氏　　名	高橋　真保子
学位（専攻分野）	博士（理学）
学 位 記 番 号	総研大甲第 1431 号
学位授与の日付	平成２３年３月２４日
学位授与の要件	生命科学研究科　遺伝学専攻 学位規則第６条第１項該当
学 位 論 文 題 目	Identification and Characterization of Lineage-specific Highly Conserved Noncoding Sequences in Mammalian Genomes
論 文 審 査 委 員	主　　査　　教授　　明石　裕 教授　　小林　武彦 教授　　五條堀　孝 准教授　高野　敏行 教授　　植田　信太郎　東京大学

論文内容の要旨

Living organisms have various characteristics that define lineages. The change in regulatory elements is thought to play a major role in development of these lineage specific characteristics. From the inception of molecular evolutionary studies, noncoding regions were suspected to be involved in gene regulation. Recent studies of genome comparisons among diverged species revealed that there are many highly conserved noncoding sequences (HCNSs) in vertebrates, and many of them actually contain regulatory elements. Based on the observations, one of the candidates for regulatory elements which contributed to the lineage specific evolution is the HCNSs conserved only in one lineage because these lineage specific HCNSs may have gained new functions during the evolution of the lineage. However, unlike the HCNSs conserved in the large lineage such as vertebrates, HCNSs conserved only in a small lineage comprised of closely related species such as primates and rodents have not been well studied. That is, identification of lineage specific HCNSs provides a new insight for the evolution of the corresponding lineage of organisms.

I first analyzed human-macaque and mouse-rat pairwise noncoding alignments, and determined to use the 250bp window which was a minimum length to detect significant conservations in the closely related species. The thresholds for the conserved sequences in human-macaque and mouse-rat were ≥98.4% and ≥97.2% identity, respectively. As the first filtering for identification of lineage specific HCNSs, I extracted conserved sequences with the thresholds as primate and rodent specific HCNS candidates from the human-macaque and mouse-rat pairwise alignments. Using the extracted primate and rodent specific HCNS candidates as queries, I performed MegaBLAST search against 8 vertebrate genomes, and removed all HCNSs that were also conserved in non-primate or non-rodent vertebrate genomes. As the second filtering for false positive HCNSs, I further compared substitution numbers between each HCNS and its flanking region and extracted HCNSs that had significantly smaller substitution numbers than those of the flaking regions. After these filtering processes, I finally obtained 192 primate- and 331 rodent-specific HCNSs.

The SNP densities in primate and rodent specific HCNSs were significantly lower than those of genome averages. I analyzed the derived allele frequency (DAF) within the primate specific HCNSs. The purifying selection which is observed as the region with DAF $\leq$ 0.1 acting on HCNSs was stronger than that of corresponding DAF level of the entire genome. Although this increase of rarer allele fraction was not significant due to the small SNP observation on HCNSs (only 44 SNPs), this tendency is consistent with previously published results on vertebrate HCNSs. Given that lineage specific HCNSs have small numbers of SNPs and substitutions as well as not low level of DAF $\leq$ 0.1 compared to the genome average, it is more likely that lineage specific HCNSs are under constraint. This suggests that lineage specific HCNSs tend to be under purifying selection, implying that primate and rodent specific HCNSs harbor important functions.

I also examined whether there is any differences in the distributions of lineage specific HCNSs and ultraconserved elements (UCEs) because the UCE is an extreme example of highly conserved vertebrate HCNSs. The distributions of primate and rodent specific HCNSs and vertebrate HCNSs were completely different in the genomes, suggesting that these lineage specific HCNSs and vertebrate HCNSs are independently evolved sets.

