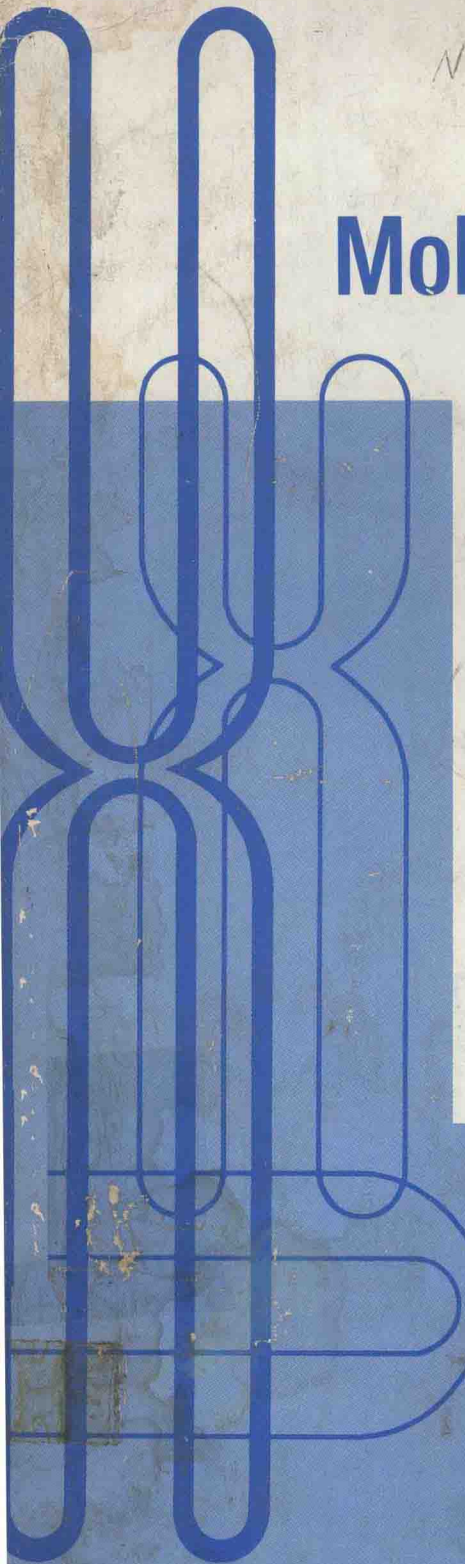
Chromosomes in Biology and Medicine

Molecular Structure of Human Chromosomes



edited by

JORGE J. YUNIS, M.D.

ACADEMIC PRESS, INC. A Subsidiary of Harcourt Brace Jovanovich, Publishers

Molecular Structure of Human Chromosomes

Edited by

JORGE J. YUNIS, M.D.

Medical Genetics Division
Department of Laboratory Medicine and Pathology
University of Minnesota, Medical School
Minneapolis, Minnesota

ACADEMIC PRESS New York San Francisco London 1977

A Subsidiary of Harcourt Brace Jovanovich, Publishers

ACADEMIC PRESS, INC.
111 Fifth Avenue, New York, New York 10003

United Kingdom Edition published by
ACADEMIC PRESS, INC. (LONDON) LTD.
24/28 Oval Road, London NW1

Library of Congress Cataloging in Publication Data

Main entry under title:

Molecular structure of human chromosomes.

(Chromosomes in biology and medicine series)
Bibliography: p.
Includes index.
1. Human chromosomes. 2. Molecular genetics. I. Yunis, Jorge J. II. Series: Chromosomes in biology and medicine. [DNLM: 1. Cytogenetics. 2. Chromosomes–Ultrastructure. 3. Molecular biology. QH600 M718]
QH600.M64 599'.9'048732 76-50672
ISBN 0–12–775168–8

PRINTED IN THE UNITED STATES OF AMERICA

Molecular Structure of Human Chromosomes

CHROMOSOMES IN BIOLOGY AND MEDICINE

Edited by

JORGE J. YUNIS, M.D.

