World of Knowledge

What Makes Us Human?
Introduction to Insertion/Deletion Events between Human and Chimpanzee

Trees-Juen Chuang

Associate Research Fellow

Genomics Research Center

Introduction

The Family Hominidae includes human (*Homo sapiens*) and great apes: chimpanzees (or common chimpanzees), bonobos (or pygmy chimpanzee), gorillas, and orangutans. The approximate divergence times of these primates are shown in Figure 1. In this article, "chimpanzee" refers to common chimpanzee (*Pan troglodytes*) because it is the species in most *Homo-Pan* comparative studies. Human and chimpanzee diverged from each other 4~6 million years ago (MYA). This evolutionarily short time, though resulting in a small genetic distance, has conferred large phenotypic differences between the two species. The studies on human-chimpanzee genetic differences are fascinating because they can, at least in part, help answer the ultimate question: what makes us human? The recent publication of the chimpanzee genome draft (reported by The Chimpanzee Genome Sequencing and Analysis Consortium 2005 or TCGSAC 2005) has brought unprecedented opportunities for investigating the genetic basis of the morphological and behavior differences between human and chimpanzee. TCGSAC reported that the two species may have 98~99% of DNA sequences in common. The limited genetic distance appears doubtful given the "large" phenotypic divergence.

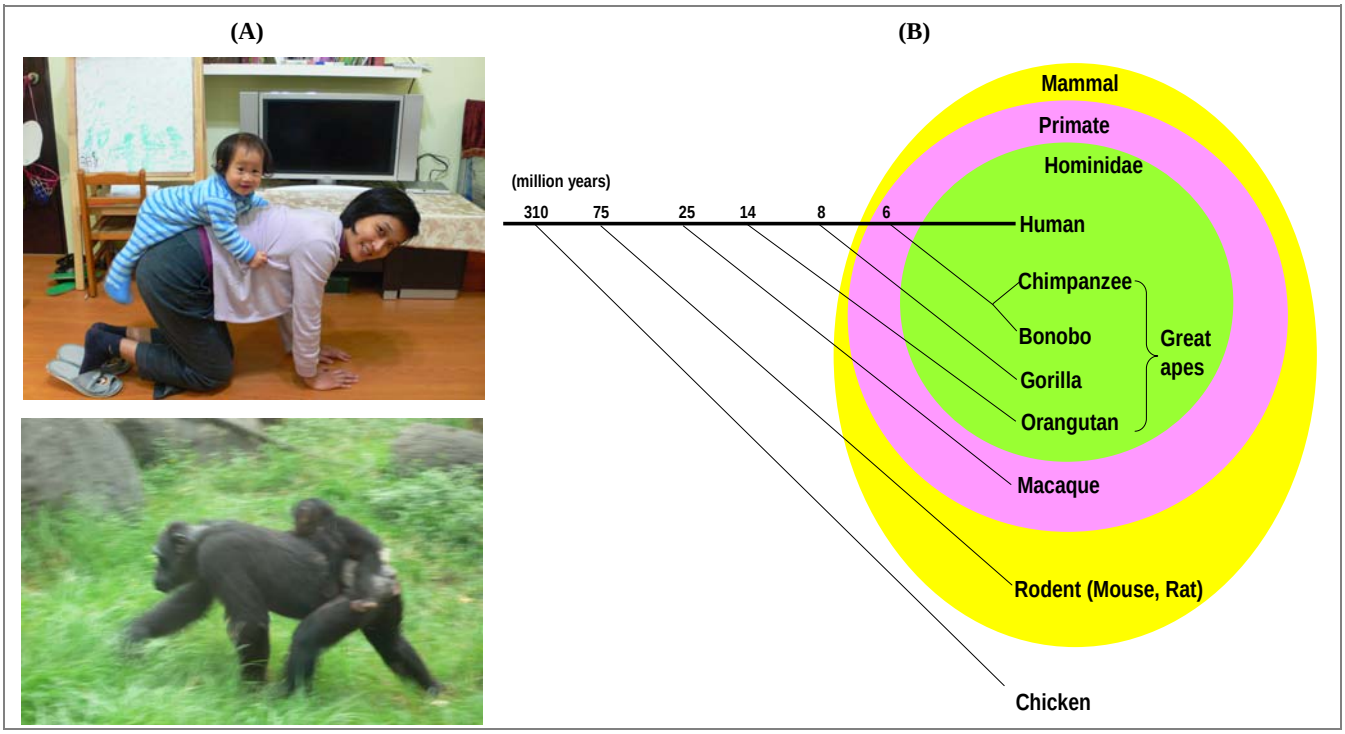

Figure 1. (A) Human and chimpanzee; (B) Evolutionary relationships among humans, great apes, and other animals. The approximate divergence times are illustrated (the branch lengths are not to the scale).

