence. Alignment columns corresponding to the TTAAAA consensus are shown in the gray area. In the first two columns of the gray area, insertion of nucleotide T and of other nucleotides (A, G, and C) have scores of +200 and –200, respectively. In the last four columns, an A-A nucleotide match and the other aligned nucleotide pairs have scores of +200 and –200, respectively (see supplemental Methods S3.1 about the reason for using these scores). Scores assigned to columns outside the gray area are the same as in Figure 3.2a. In the case of Figure 3.2b, the total score is 1352.

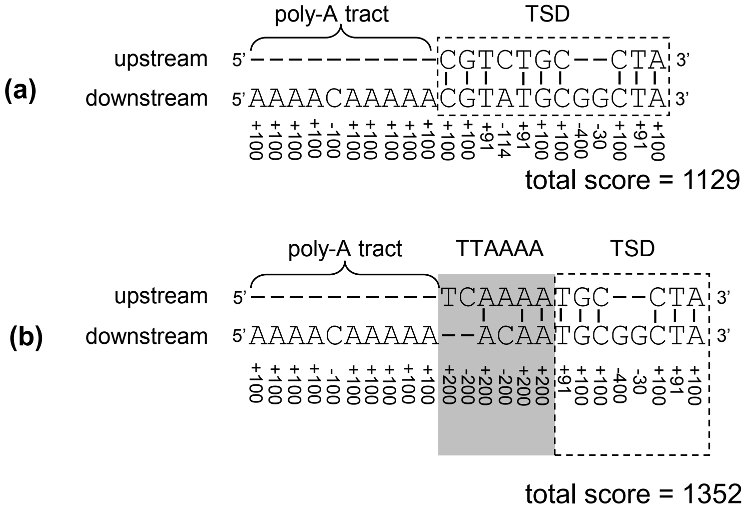


Figure 3.2 Examples of how scores are assigned to each potential pseudogene. Examples of alignments containing (a) a poly-A tract and TSD and (b) containing a TTAAAA consensus sequence, poly-A tract, and TSD. See text for details.

Next, we consider the positions of the poly-A tract and TSDs. The poly-A tract and TSDs are inserted immediately outside of a retrotransposed element (Figure 3.1). Therefore, poly-A tracts and TSDs that are distant from a pseudogene should be penalized. In TSDscan, each nucleotide insertion between u-TSD and the 5'-end of the pseudogene, and between the 3'-end of the pseudogene and a poly-A tract, has a score of –50.

In TSDscan, the alignment maximizing the total score is detected with a dynamic programming algorithm. Details of the algorithm are provided in supplemental Methods S3.2. In addition, the source code of TSDscan (perl and C++ versions) is available at http://www.ncrna.org/software/tsdscan/, which may help with understanding the algorithmic details.

3.2.2 A pipeline for the detection of pseudogenes

Deleting repeats from a genomic sequence

If *Alu* or LINE-1 were inserted within a pseudogene, the pseudogene would be disrupted. Determining both ends of such a pseudogene becomes a bit complicated, and we needed a way to cope with this complexity, because, in our method, both ends of pseudogenes need to be determined fairly precisely. To circumvent the complexity, we created a genomic sequence from which *Alu* and LINE-1 are deleted. In addition, tandem repeat sequences detected by tandem repeats finder (TRF) (Benson 1999) were masked by 'N'. Hereafter, the genomic sequence thus created is called the "processed genome".

Searching homologous regions of mRNAs

We obtained the nucleotide sequences of all human mRNAs from the 'Human mRNAs' track of the human genome (version hg17) in the UCSC genome browser (Kuhn et al. 1990). mRNAs with 3' UTR annotation were excluded from further study. Then we deleted *Alu* and LINE-1 from the mRNA sequences and masked tandem repeat sequences detected by TRF. Using these mRNAs as queries, we searched the processed genome using blastz with the default parameters. Among blastz hits, we excluded those that did not contain the 3' UTR of query mRNAs, because a pseudogene of an mRNA should contain the 3' UTR.

Excluding overlap with known genes

We converted positions of blastz hits in the processed genome into positions in the original genome. Then we excluded blastz hits that overlapped with exons of human mRNAs.

Excluding redundancy of blastz hits

Blastz hits often overlap with each other. In such cases, we excluded the blastz hit with the lower score.

Applying TSDscan

We applied TSDscan to the regions 100 bp upstream and downstream of blastz hits. Then we extracted blastz hits with a TSDscan score of 1100 or more. Among 10,000 genomic regions that we randomly selected, only about 3% had a score of 1100 or more (Figure S3.1). We excluded blastz hits having TSDs or poly-A tracts of more than 30 bp even if the score was at least 1100 because such long TSDs and poly-A tracts were rarely seen for LINE-1 and *Alu* (Figures S3.2 and S3.3), and thus they may be false positives.

3.2.3 Evaluating the accuracy of TSDscan

Because TSDscan uses general sequence signatures surrounding retrotransposed elements, it is able to detect not only mRNA pseudogenes but also sequence elements such as LINE-1 and *Alu*. For our evaluation study, we designated two types of LINE-1 subfamilies (L1P: primate specific LINE-1 and L1M: mammalian-wide LINE-1) and three types of *Alu* subfamilies (*Alu*Y, *Alu*S, and *Alu*J) as positive samples, and randomly selected genomic regions as negative samples. The detection accuracy is measured by the ACC score, which is the average of sensitivity and specificity. Sensitivity and specificity are defined as:

Sensitivity = TP/(TP+FN)

Specificity = TN/(FP+TN),

where TP, FP, TN, and FN are the number of true positives, false positives, true negatives, and false negatives, respectively.

TSDfinder (Szak et al. 2002) is a program that defines the boundaries of a retrotransposed element based on the presence of TSDs. The TSDfinder program consists of several steps, including merging and determining the boundaries of a retrotransposed element, obtaining sequences surrounding the retrotransposed element, detecting potential TSDs, and scoring TSDs. To evaluate TSDfinder using exactly the same test data as we used for TSDscan, we modified the TSDfinder program such that we could input sequence data directly.

Sequence data of *Alu*, LINE-1, and random genomic regions

We used RepeatMasker (http://www.repeatmasker.org/) to detect sequence data for *Alu* and LINE-1. RepeatMasker often detects poly-A tracts in the 3'-end of *Alu* and LINE-1 as a part of repetitive sequences. To avoid this, we excluded poly-A tracts from the 3'-end of the consensus sequences of LINE-1 and *Alu*, and ran RepeatMasker using the truncated consensus sequences as queries. Among the *Alu* and LINE-1 sequences detected by RepeatMasker, we excluded those which lacked 5 bp or more of the 3'-end, because *Alu* and LINE-1 should contain the 3'-end of their original transcripts at the time when they were retrotransposed. We also excluded *Alu* and LINE-1 if their upstream and downstream 100 bp contained other repetitive sequences detected by RepeatMasker. Among the remaining *Alu* and LINE-1 sequences, we randomly selected 1000 samples for each *Alu* and LINE-1 subfamily (L1P, L1M, *Alu*Y, *Alu*S, and *Alu*J).

Data for random genomic regions were generated by sampling 10,000 genomic regions that did not overlap with repetitive sequences identified by RepeatMasker and TRF.

3.3. Results and Discussion

3.3.1 Comprehensive detection of pseudogenes in the human genome

We obtained 84,332 mRNA sequences with a 3' UTR annotation from the 'Human mRNAs' track in the UCSC genome browser, human genome version hg17 (Kuhn et al. 2009). Using these mRNA sequences as queries, we performed homology searches against the human genome by using blastz (Schwartz et al. 2003). We excluded blastz hits that did not have homology to the 3' UTR, because, as shown in Figure 3.1, reverse transcription starts from the 3' end of the mRNA, and an mRNA pseudogene should contain the 3' UTR. After also excluding overlapping blastz hits, we obtained 27,465 hits, which we considered candidate pseudogenes. We applied TSDscan to the 100-bp upstream and downstream regions of these pseudogene candidates and extracted those with a score of 1100 or higher. Among 27,465 candidate pseudogenes, 6,982 passed the score threshold. Then, we excluded candidate pseudogenes having TSDs or poly-A tracts of more than 30 bp even if the score was at least 1100. Among the 6,982 high scoring candidates, 4,464 passed the poly-A tract and TSD length threshold, and we considered these to be mRNA pseudogenes.