New Chromosomal Syndromes, 1977

Molecular Structure of Human Chromosomes, 1977

List of Contributors

Numbers in parentheses indicate the pages on which authors' contributions begin.

G. F. BAHR (143), Department of Cellular Pathology, Armed Forces Institute of Pathology, Washington, D. C.

GIORGIO BERNARDI (35), Laboratoire de Génétique Moléculaire, Institut de Biologie Moléculaire, Université de Paris, Paris, France

RICHARD P. CREAGAN* (89), Department of Biology, Yale University, New Haven, Connecticut

BERNARD DUTRILLAUX (233), Institut de Progénèse, l'Ecole de Médecine, Paris, France

K. W. JONES (295), Institute of Animal Genetics, Edinburgh, Scotland

GABRIEL MACAYA (35), Laboratoire de Génétique Moléculaire, Institut de Biologie Moléculaire, Université de Paris, Paris, France

P. L. PEARSON (267), Instituut voor Anthropogenetica, Leiden, Holland

POTU N. RAO (205), Department of Developmental Therapeutics, The University of Texas System Cancer Center, M. D. Anderson Hospital and Tumor Institute, Houston, Texas

FRANK H. RUDDLE (89), Department of Biology, Yale University, New Haven, Connecticut

DALE M. STEFFENSEN (59), Department of Genetics and Development, University of Illinois, Urbana, Illinois

JEAN-PAUL THIERY† (35), Laboratoire de Génétique Moléculaire, Institut de Biologie Moléculaire, Université de Paris, Paris, France

MICHAEL Y. TSAI (1), Medical Genetics Division, Department of Laboratory Medicine and Pathology, University of Minnesota Medical School, Minneapolis, Minnesota

*Present address: Department of Physiology, School of Medicine, University of Connecticut Health Center, Farmington, Connecticut

†Present address: The Rockefeller University, New York, New York.

ANN M. WILLEY (1), Medical Genetics Division, Department of Laboratory Medicine and Pathology, University of Minnesota Medical School, Minneapolis, Minnesota

JORGE J. YUNIS (1), Medical Genetics Division, Department of Laboratory Medicine and Pathology, University of Minnesota Medical School, Minneapolis, Minnesota

Preface

In recent years, a number of important developments have led to a dramatic increase in our understanding of chromosomes in biology and medicine. The sequence arrangement, transcriptional capacity, and functional significance of repeated and unique DNA sequences have been extensively studied. Refined investigations of chromosome and chromatin structure have been made possible with improvements in electron microscopic techniques, the use of premature chromosome condensation, and *in situ* hybridization. By the use of various dyes and treatments, mitotic chromosomes can now be visualized as having characteristic banding patterns, facilitating identification of individual chromosomes. Relationships between the human karyotype and those of other primates have been determined using differential staining techniques and comparative analysis of repeated DNA sequences. Gene mapping has been greatly accelerated with the use of rodent – human somatic hybrids.

Reviews of advances on the organization of the eukaryotic genome have been available for some time, yet they have not focused on the genome of greatest interest to us, that of man. The aim of this volume is to fill this gap by bringing together authoritative contributions encompassing much of the knowledge available on the fine structure and molecular organization of the human genome.

Geneticists, molecular biologists, and cytogeneticists, particularly those interested in the human genome, will find this book to be an invaluable guide.

Jorge J. Yunis

Contents

1

Molecular Organization and Function of the Human Genome

JORGE J. YUNIS, MICHAEL Y. TSAI,
and ANN M. WILLEY

I. INTRODUCTION

A major obstacle encountered in studying the genome of higher eukaryotes including man has centered around the large amount of DNA present in the nuclei of these cells. Compared to the extensively studied prokaryotic genome of *Escherichia coli*, consisting of 3.2×10^6 nucleotide pairs (Cairns, 1963), the genome of the lower eukaryote *Drosophila melanogaster* contains approximately

50 times that amount of DNA (Laird, 1971); while the human genome contains 1000 times the amount of DNA of *E. coli*, or 2.7×10^9 nucleotide pairs (Sober, 1970; Rees and Jones, 1972). Most of the prokaryotic genome is transcribed (McCarthy and Bolton, 1964; Kennell, 1968; Grouse *et al.*, 1972) and codes for approximately 3000 informational genes, while the *Drosophila* and human genomes have enough DNA to encode over 150,000 and 3 million such genes, respectively. Since man and other eukaryotes probably do not need such a large number of genes, explanations for the large amounts of DNA in eukaryotes have been sought.

