

INTRODUCTION

1.1 Overview

Determination of the difference in the rates of molecular evolution between human and mouse genome has been a subject of scientific investigations for over 40 years (Kumar 2005 NRG August issue). It began with Laird et al. (1969) and Kohne et al. (1970) who found a greater than 10-fold difference between primates and rodents based on the extent of sequence divergence in the non-repetitive DNA between mouse and rat and between humans and primates. Following the advent of sequencing technology a decade later, the difference between mouse and human genes was reported to be less than two-times when comparing rates of synonymous sequence evolution in 24 protein coding genes (Wu and Li 1985). (Synonymous mutations are expected to fix in a population with a rate equal to the mutation rate under the strict neutrality, so these rates are also referred to as the mutation rates.) Over the next decade, significant controversies existed about relative rates of evolution in primates and rodents and the associated hypotheses (e.g., generation time hypothesis) proposed to explain the observed difference (see review in Easteal et al. 1995). More recently, Kumar and Subramanian (2002) reported a much lower (18% – 68%) rate difference between primates and rodents in an analysis of fourfold degenerate (strictly neutral) positions of protein coding genes which are third position in a codon. With the sequencing of the human and mouse genome, the focus shifted back to large scale analysis of noncoding DNA, as Waterston et al. (2002) ingeniously used the descendents of the various types of interspersed repeat present in the most recent common ancestor of human and mouse (ancestral repeats) to obtain relative rates of evolution between human and mouse genomes. They report a two times higher rate of point mutations in mouse as compared to humans. Under the neutral theory, we expect the mutation rates at fourfold-degenerate sites in codons to be equal to that in the non-coding regions, as long as all mutations in these positions are strictly neutral (Kimura 1983). While the presence of weak selection in coding regions (e.g., codon usage bias) and functional constraints in noncoding DNA may distort the expected relationship, it is important to investigate whether the coding and noncoding DNA indeed show different patterns before invoking selectionist scenarios. Establishing a genome-wide pattern

across coding and non-coding regions is also needed to ascertain whether one needs to invoke generation time and other hypotheses.

Unlike the analysis of protein coding genes where outgroup species are available to estimate the relative rate of mutation (**Figure 4a**), Waterston et al. (2002) used a consensus ancestral sequence (Smit et al. 1995) for each repeat family and estimated the amount of sequence divergence in human and mouse descendents (**Figure 4b**). Waterston et al. (2002) aligned large homologous segments between human and mouse and treated aligned repeat elements between human and mouse repeats of the same family to have shared a most recent common ancestor with each other than any other member of that repeat family. This is equivalent to determining sequence orthologs between species based on the conserved synteny. It is important to identify orthologous repeats sequences, because a large number of homologous-repeats between species are needed to obtain reliable estimates of neutral sequence divergence and because the species divergence time is used to estimate mutation rates. However, it is well known that human and mouse genomes have undergone a large number of chromosomal rearrangements since their divergence from a common ancestor over 90 million years ago (Bourque et al. 2004; Kumar et al. 2001; Waterston et al. 2002). This fast rate of rearrangement coupled with the long-time of species divergence may negatively impact the assumption of orthology-by-syntenly made in Waterston et al. (2002).

Therefore, we have employed an alternative approach in which reciprocal best BLAST-hits are used to identify putative orthologous repeats and the possible outgroups. This process is explained in **Figure 5** and yields a pair of human-mouse repeat elements along with either a human or a mouse repeat element outgroup, which we refer to as a repeat-triplet. In order to estimate sequence divergence using the sophisticated GTR model, we needed to construct a sequence alignment for each repeat-triplet. Because human repeats are usually larger than the mouse repeats by about 100 bases (Table 2), we used a profile alignment to preclude misled alignment by longer sequences; align a pair of orthologous repeats first and then align them with a outgroup by profile alignment. In addition, 72 young repeat families that do not have enough number of repeat-triplets with consistent outgroups (see Methods) were analyzed by using the consensus sequence and the results of this (Table 3) show the consistency between our estimates and other

estimates from Waterston et al. (2002). Unlike 30 largest repeat families, variation in relative rate of evolution is observed in 72 young repeat families (i.e. 1.11 ~ 1.72).

1.2 Statement of the Problem

The difference in mutation rate between human and mouse have been examined since 1969 (**Table1**). In very beginning work, DNA-DNA hybridization methods were used to estimate the rate of evolution because of the dearth of sequencing technology. DNA-DNA hybridization technique is not enough sophisticated to extract evolutionary information from relatively distant sequences (i.e. human and mouse). Multiple hits of mutation at same site can not be accounted in this method and repetitive sequences exist in sequences mislead us to get wrong estimation of divergence. Decade later advent of sequencing technology enables us to deal with sequences directly. From this point, mathematical models have been developed and applied to estimation of divergence between sequences. However, estimation of evolutionary rate still suffered from deficit of sequence data. For instances, Wu and Li (1985) used only 24 protein coding genes. Even after completion of human genome project (2001), Kumar and Subramanian (2002) still used very small number of sequences to estimate evolutionary divergence (i.e. 11 genes) because of lack of sequences from reference species for relative rate test. The four-fold degenerate sites in coding region (i.e. third codon position) were used in Wu and Li (1985) and Kumar and Subramanian (2002). These sites may be under weak selection (e.g., codon usage bias) and this may distort the estimation. Waterston et al. (2002) used ancestral repeat families, which existed in both human and mouse genomes. They used a consensus sequences for each repeat family and estimated the amount of sequence divergence in human and mouse descendants of interspersed repeats (**Figure 4b**). This methodology may have also a problem which is already mentioned above.

The rate of spontaneous mutation rate is very difficult to determine directly due to the rarity of mutations and multiple hits at same site. In addition, different types of genome region, such as exons, introns, regulatory regions, and repetitive sequences, have different pattern of mutational changes (Nei and Kumar 2000). In these respects, estimation of rate of evolution is not an easy task. We used ancestral repeats and

alternative approach to identify orthologs in order to estimate relative rate of evolution between human and mouse rather than the mutation rate.

Table 1 Some estimates of relative rate of evolution.

Authors	Year	Data Set	Way to estimate	Relative rate*
Laird et. al and Kohne et. al	1969, 1970	coding sequences of a, b chain of hemoglobin	DNA-DNA hybridization	>10.0
Wu and Li	1985	4-fold degenerate sites of 24 protein coding gene	Relative rate test	~ 2.0
Kumar and Subramanian	2002	4-fold degenerate sites of 11 genes	Relative rate test	1.18 ~ 1.68
Waterson et. al	2002	Ancestral Repeats of human and mouse genomes	Divergence from consensus sequence	~ 2.0

This table contains only preliminary estimates from DNA sequences.

*Relative rate = mouse / human

1.3 Significance of Study

“Evolution is process of change. At molecular level, evolution is the process of mutation with selection” (Pevsner 2003). The knowledge of the rate of mutations is therefore important to study molecular evolution. In addition, genetic novelty is arise by mutations and, in turn, genetic novelty yield genetic variation and source of genetic disease (Kumar and Subramanian 2002). Functional regions, such as coding exons, RNAs rather than mRNAs and regulatory upstream region, play a key factor involved in the fitness of organism and dose not allow mutations to survive in next generation against nature selection. In addition, mutation rate of functional region can be distorted by various factors. In these respects, it is hard to infer mutation rate directly from these sites. In contrast, the sites in the genome under no functional constraints provide clues about the mutation rate. At these sites, observed substitution rate is equal to mutation rate based on Neutral theory, proposed by Motto Kimura (Kimura, 1983). More specifically, these neutral regions include four-fold degenerate sites, introns, pseudo-genes, intergenic region and interspersed repeats. The mutations of these sites are considered to be neutral substitution because no mutations at these sites affect the phenotype. Therefore, estimation of substitution rate at these sites gives us an estimate about the mutation rate that, in turn, promotes our understanding of human genome and genetic diseases which are caused by mutations such as diabetes, Alzheimer’s disease, schizophrenia, asthma and

so on. Furthermore, the knowledge of difference in mutation between human and mouse provides us opportunity to compare between human and mouse, which is one of the best methods to extract useful information from human genomes because of the wealth of experimental results coming out from the laboratory mouse, *Mus musculus*. The study of human genomes can be enhanced by experiments on mouse (Waterston et al. 2002).

BACKGROUND

2.1 Phylogenetic Analysis from Molecular Data

Diversity of life on Earth has been an interesting topic since human history recorded (Mayr 1982) and all living organisms share a common ancestor (Darwin 1859). The object of phylogenetic analysis is to infer the correct relationship between diverse organisms and to estimate evolutionary divergence between species.

Phylogenetic analyses historically had been based on the phenotypic (morphological or physiological) characteristics of organism and paleontological records. Recently, phylogenetic analyses also rely on genomes of organism such as DNA, RNA or protein sequences. Molecular data have several advantages over morphological and physiological data (Li 1997). First, nucleotide sequences are strictly inheritable traits. Second, molecular characters can be described unambiguously. Third, evolution of nucleotides follows a much more regular manner. Fourth, it is manageable to quantify of molecular data. Fifth, assessment of homology for sequences is easier than for morphological data. Sixth, some sequences (i.e. 16s RNA) can be used to examine the evolutionary relationship between very distantly related organisms. Finally, the amount of sequence data is much more abundant than morphological data.