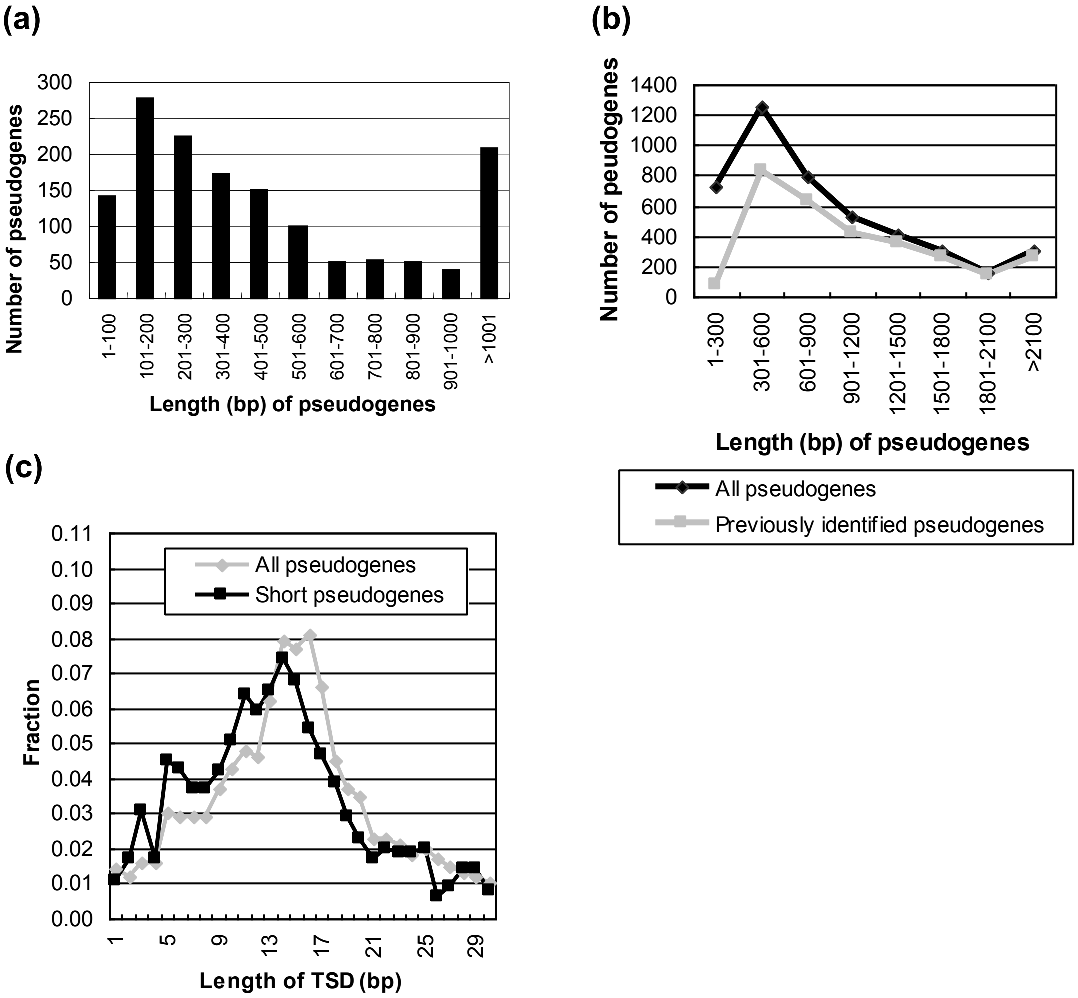


Figure 3.3 Length distribution of pseudogenes and their TSDs. (a) Length distribution of pseudogenes newly found in this study. The median length is 356 bp (mean 533.2 with s.d. 562.6 bp). (b) Comparison of length distributions of previously identified pseudogenes and all pseudogenes identified with TSDscan. Previously identified pseudogenes are the pseudogenes identified by TSDscan that overlap with the existing pseudogene annotation (Torrents et al. 2003; Zhang et al. 2003). The median length of previously identified pseudogenes is 877 bp (mean 1096.7 with s.d. 756.9 bp) and that of all pseudogenes are 691 bp (mean 911.0 with s.d. 747.3 bp). (c) Length distributions of TSDs of the 654 short pseudogenes and all pseudogenes identified with TSDscan.

3.3.2 Discovery of short pseudogenes

Most of the novel pseudogenes, that is, pseudogenes that did not overlap with the existing pseudogene annotations (Torrents et al. 2003, Zhang et al. 2003), were short (Figure 3.3a). Median length of the novel pseudogenes is 356 bp, which is much shorter than that of all pseudogenes identified in this study (691 bp). By discovering the novel pseudogenes, the length distribution of pseudogenes in the human genome changed significantly (Figure 3.3b). Our results contained 645 new pseudogenes with lengths of 300 bp or less. To verify that the new short pseudogenes were not caused by a limitation of the TSDscan algorithm, we investigated the length distribution of TSDs for short pseudogenes. The TSD length distribution of the 645 short pseudogenes, as well as that of all pseudogenes detected in this study, had a peak around 15 bp (Figure 3.3c). The TSD length distribution peak is consistent with that of LINE-1 and *Alu* (Jurka et al. 1997; Szak et al. 2002; Zingler et al. 2005; Figure S3.2), supporting the validity of the short pseudogenes we detected.

3.3.3 Relationship between parent gene and pseudogene length

Because our method could accurately predict the boundaries of a pseudogene, we could investigate the relationship between parent gene and pseudogene length. Figure 3.4a is a two-dimensional plot of parent mRNA length (X-axis) versus pseudogene coverage, i.e., the length of the pseudogene relative to that of the parent mRNA (Y-axis). A cluster can be seen along the line of pseudogene coverage = 100%, indicating that full-length pseudogene are frequent. Another cluster of plots can be seen in the lower right area, suggesting that short and truncated pseudogenes are frequently generated from long mRNAs. The fraction of full-length pseudogene gradually decreased as the length of the parent mRNA became longer, and conversely, the fractions of short and truncated pseudogenes gradually increased as the length of the parent mRNAs became longer. In addition, it can be seen that 1) most pseudogenes derived from short mRNAs are full-length; 2) most pseudogenes derived from long mRNAs are short and truncated; and 3) for medium-to-long mRNAs, both full-length and short and truncated pseudogenes are frequent. Therefore, the length distribution of pseudogenes of medium-to-long mRNAs has two peaks, similar to the bimodal length distribution already reported for LINE-1 elements (Boissinot et al. 2000; Pavlícek et al. 2002; Myers et al. 2002; Salem et al. 2003; Babushok et al. 2006).

To explicitly show the relationship between the fractions of short (≤300 bp) pseudogenes and parent mRNA length, we divided parent mRNA length into categories and calculated the fraction of the short pseudogenes for each mRNA length category (Figure 3.4b). The boundaries of the categories are shown by the vertical dotted lines in Figure 3.4a. As can be seen in Figure 3.4b, short pseudogenes were generated more frequently from long mRNAs than from short mRNAs, and the longer a parent mRNA

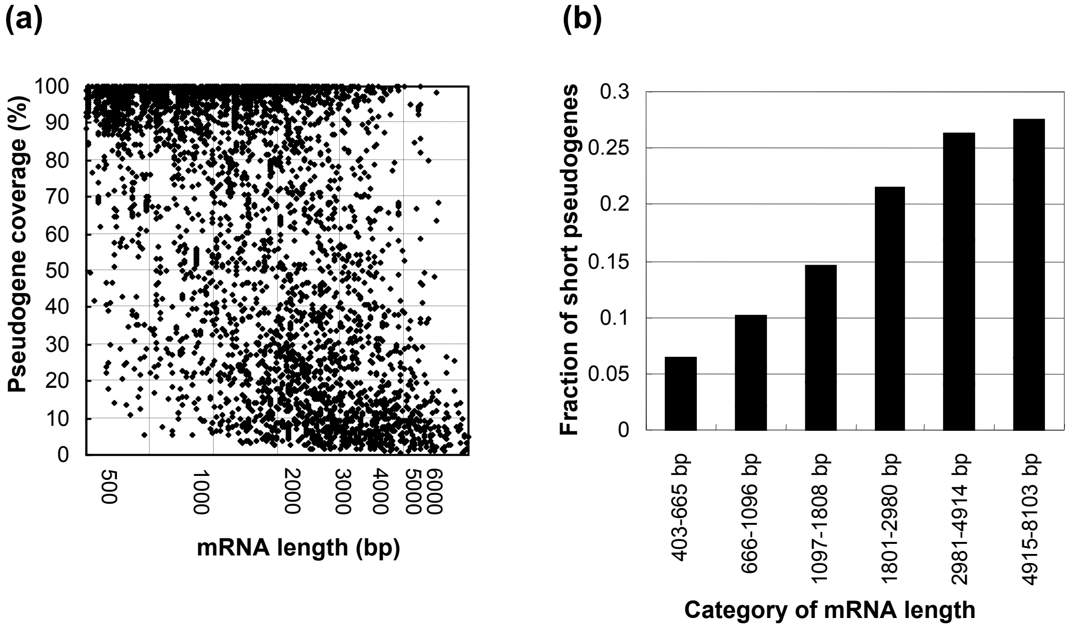


Figure 3.4 (a) Relationship between parent mRNA length and pseudogene coverage. Parent mRNA length (X-axis) is shown in a logarithmic scale. Pseudogene coverage (Y-axis) is the length of a pseudogene divided by that of its parent mRNA. (b) Fraction of short pseudogenes (≤300 bp) calculated for each mRNA length category.

was, the more frequently short pseudogenes were generated. Reverse transcription of template RNAs by L1-ORF2p is an essential step of retrotransposition, but poor processivity of L1-ORF2p alone cannot explain why long mRNAs more frequently generate short pseudogenes. Here, we hypothesize that most long mRNAs are truncated before they are reverse transcribed. Details are described later in this section.

3.3.4 Comparison of TSDscan with existing methods

Two other software packages, RTanalyzer (Lucier et al. 2007) and TSDfinder (Szak et al. 2002), are available to detect retrotransposed elements. RTanalyzer detects retrotransposed elements based on the presence of a poly-A tract and TSDs. Potential TSDs are first identified by local alignment, and the final score is calculated based on the presence of the poly-A tract and the TTAAAA consensus sequence. However, because RTanalyzer is available only through a web interface, we could not evaluate its accuracy using our large test data set and therefore could not include RTanlyzer in our comparison.

TSDfinder is a program that defines the boundaries of a retrotransposed element based on the presence of TSDs. In TSDfinder, potential TSDs are identified by aligning the upstream and downstream regions of a retrotransposed element and detecting those that have perfect nucleotide matches in at least 9 consecutive base pairs (Szak et al. 2002). The final score is calculated by considering both the TSD position and alignment score. To compare TSDscan with TSDfinder, we measured the detection accuracy by using the ACC score, an average of sensitivity and specificity scores (see Materials and Methods). Table 3.1 shows the detection accuracy of TSDscan and TSDfinder for five types of retroposons (L1P, L1M, *Alu*Y, *Alu*S, and *Alu*J). In all five types, the ACC score of TSDscan was higher than that of TSDfinder; therefore, for the purpose of detecting retrotransposed elements, TSDscan is superior to TSDfinder. The sensitivity of TSDfinder was relatively low (Table 3.1), which may be due to the stringent criterion of at least 9 perfect nucleotide matches for detecting TSDs. In contrast, our method has greater sensitivity because of its flexible requirement for detecting TSDs.

Table 3.1 Comparison of the detection accuracy of TSDscan and TSDfinder

	TSDscan			TSDfinder		
	ACC_{max} a)	sensitivity	specificity	ACC_{max} a)	sensitivity	specificity
L1P	0.926	0.914	0.937	0.628	0.263	0.994
L1M	0.834	0.821	0.848	0.525	0.056	0.995
AluY	0.991	0.991	0.991	0.788	0.582	0.994
AluS	0.975	0.968	0.981	0.679	0.365	0.994
AluJ	0.949	0.941	0.957	0.596	0.198	0.994

a) The ACC score is the average of sensitivity and specificity, and ACC_{max} is the maximum ACC score.

3.3.5 A proposed model for generating short pseudogenes

This is the first large-scale analysis of short pseudogenes derived from mRNAs in the human genome. In previous studies, retropseudogenes were detected mostly based on their lack of introns and the accumulation of random mutations in their protein-coding sequences (Torrents et al. 2003; Zhange et al. 2003; Ohshima et al. 2003). However, these methods cannot be applied to short pseudogenes, because most short pseudogenes are derived from last exons that lack protein-coding sequences, where homology searches do not work effectively. Therefore, short pseudogenes have escaped detection in previous studies. In addition to discovering novel short pseudogenes, our method accurately predicts the boundaries of pseudogenes. This enabled us to closely investigate the essential characteristics of pseudogenes.

