Sequencing the Chimpanzee Genome

David E. Reich and Eric S. Lander, Whitehead Institute/MIT Center for Genome Research; Robert Waterston, Washington University Genome Sequencing Center; Svante Pääbo, Max Planck Institute for Evolutionary Anthropology; Maryellen Ruvolo, Department of Anthropology, Harvard University; Ajit Varki, Glycobiology Research and Training Center, University of California, San Diego.

Sequencing the Chimpanzee Genome

SUMMARY: Sequence of the chimpanzee *(Pan troglodytes*) genome should be a high priority. The chimpanzee genome sequence will have a major impact on our understanding of human disease, human evolution, and human population genetics. We propose that the first phase of a project to obtain chimpanzee sequence consist of 3-4X whole-genome shotgun coverage, with the sequence obtained from multiple individuals to provide valuable information about polymorphism rates within the species.

1. Introduction

A number of scientists have recently called for the sequencing of the chimpanzee genome as a major complement to the human genome [1,2,3,4,5,6] . These articles are included as an Appendix. Because the issues have already been extensively discussed in the literature, the goal of this white paper is (i) to summarize the key points and to add a few additional points and (ii) to propose a specific sequencing plan.

NHGRI's instructions list eight "points to address" in a white paper. The chimpanzee sequence will make major contributions on five of these points: (1) Improving human health; (2) Informing human biology; (3) Informing the human sequence; (5) Expanding our understanding of basic biological processes relevant to human health; (8) Expanding our knowledge of evolutionary processes in general and human evolution in particular.

The community that will use this information is huge. It includes medical geneticists, population geneticists and evolutionary biologists. Below, we elaborate on the scientific justification and propose a sequencing strategy.

2. Scientific utility of chimpanzee sequence.

A sequence of the chimpanzee genome would be extremely valuable for a variety of reasons. Because chimpanzees are the most closely related species to humans[7,8], they provide unique types of information that are not possible to obtain from the genomes of other primates.

2.1 Identification of sites of sequence difference. The most immediate use of a chimpanzee sequence would be the identification of sequence differences between the human genome and that of our closest living relatives. Many studies (e.g. ref. 9) have confirmed that the two genomes differ by about 1.2% – corresponding to a total of about 40 million (M) single base pair substitutions, together with some insertions, deletions and rearrangements.

Of course, it is no simple matter to identify which (or even how many) of the 40 M differences are functionally important. There are no comprehensive methods for recognizing all functionally important sites. Nonetheless, there are productive starting points that can be applied and will in turn provoke the development of improved methods. Moreover, the availability of the inventory of sequence differences will stimulate research into this important topic. We note some of the most obviously productive lines of research:

- **Newly created genes in one of the specie**s. New genes can arise from local duplication. An example is the zeta-globin gene in human[5].
- **Pseudogene in one of the two specie**s. Pseudogenes can arise in one of the two species by recent mutation. An example is the CMP-sialic acid hydroxylase (functional in chimpanzee, pseudogene in human), as shown by Varki and colleagues[10].
- **Deletions in one of the specie**s, leading to gene loss.
- **Non-conservative amino acid change**s, especially at positions that are highly conserved in evolution with more distantly related mammals or at known functional sites[10].
- **Changes in regulatory sequence**s. Considerable progress is being made in using interspecies comparisons (e.g., among human, mouse and rat) to identify important regulatory regions in mammalian genomes[11,12]. As regulatory regions are identified, it will then be possible to recognize differences that alter the sequences in functionally significant ways.

About 1-2% of the human genome is known to be in the coding regions of genes[13]. In addition, the subset of the genome which is in regulatory regions—which can be identified by comparison of human and mouse genomes—is not likely to be much more extensive than coding regions[11,12]. This suggests that approximately 3% of the human genome, containing fewer than 1 M sequence differences, is likely to harbor a large proportion of the functionally important sites and the ones that mattered in human evolution.

Some of the sequence differences may be of obvious phenotypic importance. In many other cases, the sequence information can be combined with other experimental, medical and population genetic information to hone in on those changes that are important.