Broadly, phylogenetic analysis has five steps; (1) obtaining homologous sequences for analysis, (2) multiple sequence alignment to identify the homologous positions, (3) tree-building, (4) measure of distance, and (5) tree evaluation.

2.1.1 Homologous sequences

The selection of homologous sequences, which are descended from a common ancestor, is indispensable for inferring a phylogeny. We can determine experimentally whether given sequences are homologous or not. For instance, nucleotide hybridization and immunochemical cross-reactivity are usually used for this purpose. At sequence level, the test of homologous of sequences is equivalent to examine a degree of similarity. If the degree of similarity between sequences is higher than by chance, those sequences may be considered as homologous (Beanland and Howe 1992). It is better to exclude the sequence that

can not be determined as homologous clearly from analysis (Olsen 1988). Homologous sequences may be **orthologous** or **paralogous**. Walter Fitch (1970) defined these terms. He stated that “there should be two subclasses of homology. Where the homology is the result of gene duplication so that both copies have descended side by side during the history of an organism (for example, α and β hemoglobin) the genes should be called paralogous (para = in parallel). Where the homology is the result of speciation so that the history of the gene reflects the history of the species (for example α hemoglobin in man and mouse) the genes should be called orthologous (ortho = exact).”

2.1.2. Multiple sequence alignment

A nucleotide in one sequence will not be compared with an arbitrary nucleotide in another sequence. Homologous sites in an evolutionary sense will be compared between given sequences. Gray Olsen (1988) clearly mentioned that defining of counterpart position in one sequence of a position in another sequence is prerequisite for comparing sequence. Sequence alignment is designed to conduct this work. By sequence alignments, homologous residues are aligned in column.

The most commonly used algorithm is derived from the progressive sequence alignment, which was put forwarded by Feng and Doolittle (1987). CLUSTALW (Thompson et al. 1994) is the one of available programs to generate multiple sequence alignments.

2.1.3 Tree building

There are three common methods to build a phylogenetic tree; Maximum parsimony (MP), Maximum-likelihood (ML), and Distance methods. MP and ML are character-based methods. The fundamental idea of MP is to have the phylogenetic tree to contain minimum number of evolutionary events. Maximum-likelihood relies on particular probabilistic models of evolution and seeks for the tree with the highest maximum likelihood under these models. Distance methods are based on the matrix of evolutionary distances. Evolutionary distances are

computed for all pairs of taxa and a phylogenetic tree is constructed by considering the relationships among these distances.

2.1.4 Measure of distance

Measurement of divergence between sequences is complicated by several factors, which are multiple hits of mutations at a same site, back mutation, and different mutation rates (Figure 1). Therefore, we can not simply count the identical residues between sequences to estimate divergence. There are model to get reliable estimation of divergence between sequences. They will be described later (see 2.2).

A	A → C	Single substitution
A → T → C	A	Sequential substitution
A → T	C → T	Convergent substitution
G	G → C → G	Back substitution
Sequence 1	Sequence 2	

Figure1 Various types of substitutions.

2.1.5 Tree evaluation

It is important to assess the accuracy of a created phylogenetic tree. Two types of errors may occur in a phylogenetic tree; topological errors and branch length errors. The reliability of a phylogenetic tree is usually evaluated by consistency and robustness. The most common procedure is Felsenstein's (1985) bootstrap. The general idea of his test is resampling technique.

2.2 Models of DNA evolution

The primary process of DNA evolution is the substitution of a nucleotide for another one over evolutionary period. Change in a DNA sequence is a source to compute the rate of

evolution by comparing homologous sequences that have descended from a common ancestral sequence. Models of DNA evolution are necessary for this sort of comparison.

2.2.1 *p* distance

This is the simple measurement of sequence divergence by computing the proportion of different nucleotide between two sequences.

$$p = n_d / n$$

Where n_d is the number of different nucleotide and n is the number of total nucleotides. When p is small (i.e. $p \leq 0.1$), it represents the divergence between sequences. However, when p is big, it gives an underestimate due to backward and parallel substitutions.

2.2.2 Juke-Cantor Model (1969)

Each nucleotide has a same probability of changing in this model. In other word, this is a one parameter model.

	A	C	G	T
A	-	α	α	α
C	α	-	α	α
G	α	α	-	α
T	α	α	α	-

2.2.3. Kimura's two-parameter model (1980)

This model considers that the rate of transition (i.e. $A \leftrightarrow G$ and $C \leftrightarrow T$) is usually higher than that of transverstion (i.e. $A \leftrightarrow T$, $A \leftrightarrow C$, $G \leftrightarrow T$, and $G \leftrightarrow C$). Therefore, this is a two-parameter model.

	A	C	G	T
A	$1-\alpha-2\beta$	β	α	β
C	β	$1-\alpha-2\beta$	β	α
G	α	β	$1-\alpha-2\beta$	β
T	β	α	β	$1-\alpha-2\beta$

2.2.4 Tajima and Nei's method (1984)

General idea of this method is equivalent to the Juke-Cantor method except considering the base composition. It relies on two assumptions; substitution occurs at any site with equal probability and probability of a given nucleotide is identical,

irrespective of the original nucleotide (i.e. that a change from G to A is same as the change from T to A).

	A	C	G	T
A	-	αg_C	αg_G	αg_T
C	αg_A	-	αg_G	αg_T
G	αg_A	α	-	αg_T
T	αg_A	αg_C	αg_G	-

Where g_A , g_C , g_G , and g_T are the frequency of nucleotides.

2.2.5 Tamura's method (1992)

This model is an extension of Kimura 2-parameters model by accounting deviation of GC content.

	A	C	G	T
A	-	$\beta\theta_1$	$\alpha\theta_1$	$\beta\theta_2$
C	$\beta\theta_2$	-	$\beta\theta_1$	$\alpha\theta_2$
G	$\alpha\theta_2$	$\beta\theta_1$	-	$\beta\theta_2$
T	$\beta\theta_2$	$\alpha\theta_1$	$\beta\theta_1$	-

Where $\theta_1 = g_C + g_G$ and $\theta_2 = g_A + g_T$

2.2.6 Tamura and Nei's method (1993)

This model assumes different rates for transition (i.e. $A \leftrightarrow G$ versus $C \leftrightarrow T$). It also considers the frequency of nucleotides.

	A	C	G	T
A	-	βg_C	$\alpha 1 g_G$	βg_T
C	βg_A	-	βg_G	$\alpha 2 g_T$
G	$\alpha 1 g_A$	βg_C	-	βg_T
T	βg_A	$\alpha 2 g_C$	βg_G	-

Where g_A , g_C , g_G , and g_T are the frequency of nucleotides.

2.2.7 General time-reversible model (Lanave et al. 1984)

This model assumes a symmetric substitution rates (thus time reversible). In this model, each pair of nucleotide substitution has a different rate and the frequencies of nucleotides do not need to be equal.

	A	C	G	T
A	-	$\pi_C\beta$	$\pi_G\alpha$	$\pi_T\gamma$
C	$\pi_A\beta$	-	$\pi_G\delta$	$\pi_T\epsilon$
G	$\pi_A\alpha$	$\pi_C\delta$	-	$\pi_T\eta$
T	$\pi_A\gamma$	$\pi_C\eta$	$\pi_G\epsilon$	-

There are other models which are not mentioned here such as Hasegawa-Kishino-Yano (1985), F84 (Felsenstein 1984), general time-reversible model (Lanave et al. 1984).

2.3 Neutral Substitution

Mutations of DNA affect the fitness of organism by three different ways; negative (or purifying), positive and neutral. First, they may be harmful, reducing the chance of surviving or reproducing of progeny. They are not even allowed to fix in population at all when they occur at the regions that have critical functions for organisms. In some cases, mutations may enhance fitness by making organisms to adapt to changes in the environment. Finally, neutral mutations are the mutations which do not affect fitness because they do not cause the change of phenotypes.

The sites in the genome under no functional constraints provide clues about the mutation rate while it is hard to infer mutation rate from selected mutations, either negative or positive ones. At neutral sites, observed substitution rate is equal to mutation rate based on Neutral theory, proposed by Motoo Kimura (Kimura, 1983). More specifically, these regions include four-fold degenerate sites, introns, pseudo-genes, intergenic region and interspersed repeats. The mutations of these sites are considered to be neutral substitution because all mutations at these sites do not affect the phenotype. Therefore, estimation of substitution rate at these sites gives us an idea about the mutation rate.

The following formula shows that neutral substitution rate equals the mutation rate.

- Rate of substitution = mutation rate x probability of fixation
- Rate of neutral substitution = $2N\mu \times (1/2N) = \mu$,

where N is the effective population size of diploid individuals, μ is mutation rate and $1/2N$ is the probability that new arisen mutation will fix in next generation.

2.4 Interspersed Repeats

Large quantity of repetitive sequences is the characteristic of mammalian genomes. There are five main classes of repetitive DNA in human genomes (Lander et al. 2001);

Interspersed repeats (transposon-derived repeats), processed pseudogenes, simple sequences repeats, segmental duplication, and blocks of tandemly repeated sequences. Most of repetitive DNA is interspersed repeats, which are transposable elements. Nearly 46% of human genome and 37.5% of mouse genome consists of interspersed repeats (Waterston et al. 2002). This large amount of them provides an enhanced chance to explore the genome-wide mutation rate difference in human and mouse genome because they are considered as fossils of transposable elements, which do not have functional constraint. Among interspersed repeats, the repeats that existed in a common ancestor are called “ancestral repeats”. We examine the relative rate of evolution in human and mouse by using these ancestral repeats.

