氏　　名	原 雄一郎
学位(専攻分野)	博士(学術)
学位記番号	総研大乙第225号
学位授与の日付	平成25年3月22日
学位授与の要件	学位規則第6条第2項該当
学位論文題目	Comparative Genomics of Closely Related Species – Uncovering the Evolutionary Process of Shaping the Characteristics Representing the Species
論文審査委員	主　査　准教授 大田 竜也 教授 颯田 葉子 教授 五條堀 孝 専任教授 今西 規　東海大学

A huge amount of species living on the earth display extensively diversified phenotypes. During the long period of evolution from their common ancestors, characteristic shaping each species, i.e. "species-ness", has been acquired over time. One of the biggest issues in biology, since the age of Darwin, is to unveil the processes of acquiring species-ness in the course of evolution. Several approaches have been developed to infer the states of ancestral species by comparing morphological and ecological characteristics and/or the genomic contents of extant species. Despite that much evidence has been accumulated, the processes of the evolution of shaping species-specific characteristics are still limitedly understood. This would be because each inferred state at the ancestor only illustrates a point of time in the ancestral lineage, which would be hard to reconstruct a whole picture of evolutionary processes from ancestors to extant species. Today, it becomes easier to infer the ancestral genomes from those of extant species, and much more information can be extracted because of the rapid increase of number of species whose whole genome has been sequenced and rapid accumulation of knowledge on biological function of genes, which enable the inference of phenotype to genotype. If the whole pictures of ancestral genomes are provided with high resolution in timeline of evolution, the continuous process of evolution could be reconstructed. This can be most easily attained for the group of closely related species whose genomes are available and the genomes of common ancestors are almost wholly inferred with high accuracy.

My aim in this study is to uncover the process of acquisition of "species-ness", the specific characteristics of species, comparing the whole genomes of closely related extant species as well as inference of ancestors. To achieve it, I designed an approach consisting of three integrative and sequential analyses based on the reconstruction of the genomes of ancestral species. The first is to distinguish different processes of genomic changes such as single point mutation, indels, and inversion to reveal the detail of structural evolution of genomes with finer precision. The second is to reconstruct the history of species based on refined genomic comparison attained by the first procedure. The third is to identify the species-specific genomic content that provides species to have each own species-ness. This can be carried out once the evolutionary history of species is firmly established by the second procedure.

In order to examine feasibility of above approaches, I conducted comparative genomic analyses of closely related species by selecting specific examples, some of which are currently controversial and much debated, as follows: (i) Identification of ultramicro inversions within local alignments between closely related species, (ii) Reconstruction of the demographic histories of the human lineage using whole genome alignments, and (iii) Identification of the species-specific characteristics involved in the pathogenicity and adaptation to the host environments in *Theileria* parasites.

(i) *Identification of ultramicro inversions within local alignments between closely related species.*

Inversion is one of the major mechanisms for generating genomic diversity in evolution. While the inversions of large size have been well investigated since the early 20th century, little is understood about the minimal size of inversions, which would have useful information for clarifying the minute structural changes of genomes. I developed an efficient method for identifying minimal-sized inversions that I call "Ultramicro Inversions" of 5-125 bp buried in nucleotide alignments, and identified 3,330 ultramicro

inversions within the human-chimpanzee genome alignments. Around 26% of the ultramicro inversions consisted of adenine (A) and thymine (T) only, and the ultramicro inversions were also frequently found in chromosome Y and regions close to transposable elements. These observations suggested that the ultramicro inversions are related to instability of the genomic structures. Ninety ultramicro inversions were found in gene regions, and 28 out of 90 were in the coding regions, indicating that some parts of the ultramicro inversions may contribute to gene evolution. At least 31% of the ultramicro inversions in the human-chimpanzee alignments were bounded by inverted repeats, suggesting that such ultramicro inversions involved in the chromosomal recombinations via DNA stem-loops. In addition, I identified ultramicro inversions in various lineages other than primates: 1,285, 40, and 62 ultramicro inversions in the fly, fungi, and rice genomes, respectively, and 20 on average in the genomes of four and two lineages of eubacteria and archaea, respectively. This observation indicates that ultramicro inversions are ubiquitous across the three domains of the living world. While frequencies of the ultramicro inversions were up to seven times different between the lineages, the mechanisms of ultramicro inversions seemed to be more various across the lineages. The fractions of AT-exclusive and stem-loop type ultramicro inversions were much more different across the lineages. Identification of ultramicro inversion hotspots in silico would be helpful for capturing the inversions in experiments and clarifying the mechanisms of minute genome structural evolution. Our inversion-identification method is also applicable in the fine-tuning of genome alignments by distinguishing ultramicro inversions from simple point mutations and indels.