To investigate the biological impact on the lineage specific HCNS on the evolution, I next examined the function of lineage specific HCNS-flanking genes (LHF genes). The statistically overrepresented functions of primate and rodent LHF genes were "anatomical development" and "transcriptional regulation", which was consistent with the characteristics of known vertebrate HCNSs. Notably, the synonymous (*dS*) substitution of primate and rodent LHF genes were significantly smaller than those of genome wide genes, as well as the non-synonymous (*dN*) and *dN*/*dS* ratio. I also found that UCE-flanking genes showed significantly smaller *dS* values than those of genome wide genes. This indicates that there are stronger constraints on the LHF genes and UCE-flanking genes at nucleotide level compared to genes that are not associated with HCNSs. Indeed, orthologs of primate/rodent LHF genes in rodents/primates, the majority of which have no HCNSs, showed the same level of *dS* values with genome wide genes. This strongly suggests that there is a correlation between HCNSs and low *dS* genes. Given that the functions of LHF gene are important in development, the strong constraint on LHF genes at nucleotide level may be a result of tight regulation of the gene expression. For instance, many regulatory proteins bind to the LHF genes to regulate the gene expression by interacting with HCNSs.

Interestingly, even though primate and rodent LHF genes showed similar functions to UCE-flanking genes, the majority of both LHF genes were different from the UCE-flanking genes. This suggests that independent sets of genes may have contributed to develop lineage specific characteristics. Conversely, the number of LHF genes which were shared by UCE-flanking genes was small but significantly larger than expected, and many of them were involved in nervous system development as transcriptional regulators. This suggests that certain groups of genes recruited new HCNSs in addition to old HCNSs which are conserved among vertebrates.

Based on the results in this study, I propose a possibility that the lineage specific evolution occurred through the creation of new lineage specific HCNSs near two categories of genes. The first category is lineage specific sets of LHF genes. The creation of lineage specific HCNSs expands the set of LHF genes which are involved in development, but different from that of ancestral (vertebrate) HCNSs. The second category is particular groups of ancestral HCNS-flanking genes. One of the major gene groups codes transcriptional regulators which are involved in nervous system development. The results in this study provide new insights into the lineage specific evolution through interactions between HCNSs and their LHF genes.

博士論文の審査結果の要旨

Detecting functional regions of genome sequences is a central challenge in bioinformatics. Nucleotide conservation among distantly related species has been used to identify candidates for functional DNA elements but few studies have attempted to find genomic regions that are strongly conserved within one taxa but diverged in others. Such regions may be newly evolved functional elements and are candidates for taxa-specific phenotypic characteristics. Takahashi's doctoral dissertation research focused on detecting primate- and rodent-specific conserved regions within non-coding DNA.

The abundance of mammal and vertebrate genome sequences allows identification of regions conserved in particular taxa. Takahashi used human-macaque and mouse-rat genome alignments to identify conserved 250bp blocks of non-repetitive DNA outside of known coding regions. This block size was chosen to be sufficient to allow detection of significant conservation but small enough to identify functional elements of limited size. Blocks that showed significant conservation outside of the taxa of interest (*i.e.*, among mammals or distantly related vertebrates) were removed from the "lineage-specific highly conserved non-coding sequence (HCNS)" group. This pipeline led to a set of roughly 200 primate- and 300 rodent-specific HCNSs.

Sequence conservation can reflect purifying natural selection or low region-specific mutation rate. Takahashi employed human DNA polymorphism data to test whether the frequency distributions of single nucleotide polymorphisms (SNPs) differ between lineage-specific HCNSs and other genomic regions. Similar frequencies of segregating mutations in these classes are consistent with either a low mutation rate and/or strong negative selection on new mutations within HCNSs. This finding suggests that weak purifying selection does not play a major role in lineage-specific HCNS evolution. In addition, reduced synonymous DNA divergence in protein-coding genes located close to HCNSs is consistent with region-specific reductions in mutation rates or shared constraints between non-coding elements and silent sites within particular protein coding genes.

Lineage-specific HCNS regions may harbor regulatory regions and Takahashi showed that "anatomical development" and "transcriptional regulation" are over-represented functional annotations of genes that flank these regions. Although some genes are located close to both lineage specific HCNSs and ultraconserved elements (UCEs), many genes are flanked by only lineage specific

HCNSs. Overall, this research has addressed a relatively unexplored aspect of genome evolution, taxa-specific conservation of small blocks of non-coding DNA, and has revealed a number of candidate functional non-coding regions for further experimental study. Rigorous identification of such regions will be a significant contribution to comparative genomics.