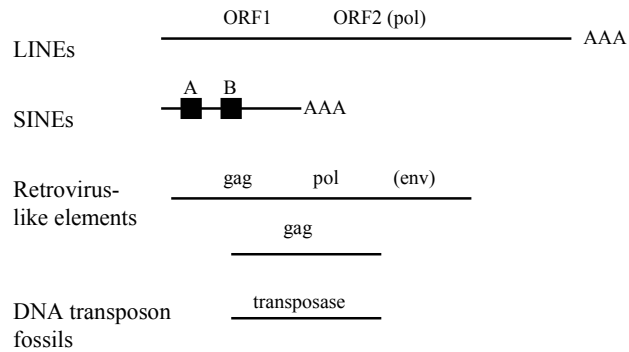


Figure 2 Classes of interspersed repeat in the human genome.

METHODOLOGY

3.1 Overview

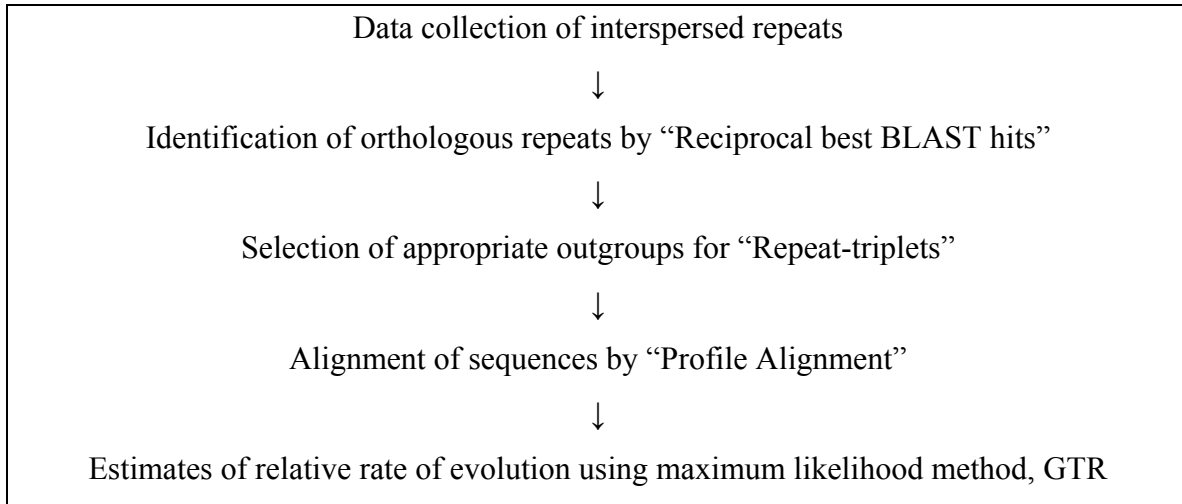


Figure 3 Brief outline of procedure for the estimating the relative rate of evolution between human and mouse

Figure 3 shows the outline of our procedure. First, we extracted interspersed repeat sequences from human and mouse genomes. We need to secure homologous sequences to estimate the rate of evolution. One pair of orthologous, which were identified by reciprocal best blast analysis, and a paralogous sequence as an outgroup are used in this study. Using these three sequences, referred to "repeat-triplet", we estimated relative rate of evolution between human and mouse. We computed the divergence of each repeat family from the consensus sequence (Jurka 2000) by counting the mismatch (p -distance). The most of 30 largest repeat families that were selected for analysis are old repeat families (divergence > 0.5 substitutions per site). Among 135 ancestral repeat families, 72 repeat families (0.4 substitutions per site $<$ divergence $<$ 0.5 substitutions per site) were defined as young. They, unfortunately, did not have enough repeat-triplet with consistent outgroups to apply our method. Therefore, we estimated the divergence by using consensus sequences, which used in Waterston et al. (2002), in order to compare our estimates with other estimates from Waterston et al. (2002).

3.2 Data Collection

We obtained the list of ancestral repeat families, which exist in both the human and mouse genome, reported in Waterson et. al (2002). Our list includes 135 ancestral repeat families (Appendix A). The sequences of interspersed REs in ancestral repeat families were obtained from UCSC genome browser (Kent et al. 2002). All interspersed REs of human and mouse used in this study were extracted from the human (Build 35) and mouse genome (Build 33) using the files that contained all information of repeats (called 'chromOut' in release hg17 and mm5) from UCSC genome browser (Kent et al. 2002). All REs in these files were identified using RepeatMasker (<http://repeatmasker.org>). Initially, REs whose length was less than 50 bases were excluded from our analysis. All extracted human and mouse REs were clustered into their families. We selected 30 largest ancestral repeat families of human genomes in the respect of the number of REs (Table 2). This produced a data set consisting of 1300457 and 266282 REs in human and mouse genomes respectively. They covered 342 mega base pairs in the human genomes and 51 mega base pairs in the mouse genomes. There is usually five-fold larger number of REs in 30 repeat families of the human genomes as compared to the mouse genomes. In addition, the average length of human RE (i.e. 220 base pairs) is larger than that of mouse RE (i.e. 120 base pairs) by about 100 base pairs. Figure 4a shows the scheme to examine the number of nucleotide substitution in human and mouse lineage. The basic concept of this scheme came from the relative rate approach (Sarich and Wilson 1967; Wu and Li 1985). It is an assessment of relative rate of evolution that is independent of the divergence times. This method requires three sequences. We used a pair of orthologous REs and a paralogous RE as an outgroup (or reference). Among these REs from 30 repeat families, the list of putative orthologous REs were secured from reciprocal best BLAST hits. Possible outgroups also were selected based upon the BLAST hits (see below). This resulted in repeat-triplets which contained an orthologous pair of human-mouse repeat elements along with either a human or a mouse RE outgroup. Finally, we procured the list consisted of 1756 orthologous repeat pairs with consistent outgroups, which regarded as appropriate outgroups.

Repeat families whose divergence between human and mouse is between 0.4 (substitutions per site) and 0.5 (substitutions per site) are regarded as relatively young.

Unfortunately, the number of repeat-triplets with consistent outgroups from young repeat families was so small (i.e. less than three or even zero) that we could not apply our method to estimate relative ratio. Therefore, we just identified orthologs using reciprocal best BLAST analysis. 3111 orthologous pairs were identified and they covered 13980470 and 8595874 bases in the human and mouse genome respectively.

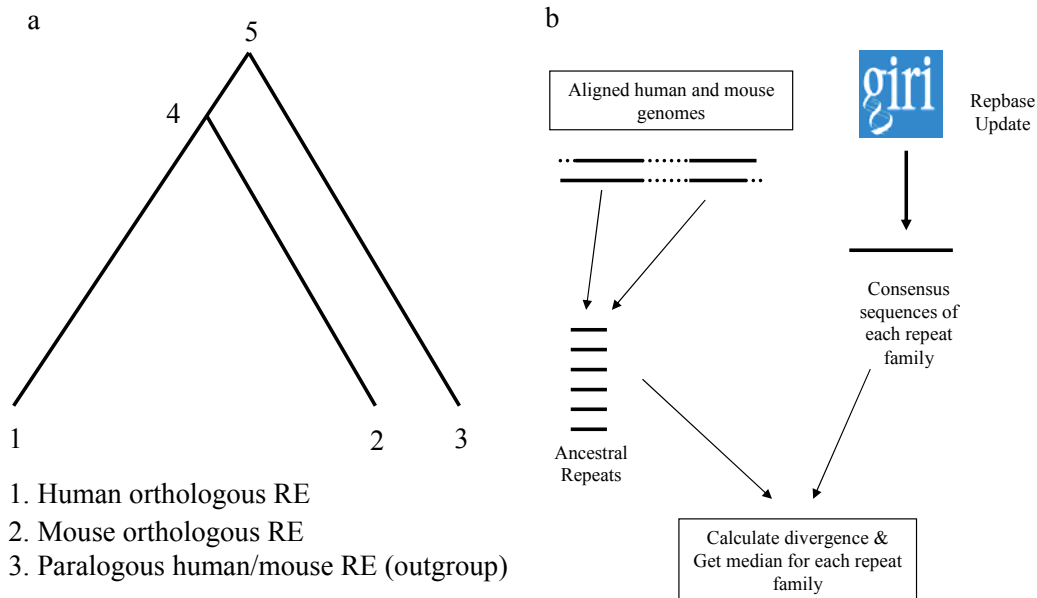


Figure 4 a) Our scheme to estimate the relative rate of evolution. Schematic shows the topology of repeat-triplet to estimate the number of substitutions in human and mouse lineage b) Identification of repeats as ancestral is based on aligned human-mouse genomes. Waterston et al. estimated the divergence of ancestral repeats using consensus sequences of each repeat family.