2.2 Comparative expression analysis in human and chimpanzee. Recently, several groups have initiated comparative expression analysis between human and chimpanzee by using nucleic acid-microarrays to compare mRNA profiles of different tissues and cell types. The observed gene expression differences between humans and chimpanzees all indicate possibly important physiological changes. The availability of the comparative sequence will greatly facilitate the investigation of the genomic basis of such changes. It will allow not simply study of individual genes, but the detection of coordinated evolution of regulatory regions (for example, those that give rise to neotenic traits in the human lineage).

In addition, the availability of the chimpanzee sequence will technically facilitate such human-chimpanzee expression comparison – by allowing the design of microarrays to control for differential hybridization (for example, by using oligonucleotides that are constant in both species or by using both the human and chimpanzee sequences in a symmetrical fashion).

The information obtained will primarily be used in investigations using human material. Of course, any experiments using chimpanzees that are potentiated by the availability of the sequence would have to be done in conformity with humane practices and existing guidelines, recognizing the uniqueness of this endangered species as the closest living relative of humans. In the United States, NIH policies provide an appropriate set of guidelines for conducting chimpanzee research.

2.3 Comparative medical analysis between human and chimpanzee. One of the most intriguing and important comparisons between humans and chimpanzees is in differences in physiology, anatomy, and pathology caused by the modest sequence difference. Examples include dramatic differences in reproductive biology[6], unique vertebral column structure, unusually high human susceptibility to *falciparum* malaria[14], suggestive evidence of different rates of epithelial cancers[15,16], Alzheimer's disease[17], and HIV progression to AIDS[18]. Comparative analysis of the human and chimpanzee sequences might help to identify the human genes involved in these processes, and studies of such genes could shed light on the biology underlying such physiology and disease.

2.4 Evolutionary and population genetic studies of humans. One of the most important issues is to detect the signatures of selection that occurred in early human evolution or in the more recent human population. Some human genes will have changed dramatically compared to their counterparts in the chimpanzee genome due to the action of natural selection. Identifying such genes is particularly useful because they are the ones that confer human-unique medical conditions, such as those listed previously. Also, these intensively selected human genes will be part of biological systems for which the use of animal models will not be appropriate or would be misleading. Conversely, knowing which genes are part of biological systems that have not undergone especially strong selection in humans means that we can have increased confidence in using animal models to study those systems. Identifying the substrates of natural selection is intellectually interesting, finally, because it will point to genomic regions in which alleles have been recently selected for resistance to parasites or epidemic diseases. An example is recent work showing that the *BRCA1* gene has been under positive selection in humans and chimpanzees[19].

The chimpanzee sequence can be used, in combination with human polymorphism information, in several ways:

*(i) Regions of rapid evolutio*n. Population genetic tests for natural selection that occurred in the history of human populations usually require an 'outgroup' species—used

to determine the ancestral allele at a site in the genome that is polymorphic in humans, or the mutation rate over a stretch of sequence. The ideal outgroup is a species that is as closely related to the studied population as possible. The population genetic tests that can benefit from a chimpanzee genome sequence include the HKA test[20] and the McDonald-Kreitman test[21].

To detect more ancient selection — for example, selection that occurred over the time period since the divergence of humans from chimpanzees — another class of tests is available. Such tests consider homologous sequences from the coding region of the same gene in human and chimpanzee[22]. In the K_A/K_S test, if the number of protein-coding changes in the gene is substantially more (or less) than would be expected given the number of non-protein-coding changes, there is evidence for positive (or alternatively purifying) selection since the divergence of the two species. Similar tests can be performed by comparison with surrounding non-coding sequence. Although the human and chimpanzee sequences are too similar to make this approach reliable for *all* genes, some genes are likely to stand out as extreme cases and are likely to be sites of selection. Examples of such rapid evolution in genes have been recently identified by Eichler and colleagues[23].

*(ii) Non-ancestral alleles with high frequency in the human populatio*n. In general, the most frequent allele in a population is the ancestral allele. Regions of the human genome in which the most frequent alleles are not the ancestral alleles (at a string of contiguous loci) may indicate the presence of a derived allele that has risen to high frequency in the human population by recent selection.