In fact, the evolution of hominoid genomes is far more complex than simple nucleotide substitutions. As we will discuss below, the human and chimpanzee genomes actually have undergone extensive structural rearrangements, gene/segmental duplications, and insertion/deletion (indel) events in addition to nucleotide substitutions. Furthermore, inter-species differences can also occur at other levels, including transcriptome, proteome, epigenome, metabolome, and interactome, all of which could have contributed significantly to functional divergence. Therefore, *Homo-Pan* divergence at the DNA level represents but one piece, although an important one, of the puzzle. In this article, we mainly introduce our recent findings in insertion/deletion events between these two primates.

Insertions and Deletions

The TCGSAC has identified ~ 5 million indel events between the human and chimpanzee genomes. These indels comprise ~ 90 Mb of genomic sequences (40-45 Mb euchromatic sequence in each species) and correspond to a markedly larger *Homo-Pan* genetic distance than that caused by nucleotide substitutions (~ 3% vs. ~ 1.23%). The vast majority of these identified indels are small, with 45% covering only 1 bp, 96% being < 20 bp, and only 1.4% being ≥ 80 bp. Although the number of indels ≥ 80 bp is relatively small (~70,000), they comprise as high as 73% of all indels in terms of length. The majority of these indels are caused by the activities of repetitive elements (REs), which are DNA sequences that can move to different genomic regions. The major RE classes include satellites/microsatellites and four retrotransposons: short interspersed repetitive elements (SINEs), long interspersed repetitive elements (LINEs), SVE elements, endogenous retroviruses (ERVs). Among these REs, *Alu*, LINE-1 (L1), and SVA insertions account for over 95% of the transposon-derived insertions in the human and chimpanzee genomes. In sum, retrotransposons are abundant in the primate genomes. It has been reported that about 45% of the human genome is composed of retrotransposon sequences.

Human-specific Indels

So far, we have discussed indels that occur between the human and chimpanzee genomes. Most of these indels were identified based on pair-wise comparisons, which cannot distinguish between insertions and deletions. Furthermore, the TCGSAC-published chimpanzee genome in 2005 was a draft version. *Homo-Pan* indels inferred based on this draft actually include a fairly high false positive rate (estimated to be 15~19%). Inclusion of highly accurate outgroup genomic sequences such as the mouse genome can help resolve both of these problems, and indicate species-specificity for the identified indels. Human-specific (HS) indels can thus be inferred from multiple sequence alignments. On the basis of a five-way comparison that includes the human, chimpanzee, dog, mouse, and rat genomes, more than 840,000 HS small indels (i.e., indels < 100 bp) have been identified, which account for ~0.21% of the *Homo-Pan* genomic divergence. Note that in coding sequences (CDSs) and pseudogenes, HS indels result in 0.03% and 1.4% of sequence divergence, respectively. Therefore, most of the HS indels, particularly the CDS HS indels, may have been eliminated by natural selection. As expected, the distribution of HS indels among different genomic regions reflects the levels of selection pressure imposed on these regions. Most of the HS indels occur in intergenic regions (> 512,000 events),

followed by intronic regions (> 318,000), 3'UTR (3' untranslated region; > 9,700), 5'UTR (5' untranslated region;> 2,300), and lastly by CDS (> 1,700). It is noteworthy that the numbers of intergenic and intronic HS indels are underestimated because these regions are underrepresented in multiple sequence alignments. Collectively, the exonic indels (CDS+3'UTR+5'UTR) affect more than 7,000 UCSC-annotated human genes (> 11,000 transcripts). This large number of HS indel-affected genes is unexpected because it represents ~37% of the UCSC-annotated genes. Also surprising is that as many as ~55% of the CDS HS indels have lengths not divisible by 3. However, this high percentage can be an overestimate because of the relatively low quality of the chimpanzee genomic sequences. Functional analysis reveals that HS indels are enriched in genes related to transcription regulatory activity, translation regulatory activity, and viral life cycle, while underrepresented in catalytic activity and transporter activity. Therefore, HS indels may have caused *Homo-Pan* divergence through regulations of gene expression and translation. Furthermore, considering that human and chimpanzee respond differently to viral infections (e.g. HIV and hepatitis B/C virus infections), the HS indels that affect viral life cycle-related genes may be worth further functional studies.