3.3 Identification of Orthologs and Possible Outgroups; Repeat-Triplets

We obtained the list of putative human-mouse orthologous REs within a family, identified through reciprocal best hit using BLAST (Fig. 5a). Stand-alone BLAST program downloaded from NCBI web site was used in this study (Altschul et al. 1990). BLAST search to identify orthologs was executed within each repeat family. In a reciprocal BLAST analysis, the pinpoint of orthologous REs was implemented in two iterations. In the first iteration, each human RE in a repeat family was subject to BLAST

search against a database that consisted of all human and mouse RE in each repeat family to identify homologous mouse REs. The best BLAST hits of mouse RE (with $e < 0.001$) was selected as a query for BLAST search in second iteration. Human REs with higher score than best BLAST hits of mouse RE were regarded as closer paralogs (i.e. closer paralogs might duplicate after speciation). In the second iteration, the selected mouse RE, which was best BLAST hit among mouse REs, was being investigated for homologous human REs using BLAST. The best BLAST hits of human RE (with $e\text{-value} < 0.001$) was considered as the homologous REs of query mouse RE. A doublet of homologous REs that was found to repeat in both iterations (i.e., both A-B and B-A pairs were found) was considered as orthologs. Our reciprocal BLAST analysis used the database consisted of both human and mouse REs instead of database consisted of either only human or mouse REs. This made us to select possible outgroups based on BLAST hits for 30 largest repeat families (Fig. 5b). Each human and mouse RE in a pair of ortholog provided their own sets of BLAST hits. It means that possible outgroup can be selected from both BLAST hits made by using either human or mouse REs as a query (called respectively human BLAST hits and mouse BLAST hits). The best matching RE right (with $e < 0.01$) behind orthologous RE in a BLAST hit was selected as an outgroup for this orthologous pair. It can be either human or mouse RE. If selected outgroups from both BLAST hits were identical, they were referred as 'Consistent outgroups'. If not, they were called 'Inconsistent outgroups'. Eventually, 'Repeat-triplets' consist of three sequences; one pair of orthologous RE and an outgroup.

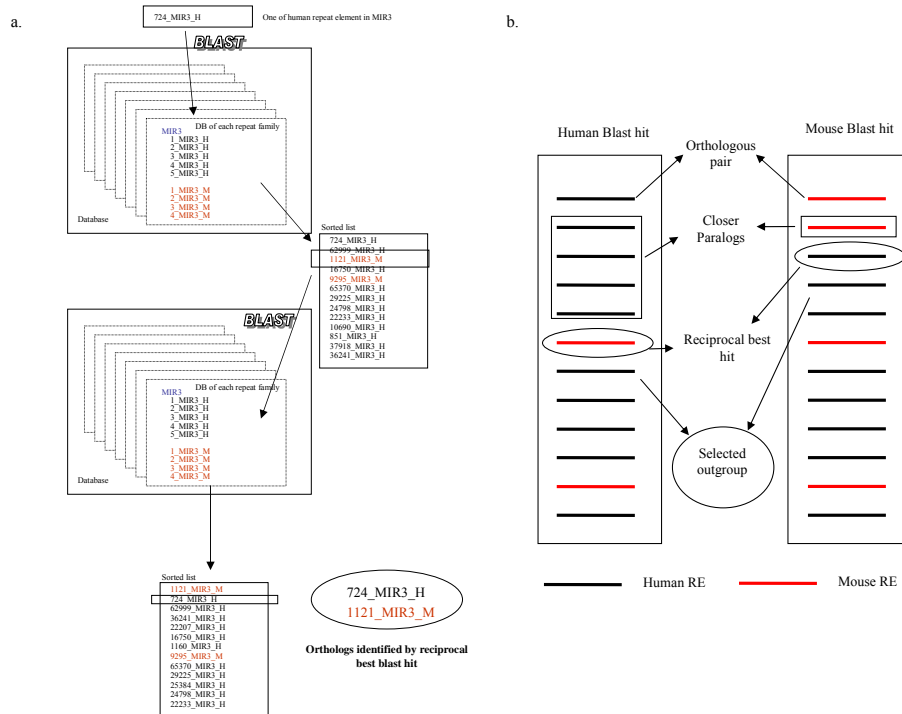


Figure 5 Detailed reciprocal best BLAST analysis.

These figures show how to identify orthologs by reciprocal best BLAST hits and to select appropriate outgroups for relative rate approach. Black lines indicate human repeat elements and Red lines indicate mouse repeat elements. Every repeat element has an assigned number based on the order of their positions on chromosomes. a) The identification of orthologs by reciprocal best BLAST hits. Putative orthologous human-mouse repeat elements are paired by reciprocal best BLAST hits. Every single human repeat element is used as a query and database consists of all human and mouse repeat elements in a repeat family. BLAST searches are conducted within a repeat family. b) The selection of possible outgroups. Every pair of orthologous repeat elements contains two BLAST hits using either human or mouse repeat elements as a query. Therefore, possible outgroups can be selected from both BLAST hits. Best matching repeat elements right behind orthologs are selected as an outgroup. Same lineage repeat elements with higher scores than orthologs are regarded as closer paralogs (i.e. duplication after speciation). When these two possible outgroups are identical, they are called consistent outgroups. If not, they are called inconsistent outgroups.

3.4 Estimation of Relative Rate of Evolution

For a given repeat-triplet with consistent outgroups, three sequences were aligned by CLUSTALW (Thompson et al. 1994) with modified settings, 'Profile alignment'. Human orthologous RE was usually aligned with human RE outgroup first, then aligned with mouse orthologous RE, because the length of human REs is usually longer than that of mouse REs. Profile alignment aligns a pair of orthologs first, and then aligns them with an outgroup. Finally, the number of nucleotide substitutions in human and mouse lineage

were computed by using the general time-reversible (GTR) method (Lanave et al. 1984; Rodriguez et al. 1990). These computations were implanted in PAUP* (Swofford 1998). Relative ratio of evolution rate is D_{24} / D_{14} in Fig. 1 (D_{ij} is the evolutionary divergence between two points, i and j). The relative rate of each repeat family is calculated by following ways; first, we estimated all D_{14} and D_{24} in a repeat family and took average of D_{14} and D_{24} over all orthologous RE and finally the average of D_{24} was divided by the average of D_{14} . Genomic average was computed by taking average of D_{14} and D_{24} over all orthologous RE from 30 repeat families and then relative rate in genomics average was calculated by average of $D_{24} /$ average of D_{14} .

In the case of 72 repeat families, we could not use our method to estimate the relative ratio between human and mouse. Therefore, we used consensus sequences to estimate the divergence of REs from ancestral repeat elements and then calculated the relative ratio. We took average of all REs in each repeat family for computing divergence.

RESULTS

4.1 Number of orthologs

We identified orthologous repeat elements (REs) between human and mouse genomes using all REs of 30 largest ancestral repeat families and 72 young repeat families in our list (see Methods). All REs in a repeat family are subject to BLAST search to identify orthologous RE within a repeat family. Orthologous REs were defined by reciprocal best BLAST hits and shown in Table 2 and 3. The number of orthologs was significantly small when comparing to total number of REs in human and mouse genomes. It indicates that reciprocal best BLAST can not catch all orthologs because of the presence of close paralogs. However, at least, the identified orthologous repeats are guaranteed to be orthologous and the number of identified orthologous REs was quite large. The count of all orthologous REs was 76723 in 30 largest repeat families and 3111 in 72 young repeat families respectively (Table 2 and 3).

Appropriate outgroups also need to be acquired for each orthologous pairs to estimate the number of substitution in the human and mouse lineage by the scheme shown in Fig. 1. The outgroup is appropriate when outgroup is not closer to human or mouse orthologs than another orthologs. Among these orthologs, we selected only orthologs have appropriate outgroups, which are called “Consistent outgroup”, in order to get reliable estimation of relative rate of evolution (see Methods). In fact, the discrepancy in estimates of evolution rate was observed from orthologs have “inconsistent outgroup” (see Methods; Appendix B). I explained the reason that consistent outgroups are appropriate in Discussion (see 5.1). Table 2 shows that very small portion of orthologs obtain consistent outgroup, which was regarded as a reciprocal best outgroup. This further suffers by the presence of numerous duplications of interspersed repeats. The number of orthologs with consistent outgroups from 30 ancestral repeat families is 1756 and they cover 427020 base pairs and 288607 base pairs in human and mouse genomes respectively. The number of human and mouse RE outgroups is respectively 1590 and 166 among 1756 repeat-triplets. Only selection of repeat-triplets with consistent outgroups reduces the number of orthologous RE substantially, but still the number of orthologous RE (i.e. 1756) analyzed was quite large. Few mouse RE outgroups may

indicate the relatively more active duplication of interspersed repeats in human genomes since divergence from their common ancestor. However, we could not obtain the orthologous pair with consistent outgroups from 72 young repeat families. As I have already mentioned, reciprocal best BLAST analysis may miss real orthologs because of closer paralogs and the number of repeat elements in 72 young repeat families is quite small comparing with 30 largest repeat families. This drawback of reciprocal best BLAST analysis is even worse in the case of consistent outgroups. Therefore, we could not enough number of orthologs with consistent outgroups from young repeat families.

