Re: Proposal for Construction a Human Haploid BAC library from Hydaditiform Mole Source Material.

Date: Oct. 10th, 2002

From: Evan Eichler, Ph.D., Dept. of Genetics, Case Western Reserve University
Urvashi Surti, Ph.D., Director, Pittsburgh Cytogenetics Laboratory, University of Pittsburgh.
Roel Ophoff, Ph.D., Dept of Human Genetics/Center for Neurobehavioral Genetics, UCLA

To: BAC Library Resource Network, National Human Genome Research Institute

Importance: The central task of human genetics is the correlation of phenotype and genotype. Much of this effort depends on our ability to track unique DNA by association or linkage with phenotype. The revelation that a significant fraction (~5%) of our genome is composed of recent segmental duplications has a serious impact on the work of human geneticist and the final assembly of the human genome (Bailey et al. 2001; Eichler 2001). Segmental duplications may span large distances of genomic sequence (in some cases 100's of kb), share a high degree of sequence identity (>99%), can harbor genes, and, unlike, other classes of repetitive sequences can not be distinguished as such, *a priori*. In essence, these properties have made a portion of our genome intractable by the standard fare of molecular techniques applied within our field. The potential for such regions to rearrange and create structural polymorphisms (Giglio et al. 2001; Giglio et al. 2002; Osborne et al. 2001) has further confounded traditional linkage analysis in these regions (Nelson Freimer and Leena Peltonen, personal communications). The development of human SNP maps is similarly hampered leading to misleadingly high density of SNPs over duplicated regions-where collapse of SNPs and paralogous sequence variants occur (Bailey et al. 2002; Estivill et al. 2002). Duplicated segments pose serious problems for the assembly and annotation of the human genome. Even among chromosomes that are near completion, there are still large (>100 kb) gaps which will require specialized efforts to fill or regions in which the present assembly is suspect. Many of these gaps lie within highly duplicated regions that are not necessarily refractory to subcloning. Instead, these same regions contain many highly duplicated segments in which the degree of sequence variation among duplicated loci (paralogous sequence variation) approaches levels of allelic variation. Finally, it has become increasingly apparent that the segmental duplications themselves provide the molecular basis for many human genetic disorders, including complex genetic disease traits (Gratacos et al. 2001; Stankiewicz and Lupski 2002). Their biological resolution of these regions is therefore essential for a complete understanding of the genetic basis of the human disease. First and foremost, however, it is essential that such highly paralogous regions be identified, their locations refined and their sequence correctly assembled into the human reference genome.

One of the key impediments in resolving the complexity of these regions is the diploid and polymorphic nature of the human genome. In the past, the distinction between allelic versus polymorphic variation has been successfully circumvented by the use of genetic material of haploid complexity. This has included the use of monchromosomal source material such as genomic cosmid libraries and/or monochromosomal somatic cell hybrids (Horvath et al. 2000; Johnson et al. 2001). Such resources, while helpful, are not satisfactory for establishing continuity across large (>100 kb) regions of segmental duplication at a genome-wide level. The final sequence and assembly of the Y chromosome (which is unusually enriched for segmental duplications) was achieved in large part due to the fact that all the "BAC clones [came] from one man's Y chromosome" (Kuroda-Kawaguchi et al. 2001). Sequence assembly was therefore not impaired by polymorphism and all sequence variants represented distinct copies of paralogous sequences.

We propose the development of a full genome, haploid BAC library from hydatidiform mole DNA to resolve the genomic structure of autosomal segmental duplications. Hydatiform moles are conception abnormalities that most often arise from the fertilization of an ovum by a single X-bearing sperm (Kajii and Ohama 1977). Subsequent diploidization results in a 46 XX karyotype in which all allelic variation has been eliminated allowing the unambiguous delineation of duplicated DNA as well as haplotype characterization (Fan et al. 2002). Formulation of this proposal was stimulated by extensive discussion with members of the human genetics community and genomics community (represented by Roel Ophoff and Evan Eichler) who has been frustrated by attempts to resolve such areas as part of disease or genomic analyses. Dr. Urvashi Surti, an expert in hydatidiform moles and characterization of conception abnormalities, was consulted extensively to provide practical advice on the use of hydatidiform cell lines for the purpose of BAC library construction. Her participation is essential for the successful construction of such a resource..