(ii) *Reconstruction of the demographic histories of the human lineage using whole genome alignments.*
The demographic history of human would provide helpful information for identifying the evolutionary events that shaped the humanity but remains controversial even in the genomic era. In order to settle the controversies, I inferred the evolutionary history of human and great apes based on an estimation of the speciation times (T) and ancestral population sizes (N) in the lineage leading to human and great apes using the whole-genome alignments. A coalescence simulation determined the sizes of alignment blocks and intervals between them required to obtain recombination-free blocks with a high frequency. This simulation revealed that the size of the alignment block strongly affects the parameter inference, indicating that recombination is an important factor for achieving optimum parameter inference and that this simulation is helpful for the optimum data collection. From the whole genome alignments (1.9 giga-bases) of human (H), chimpanzee (C), gorilla (G), and orangutan, and the small-sized regions subject to the genomic changes by the other mechanisms than point mutations, such as CpG sites and ultramicro inversions, were excluded. 100-bp alignment blocks separated by ≥5-kb intervals were sampled from the alignments and subjected to estimate $\tau=\mu T$ and $\theta=4\mu gN$ using the MCMC method, where μ is the mutation rate and g is the generation time. Although the estimated τHC differed across chromosomes, τHC and τHCG were strongly correlated across chromosomes, indicating that variation in τ is subject to variation in μ across the lineages, rather than T, and thus, all chromosomes are likely to share a single speciation time. Subsequently, I estimated Ts of the human lineage from chimpanzee, gorilla, and orangutan to be 6.0-7.6, 7.6-9.7, and 15-19 MYA, respectively, assuming variable μ across lineages and chromosomes. These speciation times were consistent with the fossil records. I conclude that the speciation times in our recombination-free analysis would be conclusive and the speciation between human and chimpanzee was a single event.

(iii) *Identification of the species-specific characteristics involved in the pathogenicity and adaptation to the host environments in Theileria parasites.*

Theileria is a tick-born apicomplexan group causing parasitosis in livestock. Some theilerias are parasitic to cattle, but the relationship between the theileria and cattle seem to have evolved specifically in each lineage. While *T. annulata* and *T. parva* (transforming theileria) induce abnormal proliferation of infected cells of lymphocyte or macrophage/monocyte lineages and are severely pathogenic, *T. orientalis* does not induce such transformation and shows moderate pathogenicity. Here, in order to clarify the process of acquiring the high pathogenicity and diverged systems infecting the hosts, I reconstructed the evolutionary history of theileria based on the comparative genomics of the almost whole genomes. While synteny across the chromosomes of the three theilerias was well conserved, subtelomeric regions were largely different: *T. orientalis* lacks the large tandemly arrayed subtelomere-encoded variable secreted protein-encoding gene family. Through the orthologue clustering, in addition, I found that duplication and deletion rates in the transforming theileria lineage were 1.66 and 1.95 times faster than those in the *T. orientalis* lineages, respectively. Expansion of particular gene families by gene duplication was found specifically in the two transforming theileria species. One of the most notable families is the TashAT/TpHN gene family, which is considered to be involved in transformation and abnormal proliferation of host leukocytes. The transforming theileria possessed around 20 TashAT/TpHN members, while only one member was identified in *T. orientalis*, and no homologues were found in a babesia and plasmodiums. I also found the gene families expanded specifically in *T. orientalis* lineages such as ABC transporters, implying species-specific strategies against host systems. Differences between the genome sequences of theileria species illustrated different tempo and mode of gene duplication and deletion between transforming theilerias and *T. orientalis*. It is implied, moreover, that such differences in evolutionary modes resulted in the novel abilities to transform and immortalize bovine leukocytes. The genomic changes between close relatives will provide insight into proteins and mechanisms that have evolved to induce and regulate this process.

In the above studies, I have examined and demonstrated effectiveness of the three steps of integrative and sequential approach for clarifying the evolutionary processes to attain "species-ness" at the genome level by reconstructing the genomes of ancestral species from closely related extant species. Even though the current approaches are based on the well-established fields of genomics, population genetics, and molecular phylogenetics, the integration of the approaches, as shown here, is innovative in the field of in silico genomic analysis and provides new insight on evolutionary biology. In the near future, comparative analysis of closely related species will be expanded for the genomes of species suitable to solve a particular biological issue. The integrative approach provided here would become one of de-facto standard for such analyses.