Using TSDscan, we made the novel discovery that long mRNAs tend to produce a higher percentage of short pseudogenes than do short mRNAs. Although target-site-primed reverse transcription (TPRT) is the currently accepted mechanism of retrotransposition (Ostertag et al. 2001), it does not explain this length-dependent phenomenon. Here, we propose that two in vivo processes generate pseudogenes in a length-dependent manner. We hypothesize that most long mRNAs are truncated before they are reverse transcribed (Figure 3.5a). Because RNAs without a 5' cap are rapidly digested (Newbury et al. 2006), the template RNA may be removed during reverse transcription. If the template RNA is digested before the completion of reverse transcription, a single-stranded cDNA is exposed, which is integrated into the genome by the microhomology-mediated mechanism proposed by Zingler et al. (2005). If reverse transcription is completed before digestion of the template RNA, L1-ORF2p moves to a genomic 3' overhang via template jumping (Bibillo et al. 2004). After the genomic 3' overhang region is reverse transcribed, removal of the template RNA and synthesis of the remaining strand occur. In contrast, when mRNAs are short, they are rarely truncated (Figure 3.5b). The 5' cap of an mRNA protects it from digestion, giving L1-ORF2p a good chance to complete reverse transcription. Subsequently, L1-ORF2p moves to a genomic 3' overhang via template jumping (Bibillo et al. 2004). After the genomic 3' overhang region is reverse transcribed, removal of the template RNA and synthesis of the remaining strand occur to generate a full-length pseudogene. Although the role of the 5' cap structure in retrotransposition has not been studied, it has been strongly suggested that LINE-1 RNAs also have the 5' cap structure because of the frequent guanines at the 5'-end of full-length LINE-1 elements (Lavie et al. 2004). Our hypothesis (Figure 3.5) can explain the bimodal length distribution of LINE-1 elements, which has been reported by many researchers (Boissinot et al. 2000; Pavlícek et al. 2002; Myers et al. 2002; Salem et al. 2003;, Babushok et al. 2006), and which cannot be explained by the TPRT mechanism alone.

If our hypothesis is true, how are RNAs truncated in a length-dependent manner? We infer that each nucleotide in all RNAs is cleaved with roughly equal probability, and thereby long mRNAs are more likely to be truncated. Assuming that the cleavage of each nucleotide is a rare event and occurs with the same probability, λ, the number of

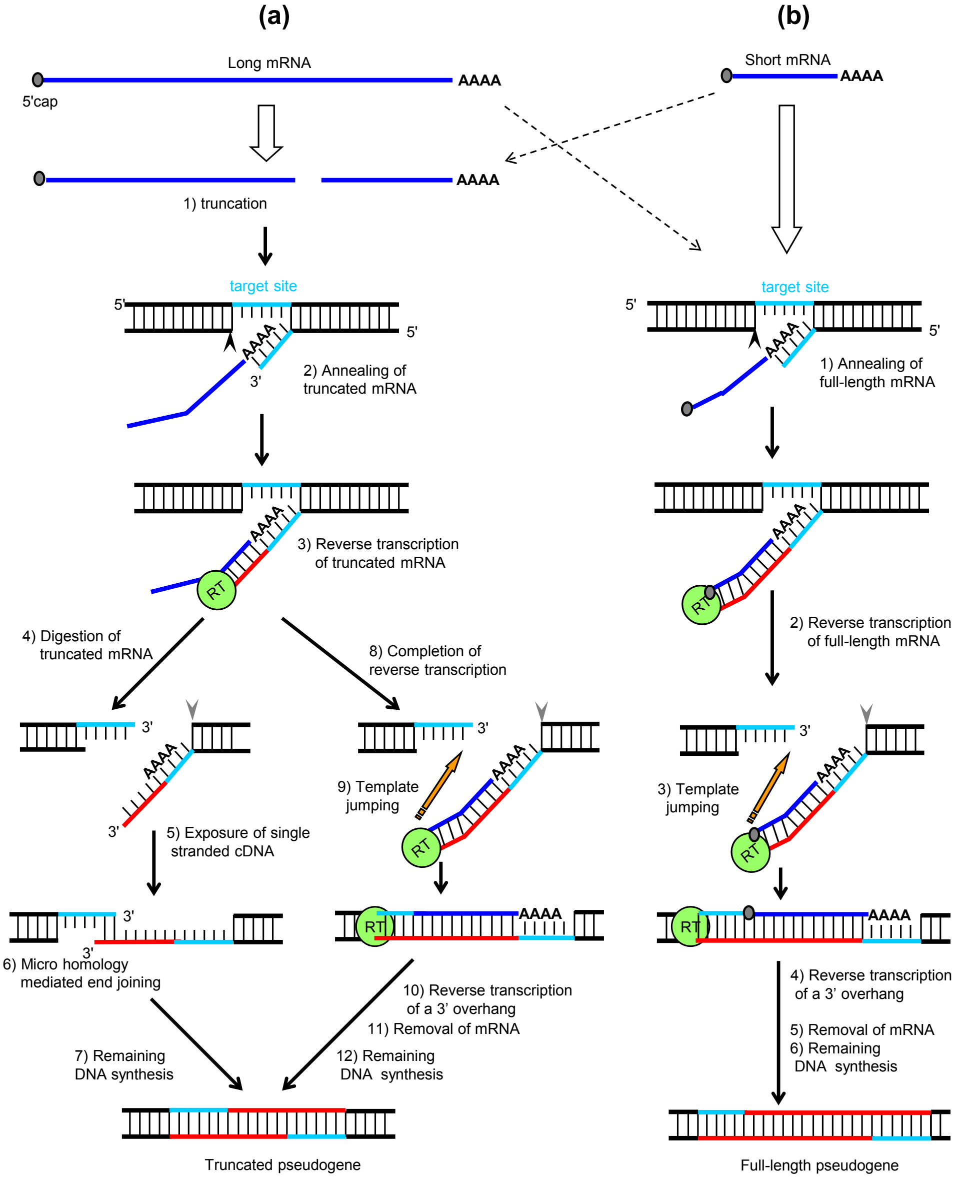

(a)
(b)
Long mRNA
AAAA
5'cap
Short mRNA
AAAA
1) truncation
target site
5'
5'
3'
2) Annealing of truncated mRNA
1) Annealing of full-length mRNA
3) Reverse transcription of truncated mRNA
RT
2) Reverse transcription of full-length mRNA
4) Digestion of truncated mRNA
8) Completion of reverse transcription
9) Template jumping
3) Template jumping
5) Exposure of single stranded cDNA
6) Micro homology mediated end joining
10) Reverse transcription of a 3' overhang
11) Removal of mRNA
4) Reverse transcription of a 3' overhang
5) Removal of mRNA
6) Remaining DNA synthesis
7) Remaining DNA synthesis
12) Remaining DNA synthesis
Truncated pseudogene
Full-length pseudogene

Figure 3.5 Hypothetical mechanism for the generation of short pseudogenes. (a) Generation of short truncated pseudogenes from long mRNAs. 1) Most long mRNAs are truncated before they are reverse transcribed. 2) After first strand cleavage (indicated by a black arrowhead), truncated mRNAs are annealed at the nick. 3) Reverse transcription of truncated mRNAs by L1-ORF2p proceeds. 4) Truncated mRNAs are digested before the completion of reverse transcription. 5) After second-strand cleavage (indicated by a gray arrowhead), a single-stranded cDNA is exposed, and 6) it base-pairs with a genomic 3' overhang (microhomology-mediated end joining) (Zingler et al. 2005). Finally, 7) the remaining DNA synthesis is completed. 8) If reverse transcription is completed before digestion of the template RNA, 9) the L1-ORF2p jumps from the template mRNA onto a genomic 3' overhang. 10) After the genomic 3' overhang region is reverse transcribed, 11) removal of the template RNA and 12) synthesis of the remaining strand occur. (b) Generation of full-length pseudogenes from short mRNAs. 1) Full-length mRNAs are annealed at the nick. 2) Reverse transcription of full-length mRNAs by L1-ORF2p proceeds. 3) After reverse transcription is completed, the L1-ORF2p jumps from the template mRNA onto a genomic 3' overhang. 4) After reverse transcription of the genomic 3' overhang is completed, 5) the template mRNA is removed, and 6) the remaining DNA synthesis is completed.

cleaved nucleotides in each RNA molecule should follow a Poisson distribution. The probability that there is no cleaved nucleotide in a given RNA is expressed as follows:

$$P^{Full-length}(L) = e^{-\lambda L},$$

where L is the length of the RNA. By taking logarithms on both sides, we obtain the following simple equation:

$$Y = -\lambda L,$$

where Y is $\log_e[\mathrm{P}^{\mathrm{full\text{-}length}}(L)]$. The fractions of full-length pseudogenes we found are well fitted by the above expression ($P = 1.95 \cdot 10^{-5}$ by F-test; Figure 3.6), supporting our inference of equiprobable nucleotide cleavage in the RNAs being retrotransposed.

3.4 Conclusions

In this study, we developed a novel method for detecting pseudogenes and found that the human genome contains many previously unidentified short pseudogenes generated by retrotransposition of mRNAs, which gives more complete view of pseudogenes in the human genome. By utilizing our findings, we performed comprehensive analyses of pseudogenes and their parent mRNAs, which presented interesting propensities: short pseudogenes are more likely sourced from long mRNAs than short mRNAs. Importantly, this length dependent phenomenon cannot be explained by the currently accepted mechanism of retrotransposition alone. Therefore, in order to explain this phenomenon, we propose a novel mechanism in which two different in vivo processes, previously reported to be associated with retrotransposition, are involved in the generation of pseudogenes. The findings we presented here provide important insights into the mechanism of retrotransposition.

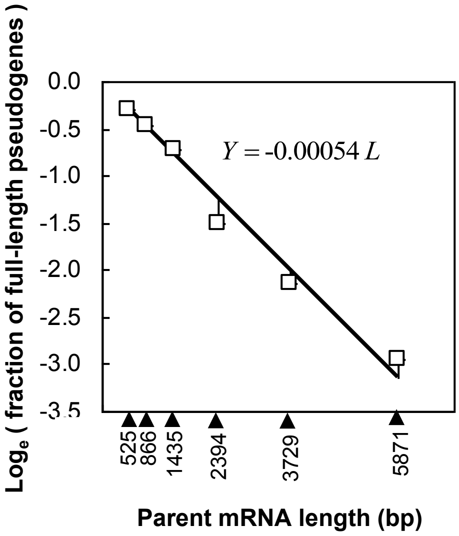


Figure 3.6 Relationship between the fraction of full-length pseudogenes and parent mRNA length. White boxes represent natural logarithms of the fractions of full-length pseudogenes. Pseudogenes were considered to be full-length if they were longer than 90% of the parent mRNA length. Black arrowheads at the bottom of the figure indicate mean lengths of mRNAs in each mRNA length category. The regression line was obtained by the least squares method.

