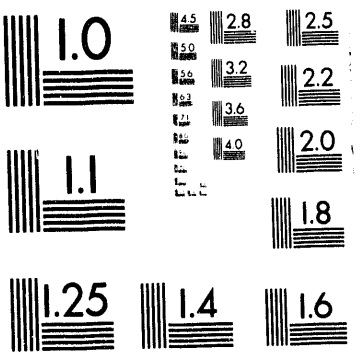

1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
6.3
7.1

1 of 1

2

Conf-940166--1

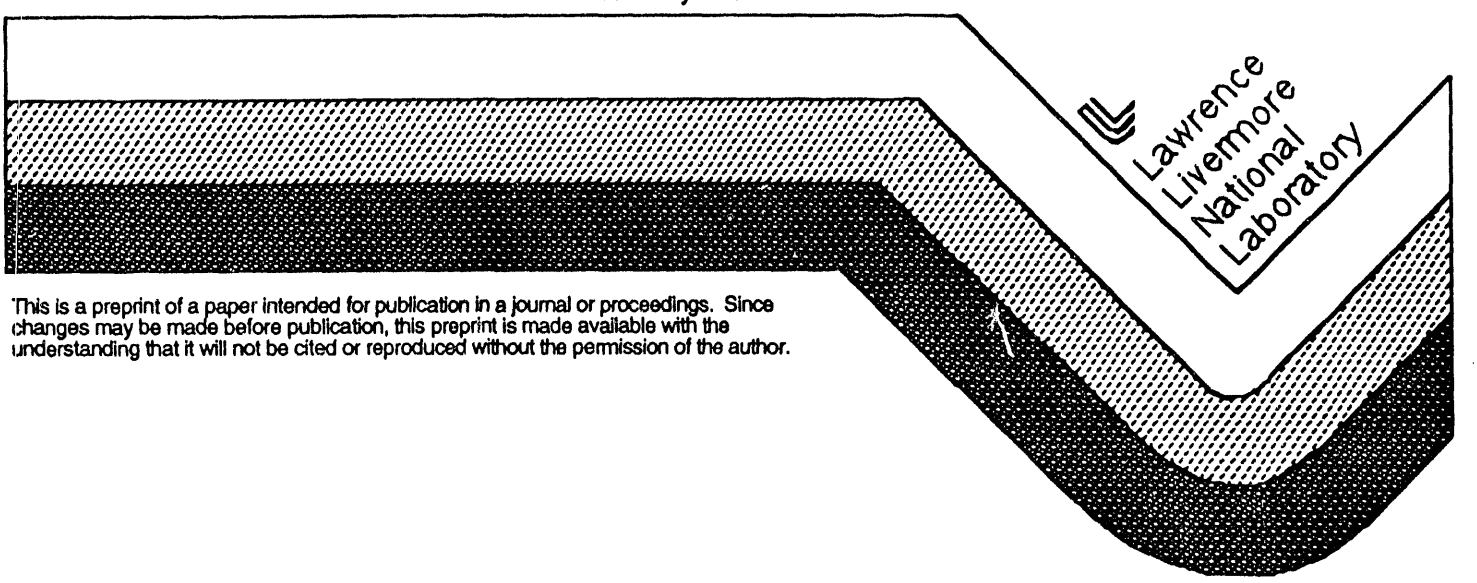
UCRL-JC-115966
PREPRINT

An overview of the human genome project

Mark A. Batzer

This paper was prepared for submittal to the First Forensic Experts Conference,
Dubai,United Araf Emirates, January 8-10, 1994

January 1994



DISCLAIMER

An Overview of the Human Genome Project

Mark A. Batzer

Human Genome Center, L-452
Biology and Biotechnology Research Program
Lawrence Livermore National Laboratory
P.O. Box 808
Livermore, CA 94551