Usage:

1) Gap closure within the human genome. The main use of this resource would be to provide sequence closure within the highly duplicated regions of the human genome. The sequence of such regions is currently under-represented within GenBank and/or the assembly of such areas is suspect. Owing in large part to its duplicated nature,,paralogous BACs are occasionally collapsed during sequence assembly.. In other cases, duplicated copies have not been sequenced—projects were terminated because they were deemed redundant, representing allelic variants of a previously sequenced locus. While the sequence and assembly of the human genome has improved substantially, significant ambiguities remain and will remain after its tentative completion date, April 2003. On some chromosomes (chromosomes 16 and 17 for example), more than 30% of type III gaps (reference) map to large blocks of segmental duplication (Eichler unpublished, Build 30, June). The presence of a BAC library of haploid complexity offers considerable advantage in resolving these regions. Sequence family variants or paralogous sequence variants can be used to distinguish copies and to provide genomic continuity across such regions. The large-insert size of the BAC allows large regions of segmental duplication (>300 kb) to be rapidly traversed in a few walking steps. We propose to construct this resource from an immortalized hydatidiform mole cell line (see below). Thus, unlike RPCI-11 (the major source of HGP sequence) for which no cell line exists, it will be possible to validate the structure and organization of duplicated loci cytogenetically prior to sequencing.

2) Critical regions of genetic disease. Filling sequence gaps of these regions will have further-reaching implications relevant not only with respect to human genome assembly but human genetic disease. Linkage analyses of a number of genetic disorders (i.e. multiple sclerosis, bipolar disorder, etc) have been hampered by the inability to resolve the structure of these regions with respect to flanking unique STS. The construction of a BAC library from a haploid source material will provide the necessary resource to order such markers within a single haplotype (Taillon-Miller et al. 1997; Taillon-Miller et al. 1999). Using BAC clones that transition between unique and duplicated regions, it will be possible to delineate unambiguously marker arrangement in order to design assays for both linkage and association studies. Such studies will likely provide the necessary impetus to resolve ambiguities within the current genome assembly (#1) but require a high quality resource be available to facilitate such efforts.

3) Chromosomal Structural Polymorphisms. Paralogous sequences have been shown to mediate large-scale inversions within the human population (Giglio et al. 2001; Giglio et al. 2002; Osborne et al. 2001). Analysis of diploid BAC libraries used for HGP sequencing for 8p, for example, show contradicting genomic structures within these regions (Ophoff, unpublished. All the sequences within the genomic library are correct but represent different configurations within this area. The two configurations can not be readily resolved. A haploid BAC library will be essential in helping to reconstruct at least one

configuration of such regions and will serve as starting point for understanding genomic variability and genomic history of each of these loci.

4) Genome Annotation. It has been estimated that 6% of genes map to the highly duplicated regions of the genome (Bailey et al. 2002). In regions of high degree of sequence identity, assignment of genes has been problematic resulting in genes being assigned to more than one location (www.genome.ucsc.edu). Similarly, the density of SNPs within duplicated regions has been calculated as twice that of the average from unique regions (Bailey et al. 2002; Estivill et al. 2002). The construction of a BAC library resource from a haploid genome should dramatically improve the quality of SNP maps and gene annotation since paralogous and allelic variation can be readily distinguished. Large-insert genomic libraries, as opposed to genomic DNA or smaller-insert libraries, offer the advantage that the phase of the variants can be readily assessed.

Research Community*:* Interest in the development of a hydatidiform BAC library is restricted to those members of the human genetics, genomics and evolutionary community which have a direct interest in segmental duplications or by virtue of genome-wide analyses/specific human genetic disease studies have been forced to resolve the structure of these regions. This includes researchers interested in recapitulating the evolutionary history of human haplotypes (eg. Aravinda Chakravarti, Svante Paabo, Maynard Olson and Wen-Hsiung Li), those studying recently duplicated regions associated with human genomic disorders and rearrangements (eg. Jim Lupski, Evan Eichler, Xavier Estivill, Barbara Trask and Tamim Shaikh), and researchers focused on linkage and association studies in the vicinity of duplicated regions (Leena Peltonen, Nelson Freimer, Roel Ophoff). It should be emphasized that resolving the structure of such regions among other primates requires that these regions be well-defined within the human genome. Therefore, it is likely that targeted comparative genomic projects within duplicated regions (Eric Green, Ken Dewar, Bruce Roe) as well as whole genome projects (chimpanzee and macaque) will benefit greatly from a human hydatidiform mole BAC library resource

Has the organism been proposed to NHGRI or another publicly funded agency for BAC-based genomic sequence? Multiple *diploid* human libraries have been constructed (CaltechD, RPCI-11, RPCI-13, etc). No formal request, to our knowledge, has been made for the construction of a human BAC library from *a haploid source*. The only comparable resource currently available is the chromosome-specific cosmid and lamda libraries generated as part of the DOE National Laboratory Gene Library Project. Based on extensive characterization of duplicated loci within these chromosome-specific libraries we have found that existing inserts are either too small and/or that the libraries are sufficiently non-random to be effectively used as a tool for resolving these regions.