Chapter 4

Conclusions and future directions

Conclusions

In the research presented in this thesis, we developed two methods for predicting specific types of genomic elements. The first one, miRRim2, detects conserved miRNA genes. It was shown that, in humans, the genome-wide prediction result obtained by our method was more accurate than other existing prediction methods. Moreover, by applying miRRim2 to the less well-characterized species *Ciona intestinalis*, we found several promising candidates, indicating that this method can also be useful for species for which only a small amount of comparative genomic resources are available.

The second method presented here, TSDscan, detects pseudogenes. By using this method, it was determined that the human genome contains many previously unidentified short pseudogenes. Comprehensive analyses of the identified pseudogenes and their parent genes revealed that pseudogene length depends on the length of the parent gene, where long genes generate more short pseudogenes than do short genes. This observation led to a new hypothesis: Most long gene transcripts are truncated before they are reverse-transcribed. Truncated gene transcripts would be degraded rapidly during reverse transcription, resulting in the generation of short pseudogenes.

Our prediction results provide a more accurate and comprehensive view of these two types of genomic elements and contribute to a better understanding of the genome.

Future directions

The recent innovation of deep-sequencing technology prompted the discovery of novel miRNA genes in many eukaryotes. While extremely useful, this technology often produces noisy data. Therefore, the task of identifying miRNAs from deep-sequencing data is not straightforward and computational methods that can detect miRNAs in deep-sequencing data need to be developed. It is important to modify miRRim2 in this direction. By using the mapping results for short reads as an additional feature, miRRim2 can naturally integrate deep-sequencing data. This modification will make miRRim2 more valuable for the study of miRNA genes.

At present, miRRim2 is not provided as a downloadable or web-based program. For miRRim2 to be used widely, it should be provided as easy-to-use software. We are planning to develop a software pipeline, in which genomic sequences and deep-sequencing data are used as input and the prediction results of miRRim2 are generated as output.

In Chapter 3, we used TSDscan to detect pseudogenes derived from known protein-coding genes. It should be noted that TSDscan can detect pseudogenes even when their parent genes are not known. For example, suppose a particular genomic region G1 is similar to another region G2. If G2 has sequence signatures of pseudogenes, G2 may be a pseudogene of G1. In this case, G1 is not necessarily a known gene. It may be possible to find a novel gene based on the presence of its pseudogenes because the presence of pseudogenes is evidence that their parent gene was surely transcribed. We are now exploring the feasibility of this pseudogene-driven approach to identify novel genes. The target of this approach does not have to be restricted to protein-coding genes because it is known that many ncRNA genes have pseudogenes.

References

Agarwal S, Vaz C, Bhattacharya A, Srinivasan A (2010) Prediction of novel precursor miRNAs using a context-sensitive hidden Markov model (CSHMM). *BMC Bioinformatics* (Suppl 1), S29.

Ambros V. (2001) microRNAs: tiny regulators with great potential. *Cell* **107**, 823–826.

Ambros V, Bartel B, Bartel DP, Burge CB, Carrington JC, Chen X, Dreyfuss G, Eddy SR, Griffiths-Jones S, Marshall M et al. A uniform system for microRNA annotation. *RNA* **9**, 277-279.

Babushok DV, Ostertag EM, Courtney CE, Choi JM, Kazazian HH Jr (2006) L1 integration in a transgenic mouse model. *Genome Res.* **16**, 240-250.

Babushok DV, Ostertag EM, Kazazian HH Jr (2007) Current topics in genome evolution: molecular mechanisms of new gene formation. *Cell Mol. Life Sci.*, **64**, 542-554.

Batzer MA, Deininger PL (2002) Alu repeats and human genomic diversity. *Nat Rev Genet.* **3**, 370-279.

Baum LE, Egon,J.A. (1967) An inequality with applications to statistical estimation for probabilistic functions of a Markov process and to a model for ecology. *Bull. Am. Meteorol. Soc.* **73**, 360-363.

Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573-80.

Berezikov E, Guryev V, van de Belt J, Wienholds E, Plasterk RH, Cuppen E (2005) Phylogenetic shadowing and computational identification of human microRNA genes. *Cell* **120**, 21-24.

Berezikov E, Chung WJ, Willis J, Cuppen E, Lai EC (2007) Mammalian mirtron genes. *Mol. Cell.* **28**, 328-336.

Bibillo A, Eickbush TH (2004) End-to-end template jumping by the reverse transcriptase encoded by the R2 retrotransposon. *J. Biol. Chem.* **279**, 14945-1453.

Boissinot S, Chevret P, Furano AV (2000) L1 (LINE-1) retrotransposon evolution and amplification in recent human history. *Mol. Biol. Evol.*, **17**, 915-928.

Bourque G, Leong B, Vega VB, Chen X, Lee YL, Srinivasan KG, Chew JL, Ruan Y, Wei CL, Ng HH et al. (2008) Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res.* **18**, 1752-1762.

Brameier M, Wiuf C (2007) Ab initio identification of human microRNAs based on structure motifs. *BMC Bioinformatics* **8**, 478.

Buzdin A, Gogvadze E, Kovalskaya E, Volchkov P, Ustyugova S, Illarionova A, Fushan A, Vinogradova T, Sverdlov E (2003) The human genome contains many types of chimeric retrogenes generated through in vivo RNA recombination. *Nucleic Acids Res.* **31**, 4385-4390.

Chiaromonte F, Yap VB, Miller W (2002) Scoring pairwise genomic sequence alignments. *Pac. Symp. Biocomput.* 115-126.

Claverie JM. (2001) Gene number. What if there are only 30,000 human genes? *Science*, **291**, 1255-1257.

Cost GJ, Feng Q, Jacquier A, Boeke JD (2002) Human L1 element target-primed reverse transcription in vitro. *EMBO J.* **21**, 5899-5910.

Cummins JM, He Y, Leary RJ, Pagliarini R, Diaz LA Jr, Sjoblom T, Barad O, Bentwich Z, Szafranska AE, Labourier E, et al. (2006). The colorectal microRNAome. *Proc Natl Acad Sci U S A* **103**, 3687-3692.

DeCaprio D, Vinson JP, Pearson MD, Montgomery P, Doherty M, Galagan JE (2007) Conrad: gene prediction using conditional random fields. *Genome Res.* **17**, 1389-1398.

Deininger PL, Moran JV, Batzer MA, Kazazian HH Jr. (2003) Mobile elements and mammalian genome evolution. *Curr Opin Genet Dev.* **13**. 651-658.

Dewannieux M, Esnault C, Heidmann T (2003) LINE-mediated retrotransposition of marked *Alu* sequences. *Nat. Genet.* **35**, 41-48

Do CB, Woods DA, Batzoglou S (2006) CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics* **22**, e90–98.

Do CB, Gross SS, Batzoglou S (2006) CONTRAlign: Discriminative Training for Protein Sequence Alignment. In Proceedings of the Tenth Annual International Conference on Computational Molecular Biology (RECOMB 2006) 160–174.

Durbin R, Eddy S, Krogh A, Mitchison G (1998) Pairwise alignment. In Biological sequence analysis: Probabilistic models of proteins and nucleic acids. Cambridge University Press pp.12-45.

Elisaphenko EA, Kolesnikov NN, Shevchenko AI, Rogozin IB, Nesterova TB, Brockdorff N, Zakian SM. (2008) A dual origin of the Xist gene from a protein-coding gene and a set of transposable elements. PLoS One **3**, e2521.

Esnault C., Maestre J, Heidmann T (2000) Human LINE retrotransposons generate processed pseudogenes. *Nat. Genet.* **24**, 363-367.

Esquela-Kerscher A, Slack FJ (2006) Oncomirs - microRNAs with a role in cancer. *Nat. Rev. Cancer* **6**, 259-269.

Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, Irvine KM, Schroder K, Cloonan N, Steptoe AL, Lassmann T et al (2009) The regulated retrotransposon transcriptome of mammalian cells. *Nat. Genet.*, **41**, 563-571

Feng Q, Moran JV, Kazazian HH Jr, Boeke JD (1996) Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell*, **87**, 905-916.

Friedman RC, Farh KK, Burge CB, Bartel DP (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.* **19**, 92–105.

Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, Cline MS, Goldman M, Barber GP, Clawson H, Coelho A et al. (2010) The UCSC genome browser database: update 2011. *Nucleic Acids Res.* **18**, D876-882

Gkirtzou K, Tsamardinos I, Tsakalides P, Poirazi P (2010) MatureBayes: a probabilistic algorithm for identifying the mature miRNA within novel precursors. *PLoS One* **5**, e11843.

Gregory RI, Chendrimada TP, Cooch N, Shiekhattar R (2005) Human RISC couples microRNA biogenesis and posttranscriptional gene silencing. *Cell* **123**, 631-640.

Han J, Lee Y, Yeom KH, Nam JW, Heo I, Rhee JK, Sohn SY, Cho Y, Zhang BT, Kim VN (2006) Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. *Cell* **125**, 887-901.

Helvik SA, Snove OJ, Saetrom P. (2007) Reliable prediction of Drosha processing sites improves microRNA gene prediction. *Bioinformatics* **23**, 142–149.

Hendrix,D., Levine,M. and Shi,W. (2010) miRTRAP, a computational method for the systematic identification of miRNAs from high throughput sequencing data. *Genome Biol.* **11**, R39.

Hertel J, Stadler PF (2006) Hairpins in a haystack: recognizing microRNA precursors in comparative genomics data. *Bioinformatics* **22**, e197-e202.

Hutvagner G, McLachlan J, Pasquinelli AE, Balint E, Tuschl T, Zamore PD (2001) A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science* **293**, 834-838.

Jiang P, Wu H, Wang W, Ma W, Sun X, Liu Z, (2007) MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Res.* **35**, W339–344.

Jurka,J (1997) Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 1872-1877.

Khvorova A, Reynolds A, Jayasena SD (2003) Functional siRNAs and miRNAs exhibit strand bias. *Cell* **115**, 209-216.

Kiryu H, Kin T, Asai K (2008) Rfold: an exact algorithm for computing local base pairing probabilities. *Bioinformatics* **24**, 367-373.

Kozomara A, Griffiths-Jones S (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.* **39**, D152-157.

Krol J, Sobczak K, Wilczynska U, Drath M, Jasinska A, Kaczynska D, Krzyzosiak WJ (2004) Structural features of microRNA (miRNA) precursors and their relevance to miRNA biogenesis and small interfering RNA/short hairpin RNA design. *J. Biol. Chem.* **279**, 42230-42239.

Kuhn RM, Karolchik D, Zweig AS, Wang T, Smith KE, Rosenbloom KR, Rhead B, Raney BJ, Pohl A, Pheasant M, et al. (2009) The UCSC Genome Browser Database: update 2009. *Nucleic Acids Res.* **37**, D755-D761.