*(iii) Differences in polymorphism rate between human and chimpanze*e. A number of population genetic tests rely on comparing the polymorphism rate in one species to the polymorphism rate in another, to highlight regions where one species has a very *different* level of diversity than another. Such regions may have been subject to intensive selective ‘sweeps’ (purifying selection) that may be of interest especially in studying the history of humans and the evolution of resistance to disease.

For this purpose, it would be valuable to have polymorphism information about the chimpanzee population. In fact, this can be easily obtained by generating the proposed shotgun sequence not from a single chimpanzee, but rather from a collection of several chimpanzees (see below).

Moreover, this would yield a SNP map of chimpanzees. Such a map would also provide the tools necessary to carry out haplotype studies and studies of recombination and population history in our closest living relative.

An example of how these approaches can suggest potentially important sequence changes is provided by the story of the *FOX*P2 gene. Human geneticists identified this gene as one that, when mutated, caused a language deficit[24]. Follow-up tests for selection by Pääbo and colleagues (in preparation) included one that showed an excess of protein sequence-changing differences in the gene (comparing chimpanzees to humans) to synonymous changes[22], suggesting that natural selection has occurred at the gene during human evolution.

In the case of *FOXP2* Pääbo and colleagues began with a candidate gene that seemed (because of its association with language) to be a reasonable candidate for selection in humans. The availability of a chimpanzee genome would mean that the list of all genes with extensive protein-coding changes in humans compared to chimpanzees could be identified. Hence, similar analyses could be applied to other genes without explicitly starting with a set of candidate genes.

Various tests are summarized in Table 1. In some cases, it will also be useful to have a sequence from a more distant primate. We would favor obtaining sequence from at least one additional primate as a subsequent priority.

TABLE 1

Classical tests for selection	Detects selection in what time period?	How many new sequences required?
Search for gross genomic differences	Hominid lineage	Chimpanzee
K_A/K_S	Hominid lineage	Chimpanzee
Selection in regulatory regions	Hominid lineage	Chimpanzee and mouse
HKA	Human population history	Chimpanzee
McDonald-Kreitman	Human population history	Chimpanzee
Ancestral alleles	Human population history	Chimpanzee

2.5 Identify regions with genomic rearrangements in chimpanzees compared to humans. With a chimpanzee genome, it will be possible to identify large-scale differences in genome structure not already identified in cytogenetic studies[9]. Such structural variations in the genome are likely to be interesting, and worth exploring. For example, having the chimpanzee genome sequence would make it possible to examine genetic elements that are newly-juxtaposed in the human genome and may, for example, bring some genes under the control of newly positioned regulatory elements.

2.6 Filling gaps in human genome sequence. Because of the very high degree of sequence and structural homology between the chimpanzee genome sequence and our own, chimpanzee sequence may help fill gaps in the human sequence. More broadly, it will act as a check on the human sequence.

3. Sequencing Strategy

We propose the following sequencing strategy.

3.1 Coverage. Genomic sequence would be obtained from 3–4× whole-genome shotgun (WGS) coverage from paired-end reads. Given the extremely high degree of sequence identity between human and chimpanzee, it is straightforward to align chimpanzee sequence directly to finished human sequence – with two exceptions. (1) The first exception is the small proportion of the human genome consisting of sequence that is duplicated with extremely high fidelity (for both sequences from the paired ends) and (2) the second exception is any region of the chimpanzee genome not present in the human[9]. A total of 3× coverage will cover 95% of the chimpanzee genome in theory and is likely to cover at least 90% in practice (allowing for cloning bias).

3.2 Vectors. The paired reads would be obtained from three types of vectors:

(i) Plasmids (4 kb).

(ii) Fosmids (40 kb).

(iii) BACs (~175 kb). These would include the chimpanzee BAC library that has already been fingerprinted at Washington University Genome Center. (Chimpanzee BAC end sequences covering approximately 1% of the chimpanzee genome have also recently been sequenced by a Japanese group[9].)

The fosmids and BACs are valuable for identifying regions in which significant insertions, deletions or rearrangements have occurred. The fosmids have a tight size range, owing to the constraints imposed by cloning. The BACs have variable insert size, but the size is known from fingerprinting.