ABSTRACT

The human genome project is one of the most ambitious scientific projects to date, with the ultimate goal being a nucleotide sequence for all four billion bases of human DNA. In the process of determining the nucleotide sequence for each base, the location, function, and regulatory regions from the estimated 100,000 human genes will be identified. The genome project itself relies upon maps of the human genetic code derived from several different levels of resolution. Genetic linkage analysis provides a low resolution genome map. The information for genetic linkage maps is derived from the analysis of chromosome specific markers such as Sequence Tagged Sites (STSs), Variable Number of Tandem Repeats (VNTRs) or other polymorphic (highly informative) loci in a number of different families. Using this information the location of an unknown disease gene can be limited to a region comprised of one million base pairs of DNA or less. After this point, one must construct or have access to a physical map of the region of interest. Physical mapping involves the construction of an ordered overlapping (contiguous) set of recombinant DNA clones. These clones may be derived from a number of different vectors including cosmids, Bacterial Artificial Chromosomes (BACs), P1 derived Artificial Chromosomes (PACs), somatic cell hybrids, or Yeast Artificial Chromosomes (YACs). The ultimate goal for physical mapping is to establish a completely overlapping (contiguous) set of clones for the entire genome. After a gene or region of interest has been localized using physical mapping the nucleotide sequence is determined. The overlap between genetic mapping, physical mapping and DNA sequencing has proven to be a powerful tool for the isolation of disease genes through positional cloning.

Humans are comprised of trillions of individual cells. Almost all of these cells contain the a nucleus within which resides the information to determine each and every feature of a human. This material called deoxyribonucleic acid or DNA is contained in a series of discreet bundles termed chromosomes. Humans have 23 different chromosomes which are encoded by four billion bases of DNA. The ultimate goal of the human genome project is to determine the nucleotide sequence for each and every base of DNA . Modern humans (*Homo sapiens*) differ from our nearest non-human primate relative the chimpanzee (*Pan troglodytes*) by approximately 40 million single base pair substitutions which have occurred in the last 4-6 million years (Miyamoto *et al.*, 1987). Individual humans differ by only 10 million point mutations. The construction of a DNA sequence map of the human genome will allow us to determine the location and nature of all one hundred thousand human genes and also allow us to determine exactly what makes each and every human unique.

In the process of determining the DNA sequence information for the human genome a number of intermediate maps will be constructed. These intermediate maps can be divided into two general classes, genetic and physical maps. Genetic maps define which chromosome a particular gene is located on through the genetic linkage analysis of families with highly polymorphic (informative) markers (Ott, 1991). These markers are generally either restriction fragment length polymorphisms (RFLPs) which are assayed on the basis of the presence or absence of restriction endonuclease cleavage sites (Botstein *et al.*, 1980) or variable number of tandem repeats (VNTRs) (Nakamura *et al.*, 1987) loci which are commonly assayed using the polymerase chain reaction (PCR). One of the intermediate goals for the human genome project is the establishment of a high resolution genetic map of each and every human chromosome. Physical maps involve the localization of genes or sequences of interest to individual DNA segments or clones. The process of cloning involves the replication and maintenance of segments of human DNA joined to a vector (e.g. plasmid or bacteriophage) within a host (generally bacteria) organism. These segments of DNA are then replicated by the host as individual clones. Physical mapping the genome involves the construction of large stretches of overlapping (contiguous) segments of DNA using the cloned human DNA segments.

A number of popular cloning systems have been used for physical mapping. Each of these systems differs in the amount of exogenous (human) DNA that can be faithfully propagated. One of the more traditional systems for cloning/mapping involves the use of bacteriophage lambda (Murray and Murray, 1974). The bacteriophage lambda vectors can faithfully maintain 9-23 kilo base pairs (kb) of exogenous DNA. However, individual bacteriophage clones lyse their bacterial hosts and are not suitable for the production of high density bacterial or yeast colony filters (Olsen

et al., 1993). Cosmid vectors take advantage of the bacterial packaging feature of lambda phages but do not include the majority of the viral genome and can propagate exogenous segments of DNA with an average size of 40 kb (de Jong *et al.*, 1989). The cosmid vectors replicate as large high-copy number plasmids and can be arrayed in high-density bacterial colony grids for analysis. Cloning with cosmid vectors is so efficient that chromosome specific cosmid libraries have been constructed for the majority of human chromosomes by Los Alamos and Livermore National Laboratories. These libraries are constructed from individual flow sorted human chromosomes derived from human/rodent hybrid cell lines which contain a single human chromosome. Recently, larger clones that are replicated in bacterial or yeast hosts have also been constructed.