Lafferty J, McCallum A, Pereira F (2001) Conditional random fields: probabilistic models for segmenting and labeling sequence data. *In Proc. 18th Int. Conf. on Machine Learning, William College, MA* pp. 282-289

Lagos-Quintana M, Rauhut R, Yalcin A, Meyer J, Lendeckel W, Tuschl T. (2002) Identification of tissue-specific microRNAs from mouse. *Curr Biol.* **12**, 735-739.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. (2001) Initial Sequencing and Analysis of the Human Genome. *Nature* **409**, 860-921.

Landgraf P, Rusu M, Sheridan R, Sewer A, Iovino N, Aravin A, Pfeffer S, Rice A, Kamphorst AO, Landthaler M, et al. (2007) A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell* **129**, 1401-1414.

Lavie L, Maldener E, Brouha B, Meese EU, Mayer J (2004) The human L1 promoter: variable transcription initiation sites and a major impact of upstream flanking sequence on promoter activity. *Genome Res.* **14**, 2253-2260.

Lee Y, Ahn C, Han J, Choi H, Kim J, Yim J, Lee J, Provost P, Radmark O, Kim S et al. (2003) The nuclear RNase III Drosha initiates microRNA processing. *Nature* **425**, 415-419.

Lewis BP, Burge CB, Bartel DP (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**, 15–20.

Liang H, Li WH (2009) Lowly expressed human microRNA genes evolve rapidly. *Mol. Biol. Evol.* **26**, 1195-1198.

Lim LP, Lau NC, Weinstein EG, Abdelhakim A, Yekta S, Rhoades MW, Burge CB, Bartel DP (2003a) The microRNAs of Caenorhabditis elegans. *Genes Dev.* **17**, 991-1008.

Lim LP, Glasner ME, Yekta S, Burge CB, Bartel DP. (2003b) Vertebrate microRNA genes. *Science*. **299**, 1540.

Lindsay MA (2008) microRNAs and the immune response. *Trends Immunol.* **29**, 343-351.

Liu DC, Nocedal J (1989) On the limited memory method for large scale optimization. *Math. Programming* **45**, 503-528.

Liu X, He S, Skogerbø G, Gong F, Chen R (2012) Integrated Sequence-Structure Motifs Suffice to Identify microRNA Precursors. *PLoS One* **7**, e32797.

Lu J, Shen Y, Wu Q, Kumar S, He B, Shi S, Carthew RW, Wang SM, Wu CI (2008) The birth and death of microRNA genes in Drosophila. *Nat. Genet.* **40**: 351–355.

Luan DD, Korman MH, Jakubczak JL and Eickbush TH. (1993) Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell*, **72**, 595-605

Lucier JF, Perreault J, Noël JF, Boire G, Perreault JP (2007) RTAnalyzer: a web application for finding new retrotransposons and detecting L1 retrotransposition signatures. *Nucleic Acids Res.*, **35**, W269-274.

Lund E, Güttinger S, Calado A, Dahlberg JE, Kutay U (2004) Nuclear export of microRNA precursors. *Science.* **303,** 95-98.

Lyon MF (1961) Gene action in the X-chromosome of the mouse (Mus musculus L.). *Nature* **190**, 372-373.

McCaskill JS (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* **29**, 1105-1119.

Mercer TR, Dinger ME, Mattick JS (2009) Long non-coding RNAs: insights into functions. *Nat. Rev. Genet.* **10**, 155-159.

Mituyama T, Yamada K, Hattori E, Okida H, Ono Y, Terai G, Yoshizawa A, Komori T, Asai K (2009) The Functional RNA Database 3.0: databases to support mining and annotation of functional RNAs. *Nucleic Acids Res.* **37**, D89-92.

Moran JV, Holmes SE, Naas TP, DeBerardinis RJ, Boeke JD, Kazazian HH Jr (1996) High frequency retrotransposition in cultured mammalian cells. *Cell*, **87**, 917-927.

Moran JV, DeBerardinis RJ, Kazazian HH Jr (1999) Exon shuffling by L1 retrotransposition *Science* **283**, 1530-1534.

Myers JS, Vincent BJ, Udall H, Watkins WS, Morrish TA, Kilroy GE, Swergold GD, Henke J, Henke L, Moran JV et al. (2002) A comprehensive analysis of recently integrated human Ta L1 elements. *Am. J. Hum. Genet.* **71**, 312-326.

Nam JW, Shin KR, Han J, Lee Y, Kim VN et al. (2005) Human microRNA prediction through a probabilistic co-learning model of sequence and structure. *Nucleic Acids Research* **33**, 3570-3581.

Newbury SF (2006) Control of mRNA stability in eukaryotes. *Biochem. Soc. Trans.* **34**, 30-34.

Norden-Krichmar TM, Holtz J, Pasquinelli AE, Gaasterland T (2007) Computational prediction and experimental validation of Ciona intestinalis microRNA genes. *BMC Genomics* **8**, 445.

Ohshima K, Hattori M, Yada T, Gojobori T, Sakaki Y, Okada N (2003) Whole-genome screening indicates a possible burst of formation of processed pseudogenes and *Alu* repeats by particular L1 subfamilies in ancestral primates. *Genome Biol.* **4**, R74.

Ostertag EM, Kazazian HH Jr (2001) Biology of mammalian L1 retrotransposons. *Annu. Rev. Genet.* **35**, 501-538.

Oulas A, Boutla A, Gkirtzou K, Reczko M, Kalantidis K, Poirazi P (2009) Prediction of novel microRNA genes in cancer-associated genomic regions--a combined computational and experimental approach. *Nucleic Acids Res.* **37**, 3276-3287.

Pavlícek A, Paces J, Zíka R, Hejnar J (2002) Length distribution of long interspersed nucleotide elements (LINEs) and processed pseudogenes of human endogenous retroviruses: implications for retrotransposition and pseudogene detection. *Gene* **300**,189-194.

Pei B, Sisu C, Frankish A, Howald C, Habegger L, Mu XJ, Harte R, Balasubramanian S, Tanzer A, Diekhans M et al. (2012) The GENCODE pseudogene resource. *Genome Biol.* **13**, R51.

Perreault J, Noël JF, Brière F, Cousineau B, Lucier JF, Perreault JP, Boire G (2005) Retropseudogenes derived from the human Ro/SS-A autoantigen-associated hY RNAs. *Nucleic Acids Res.*, **33**, 2032-2041.

Prasanth KV, Spector DL (2007) Eukaryotic regulatory RNAs: an answer to the 'genome complexity' conundrum. *Genes Dev.* **21**, 11-42.

Ponting CP, Oliver PL, Reik W (2009) Evolution and functions of long noncoding RNAs. *Cell* **136**, 629-641.

Ruby JG, Stark A, Johnston WK, Kellis M, Bartel DP, Lai EC (2007) Evolution, biogenesis, expression, and target predictions of a substantially expanded set of Drosophila microRNAs. *Genome Res.* **17**, 1850-1864.

Saetrom P, Snove O, Nedland M, Grunfeld TB, Lin Y, Bass MB, Canon JR (2006) Conserved microRNA characteristics in mammals. *Oligonucleotides* **16**, 115-144.

Sato K, Sakakibara Y (2005) RNA secondary structural alignment with conditional random fields. *Bioinformatics* (Suppl 2), ii237–242.

Salem AH, Myers JS, Otieno AC, Watkins WS, Jorde LB, Batzer MA (2003) LINE-1 preTa elements in the human genome. *J. Mol. Biol.* **326**, 1127-1146.

Sayah DM, Sokolskaja E, Berthoux L, Luban J. (2004) Cyclophilin A retrotransposition into TRIM5 explains owl monkey resistance to HIV-1. *Nature*. **430**, 569-573.

Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison R, Haussler D, Miller W (2003) Human-Mouse Alignments with BLASTZ. *Genome Res.* **13**, 103-107.

Sheng Y, Engstrom PG, Lenhard B (2007) Mammalian MicroRNA Prediction through a Support Vector Machine Model of Sequence and Structure. PloS ONE **2**, e946.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034-1050.

Siepel A, Pollard KS, Haussler D (2006) New methods for detecting lineage-specific selection. *Proc. 10th Int. Conf. on Research in Computational Molecular Biology (RECOMB 2006)* pp. 190-205.

Szak ST, Pickeral OK, Makalowski W, Boguski MS, Landsman D and Boeke JD (2002) Molecular archeology of L1 insertions in the human genome. *Genome Biol.* **3**, research0052.

Terai G, Komori T, Asai K, Kin T (2007) miRRim: a novel system to find conserved miRNAs with high sensitivity and specificity. *RNA* **13**, 2081-2090.

Torrents D, Suyama M, Zdobnov E, Bork P (2003) A genome-wide survey of human pseudogenes. *Genome Res.* **13**, 2559-2567

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA et al. (2001) The Sequence of the Human Genome. *Science* **291**, 1304-1351.

Watanabe T, Totoki Y, Toyoda A, Kaneda M, Kuramochi-Miyagawa S, Obata Y, Chiba H, Kohara Y, Kono T, Nakano T et al. (2008). Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature* **453**, 539-453.

Wei W, Gilbert N, Ooi SL, Lawler JF, Ostertag EM, Kazazian HH, Boeke JD, Moran JV (2001) Human L1 retrotransposition: cis preference versus trans complementation. *Mol. Cell. Biol.* **21**, 1429–1439.

Wienholds E, Plasterk RH (2005) MicroRNA function in animal development. *FEBS Lett.* **579**, 5911-5922.

Yousef M, Nebozhyn M, Shatkay H, Kanterakis S, Showe LC, Showe MK (2006) Combining multi-species genomic data for microRNA identification using a Naive Bayes classifier. *Bioinformatics* **22**, 1325–1334.

Zhang Z, Harrison PM, Liu Y, Gerstein M (2003) Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. *Genome Res.* **13**, 2541-2558.

Zingler N, Willhoeft U, Brose HP, Schoder V, Jahns T, Hanschmann KM, Morrish TA, Löwer J, Schumann GG (2005) Analysis of 5' junctions of human LINE-1 and Alu retrotransposons suggests an alternative model for 5'-end attachment requiring microhomology-mediated end-joining. *Genome Res.* **15**, 780-789.

