Human-Chimp Speciation
Dec. 5, 2006

Relevant Literature

- **Inferring the Mode of Speciation From Genomic Data: A Study of the Great Apes**, Naoki Osada and Chung-I Wu
- **Genetic evidence for complex speciation of humans and chimpanzees**, Nick Patterson, Daniel J. Richter, Sante Gnerre, Eric S. Lander & David Reich

Outline

1 Terms
2 First Paper
   2.1 Main contribution
   2.2 Allopatric vs parapatric speciation
   2.3 Maximum likelihood estimation
   2.4 Results
3 Second Paper
   3.1 Main contribution
   3.2 Inferring ancient speciation from genetic data

1 Terms

*Species*: Species are groups of actually or potentially interbreeding populations, which are reproductively isolated from other such groups.” (Mayr, 1942)

*Speciation:*
The evolutionary process by which new biological species arise. There are three modes of speciation:

1. **Allopatric (Greek, from *allos*, other):**
This kind of speciation happen when a geographical barrier separates a population A into two isolated subpopulations A1 and A2 (where no gene flow can occur between A1 and A2 as a result of the barrier) and these subpopulations evolve into new populations B and
C where they are so different from each other due to evolution that they no longer can interbreed when the barrier is lifted.

2. Parapatric (Greek from para, beside):
This kind of speciation happens when a population spreads into neighboring area where they have not yet adapted to. If the environmental conditions in the neighboring areas are significantly different, it will introduce new selective pressure criteria on the population in that area. As a consequence, the two populations evolve side by side to significantly different species. The final population of these species is continuous; however, the species within the population will have preferential mating with their own kind.

3. Sympatric (Greek sun-, from sun together):
In sympatric speciation, a population of species splits into two (closely related species) without any geographical separation of the ancestral species. One explanation is that within a population there will be a fitness difference between species for a particular environmental condition that will cause the two species kind to evolve into two non-interbreeding species.

*Divergence*:
Divergence time is shown in the following figure:
This document is organized such that we first mention the main contribution of each paper that the presenter discussed and then continue following his steps to convey the contents of the paper.

2 First paper: Inferring the Mode of Speciation from Genomic Data: A Study of the Great Apes

2.1 Main contribution
In this paper, the authors investigate the pattern of divergence in order to test the allopatric model of speciation. Uniformity in the divergence time across genomic regions is a strong predictor of allopatric model of speciation. Using 345 coding sequences (CDS) and 143 intergenic sequences (IGS) from the African great apes, the authors reject the null hypothesis that the divergence time in the CDS and IGS is the same between human and chimpanzee. Their conclusion is that this difference between convergence time suggests that there has been a prolonged period of genetic exchange during the formation of these two species.

2.2 Allopatric vs. parapatric speciation
In order to understand the null and alternative hypothesis of the paper, let us revisit allopatric and parapatric speciation. In allopatry, there is no gene flow beyond the time of separation. All genes have diverged for a fixed time $t$ and further coalesce with an average length of $2N_e$ generations. Under the parapatric model, gene flow between nascent species is possible over a period of time. In the figure below the intensity of shade indicates the strength of the barrier to gene flow.

The underlying assumption is that for genomic regions associated with reproductive incompatibility and hence speciation (such as CDSs), it is more likely that gene flow ceased early. For regions free of such association (including most IGSs), gene flow may continue until relatively late.
2.3 Maximum-likelihood estimate of divergence time and ancestral population size

Let $k_i$ be the number of synonymous changes for the $i$th sequence in either coding sequence (CDS) or intergenic sequence (IGS). The probability of observing $k_i$ is given by a mixture of Poisson distribution in the divergence portion ($\tau_i$) and the mismatch distribution in the coalescence portion ($\theta_i$).

$$P(k_i) = \frac{e^{-l_i\tau_i}}{1 + l_i\theta_i} \sum_{d=0}^{k_i} \frac{(l_i\tau_i)^d}{d!} \left( \frac{l_i\theta_i}{1 + l_i\theta_i} \right)^{k_i-d},$$

where $\tau_i = 2tu_i$ and $\theta_i = 4Neu_i$. $l_i$ is the length of sequence $i$ and $u_i$ is the per/nucleotide substitution rate for the $i^{th}$ sequence.

The joint likelihood of divergence time and ancestral population size is found by the following:

$$L(\tau, \theta) = \ln \prod_{i=1}^{m} P(k_i)$$

Maximum Likelihood estimates of the two parameters were calculated using numerical iteration.

