Justification and Rationale for Investigating Structural Variation in *Pan paniscus*

Devin Locke, Elaine Mardis, Richard K. Wilson

Introduction

Of the great ape species for which genome projects have been assigned, initiated, or completed, only the bonobo (*Pan paniscus*) has escaped scrutiny at a genome-wide level. The reason, at first glance, is relatively straightforward as the bonobo is closely related to the common chimpanzee (referred to simply as "chimpanzee" for the remainder of this document) for which genome sequence exists; however, there are compelling reasons for pursuing the characteristics that differentiate bonobos from humans and chimpanzees. It is therefore our proposal that a project be initiated to study the genome of the bonobo using an approach less extensive than that of a full genome project, yet capable of capturing important structural variation and also informing studies of nucleotide-level variation.

Background – Bonobo Biology

The bonobo, also known as the pygmy chimpanzee, was recognized as a distinct species of the *Pan* genus in the early 20th century [1,2]. The modern range of the bonobo is bound by the Congo River to the north and the Kasai River to the south, in the central forest basin of the Democratic Republic of Congo. The range of the bonobo is not shared by any other great ape species, including chimpanzees, which are found in adjacent territories. The geographic distribution of chimpanzees and bonobos within the region supports a very strong role for rivers as barriers maintaining separation between great ape populations. In fact, intra-species variation among bonobo subpopulations has been correlated more highly with location defined by rivers as opposed to linear distance between populations [3].

Interestingly, the paleogeographic history of the region during the glaciations of the Pleistocene suggests that the basin inhabited by the bonobo was a refuge from disturbance; while to the contrary, chimpanzee populations may have been forced to disperse during this time [4,5]. This dispersal may be reflected in the divergence and gene flow observed among chimpanzee populations and subspecies from the western, central and eastern regions of central Africa [6,7]. Bonobo-chimpanzee divergence time estimates based on sampling of the mitochondrial and nuclear genomes indicate bonobos diverged from a common ancestor from 900,000 to 3 million years ago [8-12]. A recent meta-analysis using a compilation of previously published sequence data and more recent Markov chain Monte Carlo methods strongly supported the lower time estimate at approximately 860,000-900,000 years ago, and suggested little gene flow had occurred between bonobo and chimpanzee populations since that time [7].

While bonobos are clearly geographically distinct from chimpanzees today, there are also several measurable biological characteristics that have been used to distinguish the bonobo from the chimpanzee in the great ape phylogeny including dental morphology

and cranial geometry [13,14]. At the cytogenetic level, several rearrangements have been observed in the bonobo lineage, and these differences indicate the bonobo karyotype may be further derived from the ancestral great ape karyotype than either the chimpanzee or human karyotype [15].

Additionally, in contrast to overtly aggressive patriarchal chimpanzee social dynamics, which often involve violent conflict, matriarchal bonobos are significantly less aggressive and solve conflict through sexual contact [16]. To distill the extreme divergence of social behavior of chimpanzees and bonobos to an anecdote: bonobos are the lovers, and chimpanzees the fighters. In fact the structure of chimpanzee and bonobo societies are polar opposites in a great many respects, which may have lead to significant differences in population structure at the genetic level. The extreme contrast in behavior between two species so closely related genetically and geographically may provide a unique opportunity to discover genetic variants which are associated with such behavioral polarity. Additionally, the frequent sexual contact that occurs in bonobo societies may have had a substantial impact on the evolution of the bonobo immune system.

Sequencing and Analysis Strategy

To survey structural variation specific to the bonobo genome we propose to implement the methods established by Tuzun et al. (2005) that employed mapping fosmid paired-end sequences to detect structural variation. This method relies on deep fosmid clone coverage and a narrow distribution of fosmid insert size to detect variants with high resolution and confidence. An additional benefit of such an approach is that sites of variation are, by the nature of the assay, captured in the fosmid library and representative variant fosmids can be sequenced fully.

The project consists of four components:

1) We first propose the production of a fosmid library containing 1.5 million clones, with an average insert size of 40 kb, representing ~20x clone coverage of the bonobo genome. The library would be derived from a single female individual to provide coverage of the X chromosome equal to the level of the autosomes. Due to the geographic constraint of the bonobo population to central Africa and the lower level of intra-species divergence compared to the chimpanzee, one individual will likely contain the majority of inter-specific structural variation. The material required for library production will be obtained through a collaboration with Dr. Oliver Ryder at the San Diego Zoo, and this effort will adhere to all proper regulatory protocols for handling material derived from an endangered species.

2) Sequencing of the fosmid ends will require approximately 3.6M attempted sequence reactions, assuming an 85% sequence pass rate, which is realistic in comparison to recent fosmid end sequencing efforts at the WashU GSC (i.e. 0.5M maize fosmid end sequences). The resulting sequence represents 0.5x genome coverage, with an average read length of 500 bp, which is a conservative estimate based on current read length averages using the ABI 3730XL sequencing pipeline. It is important to recognize that

read length influences the ability of reads to be uniquely mapped within repeat-rich genomes and thus the optimal approach at this time utilizes a conventional Sanger sequencing pipeline.

3) The fosmid end sequences will then be mapped to the human genome using established methods and parameters and the resulting regions of variation identified[17,18]. Comparison of this data set to a previous analysis of structural variation in the common chimpanzee genome will determine whether variants are lineage specific or pre-date the divergence of bonobos and chimpanzees [18]. These data will also be compared to the growing database of human structural variants to ascertain whether instability at such loci are a fundamental component of great ape and/or *Pan* and *Homo* genomes. In addition, the sequence produced will be available for analyses that are not dependent upon physically linked paired-ends such as estimates of divergence and phylogenetics.