Complementary Genomic Resources. A collection of eight hydatidiform cell lines have been recently characterized by Dr. Surti and colleagues. Two of these have been distributed to Coriell and are available to the academic community.

Proposed Strain Selection. A complete hydatidiform mole, as opposed to a partial hydatidiform which contains diploid material or a mole which has formed as a result of dispermy, is required for this purpose. It is essential that that all bci be haploid and the source material be well-characterized. There are two possible sources: primary cell culture or a transformed cell line. Dr. Urvashi Surti has performed a number of analyses to ensure the high quality of a cell line or primary cell culture that would be used in this study. This includes a) karyotype analysis of the cell lines at different stages during passaging and b) extensive SNP genotyping with 1494 SNP markers to ensure haploid content.(Fan et al. 2002). A major difficulty in

using hydatidiform moles has been the construction of stable immortal cell lines in order to provide sufficient quality and quantity of DNA. This impasse has recently been overcome by the use of human telomerase during transformation (U. Surti, personal communication). Since such cell lines have only recently been established and their longevity has not been tested, Dr. Surti has agreed to provide primary cell culture material from a single complete hydatidiform mole as an alternate.

Genome Size: Genome size has been estimated as 2,90 Mbp (Consortium 2001).

Source DNA: The construction of a BAC library requires typically 1-3 X 10^8 cells (or the equivalent of 600—1500 ug of high quality DNA). Most complete hydatidiform moles are diagnosed early during pregnancy (10-16 weeks) and are too small (grape-like in size) to generate sufficient DNA for the purpose of BAC library construction. Further, evacuation of the primary mole during the second trimester often results in considerable contamination with maternal tissue. In order to obtain sufficient amount of DNA and of suitable quality, it will be necessary to use a well-established primary cell culture or one of 8 cell lines that have been isolated, transformed and characterized (genotyped and karyotyped by Dr. Urvashi Surti). Two cell lines have been deposited into the Coriell repository, have been anonymized and will be available for academic purposes.

Specifications: A random BAC library (*EcoR*1 partial digest, insert size (not less than 150 kb), 10-fold genomic redundancy) should be considered using standard cloning vector (i.e. BACe3.6). Libraries significantly less than 150 kb insert will complicate the analysis of duplicated regions as these regions are often in excess of 100 kb. Regions of rearrangement when compared to a human reference sequence require multiple dones to eliminate the possibility of artifacts due to chimeric BACs. Consequently sufficient depth (10X) is requested.

Time frame: Considering the timeline of the Human Genome Project, the need for this library is immediate and should be considered a high priority. It is unlikely that many of these clones, however, will contribute to the April, 2003 version of the human genome. This resource, however, will become much more valuable after the "completion" of the genome as ambiguities within the assembly of the sequence become apparent.

Other Support: No other support is available for the construction of this BAC library.

Other Relevant Information: We have worked previously with Dr. Pieter de Jong in obtaining tissue material from primates (chimpanzee, gorilla, orangutan, vervet monkey, marmoset, owl monkey, mouse lemur and galago) for BAC library construction. Dr. de Jong has expressed an interest in developing a BAC library from hydatidiform source material. Based on our demonstrated working-relationship, we would prefer that the library if approved would be assigned to Dr. de Jong. The construction of a BAC library from transformed cell line material poses no technical difficulty as it has been performed within Dr. De Jong's laboratory previously. It should also be noted that Dr. Urvashi Surti has maintained careful records on the karyotype of the cell lines both prior and after transformation. A cytogenetic abnormality was observed in only 1/5 transformed cell lines..

References.

Bailey, J.A., Z. Gu, R.A. Clark, K. Reinert, R.V. Samonte, S. Schwartz, M.D. Adams, E.W. Myers, P.W. Li, and E.E. Eichler. 2002. Recent segmental duplications in the human genome. *Science* **297:** 1003-1007.