4.2 30 largest repeat families

4.2.1 Relative Rates of Evolution

The schematic in Fig. 1a shows the topology of orthologs with consistent outgroups to estimate the number of substitutions in the human and mouse lineage. The number of nucleotide substitutions was computed by three maximum likelihood methods. We eventually employed general time-reversible (Lanave et al. 1984; Rodriguez et al. 1990), GTR, model for estimation of evolution rate after likelihood ratio test (Goldman 1993; Yang et al. 1994). Likelihood ratio test showed the favor of GTR instead of HKY (Hasegawa et al. 1985) and GTR incorporating a gamma distribution (Uzzell and Corbin 1971), GTR + Γ . HKY showed nevertheless similar results with GTR. However, the lengths of REs are relatively short for such a sophisticated method, GTR + Γ . Table 2 shows the relative rate of 30 ancestral repeat families. The relative ratio of each repeat family was computed by taking average over all orthologous REs within each repeat family. Eventually, 1.018 (conceptually grand average) is estimated by taking average of evolutionary rate over all 1756 orthologous REs with consistent outgroups from 30 ancestral repeat families. This means that the mutation rates in the evolutionary lineage leading to mice, since their divergence from their common ancestor with humans, is only 2 % higher than that observed for the evolutionary lineage leading to humans.

Table 2 Number of orthologs and relative mutation rates between human and mouse

Interspersed Repeat		No of repeats		NO of orthologs		Divergence		Relative Rate*
Family	Class	Human	Mouse	Total	Consistent Outgroup	Human	Mouse	
L2	LINE	385405	54736	17393	492(461,31)	0.293	0.271	0.925
MIR	SINE	194742	44069	15323	394(345,49)	0.279	0.306	1.100
MIR3	SINE	69054	9531	4054	167(146,21)	0.266	0.255	0.961
L1M4	LINE	47531	19998	5081	66(58,8)	0.276	0.311	1.125
L3	LINE	44837	9549	1300	71(62,9)	0.178	0.173	0.972
L1ME4a	LINE	40046	1810	1148	43(39,4)	0.255	0.240	0.940
MER5A	DNA	33903	9163	2710	48(43,5)	0.303	0.301	0.991
L1ME	LINE	32950	9214	2851	82(67,15)	0.326	0.301	0.923
L1ME1	LINE	31635	5349	1734	19(18,1)	0.295	0.278	0.944
L1MC4	LINE	30177	5923	1921	21(21,0)	0.290	0.280	0.964
L1MC4a	LINE	29312	3096	1137	22(22,0)	0.316	0.300	0.949
L1ME3B	LINE	28905	1461	703	19(19,0)	0.310	0.400	1.289
HAL1	LINE	27247	3816	1291	30(27,3)	0.314	0.312	0.995
MER5B	DNA	22704	4785	1881	60(57,3)	0.281	0.335	1.191
L1MB7	LINE	21844	6893	1116	10(10,0)	0.173	0.437	2.532
L1MC5	LINE	20069	2664	1240	35(33,2)	0.303	0.323	1.066
L1MEc	LINE	19841	5464	1611	19(17,2)	0.282	0.338	1.199
MLT1D	LTR	19428	7986	1665	7(7,0)	0.308	0.388	1.259
MLT1C	LTR	18811	7455	1303	4(4,0)	0.181	0.262	1.447
MLT1K	LTR	18208	2900	1151	46(43,3)	0.338	0.296	0.875
L1ME2	LINE	17948	3502	1249	13(12,1)	0.306	0.331	1.079
L1	LINE	17855	11317	1300	7(4,3)	0.272	0.284	1.044
L1ME3A	LINE	17215	2049	884	13(12,1)	0.246	0.282	1.149
MLT1B	LTR	17056	9004	1189	5(4,1)	0.230	0.344	1.493
L1MB3	LINE	16498	5436	664	1(1,0)	0.171	0.310	1.811
L1MB8	LINE	16341	5973	1169	7(7,0)	0.218	0.381	1.753
MER20	DNA	15908	4669	1291	8(8,0)	0.183	0.350	1.905
L1MA9	LINE	15449	4258	673	2(2,0)	0.161	0.206	1.281
MLT1J	LTR	15097	3090	1222	37(34,3)	0.327	0.298	0.909
L1MEd	LINE	14441	1122	469	8(7,1)	0.241	0.360	1.495
SUM		1300457	266282	76723	1756(1590,166)	0.303	0.308	1.018

Count indicates the number of interspersed repeats identified by RepeatMasker in the human genome (Build 35) and mouse genome (Build 33). Orthologous repeats are identified by reciprocal best BLAST hits (see Methods). Human-mouse orthologous repeats are the best BLAST hit for each other. Possible outgroups from human and mouse BLAST hits are identical, they are referred to consistent outgroups (see Methods). The two numbers in parentheses are the number of human RE selected as an outgroup and the number of mouse RE selected as an outgroup

4.2.2 Distribution of evolutionary divergence and relative ratio

Figure 6a shows the distribution of evolutionary distance estimated by GTR method using 1756 orthologous REs with consistent outgroups. This distribution

is mostly symmetrical and bell-shaped with dispersion (mean = 0.611, variance = 0.095). This graph shows the reliability of GTR method for analyzing of our data.

The distribution in Fig. 6b contains the relative ratio of evolution rate between human and mouse repeats from all identified orthologous REs from 30 ancestral repeats. This distribution is also mostly symmetrical and bell-shaped with dispersion. This graph indicates that the relative rates of evolution in this study are subject to random chance and the highest mode is observed around 1.0, which means that substitution rate in the mouse lineage is same of that in the human lineage.

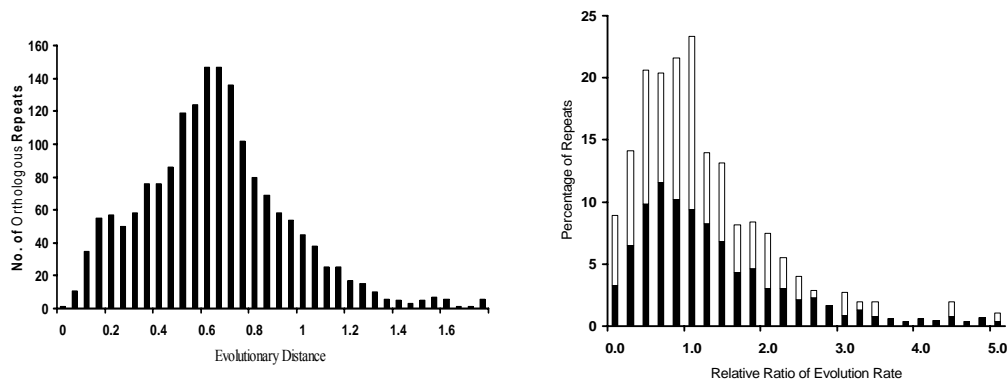


Figure 6 a) This graph shows the distribution of the evolutionary divergences for 1756 orthologous repeats of human and mouse. The evolutionary distances between orthologs were estimated by GTR method. This distribution is bell-shaped and mostly symmetrical (mean = 0.611, variance = 0.095). b) This histogram contains the relative ratio of all 1756 orthologous repeats. The black and open bars depict the relative evolutionary rates of human and mouse genomes, which were estimated by using orthologous with consistent human and mouse outgroup respectively.

4.3 72 young repeat families

4.3.1 Relative Rates of Evolution

The orthologous RE pairs of 72 young repeat families were also identified by reciprocal best BLAST analysis. We estimated divergence (mismatch) level of each orthologs from the consensus sequence of each repeat family instead of using scheme in Figure 4a due to deficit of consistent outgroups. Divergence level (i.e. p -distance) from consensus sequence is converted to Juke-Cantor distance by following formula;

$$d = - (3/4) \ln[1-(4/3)p]$$

where p is the proportion of different nucleotides between X and Y.

As known, p -distance does not produce proper estimation of evolutionary divergence between sequences when $p > 0.1$ (Nei and Kumar 2000). I could not get the ancestral repeat sequence of each repeat family from Repbase (Jurka 2000) for estimating evolutionary divergences by other evolution model of DNA such as Kimura 2-parameters, Tamura-Nei and so on. These estimates can not be converted from p -distance. Table 3 shows the estimates of divergence and relative rate of some young repeat families. Unlike 30 largest repeat families which of them most are old repeat families, 72 young repeat families show variation in relative rate of evolution (i.e. 1.11 ~ 1.72). The estimates from our orthologs by using the consensus sequences correspond closely to the estimates from Waterston et al. (2002). It indicates the validity of our approach to identify orthologs.