4) Full-insert sequencing and finishing of fosmids representing variant sites in the bonobo genome will allow precise ascertainment of the nature of each structural variation locus. Given the existence of two closely-related genome sequences, these data would allow high-confidence alignments and comparisons. A chimpanzee-human comparison using the fosmid end-sequence approach, described in Newman et al. (2005), identified 651 structural variants between chimpanzees and humans. Assuming a 6 mya divergence time for the human and chimpanzee/bonobo lineages, and applying the lower estimate of ~1 mya for the divergence of the bonobo and chimpanzee lineages, this yields an estimate of approximately 110 fosmids for full-insert sequencing.

Additional Benefits and Future Possibilities

Sequencing a bonobo fosmid library would provide several benefits beyond the scope of this initial proposal. First, placement of the fosmid end sequences would allow researchers to identify fosmids encompassing regions of interest for projects beyond the scope of structural variation (i.e. positive selection, etc). These clones then might be candidates for full length sequencing and finishing. Second, the SNPs gleaned from the 0.5x genome coverage would provide a resource for studies of chimpanzee and bonobo population genetics. Third, should the opportunity arise for full sequencing of the bonobo genome in the future, the fosmid paired-end framework could also be readily used in the assembly process. Last, the fosmid paired-end framework could be used in combination with novel high-throughput sequencing methodologies to model the effectiveness of combined cloned-insert and short-read assemblies of complex mammalian genomes.

Summary

The bonobo is positioned uniquely in the phylogeny of the great apes, having most likely diverged very recently (900k years ago) from a common ancestor with the chimpanzee. There are numerous biological and behavioral characteristics that differentiate the bonobo from the chimpanzee that are worthy of exploration at the genomic level. A fosmid paired-end sequencing approach would allow for sites of structural variation in the

bonobo genome to be identified and could shed light on the events that may have had an impact on the most recent great ape speciation event.

References

1. Schwarz, E. Das vorkommen des Schimpansen auf dem linken Kongo-Ufer. *Revue de Zoologie et de Bontanique africaines* **16**, 425-426 (1929).
2. Coolidge, H. *Pan paniscus*, pygmy chimpanzee from south of the Congo River. *American Journal of Physical Anthropology* **18**, 1-59 (1933).
3. Eriksson, J., G. Hohmann, C. Boesch and L. Vigilant. Rivers influence the population genetic structure of bonobos (*Pan paniscus*). *Molecular Ecology* **13**, 3425-3435 (2004).
4. Maley, J. The African rain forest - main characteristics of changes in vegetation and climate from the Upper Cretaceous to the Quaternary. *Proceedings of the Royal Society of Edinburgh* **104B**, 31-73 (1996).
5. Colyn, M., A. Gautier-Hion, W. Verheyen. A re-appraisal of paleoenvironmental history in central Africa: evidence for a major fluvial refuge in the Zaire Basin. *Journal of Biogeography* **18**, 403-407 (1991).
6. Gagneux, P., Gonder, M.K., Goldberg, T.L. & Morin, P.A. Gene flow in wild chimpanzee populations: what genetic data tell us about chimpanzee movement over space and time. *Philos Trans R Soc Lond B Biol Sci* **356**, 889-97 (2001).
7. Won, Y.J. & Hey, J. Divergence population genetics of chimpanzees. *Mol Biol Evol* **22**, 297-307 (2005).
8. Bailey, W.J. et al. Reexamination of the African hominoid trichotomy with additional sequences from the primate beta-globin gene cluster. *Mol Phylogenet Evol* **1**, 97-135 (1992).
9. Horai, S. et al. Man's place in Hominoidea revealed by mitochondrial DNA genealogy. *J Mol Evol* **35**, 32-43 (1992).
10. Pesole, G., Sbisa, E., Preparata, G. & Saccone, C. The evolution of the mitochondrial D-loop region and the origin of modern man. *Mol Biol Evol* **9**, 587-98 (1992).
11. Kaessmann, H., Wiebe, V. & Paabo, S. Extensive nuclear DNA sequence diversity among chimpanzees. *Science* **286**, 1159-62 (1999).
12. Yu, N. et al. Low nucleotide diversity in chimpanzees and bonobos. *Genetics* **164**, 1511-8 (2003).
13. Pilbrow, V. Population systematics of chimpanzees using molar morphometrics. *J Hum Evol* **51**, 646-62 (2006).
14. Lieberman, D.E., Carlo, J., Ponce de Leon, M. & Zollikofer, C.P. A geometric morphometric analysis of heterochrony in the cranium of chimpanzees and bonobos. *J Hum Evol* (2007).
15. Stanyon, R., Chiarelli, B., Gottlieb, K. & Patton, W.H. The phylogenetic and taxonomic status of *Pan paniscus*: a chromosomal perspective. *American Journal of Physical Anthropology* **69**, 489-498 (1986).
16. de Waal, F. & Lanting, F. *Bonobo: the forgotten ape*, (University of California Press, Berkeley and Los Angeles, 1997).

17. Tuzun, E. et al. Fine-scale structural variation of the human genome. *Nat Genet* **37**, 727-32 (2005).
18. Newman, T.L. et al. A genome-wide survey of structural variation between human and chimpanzee. *Genome Res* **15**, 1344-56 (2005).