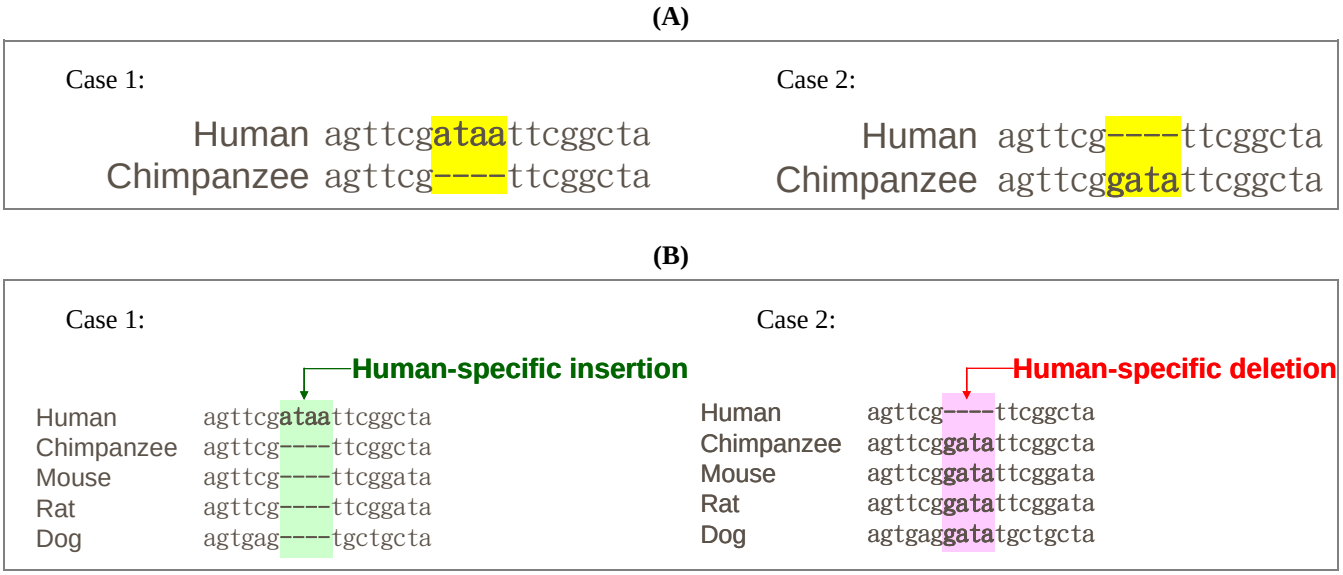

Figure 2. (A) indels identified based on pair-wise comparisons; (B) human-specific indels inferred from multiple sequence alignments.

Concluding Remarks

The "nucleotide divergence" between human and chimpanzee is in fact a mixture of various genetic changes driven by different evolutionary forces and molecular mechanisms. The differences between the two genomes are far more complex than the generally cited "1% genetic distance". This article provides an overview of indel events that contribute much more than 1% *Homo-Pan* genetic distance. However, nucleotide divergence alone rarely provides direct evidence for phenotypic divergence between the two species. Integration of multiple disciplines, including genomics, proteomics, systems biology, and gene-environment interactions will probably bring us closer to some of the answers. We sincerely welcome graduates or graduate students with **computer science or engineering** backgrounds to join us. For further details, please visit our research web site (http://www.sinica.edu.tw/~trees/) and see our related publications as follows.

1. Feng-Chi Chen and Trees-Juen Chuang* (2008). Nucleotide Sequence Divergence between Humans and Chimpanzees. Encyclopedia of Life Sciences (ELS). Invited review article.
2. Feng-Chi Chen, Chueng-Jong Chen, and Trees-Juen Chuang* (2007). INDELSCAN: a web server for comparative identification of species-specific and non-species-specific insertion/deletion events, *Nucleic Acids Research*, 35 (Web Server issue):W633-8.
3. Feng-Chi Chen, Chueng-Jong Chen, Wen-Hsiung Li*, and Trees-Juen Chuang* (2007). Human-specific insertions and deletions inferred from mammalian genome sequences. *Genome Research*, 17(1), 16-22.