Table 3 Analysis of young repeat families

Family	No. of Orthologs	<i>p</i> distance		JC		Ratio ¹
		Human	Mouse	Human	Mouse	
Charlie9	71	0.22	0.23	0.25	0.28	1.11*
L1	1364	0.22	0.28	0.26	0.35	1.38
L1MA6	485	0.16 (0.16)	0.25 (0.28)	0.18	0.31	1.72**
L1MA7	415	0.16 (0.16)	0.25 (0.28)	0.19	0.31	1.63
L1MA8	427	0.16 (0.15)	0.25 (0.27)	0.18	0.31	1.72**
L1MA9	728	0.18 (0.18)	0.26 (0.28)	0.21	0.32	1.52
L1MA10	335	0.19 (0.19)	0.27 (0.29)	0.22	0.33	1.50
L1MB1	461	0.18 (0.18)	0.27 (0.29)	0.21	0.33	1.57
L1MB2	444	0.18 (0.18)	0.26 (0.28)	0.20	0.32	1.60
L1MC1	454	0.18 (0.17)	0.27 (0.28)	0.21	0.33	1.57
MLT1A	783	0.22 (0.21)	0.29 (0.31)	0.26	0.37	1.42
MLT1A0	1086	0.20 (0.19)	0.28 (0.3)	0.23	0.36	1.57
MLT1A1	759	0.19 (0.19)	0.26 (0.29)	0.22	0.32	1.45
MLT1B	1275	0.19 (0.18)	0.28 (0.28)	0.22	0.34	1.55
MLT1C	1424	0.21 (0.21)	0.29 (0.30)	0.25	0.36	1.44
Looper	30	0.19 (0.18)	0.26 (0.28)	0.22	0.33	1.50
MER20	1372	0.20 (0.19)	0.27 (0.29)	0.23	0.33	1.43
MER33	502	0.18 (0.18)	0.25 (0.27)	0.21	0.31	1.48
MER53	378	0.17 (0.17)	0.23 (0.26)	0.19	0.27	1.42
Tigger6a	60	0.19 (0.18)	0.28 (0.29)	0.22	0.36	1.64

Among 135 ancestral repeat families, 72 repeat families are young. Their divergences (mismatch level) between human and mouse are less than 0.5 and larger than 0.4. 20 young repeat families are shown here. Mismatch level is converted to Juke-Cantor estimate. The relative ratio shows variation from 1.11 to 1.72. Number in parenthesis indicates the estimates from Waterston et al. (2002). Our estimates are quite consistent with theirs.

¹ Ratio = mouse / human

* Lowest value

** Highest value

4.3.2 Relationship between GC content and relative ratio

What cause the variation in relative rate of evolution from 72 young repeat families between the human and mouse genomes? If interspersed repeats are selectively neutral, no factor may affect the mutation rate except nucleotide compositions. Therefore, we looked at the nucleotide composition of repeat families and found the wide spread of GC content of repeat families from 0.28 to 0.48. The positive relationship between GC content and relative ratio was observed in 72 young repeat families (Figure 7), suggesting that relative ratio may

be dependent on nucleotide composition. It may indicate that fast evolving repeat families, which have high GC content, have relatively higher ratio while slow evolving repeat families that have low GC content have relatively low ratio.

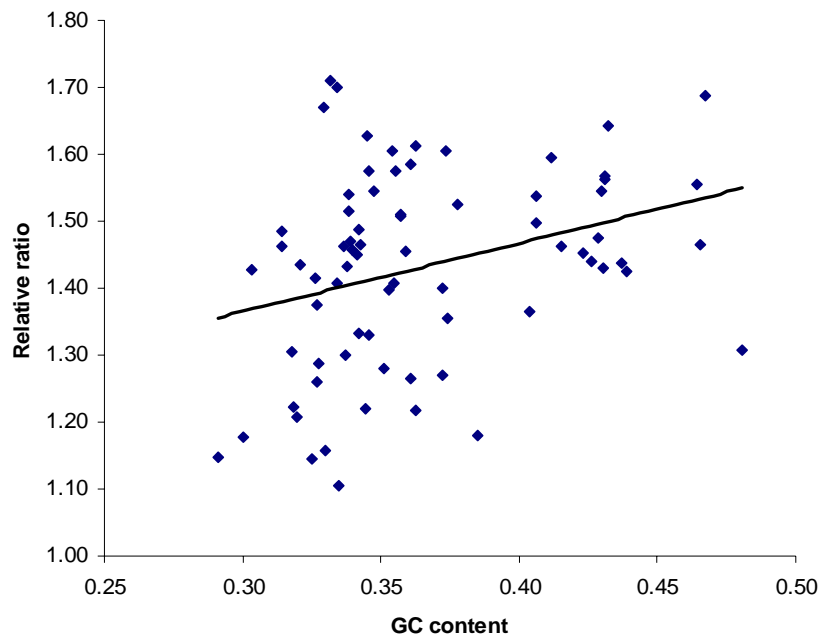


Figure 7 Relationship between GC content and relative rate from 72 young repeat families; $R^2 = 0.1$.

DISCUSSION

5.1 Discrepancy in estimates of evolution rate

The results in our study show that the evolution rate in the human and mouse lineages are different by 2% using orthologs with consistent outgroups while there is a discrepancy in estimation from repeat-triplets with inconsistent outgroups of human RE outgroups and mouse RE outgroups. The discrepancy in the estimates of relative rates of evolution from orthologs with inconsistent outgroups indicates that selection of appropriate outgroups is hard for interspersed repeats due to the presence of lots of closer paralogs, especially in human genomes. This fact may be reflected in BLAST hits. The number of closer paralogs (Fig. 3b) in BLAST hits using human RE as a query (human BLAST) much larger than that in BLAST hits using mouse RE as a query (mouse BLAST). There is still higher chance to select closer paralogs from human BLAST hits even though we pick up best matching RE right behind the orthologous RE. It disrupts the scheme shown in Figure 1. Points 2 and 3 will switch out (i.e. 2 is human RE outgroup and 3 is orthologous mouse RE) and this misleads to get inappropriate relative rate. In this case, relative rate corresponds to D_{34} / D_{14} rather than D_{24} / D_{14} (D_{ij} is the divergence between two points, i and j). Larger D_{34} (because of including common ancestor period) makes relative rate large, which observed in the estimation from repeat-triplets with inconsistent outgroups with human outgroups. In addition, closer paralogs also can be selected as mouse RE outgroups of repeat-triplets with inconsistent outgroups even though the chance is lower than that of human RE outgroups. It directs us to get lower relative rate, which observed in the estimation from repeat-triplets with inconsistent outgroups with mouse RE outgroups. In the other hand, the repeat-triplets with consistent outgroups, which are regarded as a appropriate outgroups, provides us proper sequences to estimate relative rate of evolution between human and mouse.

5.2 The advantage of our approach

Our repeat-triplets (i.e. one pair of orthologs with a possible outgroup) have advantages in estimating relative rate of evolution. First, ancestral interspersed repeats were used in this study instead of fourfold degenerate site of protein coding genes which may be under

weak selection due to codon usage bias (Li 1997). In addition, the available amount of ancestral interspersed repeats is much more than that of fourfold degenerate site. Second, repeat-triplets are generated by rigorous approach of identifying the orthologs than previous study (Waterston et al. 2002). Waterston et al. (2002) estimated the divergence of ancestral repeat in the human and mouse lineage using a consensus sequence (Jurka 2000) for each repeat family. Their ancestral repeats are comparable to syntenic orthology. However, fast chromosomal rearrangement between human and mouse genomes should not be neglected and it may disrupt their assumption. In fact, the identified orthologous REs in our study were not syntenic (data not shown). Third, selected outgroups in our study are paralogous sequences rather than orthologous sequences from other reference species. It means that our repeat-triplets are more homologous each other than three orthologous sequences. It reduces the error may occur in multiple sequence alignment and leads us to get more accurate estimations. In these respects, we can get more reliable estimation of evolution rate between human and mouse than any other studies. In addition, from the analysis of 72 young repeat families, we can compare directly our estimates with other estimates from Waterston et al. (2002). This comparison validates our way to identify orthologous. Finally, we could not construct the proper phylogenetic trees of repeat families by using all repeat elements in a repeat family. The difference in average length of human and mouse repeat elements disrupts the cluster of sequences. During constructing phylogenetic tree, the long sequences, human repeats, tend to be clustered together. The short ones, mouse repeats, do too. Therefore, it is very hard to define the orthologs based on the phylogenetic tree of repeat families. Eventually, this situation led us to employ the relative rate approach to estimate the rate of evolution in human and mouse genomes.

5.3 Implication of our results

Many studies have examined the difference in neutral mutation rate between human and mouse and have come to conflicting conclusions (Eastal 1988; Eastal 1990; Goodman et al. 1971; Hardison et al. 2003; Kohne 1970; Kumar and Subramanian 2002; Laird et al. 1969; Waterston et al. 2002; Wu and Li 1985). The significant difference reported was elucidated by generation-time effect hypothesis and different physiological attributes

between human and mouse (Laird et al. 1969; Li et al. 1996; Li et al. 1987; Wu and Li 1985). The paucity of most studies is the very small amount of data used. Almost complete sequences of the human and mouse genomes have now been reported (Lander et al. 2001; Waterston et al. 2002) and Waterston et al. (2002) used large amount of ancestral repeats to overcome this defect, but they still had problems mentioned above. Our analysis based on large amount of data, rigorous phylogenetic approach and a sophisticated evolution model of DNA (i.e. GTR) shows that difference in evolution rate between human and mouse genome is only 2%. Comparison our result with others testify the estimates of evolution rate heavily rely on the method to infer. We do not know the real answer for evolutionary history. Therefore, we need to try more reliable manner to estimate the mutation rate at least. I believe that our approach has advantages over other ways that were used in previous studies and then it may create more accurate estimates. Furthermore, the similar evolution rate between human and mouse, shown in our results, has significant meaning. It may suggest that replication-independent mutation procedure, such as recombination and repair mechanism, may more contribute to the source of mutation than expected (Drake et al. 1998; Huttley et al. 2000).