Bailey, J.A., A.M. Yavor, H.F. Massa, B.J. Trask, and E.E. Eichler. 2001. Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res* **11:** 1005-1017.

The International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860-920.

Eichler, E.E. 2001. Segmental duplications: what's missing, misassigned, and misassembled- and should we care? *Genome Res* **11:** 653-656.

Estivill, X., J. Cheung, M.A. Pujana, K. Nakabayashi, S.W. Scherer, and L.C. Tsui. 2002. Chromosomal regions containing high-density and ambiguously mapped putative single nucleotide polymorphisms (SNPs) correlate with segmental duplications in the human genome. *Hum Mol Genet* **11:** 1987-1995.

Fan, J.B., U. Surti, P. Taillon-Miller, L. Hsie, G.C. Kennedy, L. Hoffner, T. Ryder, D.G. Mutch, and P.Y. Kwok. 2002. Paternal origins of complete hydatidiform moles proven by whole genome single-nucleotide polymorphism haplotyping. *Genomics* **79:** 58-62.

Giglio, S., K.W. Broman, N. Matsumoto, V.V. Calvari, G. Gimelli, T. Neumann, H. Ohashi, L. Voullaire, D. Larizza, R. Giorda, J.L. Weber, D.H. Ledbetter, and O. Zuffardi. 2001. Olfactory Receptor-Gene Clusters, Genomic-Inversion Polymorphisms, and Common Chromosome Rearrangements. *Am J Hum Genet* **68:** 874-883.

Giglio, S., V. Calvari, G. Gregato, G. Gimelli, S. Camanini, R. Giorda, A. Ragusa, S. Guerneri, A. Selicorni, M. Stumm, H. Tonnies, M. Ventura, M. Zollino, G. Neri, J. Barber, D. Wieczorek, M. Rocchi, and O. Zuffardi. 2002. Heterozygous submicroscopic inversions involving olfactory receptor-gene clusters mediate the recurrent t(4;8)(p16;p23) translocation. *Am J Hum Genet* **71:** 276-285.

Gratacos, M., M. Nadal, R. Martin-Santos, M.A. Pujana, J. Gago, B. Peral, L. Armengol, I. Ponsa, R. Miro, A. Bulbena, and X. Estivill. 2001. A polymorphic genomic duplication on human chromosome 15 is a susceptibility factor for panic and phobic disorders. *Cell* **106:** 367-379.

Horvath, J., S. Schwartz, and E. Eichler. 2000. The mosaic structure of a 2p11 pericentromeric segment: A strategy for characterizing complex regions of the human genome. *Genome Res* **10:** 839-852.

Johnson, M.E., L. Viggiano, J.A. Bailey, M. Abdul-Rauf, G. Goodwin, M. Rocchi, and E.E. Eichler. 2001. Positive selection of a gene family during the emergence of humans and African apes. *Nature* **413:** 514-519.

Kajii, T. and K. Ohama. 1977. Androgenetic origin of hydatidiform mole. *Nature* **268:** 633-634.

Kuroda-Kawaguchi, T., H. Skaletsky, L.G. Brown, P.J. Minx, H.S. Cordum, R.H. Waterston, R.K. Wilson, S. Silber, R. Oates, S. Rozen, and D.C. Page. 2001. The AZFc region of the Y chromosome features massive palindromes and uniform recurrent deletions in infertile men. *Nat Genet* **29:** 279-286.

Osborne, L.R., M. Li, B. Pober, D. Chitayat, J. Bodurtha, A. Mandel, T. Costa, T. Grebe, S. Cox, L.C. Tsui, and S.W. Scherer. 2001. A 1.5 million-base pair inversion polymorphism in families with Williams-Beuren syndrome. *Nat Genet* **29:** 321-325.

Stankiewicz, P. and J.R. Lupski. 2002. Genomic architecture, rearrangements and genomic disorders. *Trends Genet* **18:** 74-82.

Taillon-Miller, P., I. Bauer-Sardina, H. Zakeri, L. Hillier, D.G. Mutch, and P.Y. Kwok. 1997. The homozygous complete hydatidiform mole: a unique resource for genome studies. *Genomics* **46:** 307-310.

Taillon-Miller, P., E.E. Piernot, and P.Y. Kwok. 1999. Efficient approach to unique single-nucleotide polymorphism discovery. *Genome Res* **9:** 499-505.