博士論文の審査結果の要旨

ゲノム情報や組織・器官・細胞での遺伝子発現情報が集積している現在、これらのビッグデータから適切に情報を抽出することが求められている。本博士論文では、その一つとしてゲノム情報から生物進化を研究すること、特に現存の生物がどのようにその種に特有なゲノムの特徴を獲得するに至ったかという進化過程を明らかにする手法を確立し、その進化過程の生物学的な意義を考察することを目的に研究が行われた。出願者は既存の集団遺伝学・分子進化学・ゲノム科学の手法と共に自身が改良した方法を用い、近縁種のゲノムを比較し、（1）進化の過程で生じたゲノムの変化を正確に推測する、（2）適切なデータを標本抽出・解析し生物の進化史を推測する、（3）各進化系統で獲得あるいは削除されたゲノム領域を同定しその生物学的な意味を明らかにする、という3段階で研究を行った。その詳細は次のとおりである。

（1）既存の塩基配列の整列（アライメント）法をもとに微小サイズの逆位を考慮した解析法を開発した。これにより正確に点突然変異、挿入・欠失、逆位を同定することを可能とした。コンピューター・シミュレーションで解析法の精度を確かめた上でヒト、チンパンジー、ゴリラ、オランウータンのゲノム比較に応用した。その結果、超微小サイズの逆位が霊長類の進化の中で多く生じてきたこと、それらがゲノムの比較的不安定な領域で多く観察されること、また僅かではあるが一部の遺伝子のタンパクをコードする領域にも存在することを示した。さらに同方法を他の生物グループのゲノム比較にも応用し、超微小サイズの逆位は一般に生物で観察される現象であることも明らかにした。

（2）全ゲノム配列データから生物の進化史、すなわち進化における集団動態と種の分岐年代を推定することを試みた。この推定に際しては遺伝的な連鎖の効果をできるだけ小さくする必要があり、そのための標本抽出法を開発した。例として生物学的に最も重要な課題の一つであり、多くの議論を生んできた霊長類、特にヒトとチンパンジーの種分化過程に関する解析を行った。その結果、現在のゲノムのデータはヒトとチンパンジーの分岐後の遺伝的な交流を支持するものではないこと、すなわち遺伝的な距離が染色体で異なるのは分岐時間が異なるためではなく染色体ごとで塩基置換率が異なることで主に説明できることを示した。

（3）推定された生物の系統関係にもとづいて各生物種で獲得あるいは削除されたゲノム領域を明らかにし、それぞれの種を特徴づける形質の解明を試みた。その一例として寄生性原生生物の *Theileria* のゲノム解析を行った。理由は *Theileria* のゲノムが小さく、さらに遺伝子の機能が比較的推察しやすい、あるいは近い将来に機能が明らかにされることが期待できるためである。そして *Theileria* のゲノムの進化では倍数化やゲノム重複は観察されないこと、多重遺伝子族の拡張・縮小のパターンが各系統によって異なることなどを示した。血球に感染した時期に遺伝子発現が上昇するため宿主細胞の形質転換に関与されていることが示唆されている TashAT/TpHN 遺伝子族の解析では、これらの遺伝子族が近縁種の *Babesia* などには存在せず、また低病原性の *Theileria* ではシングルコピーであるのに対し高病原性の *Theileria* の2種において数が増加していたことを示した。これ以外にも幾つかの遺伝子が高病原性の *Theileria* の2種において数が増加しており、今後の機能解析の研究対象としてあげられた。同様の多重遺伝子族の拡張・欠失は低病原性の

Theileria にも観察されており、高病原性の *Theileria* に見られる程頻度は高くないものの各々の系統に特有の形質の獲得に貢献していることも示した。

公開発表会では以上の学位申請論文の内容に沿い、主にヒトを含む霊長類のゲノム解析の結果を発表した。また公開発表、発表後の質疑応答、および口頭試問では、本研究の内容・意義、既存の研究の問題点などを論理的に解説し、その内容は的確であった。本論文で行われた研究は、それぞれの生物種が固有の形質を獲得した進化過程を明らかにするための解析方法を開発し、理論的な背景を踏まえてその方法を実際に生物学的に重要な問題に応用した点で優れている。今後ここで用いられた方法により各生物種のゲノムの特異的な特徴を明らかにすること、さらにそれぞれの遺伝子の機能情報と統合することで個々の生物がどのように独特の形質などの表現型を獲得したかを明らかにすることが可能となることが期待され、本研究は今後の進化研究に大いに貢献するものと判断された。

以上により、本論文は学位を授与するに十分な水準に達していると審査員全員一致で判断した。