CONCLUSION

We find that the mutation rates in the evolutionary lineage leading to mice, since their divergence from their common ancestor with humans, is about 2% higher than that observed for the evolutionary lineage leading to humans while there is variation in relative rate of evolution among young repeat families even though this variation might be caused by the way of estimation. In Waterston et al. (2002), identification of ancestral repeats is equivalent to determining sequence orthologs between species based on the conserved synteny. It is important to identify orthologous repeat sequences, because a large number of homologous-repeats between species are needed to obtain reliable estimates of neutral sequence divergence. However, it is well known that human and mouse genomes have undergone a large number of chromosomal rearrangements since their divergence from a common ancestor over 90 million years ago (Bourque et al. 2004; Kumar et al. 2001; Waterston et al. 2002). This fast rate of rearrangement coupled with the long-time of species divergence may negatively impact the assumption of orthology-by-synteny made in Waterston et al. (2002). Therefore, our alternative approach may provide the opportunity to get more reliable estimation of evolutionary rate in human and mouse genomes.

ACKNOWLEDGEMENTS

I would like to thank Dr. Sudhir Kumar for suggestion of project, genuine idea of approach, useful advices about my trials, practical interpretation of results and edition of text extensively and also Dr. Jeffrey Touchman and Dr. Andrzej Czygrinow for useful comments. I also appreciate to other members of the lab for their help, specially, Dr. Sankar Subramanian for helpful discussion.

Finally, I want to show thanks to Center for Evolutionary Functional Genomics, the Biodesign institute at Arizona State University for internship fund and to National Institute of Health (NIH) for research grants (to Sudhir Kumar).

REFERENCE

- Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. *J Mol Biol* **215**: 403-410.
- Beanland, T.J. and C.J. Howe. 1992. The inference of evolutionary trees from molecular data. *Comp Biochem Physiol B* **102**: 643-659.
- Bourque, G., P.A. Pevzner, and G. Tesler. 2004. Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse, and rat genomes. *Genome Res* **14**: 507-516.
- Darwin, C. 1859. *The origin of species Signet Classic*, New York, NY
- Drake, J.W., B. Charlesworth, D. Charlesworth, and J.F. Crow. 1998. Rates of spontaneous mutation. *Genetics* **148**: 1667-1686.
- Easteal, S. 1988. Rate constancy of globin gene evolution in placental mammals. *Proc Natl Acad Sci U S A* **85**: 7622-7626.
- Easteal, S. 1990. The pattern of mammalian evolution and the relative rate of molecular evolution. *Genetics* **124**: 165-173.
- Easteal, S., Collet, C., and Betty, D. 2005 *The mammalian molecular clock*. Springer-Verlag, Austin, TX.
- Goldman, N. 1993. Statistical tests of models of DNA substitution. *J Mol Evol* **36**: 182-198.
- Goodman, M., J. Barnabas, G. Matsuda, and G.W. Moore. 1971. Molecular evolution in the descent of man. *Nature* **233**: 604-613.
- Hardison, R.C., K.M. Roskin, S. Yang, M. Diekhans, W.J. Kent, R. Weber, L. Elnitski, J. Li, M. O'Connor, D. Kolbe, S. Schwartz, T.S. Furey, S. Whelan, N. Goldman, A. Smit, W. Miller, F. Chiaromonte, and D. Haussler. 2003. Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res* **13**: 13-26.
- Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* **22**: 160-174.
- Huttley, G.A., I.B. Jakobsen, S.R. Wilson, and S. Easteal. 2000. How important is DNA replication for mutagenesis? *Mol Biol Evol* **17**: 929-937.
- Jurka, J. 2000. Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet* **16**: 418-420.
- Kent, W.J., C.W. Sugnet, T.S. Furey, K.M. Roskin, T.H. Pringle, A.M. Zahler, and D. Haussler. 2002. The human genome browser at UCSC. *Genome Res* **12**: 996-1006.
- Kimura, M. 1983. *The neutral theory of molecular evolution*. Cambridge University Press, Cambridge.
- Kohne, D.E. 1970. Evolution of higher-organism DNA. *Q Rev Biophys* **3**: 327-375.
- Kumar, S., S.R. Gadagkar, A. Filipinski, and X. Gu. 2001. Determination of the number of conserved chromosomal segments between species. *Genetics* **157**: 1387-1395.
- Kumar, S. and S. Subramanian. 2002. Mutation rates in mammalian genomes. *Proc Natl Acad Sci U S A* **99**: 803-808.
- Laird, C.D., B.L. McConaughy, and B.J. McCarthy. 1969. Rate of fixation of nucleotide substitutions in evolution. *Nature* **224**: 149-154.

- Lanave, C., G. Preparata, C. Saccone, and G. Serio. 1984. A new method for calculating evolutionary substitution rates. *J Mol Evol* **20**: 86-93.
- Lander, E.S. L.M. Linton B. Birren C. Nusbaum M.C. Zody J. Baldwin K. Devon K. Dewar M. Doyle W. FitzHugh R. Funke D. Gage K. Harris A. Heaford J. Howland L. Kann J. Lehoczy R. LeVine P. McEwan K. McKernan J. Meldrim J.P. Mesirov C. Miranda W. Morris J. Naylor C. Raymond M. Rosetti R. Santos A. Sheridan C. Sougnez N. Stange-Thomann N. Stojanovic A. Subramanian D. Wyman J. Rogers J. Sulston R. Ainscough S. Beck D. Bentley J. Burton C. Clee N. Carter A. Coulson R. Deadman P. Deloukas A. Dunham I. Dunham R. Durbin L. French D. Grafham S. Gregory T. Hubbard S. Humphray A. Hunt M. Jones C. Lloyd A. McMurray L. Matthews S. Mercer S. Milne J.C. Mullikin A. Mungall R. Plumb M. Ross R. Shownkeen S. Sims R.H. Waterston R.K. Wilson L.W. Hillier J.D. McPherson M.A. Marra E.R. Mardis L.A. Fulton A.T. Chinwalla K.H. Pepin W.R. Gish S.L. Chissoe M.C. Wendl K.D. Delehaunty T.L. Miner A. Delehaunty J.B. Kramer L.L. Cook R.S. Fulton D.L. Johnson P.J. Minx S.W. Clifton T. Hawkins E. Branscomb P. Predki P. Richardson S. Wenning T. Slezak N. Doggett J.F. Cheng A. Olsen S. Lucas C. Elkin E. Uberbacher M. Frazier R.A. Gibbs D.M. Muzny S.E. Scherer J.B. Bouck E.J. Sodergren K.C. Worley C.M. Rives J.H. Gorrell M.L. Metzker S.L. Naylor R.S. Kucherlapati D.L. Nelson G.M. Weinstock Y. Sakaki A. Fujiyama M. Hattori T. Yada A. Toyoda T. Itoh C. Kawagoe H. Watanabe Y. Totoki T. Taylor J. Weissenbach R. Heilig W. Saunin F. Artiguenave P. Brottier T. Bruls E. Pelletier C. Robert P. Wincker D.R. Smith L. Doucette-Stamm M. Rubenfield K. Weinstock H.M. Lee J. Dubois A. Rosenthal M. Platzer G. Nyakatura S. Taudien A. Rump H. Yang J. Yu J. Wang G. Huang J. Gu L. Hood L. Rowen A. Madan S. Qin R.W. Davis N.A. Federspiel A.P. Abola M.J. Proctor R.M. Myers J. Schmutz M. Dickson J. Grimwood D.R. Cox M.V. Olson R. Kaul N. Shimizu K. Kawasaki S. Minoshima G.A. Evans M. Athanasiou R. Schultz B.A. Roe F. Chen H. Pan J. Ramser H. Lehrach R. Reinhardt W.R. McCombie M. de la Bastide N. Dedhia H. Blocker K. Hornischer G. Nordsiek R. Agarwala L. Aravind J.A. Bailey A. Bateman S. Batzoglu E. Birney P. Bork D.G. Brown C.B. Burge L. Cerutti H.C. Chen D. Church M. Clamp R.R. Copley T. Doerks S.R. Eddy E.E. Eichler T.S. Furey J. Galagan J.G. Gilbert C. Harmon Y. Hayashizaki D. Haussler H. Hermjakob K. Hokamp W. Jang L.S. Johnson T.A. Jones S. Kasif A. Kasprzyk S. Kennedy W.J. Kent P. Kitts E.V. Koonin I. Korf D. Kulp D. Lancet T.M. Lowe A. McLysaght T. Mikkelsen J.V. Moran N. Mulder V.J. Pollara C.P. Ponting G. Schuler J. Schultz G. Slater A.F. Smit E. Stupka J. Szustakowski D. Thierry-Mieg J. Thierry-Mieg L. Wagner J. Wallis R. Wheeler A. Williams Y.I. Wolf K.H. Wolfe S.P. Yang R.F. Yeh F. Collins M.S. Guyer J. Peterson A. Felsenfeld K.A. Wetterstrand A. Patrinos M.J. Morgan P. de Jong J.J. Catanese K. Osoegawa H. Shizuya S. Choi and Y.J. Chen. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860-921.
- Li, W.-H. 1997. *Molecular evolution*. Sinauer Associates, Sunderland, MA
- Li, W.H., D.L. Ellsworth, J. Krushkal, B.H. Chang, and D. Hewett-Emmett. 1996. Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis. *Mol Phylogenet Evol* **5**: 182-187.