Supplemental materials

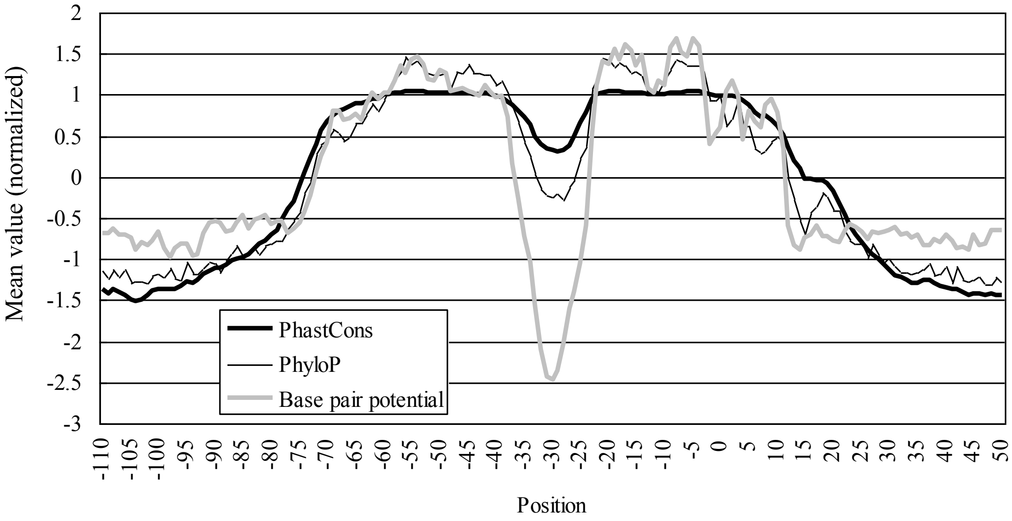


Figure S2.1. PhastCons scores, PhyloP scores, and base-pair potential averaged in each position. Position 0 indicates the 3' ends of miRNA-duplexes in the 3'-arm of miRNA hairpins. The DRB in the 3'-arm is located around position +11, where the base pair potential and PhyloP score sharply decrease.

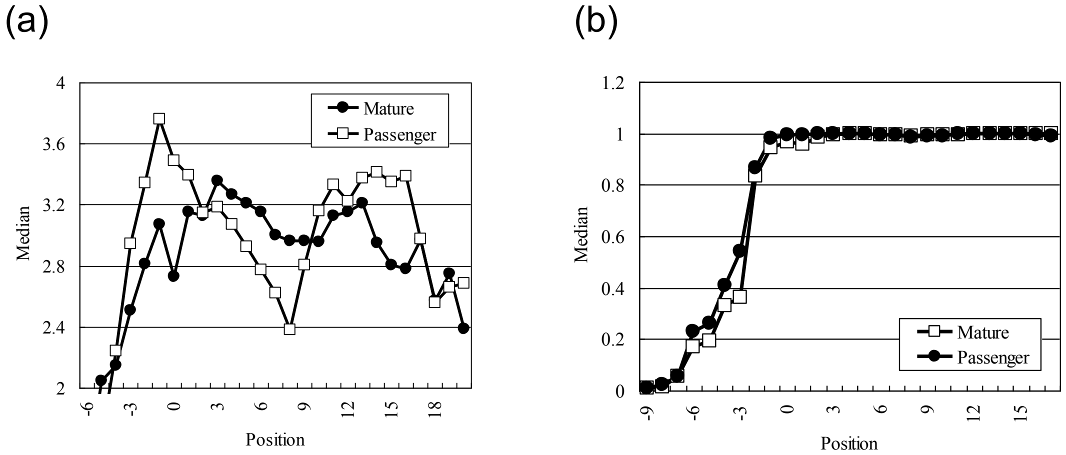


Figure S2.2. Difference between mature miRNA and passenger strand in the 3'-arm of miRNA hairpins. Median values of the (a) PhyloP score and (b) base-pair potential are shown in each position. Position 0 indicates the 5'-ends of mature miRNA or passenger strands.

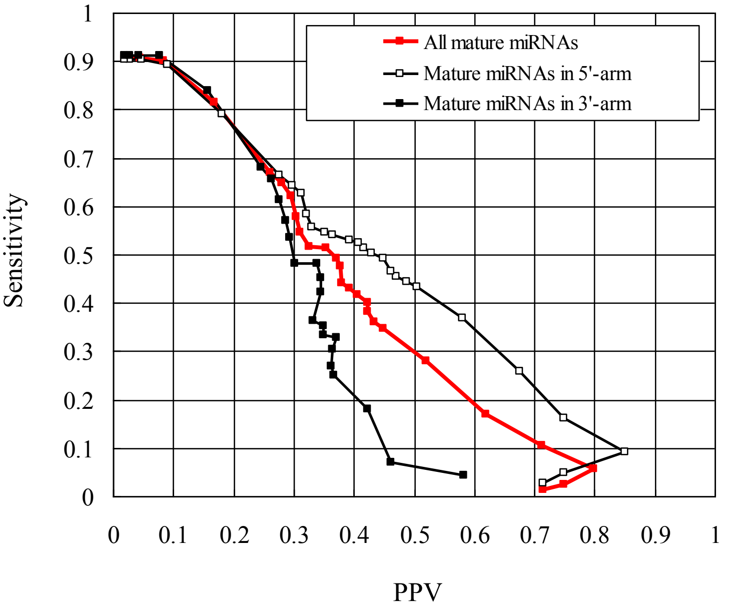


Figure S2.3. Prediction accuracy of the 5'-end of mature miRNAs. The detection accuracy of mature miRNAs in the 5'-arm is higher than in the 3'-arm strand.

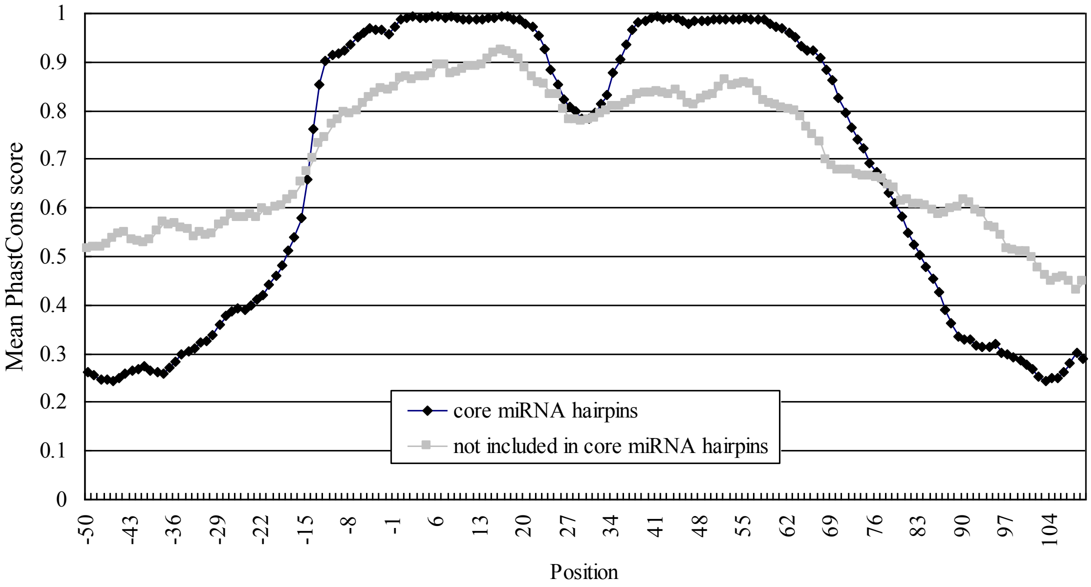


Figure S2.4. The difference of a conservation pattern between the core miRNA hairpin and non-core miRNA hairpin. Position 0 indicates the 5' ends of mature or passenger miRNAs in the 5'-arm of miRNA hairpins.

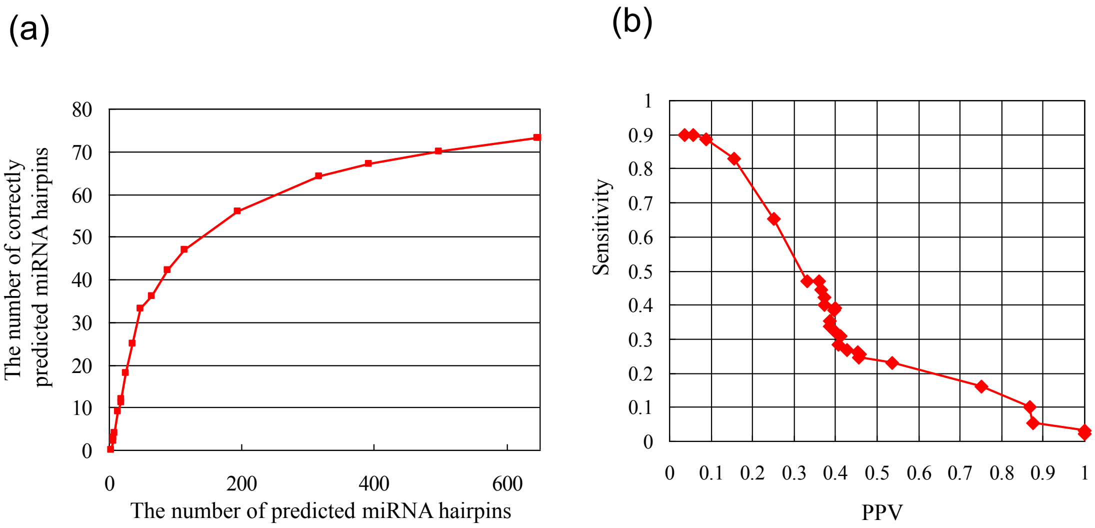


Figure S2.5. The prediction performance for the *Ciona intestinalis* genome. (a) The detection/prediction performance of miRNA hairpins. (b) Sensitivty-PPV plot for mature miRNA prediction.

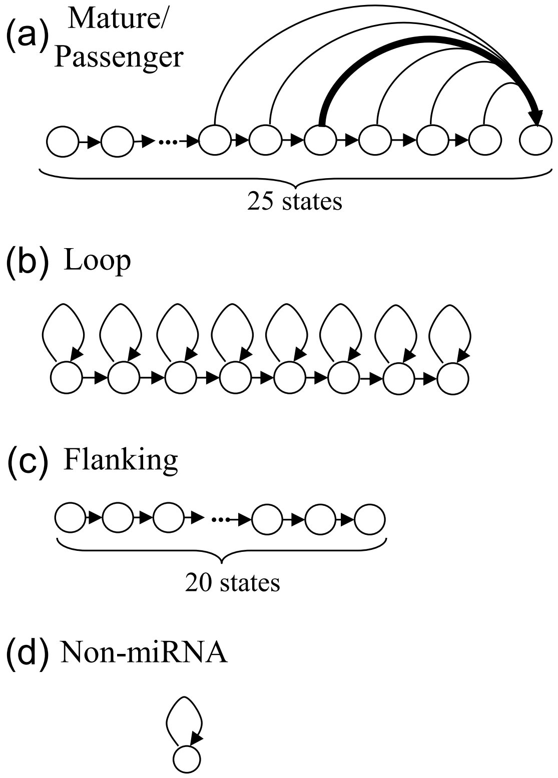


Figure S2.6 Architecture of each sub-model.