The larger bacterial clones are made with packaging extracts (bacteriophage P1) (Sternberg, 1990; Pierce *et al.*, 1992) or through electroporation of bacterial artificial chromosomes (BACs) (Shizuya et al., 1992) and bacteriophage P1 derived artificial chromosomes (PACs) (Ioannou *et al.*, in press). The bacteriophage P1 system is limited based upon the packaging extracts to an average DNA fragment size of 90 kb. Total human genomic libraries constructed from the BAC and PAC systems have an average insert size of 150 kb. In addition, each of these clones is non-chimeric (non-mixed) in structure and stable across multiple generations of growth (Shizuya *et al.*, 1992; Ioannou *et al.*, in press). One of the most popular physical mapping systems involves Yeast Artificial Chromosomes (YACs) (Burke *et al.*, 1987). The YAC cloning system will replicate very large segments of exogenous DNA (100-1000 kb).

The choice of a vector for physical mapping the human genome is complicated by a number of factors. On the surface, this choice would appear to be a simple matter based only upon the size of the cloned segment of DNA and the desired resolution of the map, with smaller segments offering higher resolution. However, each and every cloning system has some type of limitation imposed by a variety of different factors. The YAC system, for instance, suffers from a high rate of clone instability, chimerism of cloned DNA molecules (50-60 %), internally deleting clones and difficulty in handling. The construction and array of a total genomic cosmid library would require thousands of 96-well microtiter plates full of clones for a five-fold redundant representation of the entire human genome. The maintenance of clone libraries this size is very laborious and nearly impossible without robotic assistance (Wong *et al.*, in press). The ease of manipulation, defined bacterial host and size of cloned DNA molecules make BACs and PACs attractive alternatives to YACs for physical mapping the human genome (Rouquier *et al.*, in press).

After the physical map is complete, a map with the ultimate level of resolution derived from DNA sequences will be constructed. This type of map will involve determining the sequence of each and every nucleotide 4-10 times to establish the correct sequence. Currently this type of work is divided into two phases, refinements in DNA sequencing capability and production DNA sequencing. The majority of production DNA sequencing within humans has been devoted to directed projects involving protein encoding sequences (Adams *et al.*, 1991) or entire genomic regions around a variety of disease genes. Refinements in technology have involved increasing the capacity of current automated DNA sequencers as well as attempts to develop completely new approaches to DNA sequencing. At the moment, improvements in established technology appear to be much more rapid than the development of new DNA sequencing techniques.

The human genome project is still in early stages of its development. However, a number of interesting genes have already been discovered as a direct result of the genome project. These loci include myotonic dystrophy, the most prevalent form of muscular dystrophy (Aslanidis *et al.*, 1992; Brook *et al.*, 1992; Buxton *et al.*, 1992; Harley *et al.*, 1992; Mahadevan *et al.*, 1992; Mahadevan *et al.*, 1993), and Fragile-X syndrome (Verkerk *et al.*, 1991). Both of these diseases result from the expansion of a trinucleotide repeat and display genetic anticipation. The genes for Neurofibromatosis Type I (Wallace *et al.*, 1990) Neurofibromatosis Type II (Trofatter *et al.*, 1993) and Cystic Fibrosis (Rommens *et al.*, 1989) have also been identified. Over seventeen different human disease gene loci have been identified in the last four years. The identification of these genes is the first step toward any type of potential gene therapy for these diseases. The identification and the characterization of variation within an even greater number of human disease genes should continue to provide a very bright future for the human genome project.

REFERENCES

1. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CM, Wu A, Olde B, Moreno RF, Kerlavage AR, McCombi WR, Venter JC: Science 252:1651-1651, 1991

2. Aslanidis C, Jansen G, Amemiya C, Shutler G, Mahadevan M, Tsilfidis C, Alleman J, Wormskamp NGM, Vooijs M, Buxton J, Johnson K, Smeets HJM, Lennon GG, Carrano AV, Korneluk RG, Wieringa B, de Jong PJ: Nature 355:548-551, 1992

3. Botstein D, White RL, Skolnick MH, Davis RW: Amer. J Hum Genet 32:314-331, 1980

4. Brook JD, McCurrach ME, Harley HG, Buckler AJ, Church D, Aburatani H, Stanton VP, Thition JP, Hudson T, Sohn R, Zemelman B, Snell RG, Rundle SA, Crow S, Davies J, Shelbourne P, Buxton J, Jones C, Juvonen V, Johnson K, Harper PS, Shaw DJ, Housman DE: Cell 68:799-708, 1992

5. Burke DT, Carle GF, Olson MV: Science 236:806-812, 1987

6. Buxton J, Shelbourne P, Davies J, Jones C, van Tongeren T, Aslanidid C, de Jong P, Jansen G, Anvret M, Riley B, Williamson R, Johnson K: Nature 355:547-548