II. RENATURATION STUDIES OF EUKARYOTIC DNA

The first step to the solution of the apparent enigma was the demonstration by Britten and co-workers of the presence of repetitive nucleotide sequences within the DNA of eukaryotic organisms (Bolton *et al.*, 1966; Waring and Britten, 1966; Britten and Kohne, 1968). Renaturation experiments done on a number of highly sheared animal DNA's have shown relatively constant proportions of repetitive and unique sequence DNA's (Britten and Kohne, 1968; Davidson *et al.*, 1974). In mammals, for example, repetitive sequences usually account for 30–40% of the genome, and unique sequences for the remaining 60–70%. Modeling after the prokaryotic system of gene regulation, transcription and translation, it was assumed that structural gene sequences would be represented in the unique copy DNA, whereas repetitive sequences were suspected of having regulatory functions (Britten and Davidson, 1969; Georgiev, 1969). Even under this assumption, however, the question remained as to the number of informational gene sequences present in the 60–70% of the genome consisting of unique copy DNA.

Detailed characterization of the arrangement of repetitive and nonrepetitive sequences was carried out in the genome of sea urchin and *Xenopus* (Davidson *et al.*, 1973; Graham *et al.*, 1974). The techniques employed included analyses of hydroxyapatite binding as a function of fragment length after low Cot incubation with carrier DNA's, measurement of repetitive sequence length by isolation of single-strand-specific, nuclease-resistant duplex on agarose gel column, and direct measurement with electron microscopy. The results of these studies revealed that unique copy DNA averaging 1000 nucleotides alternate with repetitive sequences averaging 200–400 nucleotides in about 50% of the genome. Two other patterns were also seen: long period interspersion of unique copy DNA several thousands of nucleotides in length with a few percent of repetitive sequences in about 40% of the genome, and clustered highly repetitive sequences in 5–10% of the total DNA. Similar patterns have since been reported in other eukaryotes (Firtel and Kindle, 1975; Angerer *et al.*, 1975). The only

organism which has been shown to have a large digression from the above pattern is *Drosophila melanogaster* (Manning *et al.*, 1975). In these animals, although the highly repetitive DNA sequences show a similar uninterrupted pattern, the remainder of the genome consists of interspersed unique and repetitive sequences covering a wide range with repetitive regions ranging from 500 to 13,000 base pairs averaging 5600, and unique copy stretches ranging from 2500 to 40,000 base pairs and averaging 13,000 nucleotides.

In man, Schmid and Deininger (1975) have recently reported a pattern of interspersed repetitive, single copy and inverted repeats in 50% of the genome. The average repetitive and nonrepetitive segments were thought to be 400 and 2000 nucleotides, respectively, although measurement by either electron microscopy or isolation of single-strand, nuclease-resistant duplexes was not carried out.

Evidence supporting the idea that unique copy sequences contiguous to repetitive DNA represent structural genes has been presented by Davidson *et al.* (1975) in the sea urchin genome. These authors have shown that 80–100% of the mRNA molecules present in sea urchin embryos are transcribed from single copy DNA sequences adjacent to interspersed repetitive sequences in the genome. However, single copy DNA finely interspersed with short repetitive sequences represents 40% of the total genome. As will be described in the following section, the implication that these large amounts of DNA all represent structural genes contradicts other estimates of gene number, which conclude only 1–6% of the eukaryotic genome represents informational genes.