Supplemental methods S2.1

The null model for predicting 5'-end of mature miRNAs.

In our null model, all the Us are considered as 5'-end of mature miRNA. Each U has a penalty score, which is designed such that Us in a plausible position have low penalty score. The penalty is defined as $P = |S - l|$, where l is the length between a pair of Drosha cleavage sites, *p1* and *p2*, and S is a typical length between Drosha cleavage sites. In this study, S = 60 was used.

The *p1* and *p2* were determined based on the position of U. When a given U is located on 5'-arm (Fig. A1(a)), *p1* is the position of the U, and *p2* is deduced from predicted hairpin structures assuming the 2-bp 3'-overhang. When a given U is located on 3'-arm (see, Fig. A1(b), *p1* is a 21-bp downstream position from the position of the U. Then, *p2* is deduced from predicted hairpin structures assuming the 2-bp 3'-overhang.

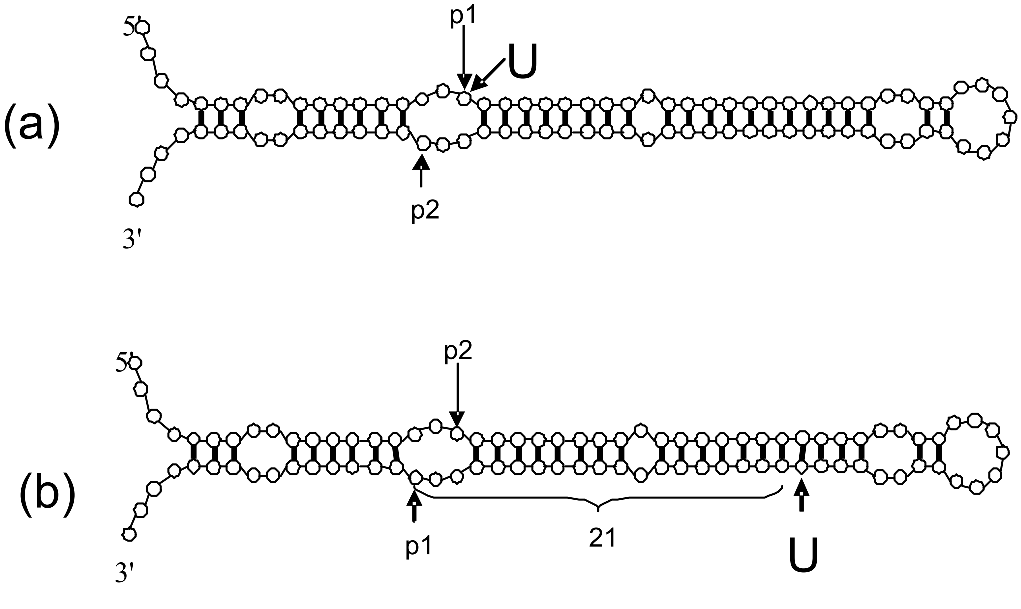
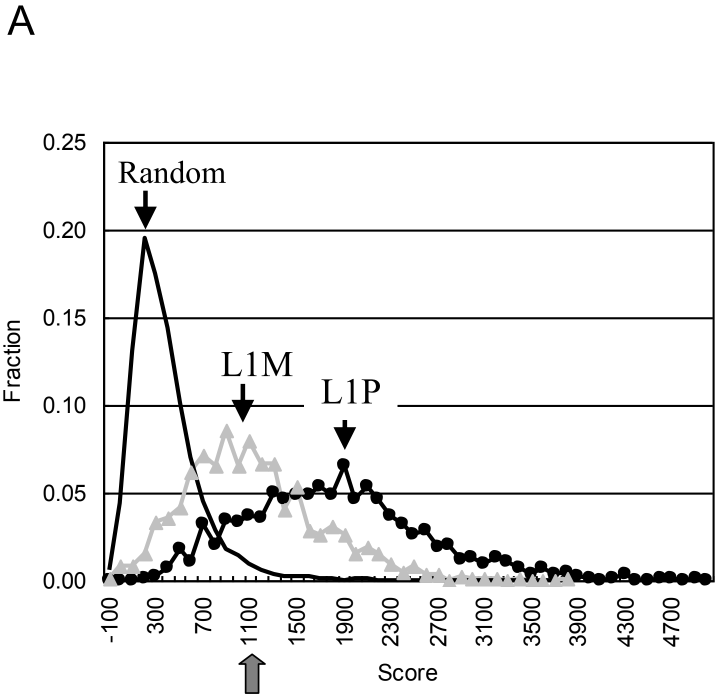


Figure A1 Position of Drosha cleavage sites (p1 and p2)

A given U is located on (a) 5'-arm and (b) 3'-arm. Red arrows indicate the position of the U. Black arrows indicate the position of deduced Drosha cleavage sites (p1 and p2).



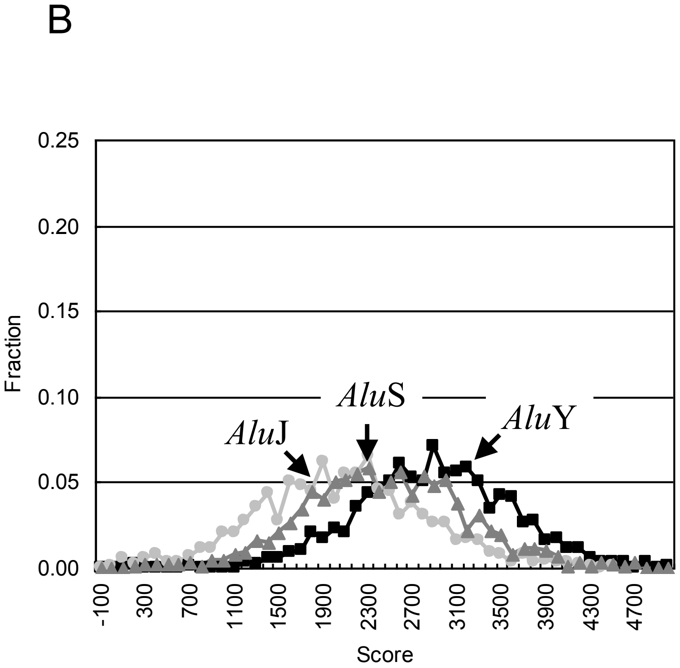


Figure S3.1 Score distributions for LINE-1, *Alu*, and randomly selected regions. (A) The TSDscan scores of 2 forms of LINE-1, mammalian-wide (L1M) and primate-specific (L1P), were compared with randomly selected regions (Random). The gray arrow indicates the score threshold used to identify pseudogenes derived from mRNA (see, text). (B) The TSDscan scores of 3 forms of *Alu* (*Alu*J, *Alu*S, and *Alu*Y) were compared.

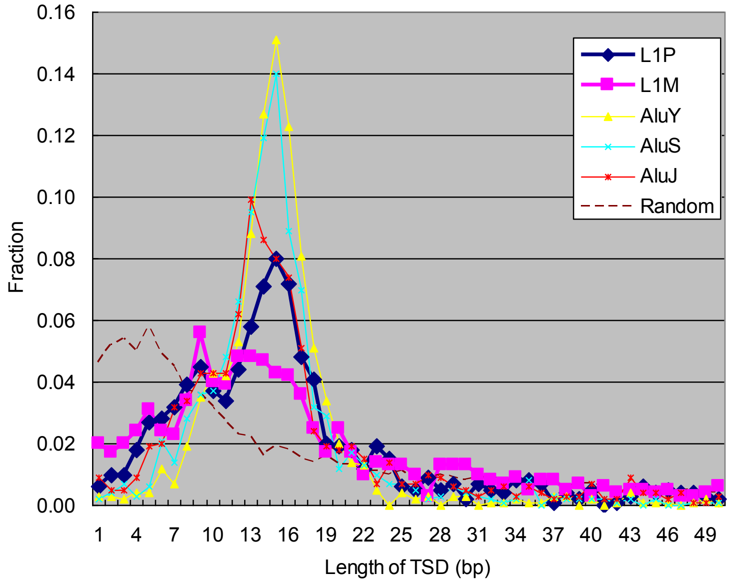


Figure S3.2 Length distribution of TSD. L1P: primate-specific LINE-1, L1M: mammalian-wide LINE-1, Random: randomly selected regions.

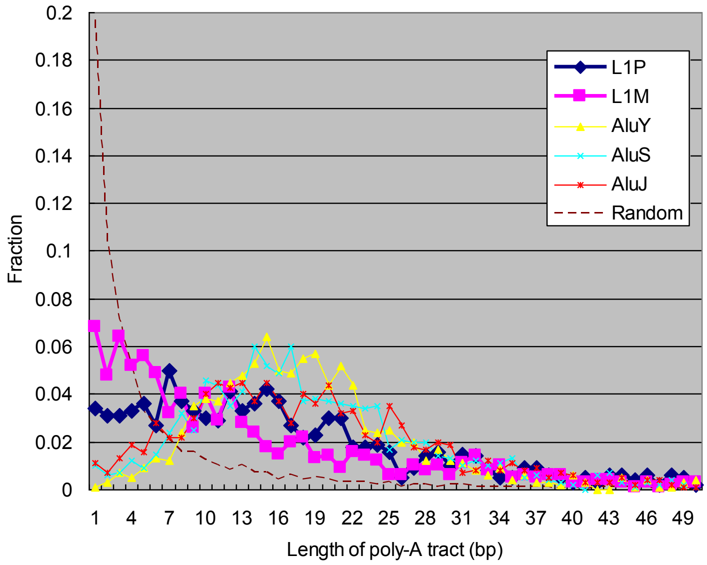


Figure S3.3 Length distribution of poly-A tract. L1P: primate-specific LINE-1, L1M: mammalian-wide LINE-1, Random: randomly selected regions.

Supplemental methods S3.1

The reason of using the scores of +200/-200 as default parameters for alignment of the TTAAAA sequence.

The default scoring parameters for the TTAAAA sequence is as follows: For the first two positions, nucleotide T and other nucleotides (A, G, and C) have scores of +200 and -200, respectively. In the last four positions, an A-A nucleotide match and the other aligned nucleotide pairs have scores of +200 and -200, respectively.

Because chromosomal target sites are only 6-bp, they easily become undetectable if only a few mutations are accumulated. To avoid this, we enlarged scoring parameters used for alignment of target sites. This is why the positive/negative scores for the TTAAAA sequence have +200/-200, which is about twice as the positive/negative scores for poly-A tail and TSDs.