7. de Jong PJ, Yokabata K, Chen C, Lohman F, Pedersen L, McNinch J, Van Dilla M: Cytogen. Cell Genet. 51:985, 1989

8. Fu YH, Pizzuti A, Fenwick RG Jr, King J, Rajnarayan S, Dunne PW, Dubel J, Nasser GA, Ashizawa T, de Jong P, Wieringa B, Korneluk R, Perryman MB, Epstein HF, Caskey CT: Science 255:1256-1258, 1992

9. Harley HG, Brook JD, Rundle SA, Crow S, Reardon W, Buckler AJ, Harper PS, Housmen DE, Shaw DJ: Nature 355:545-546, 1992

10. Ioannou PA, Amemiya CT, Garnes J, Kroisel PM, Shizuya H, Chen C, Batzer MA, de Jong PJ: Nature Genet, In Press, 1994

11. Mahadevan M, Tsilfidis C, Sabourin L, Shutler G, Amemiya C, Jansen G, Neville C, Narang M, Barcelo J, O'Hoy K, Leblond S, Earle-Macdonald J, de JOng PJ, Wieringa B, Korneluk RG: Science 255:1253-1255, 1992

12. Mahadevan MS, Amemiya C, Jansen G, Sabourin L, Baird S, Neville CE, Wormskamp N, Segers B, Batzer M, Lamerdin J, de Jong P, Wieringa B, Korneluk RG: Hum Mol Gen 2:299-304, 1993

13. Miyamoto MM, Slightom JL, Goodman M: Science 238:369-373, 1987

14. Murray NE, Murray K: Nature 251:476-481, 1974

15. Nakamura Y, Leppert M, O'Connell P, Wolff R, Holm T, Culver M, Martin C, Fujimoto E, Hoff M, Kumlin E, White R: Science 235:1616-1622, 1987

16. Olsen AS, Combs J, Garcia E, Elliot J, Amemiya C, de Jong P, Threadgill G: BioTechniques 14:116-123, 1993

17. Ott J: Analysis of human genetic linkage. Baltimore, MD, Johns Hopkins Univ Press, 1991

18. Pierce JC, Sauer B, Sternberg N: Proc Natl Acad Sci USA 89:2056-2060, 1992

19. Rommens JM, Iannuzzi MC, Kerem BS, Drumm ML, Melmer G, Dean M, Rozmahel R, Cole JL, Kennedy D, Hidaka N, Zsiga M, Buchwald M, Riordan JR, Tsui LC, Collins FS: Science 245:1059-1065, 1989

20. Rouquier S, Batzer MA, Giorgi D: Anal Biochem, In Press, 1994

21. Shizuya H, Birren B, Kim U-J, Mancino V, Slepak T, Tachiiri Y, Simon M: Proc Natl Acad Sci USA 89:8794-8797, 1992

22. Sternberg, N: Proc Natl Acad Sci USA 87:103-107, 1990

23. Trofatter JA, MacCollin MM, Rutter JL, Murrell JR, Duyao MP, Parry DM, Eldrige R, Kley N, Menon AG, Pulaski K, Haase VH, Ambrose CM, Monroe D, Bove C, Haines JL, Martuza RL, MacDonald ME, Seizinger BR, Short MP, Buckler AJ, Gusella JF: Cell 72:791-800, 1993

24. Verkerk AJMH, Pieretti M, Sutcliffe JS, Fu YH, Kuhl DPA, Pizutti A, Reiner O, Richards S, Victoria MF, Zhang F, Eussen BE, van Ommen GJB, Blonden LAJ, Riggins GJ, Chastin JL, Kunst CB, Galjaard H, Caskey CT, Nelson DL, Oostra BA, Warren ST: Cell 65:905-914, 1991

25. Wallace MR, Marchuk DA, Andersen LB, Letcher R, Odeh HM, Saulino AM, Fountain JW, Brereton A, Nicholson J, Mitchell Al, Brownstein BH, Collins FS: Science 249:181-186, 1990

26. Wong BS, de Jong PJ, Batzer MA: Anal Biochem, In Press, 1994

ACKNOWLEDGMENTS

Work at Lawrence Livermore National Laboratory was performed under the auspices of the U.S. Department of Energy contract No.W-7405-ENG-48.

DATE
FILMED
5/2/94
END