The objective of the authors to find the ML estimates were to test if $t$ (in $\tau_i = 2tu_i$) is the same between coding and intergenic regions. Since $u$ can change between the two regions, they define $\gamma = t/\theta = t/2N_e$ and compare $\gamma$ between the two regions. $\gamma$ is the relative divergence after speciation and should be constant across the two regions for the null hypothesis of allopatry (refer to figure in previous page).

2.4 Results

The results of the ML estimation show that the null hypothesis (allopatric model) has a lower likelihood than the alternative hypothesis (parapatric model). Hence, the authors reject the null hypothesis. The presenter (Frank) discussed his doubts about their result and how the likelihood difference is too small to make strong conclusion about the model.
3 Second Paper: **Genetic evidence for complex speciation of humans and chimpanzees**

### 3.1 Main contribution

Significant information regarding the timing and process of speciation is conveyed from the large variations across the genome of two or more species. In this paper a framework for studying this variation has been developed. The analysis of this paper shows that human-chimpanzee speciation happened less than 6.3 years ago. This is in conflict with some interpretations of ancient fossils. Noticeably, chromosome X shows an extremely young genetic divergence time. These controversial results can be explained if the chimpanzee and human lineages exchanged genes after their initial divergence and before separating permanently.

### 3.2 Inferring ancient speciation from genetic data

The following issues must be considered while studying the genetic divergence at any position.

- The genetic divergence has to be corrected for local variation in the neutral mutation rates across the genome
- Any estimate of local genetic divergence should be corrected for the effects of recurrent mutation.
- The variability of the estimate should be assessed by studying large enough subsets of the genome.

#### Table 2

<table>
<thead>
<tr>
<th></th>
<th>Human-chimp (n_c = 345, n_l = 143)</th>
<th>Human-gorilla (n_c = 76, n_l = 53)</th>
<th>Chimpanzee-gorilla (n_c = 76, n_l = 53)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau_{cBS}$</td>
<td>0.005855</td>
<td>0.01299</td>
<td>0.01317</td>
</tr>
<tr>
<td>$\theta_{cBS}$</td>
<td>0.00454</td>
<td>0.00500</td>
<td>0.00380</td>
</tr>
<tr>
<td>$\gamma_{cBS}$</td>
<td>0.00876</td>
<td>0.01004</td>
<td>0.01204</td>
</tr>
<tr>
<td>$\gamma_{cBS}$</td>
<td>0.00466</td>
<td>0.00421</td>
<td>0.00347</td>
</tr>
<tr>
<td>$\gamma_{cBS}$</td>
<td>1.88</td>
<td>2.60</td>
<td>3.47</td>
</tr>
<tr>
<td>$\ln L$</td>
<td>$-1093.971$</td>
<td>$-276.117$</td>
<td>$-270.641$</td>
</tr>
<tr>
<td>$\ln L$</td>
<td>$-1093.970$</td>
<td>$-276.114$</td>
<td>$-269.906$</td>
</tr>
<tr>
<td>$P$</td>
<td>0.027</td>
<td>0.950</td>
<td>0.224</td>
</tr>
</tbody>
</table>
In this study, a shotgun sequence from gorilla was compared against human, chimpanzee, orangutan and macaque. All divergent sites, places at which two alternative alleles were observed across the aligned sequences of the species, were analyzed. The divergent sites were categorized according to how they partitioned the species (Table below).

**Table 1 | Main data sets in the study**

<table>
<thead>
<tr>
<th>Class</th>
<th>Pattern</th>
<th>Species Bases</th>
<th>Autosomes</th>
<th>X chromosome</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>HCGOM</td>
<td>HCGM</td>
<td></td>
</tr>
<tr>
<td>$n_H$</td>
<td>10000</td>
<td>28,504</td>
<td>59,175</td>
<td>936</td>
</tr>
<tr>
<td>$n_C$</td>
<td>01000</td>
<td>28,495</td>
<td>59,844</td>
<td>944</td>
</tr>
<tr>
<td>$n_G$</td>
<td>00100</td>
<td>38,677</td>
<td>81,671</td>
<td>1,430</td>
</tr>
<tr>
<td>$n_{HC}$</td>
<td>11000</td>
<td>8,561</td>
<td>20,408</td>
<td>457</td>
</tr>
<tr>
<td>$n_{HG}$</td>
<td>10100</td>
<td>1,302</td>
<td>4,809</td>
<td>14</td>
</tr>
<tr>
<td>$n_{CG}$</td>
<td>01100</td>
<td>1,430</td>
<td>4,600</td>
<td>12</td>
</tr>
<tr>
<td>$n_{HCG}$</td>
<td>11100</td>
<td>41,928</td>
<td>1,493</td>
<td></td>
</tr>
<tr>
<td>$n_O$</td>
<td>00010</td>
<td>82,670</td>
<td>3,086</td>
<td></td>
</tr>
<tr>
<td>$n_M$</td>
<td>11110</td>
<td>244,270</td>
<td>596,939</td>
<td>9,621</td>
</tr>
<tr>
<td>$n_{HO}$</td>
<td>10010</td>
<td>412</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>$n_{CO}$</td>
<td>01010</td>
<td>397</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>$n_{GO}$</td>
<td>00110</td>
<td>764</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>$n_{HCO}$</td>
<td>11010</td>
<td>1,347</td>
<td>56</td>
<td></td>
</tr>
<tr>
<td>$n_{HGO}$</td>
<td>10110</td>
<td>989</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>$n_{CGO}$</td>
<td>01110</td>
<td>872</td>
<td>32</td>
<td></td>
</tr>
</tbody>
</table>