Regions of particular interest or complexity can also be subjected to deep shotgun or finished sequencing based on biological interest, by using the end-sequenced fosmids and BACs (which should be arrayed and retained).

3.3 Number of reads. We would suggest approximately 2.5 ×, 0.2× and 0.05× coverage in plasmids, fosmids and BACs, respectively. (Assuming 500 bp reads, 80% pass rate and 70% pairing rate, this would correspond to clone numbers of ~10.3 M plasmids; 0.75 M fosmids; and ~0.2 M BACs and a total of 22.5 M attempted reads. Each clone type would provide roughly 20-fold physical coverage. The total number of reads would decrease with greater read length or higher pass rates.)

The exact distribution should be re-assessed in light of continuing technology evolution.

3.4. Individuals for sequencing. We would propose that plasmid libraries be prepared from a number of different individuals – for example, five chimpanzees (i.e., 10 chromosomes) at 0.5× each. In addition, the fosmid and BAC libraries would be derived from different individuals.

By sequencing different individuals, one has the added opportunity to explore polymorphisms in the chimpanzee population. (Such a strategy might be risky with a more distant organism in which de novo assembly would be routinely required and could be complicated by polymorphism. However, this is not a serious issue for the chimpanzee.)

Given 3× coverage obtained in this fashion, about 80% of the genome would be covered to at least 2× depth and thus would be susceptible to SNP discovery. Given the reported rate of polymorphism in chimpanzee[25], this would yield a catalogue of more than 4M SNPs within the chimpanzee population, at an average density exceeding 1 per 750 bp.

The chimpanzees would be selected from the eastern, central and western populations, to maximize diversity.

4. Relationship to the sequencing of other primate genomes

4.1 Priority of Chimpanzee. There are strong reasons why a chimpanzee genome sequence should be of highest priority compared to all other primates *(contra* VandeBerg *et al.* 2000)[26] .

• **Chimpanzees have a unique role to play in studies of human genetic variation.** To identify genetic changes that occurred in human history since the human divergence from the other great apes, it is necessary to study the closest living relative of humans, the chimpanzee. Genome sequences of more distantly related species are useless in this regard.

• **Chimpanzees are the best primates to use for comparative studies of gene expression levels.** Chimpanzees would be the optimal species to study the expression levels in the population that gave rise to humans, because they are the modern species most closely related to humans. In addition, chimpanzees would be optimal for comparative expression studies because it would be easy to develop oligonucleotide arrays and other reagents for studying gene expression that apply equally well to both chimpanzees and humans. For more distantly related primates, sequences are generally too diverged (e.g., Old World monkeys are 5-10 times more different from humans than are chimpanzees[8]) to readily design oligonucleotide arrays that work equally well in both species.

However, for many reasons, chimpanzees are to be used as experimental animals in only very limited circumstances, and opportunities to use chimpanzee samples in such comparative analyses would have to be limited to chimpanzee tissues that might be obtained during normal clinical care of the animals or at autopsy.

4.2 Sequencing of other primates. Once the sequencing of the chimpanzee is complete, we advocate the sequencing of the rhesus macaque monkey *Macaca mulatta* or the olive baboon *Papio hamadryas anubis* as a subsequent priority [26].

The genome of a more distantly related primate would provide us with ancestral sequence information and thus a means to determine which of the differences between humans and chimpanzees occurred due to mutations in the human lineage and thus are truly interesting. Ideally, one would choose a primate that is sufficiently close to humans for the ancestral genetic type to be unambiguous, but sufficiently distant from humans to identify sequences conserved by natural selection. An Old World monkey such as the rhesus macaque or baboon, which have been proposed for genome sequencing because they are more commonly used in physiological studies[26], would meet these criteria.

5. Conclusion

Extensive genomic sequence from the chimpanzee offers enormous promise for human studies – including human medicine, human evolution, and human population genetics. It should be a high priority.

In addition to the chimpanzee, there is also justification for sequencing an additional primate (probably the rhesus macaque monkey *Macaca mulatta* or the olive baboon *Papio hamadryas anubi*s) as a subsequent priority[26].