By using the default parameters, TSDscan can detect the TTAAAA sequence in about 82% and 66% of AluY and L1P insertions, respectively, which seems consistent with previously reported (Jurka 1997; Babushok et al., 2006). If we make these scores to be one half (that is, T and non-T nucleotides in the first two positions have scores of +100 and -100, respectively, and an A-A nucleotide match and the other aligned nucleotide pairs in the last four positions to be +100 and -100, respectively), TSDscan can detect the target site only in 53% and 24% of AluY and L1P insertions, respectively, which seems smaller than the previous reports.

Supplemental methods S3.2

The TSDscan algorithm

The TSDscan algorithm is designed by adopting a pattern recognition algorithm with hidden Markov model to detect pseudogenes. For explanation, upstream and downstream sequences of a pseudogene is denoted by *x* and *y*, respectively. Because TSDs in *x* and *y* (hearafter xTSD and yTSD) are similar, they can be detected by performing a local alignment between *x* and *y*. To detect poly-A tract, we assigned positive scores to insertion of A-nucleotides immediately before yTSD. Figure T1A is an automaton for detecting both poly-A tract and TSD. Sequences corresponding TSD and poly-A is emitted from the M and A_y state, respectively. If we delete the A_y state from this automaton, the resultant automaton is the same as an automaton for simple local alignment.

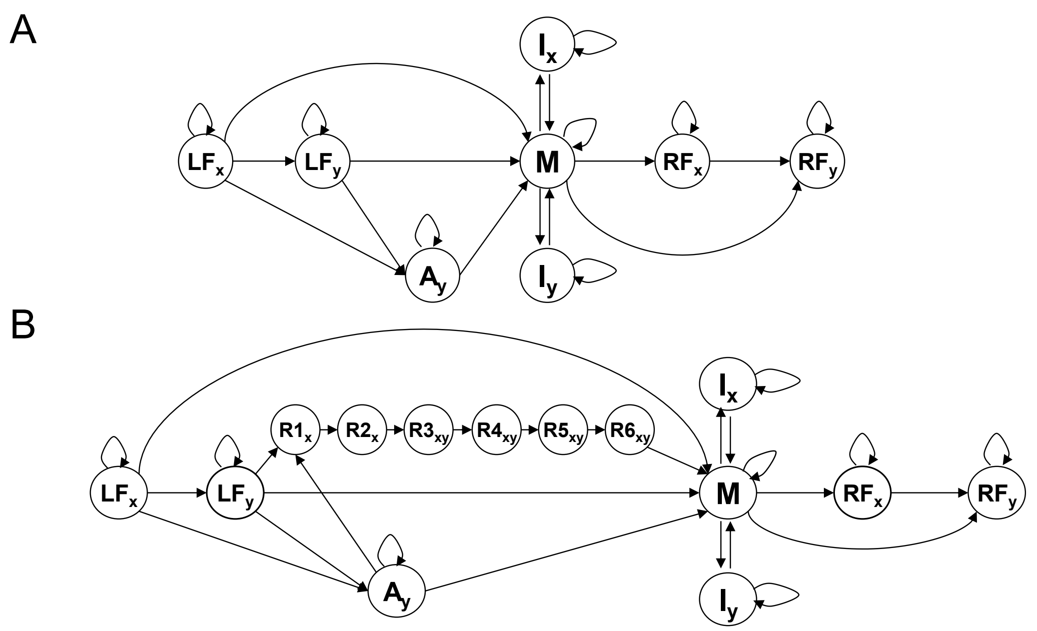


Figure T1 Automatons for alignment between upstream and downstream sequences of retrocopies. Upstream and downstream sequence is denoted as *x* and *y*, respectively. State M emits two letters to be aligned, which correspond to TSD. State I_x (I_y) emits a letter in sequence *x* (*y*) that is aligned to a gap. States LF_x and LF_y (RF_x and RF_y) emit unaligned flanking subsequences on the left (right) of the alignment. (a) an automaton that consider Poly-A tract. State A_y emits a letter in poly-A tract in sequence *y* that is aligned to a gap. (b) an automaton used in TSDscan. $R1_x$ and $R2_x$ emit the first two bases of the TTAAAA in sequence *x* that are aligned to gaps. $R3_{xy}$–$R6_{xy}$ emit two aligned letters one for each sequence, which correspond to the last four bases in the TTAAAA. For simplicity, the start and end state are omitted from the figure.

Recursion equations detecting the optimal alignment using the automaton in Figure T1A is as follows:

Initialization:

$$M(0,j)=M(i,0)=\mathbf{A_y(0,j)=A_y(i,0)=}0 \text{ (for all i and j)}$$

$$I_x(0,j)=I_y(i,0)=-\infty \text{ (for all i and j)}$$

Recursion:

$$M(i,j) = \max \begin{cases} M(i-1,j-1) + s(x_i,y_j) \\ I_x(i-1,j-1) + s(x_i,y_j) \\ I_y(i-1,j-1) + s(x_i,y_j) \\ \mathbf{A_y(i-1,j-1) + s(x_i,y_i)} \\ 0 \end{cases} \qquad \mathbf{A_y(i,j) = max} \begin{cases} \mathbf{A_y(i,j-1) + polyA(y_i)} \\ \mathbf{0} \end{cases}$$

$$I_x(i,j) = \max \begin{cases} M(i-1,j) + g \\ I_x(i-1,j) + e \end{cases} \qquad I_y(i,j) = \max \begin{cases} M(i,j-1) + g \\ I_y(i,j-1) + e \end{cases}$$

, where $s(x_i,y_i)$ is match/mismatch score in alignment of TSD, ‘g’ and ‘e’ are a gap open and extension penalty, respectively, and $polyA(y_i)$ is a score of nucleotide y_i in the A_y state. The $polyA(y_i)$ is positive when y_i = ‘A’. Equations written in bold letters are related to the A_y state. By deleting these equations, we get the same recursion equations used in a simple local alignment (Durbin et al., 1998).

The automaton shown in Figure T1B is made by adding the R1~R6 states, which emits TTAAAA consensus sequence, to the automaton shown in Figure T1A. The first two letters of the consensus, TT, is detected by assigning positive scores to the insertion of TT di-nucleotide immediately before xTSD. This can be done by assigning a positive score to T nucleotide in the $R1_x$ and $R2_x$ state. The AAAA followed by the TT di-nucleotide is detected by assigning a positive score to the A:A match followed by the TT insertion. This can be done by assigning positive scores to the A:A match in the $R3_{xy}$ ~$R6_{xy}$ state.

Next, we consider positions of poly-A tract and TSD. The poly-A tract and TSD are inserted at immediately outside of a pseudogene. Therefore, the poly-A tract and TSD which are distant from a pseudogene should be penalized. We penalize nucleotide insertions between xTSD and a 5’-end of pseudogene, and those between 3’-end of a pseudogene and poly-A tract. This can be done by assigning negative scores in any nucleotide insertion in the LF_y and RF_x state in Figure T1B. Recursion equations detecting the optimal alignment using the automaton in Figure T1B is a bit complicated, which is shown below.

Initialization:

```
for (i = 0; i <= Length(x); i++)
    $M[i][0] = I_x[i][0] = I_y[i][0] = RF_x[i][0] = R1\sim R6_x[i][0] = A_y[i][0] = -\infty$
for (j = 0; j <= Length(y); j++)
    $M[0][j] = I_x[0][j] = I_y[0][j] = RF_x[0][j] = R1\sim R6_x[0][j] = A_y[i][0] = -\infty$
for (i = 0; i <= Length(x); i++){
    for (j = 0; j <= Length(y); j++){
        $LF_y[i][j] = j * b$
    }
}
```

Recursion:

$$M[i][j] = \max \begin{cases} M[i-1][j-1] + s(x_i, y_j) \\ I_x[i-1][j-1] + s(x_i, y_j) \\ I_y[i-1][j-1] + s(x_i, y_j) \\ A_y[i-1][j-1] + s(x_i, y_i) \\ LF_y[i-1][j-1] + s(x_i, y_i) \\ R6_{xy}[i-1][j-1] + s(x_i, y_i) \end{cases}$$

$$A_y[i][j] = \max \begin{cases} A_y[i][j-1] + \mathrm{polyA}(y_i) \\ LF_y[i][j-1] + \mathrm{polyA}(y_i) \end{cases}$$

$$RF_x[i][j] = \max \begin{cases} RF_x[i][j-1] + b \\ M[i][j-1] + b \end{cases}$$

$$I_x[i][j] = \max \begin{cases} M[i-1][j] + g \\ I_x[i-1][j] + e \end{cases}$$

$$I_y[i][j] = \max \begin{cases} M[i][j-1] + g \\ I_y[i][j-1] + e \end{cases}$$

$$R1_x[i][j] = \max \begin{cases} LF_y[i-1][j] + tt(x_i) \\ A_y[i-1][j] + tt(x_i) \end{cases}$$

$$R2_x[i][j] = R1_x[i-1][j] + tt(x_i)$$
$$R3_{xy}[i][j] = R2_x[i-1][j-1] + aaaa(x_i, y_i)$$
$$R4_{xy}[i][j] = R3_{xy}[i-1][j-1] = aaaa(x_i, y_i)$$
$$R5_{xy}[i][j] = R4_{xy}[i-1][j-1] = aaaa(x_i, y_i)$$
$$R6_{xy}[i][j] = R5_{xy}[i-1][j-1] = aaaa(x_i, y_i)$$

, where $s(x_i,y_i)$ is match/mismatch score in state M, 'g' and 'e' are a gap open and extension penalty, respectively, and the polyA(y_i) is a score of y_i-nucleotide in the state A_y. The 'b' is a penalty of any nucleotide in the state LF_y or RF_x. The tt(x_i) is a score of x_i-nucleotide in $R1_x$ and $R2_x$. The aaaa(x_i, y_i) is match/mismatch scores in state $R3_{xy}$~$R6_{xy}$. The scoring parameters used in TSDscan are summarized below:

$e = -30$
$g = -400$
$b = -50$

$$\mathrm{polyA}(y_i) \begin{cases} +100 & \text{if } y_i = A \\ -100 & \text{otherwise} \end{cases}$$

$$\mathrm{tt}(x_i) \begin{cases} +200 & \text{if } x_i = T \\ -200 & \text{otherwise} \end{cases}$$

$$\mathrm{aaaa}(x_i, y_i) \begin{cases} +200 & \text{if } x_i = A,\ y_i = A \\ -200 & \text{otherwise} \end{cases}$$

$s(x_i,y_i)$ is the HOXD matrix (Chiaromonte et al., 2002) shown bellow:

	A	C	G	T
A	91	-114	-31	-123
C	-114	100	-125	-31
G	-31	-125	100	-114
T	-123	-31	-114	91