Each divergent site class is designated by a string of 0s and 1s, the bases seen in human-chimpanzee-gorilla-(orangutan)-macaque. The macaque allele is defined to have state '0'.

The proportion of each class would be strictly proportional to relative branch length of the genealogical tree, if all divergent sites were due to single historical mutations. However, the correspondence is not exact because some sites are due to more than one mutation. These alignment data showed six classes of divergent site that could not be explained by single historical mutations. One example is HO (Human and Orangutan). These recurrent mutations have a distorting effect on short branches. The effect of
recurrent mutations on short branches was estimated by using five species alignment data (HCGOM). Past studies have failed to consider recurrent mutations and thus they overestimate the proportion of the genome in which humans and chimpanzee are not most closely related.

The divergence can help define the relative branch lengths (depicted in the above figure).

**Large variation in divergence time across genome/ Large reduction in divergence time on chromosome X**

The genetic divergence time between two species varies across the genome and is always greater than or equal to the speciation time, which is the time of last gene flow between species.
In this study, subsets of the genome were selected where the divergence was different from the average. Then the authors calculate the average values of estimated average of local genetic divergence across each subset and divided by the estimated genome divergence to obtain relative age, $A$, compared with the autosomal average. The authors began by considering subsets of the genome consisting of the neighborhoods of HC sites. Then they considered the neighborhoods of HG or CG sites. The figure below shows the percentage of divergence for different sites with respect to average genome divergence.

Authors also considered the divergence along individual chromosomes, especially chromosome X. The relative divergences for the autonemies are all close to the average as shown in the figure below. The divergence is reduced along nearly the entire length of chromosome X. Human-chimpanzee genetic divergence is lower on chromosome X than on every other chromosome, even lower than the theoretical expectation. By contrast, the gorilla chromosome X comparison shows no decrease beyond the expectation from theory. The lower mutation rate on X chromosome has been seen in many comparisons however not in the same magnitude as human-chimp comparison. This discrepancy can be explained if the low divergence on X chromosome is a result of a low time divergence.

Moreover, the rate of HG and CG sites is also greatly reduced along chromosome X, which is consistent with humans and chimpanzees being most closely related essentially everywhere along chromosome X.
Implications for current models of human-chimp speciation

The assumption that human-chimpanzee genetic divergence varies over more than 4 Myr, and that genetic divergence is about 1.2 Myr less on chromosome X than the autosomes, introduces two interesting issues about human-chimpanzee speciation. Firstly, these results put an upper bound on the age of human-chimpanzee speciation which is in conflict with some inferences from the fossil record. Secondly, the properties of chromosome X shows an unusual evolutionary history around the time of human-chimpanzee ancestral speciation, which points to the explanation that the structure of the population around the time of speciation was very different from that in modern human or ape.

Possible hybridization in the human-chimp lineage

A controversial explanation of the results is that the hominin and chimpanzee lineages initially separated but then exchange genes before finally separating less than 6.3 Myr ago.
In conclusion, the presenter criticized the paper's method for measuring divergence time and how certain variables involved in measuring divergence are correlated and one cannot independently measure the dependence of divergence time on one specific variable (HC site for example). He also challenged the conclusion of the paper by pointing out a recent article by Barton:

“Such a scenario [that human lineage hybridized with the chimp lineage]— or a range of kinds of population subdivision—can indeed account for diversity among loci in divergence, but Patterson et al. do not test whether their data are consistent with the simple null model of abrupt allopatric speciation of a single well-mixed population. A simple calculation (H. Innan, personal communication) shows that their data are consistent with an ancestral effective population size of $N_e \sim 45,000$, which does not seem unreasonable, and is consistent with previous studies. Thus, there is no statistical evidence for hybridization.”(Curr Biol. 2006 Aug 22;16(16):R647-50)