References

1. McConkey, E.H. and Goodman, M. (1997) A human genome evolution project is needed. *Trends Genet.***13,** 350-351.

2. Vigilant, L. and Pääbo, S. (1999) A chimpanzee millenium. *Biol. Chem.* **380,** 1353-1354.

3. McConkey, E.H. *et a*l. (2000) Proposal for a human genome evolution project. *Mol. Phylogenet. Evol*., **15,** 1-4.

4. McConkey, E.H. and Varki, A. (2000) A primate genome project deserves high priority. *Science,* **28**9, 1295-1296.

5. Varki, A. (2000) A chimpanzee genome project is a biomedical imperative. *Genome Research* **10,** 1065-1070.

6. Gagneux, P. and Varki, A. (2001) Genetic differences between humans and great apes. *Mol. Phylogenet.Evol.,* **18,** 2-13.
7. Ruvolo, M. (1997) Molecular phylogeny of the hominoids: inferences from multiple independent DNA sequence data set*s. Mol. Biol. Evol.* **14,** 248-265.
8. Caccone, A. and Powell, J.R. (1989) DNA divergence among hominoids. *Evolution* **43,** 925-942.
9. Fujiyama, A. *et al.* (2002) Construction and analysis of a human-chimpanzee comparative clone map. *Science* **295,** 131-134.
10. Angata, T., Varki, N.M. and Varki, A. (2001) A second uniquely human mutation affecting sialic acid biology. *J. Biol. Chem.* **276,** 40282-40287.
11. Hardison, R.C. (2000) Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet.* **16,** 369-372.
12. Dehal, P. *et al.* (2001) Human chromosome 19 and related regions in mouse: conservative and lineage-specific evolution. *Science* **29**3, 104-111.
13. Lander, E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* **409,** 860-921.
14. Ollomo, B. *et al.* (1997) Lack of malaria paraiste transmission between apes and humans in Gabon. *Am.J. Trop. Med. Hyg.* **56,** 440-445.
15. McClure, H.M. (1973) Tumors in nonhuman primates: observations during a six-year period in the Yerkes primate center colony. *Am. J. Phys. Anthropol.* **38,** 425-429.
16. Schmidt, R.E. (1975) Systemic pathology of chimpanzees. *J. Med. Primatol.* **7,** 274-318.
17. Gearing, M. *et al.* (1994) Neuropathology and apolipoprotein E profile of aged chimpanzees: implications for Alzheimer's disease. *Proc. Natl. Acad. Sci. USA* **91,** 9382-9386.
18. Novembre, F.J. *et al.* (1997) Development of AIDS in a chimpanzee infected with human immunodeficiency virus type 1. *J. Virol.* **71,** 4086-4091.
19. Huttley, G.A. *et al.* (2000) Adaptive evolution of the tumour suppressor *BRCA1* in humans and
chimpanzees. Australian Breast Cancer Family Study. *Nat. Genet.* **25,** 410-413.
20. Hudson, R.R., Kreitman, M. and Aguadé, M. (1987) A test of neutral molecular evolution based on nucleotide data. *Genetics* **116,** 153-159.
21. McDonald, J.H. and Kreitman, M. (1991) Adaptive protein evolution at the Adh locus in Drosophila. *Nature* **351,** 652-654.
22. Li, W.-H. (1997) *Molecular Evolution* (Sinauer Associates, Sunderland).
23. Johnson, M.E. *et al.* Positive selection of a gene family during the emergence of humans and African apes. *Nature* **413,** 514-519.
24. Lai, C.S. *et al.* (2001) A forkhead-domain gene is mutated in a severe speech and language disorder.*Nature* **41**3, 519-523.
25. Kaessmann, H., Wiebe, V., and Pääbo, S. (1999) Extensive nuclear DNA sequence diversity among chimpanzees. *Science* **286,** 1159-1162.
26. VandeBerg, J.L, Williams-Blangero, S., Dyke, B. and Rogers, J. Examining priorities for a primate genome project. *Science* **290,** 1504-1505.