- Li, W.H., M. Tanimura, and P.M. Sharp. 1987. An evaluation of the molecular clock hypothesis using mammalian DNA sequences. *J Mol Evol* **25**: 330-342.
- Nei, M. and Kumar, S. 2000. *Molecular evolution and phylogenetics*. Oxford University Press, New York, NY.
- Olsen, G.J. 1988. Phylogenetic analysis using ribosomal RNA. *Methods Enzymol* **164**: 793-812.
- Rodriguez, F., J.L. Oliver, A. Marin, and J.R. Medina. 1990. The general stochastic model of nucleotide substitution. *J Theor Biol* **142**: 485-501.
- Sarich, V.M. and A.C. Wilson. 1967. Immunological time scale for hominid evolution. *Science* **158**: 1200-1203.
- Smit, A.F., G. Toth, A.D. Riggs, and J. Jurka. 1995. Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. *J Mol Biol* **246**: 401-417.
- Thompson, J.D., D.G. Higgins, and T.J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**: 4673-4680.
- Uzzell, T. and K.W. Corbin. 1971. Fitting discrete probability distributions to evolutionary events. *Science* **172**: 1089-1096.
- Waterston, R.H. K. Lindblad-Toh E. Birney J. Rogers J.F. Abril P. Agarwal R. Agarwala R. Ainscough M. Alexandersson P. An S.E. Antonarakis J. Attwood R. Baertsch J. Bailey K. Barlow S. Beck E. Berry B. Birren T. Bloom P. Bork M. Botcherby N. Bray M.R. Brent D.G. Brown S.D. Brown C. Bult J. Burton J. Butler R.D. Campbell P. Carninci S. Cawley F. Chiaromonte A.T. Chinwalla D.M. Church M. Clamp C. Clee F.S. Collins L.L. Cook R.R. Copley A. Coulson O. Couronne J. Cuff V. Curwen T. Cutts M. Daly R. David J. Davies K.D. Delehaunty J. Deri E.T. Dermitzakis C. Dewey N.J. Dickens M. Diekhans S. Dodge I. Dubchak D.M. Dunn S.R. Eddy L. Elnitski R.D. Emes P. Eswara E. Eyras A. Felsenfeld G.A. Fewell P. Flicek K. Foley W.N. Frankel L.A. Fulton R.S. Fulton T.S. Furey D. Gage R.A. Gibbs G. Glusman S. Gnerre N. Goldman L. Goodstadt D. Grafham T.A. Graves E.D. Green S. Gregory R. Guigo M. Guyer R.C. Hardison D. Haussler Y. Hayashizaki L.W. Hillier A. Hinrichs W. Hlavina T. Holzer F. Hsu A. Hua T. Hubbard A. Hunt I. Jackson D.B. Jaffe L.S. Johnson M. Jones T.A. Jones A. Joy M. Kamal E.K. Karlsson D. Karolchik A. Kasprzyk J. Kawai E. Keibler C. Kells W.J. Kent A. Kirby D.L. Kolbe I. Korf R.S. Kucherlapati E.J. Kulbokas D. Kulp T. Landers J.P. Leger S. Leonard I. Letunic R. Levine J. Li M. Li C. Lloyd S. Lucas B. Ma D.R. Maglott E.R. Mardis L. Matthews E. Mauceli J.H. Mayer M. McCarthy W.R. McCombie S. McLaren K. McLay J.D. McPherson J. Meldrim B. Meredith J.P. Mesirov W. Miller T.L. Miner E. Mongin K.T. Montgomery M. Morgan R. Mott J.C. Mullikin D.M. Muzny W.E. Nash J.O. Nelson M.N. Nhan R. Nicol Z. Ning C. Nusbaum M.J. O'Connor Y. Okazaki K. Oliver E. Overton-Larty L. Pachter G. Parra K.H. Pepin J. Peterson P. Pevzner R. Plumb C.S. Pohl A. Poliakov T.C. Ponce C.P. Ponting S. Potter M. Quail A. Reymond B.A. Roe K.M. Roskin E.M. Rubin A.G. Rust R. Santos V. Sapojnikov B. Schultz J. Schultz M.S. Schwartz S. Schwartz C. Scott S. Seaman S. Searle T. Sharpe A. Sheridan R. Shownkeen S. Sims J.B. Singer G. Slater A. Smit D.R. Smith B. Spencer A. Stabenau N. Stange-Thomann C. Sugnet M. Suyama G.

- Tesler J. Thompson D. Torrents E. Trevaskis J. Tromp C. Ucla A. Ureta-Vidal J.P. Vinson A.C. Von Niederhausern C.M. Wade M. Wall R.J. Weber R.B. Weiss M.C. Wendl A.P. West K. Wetterstrand R. Wheeler S. Whelan J. Wierzbowski D. Willey S. Williams R.K. Wilson E. Winter K.C. Worley D. Wyman S. Yang S.P. Yang E.M. Zdobnov M.C. Zody and E.S. Lander. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**: 520-562.
- Wu, C.I. and W.H. Li. 1985. Evidence for higher rates of nucleotide substitution in rodents than in man. *Proc Natl Acad Sci U S A* **82**: 1741-1745.
- Yang, Z., N. Goldman, and A. Friday. 1994. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Mol Biol Evol* **11**: 316-324.

WEB SITE REFERENCES

- <http://genome.ucsc.edu>; UCSC Human Genome Browser.
- <http://repeatmasker.org>; Smit, AFA & Green, P RepeatMasker.
- <http://ncbi.nlm.nih.gov>; NCBI web site.

APPENDIX

Appendix A The list of 135 ancestral repeat families.

Repeat Family	Class	Repeat Family	Class	Repeat Family	Class
5S	rRNA	L1MC4	LINE	MER81	DNA
7SK	RNA	L1MC4a	LINE	MER82	DNA
7SLRNA	srpRNA	L1MC5	LINE	MER89-int	LTR
BLACKJACK	DNA	L1MCa	LINE	MIR	SINE
Charlie1	DNA	L1MCb	LINE	MIR3	SINE
Charlie1a	DNA	L1MCc	LINE	MLT1-int	LTR
Charlie1b	DNA	L1MD1	LINE	MLT1B	LTR
Charlie2	DNA	L1MD2	LINE	MLT1C	LTR
Charlie2a	DNA	L1MD3	LINE	MLT1D	LTR
Charlie2b	DNA	L1MDa	LINE	MLT1E	LTR
Charlie4	DNA	L1MDb	LINE	MLT1E1	LTR
Charlie4a	DNA	L1ME	LINE	MLT1E2	LTR
Charlie7	DNA	L1ME1	LINE	MLT1F	LTR
Charlie8	DNA	L1ME2	LINE	MLT1F1	LTR
Charlie9	DNA	L1ME3	LINE	MLT1F2	LTR
ERVL	LTR	L1ME3A	LINE	MLT1G	LTR
ERVL-E	LTR	L1ME3B	LINE	MLT1G1	LTR
HAL1	LINE	L1ME4a	LINE	MLT1G3	LTR
HAL1b	LINE	L1MEc	LINE	MLT1H	LTR
HERV16	LTR	L1MEd	LINE	MLT1H1	LTR
HY1	scRNA	L2	LINE	MLT1I	LTR
HsTC2	DNA	L3	LINE	MLT1J	LTR
L1	LINE	L3b	LINE	MLT1J1	LTR
L1M	LINE	MARNA	DNA	MLT1J2	LTR
L1M3	LINE	MER2	DNA	MLT1J2-int	LTR
L1M3c	LINE	MER20	DNA	MLT1K	LTR
L1M3e	LINE	MER20B	DNA	MLT1L	LTR
L1M4	LINE	MER3	DNA	MLT2B1	LTR
L1M4b	LINE	MER31-int	LTR	MLT2B2	LTR
L1M4c	LINE	MER31B	LTR	MLT2C1	LTR
L1MA6	LINE	MER33	DNA	MLT2C2	LTR
L1MA7	LINE	MER34	LTR	MLT2D	LTR
L1MA8	LINE	MER50-int	LTR	MLT2E	LTR
L1MA9	LINE	MER53	DNA	MLT2F	LTR
L1MB1	LINE	MER54A	LTR	ORSL	DNA
L1MB2	LINE	MER54B	LTR	REP3	LTR
L1MB3	LINE	MER58	DNA	Tigger5	DNA
L1MB4	LINE	MER58A	DNA	Tigger6	DNA
L1MB5	LINE	MER58B	DNA	Tigger6a	DNA
L1MB7	LINE	MER58C	DNA	Tigger8	DNA
L1MB8	LINE	MER58D	DNA	U2	snRNA
L1MC/D	LINE	MER5A	DNA	U3	snRNA
L1MC1	LINE	MER5B	DNA	U5	snRNA
L1MC2	LINE	MER5C	DNA	U6	snRNA
L1MC3	LINE	MER8	DNA	Zaphod	DNA

Appendix B Discrepancy in estimates of evolution rate.

	Genomics Average	
	Consistent	Inconsistent
Human outgroup	1.020	1.252
Mouse outgroup	0.998	0.797