III. GENE NUMBER AND GENOME TRANSCRIPTION

Several lines of evidence exist in favor of the tenet that only a small percentage of the eukaryotic genome represents structural genes. Ohta and Kimura (1971) postulated that less than 6% of the mammalian genome represents structural genes based on mutation rate and the resultant genetic load. In the giant salivary chromosomes of *Drosophila*, Judd and co-workers (Judd *et al.*, 1972; Judd and Young, 1974) observed that each chromomere represents one functional genetic unit, putting the total number of informational genes in this species at about 5000. In agreement with this view, it has been found that cytoplasmic polysomal RNA's from a large number of eukaryotes, including *Drosophila* and man, transcribe from approximately 2% of the genome (Greenberg and Perry, 1971; Galau *et al.*, 1974; Lewin, 1975). Moreover, Bishop *et al.* (1975) have shown that in *Drosophila* the total number of mRNA sequences expressed within the different stages of the life cycle of the fly do not exceed the 5000 chromomeres of the polytene chromosomes.

In contrast to this low amount of DNA responsible for mRNA transcription, it is known that a large amount of the DNA of the eukaryotic genome is transcribed but not translated. On the average, assuming asymmetrical transcription, total cellular RNA is transcribed from at least 10–30% of the single copy DNA of the total genome (Davidson and Hough, 1971; Gelderman *et al.*, 1971; Grouse *et al.*, 1972, 1973; Grady and Campbell, 1973; Turner and Laird, 1973). Additive experiments with RNA's from different organs showed that there is a considerable, although not total, overlap between RNA's from different tissues. In the case of the slime mold, it was shown that, overall, 56% of the genome is represented by transcripts between the amoeba and midculmination stages (Firtel, 1972).

A partial explanation for the discrepancy found between the large amount of transcribed DNA and the small percentage of the genome believed to represent messages can be found in recent studies on heterogeneous nuclear RNA (HnRNA). In mammalian cells, HnRNA was found to have more than five times the complexity of mRNA (Getz *et al.*, 1975), and in sea urchin embryos mRNA represents 2.7% of the genome, while 28.5% of the total single copy DNA hybridizes to HnRNA (Galau *et al.*, 1974; Hough *et al.*, 1975). Since there is also evidence that a large portion of the HnRNA of the sea urchin is composed of interspersed nonrepetitive and repetitive sequences (Smith *et al.*, 1974), it is possible that a large portion of the 50% finely interspersed single copy and repetitive DNA of the sea urchin is involved in HnRNA transcription.

HnRNA's are generally characterized by their overall rapid synthesis and degradation, large molecular weight, DNA-like base composition, their presence in all eukaryotes examined, and by the fact that the bulk of this class of molecules never leaves the nucleus (Sibatani *et al.*, 1962; Georgiev and Mantieva, 1962; Scherrer *et al.*, 1963; for review, also see Georgiev, 1974; Darnell, 1975). Besides the unique copy sequences of HnRNA that are known to have from 5 to 10 times the complexity of mRNA, HnRNA's also contain sequences transcribed from DNA of different degrees of repetitiveness. Holmes and Bonner (1974a) have suggested that in rat ascites cells, HnRNA molecules contain at least one middle repetitive sequence covalently attached to a single copy sequence. In these cells, HnRNA's are transcribed from approximately 12% of the genome, of which approximately 25% comes from repetitive and 75% from single copy DNA (Holmes and Bonner, 1974a,b). One type of repetitive sequence is characterized by its resistance to pancreatic RNase and is thought to be a double-stranded region that is formed by intramolecular base pairing. When denatured, the RNA sequences from the double-stranded regions, including those from HeLa cells, hybridize to DNA at a $Cot_{1/2}$ of about 10 (Jelinek and Darnell, 1972; Ryskov *et al.*, 1973a). A second type of repetitive sequence found in HnRNA of HeLa cells is largely (about 80%) made up of uridylic acid, is about 30 nucleotides in length and also hybridizes at a $Cot_{1/2}$ of about 10 (Molloy *et al.*, 1972). This oligo(U)

segment is located "only" in large (70–90 S) HnRNA molecules, is T_1 ribonuclease-resistant and pancreatic ribonuclease-sensitive. Recently, Molloy *et al.* (1974) found in HeLa cells that poly(A)-terminated HnRNA molecules longer than 20,000 nucleotides contain 2–3 units of oligo(U) nucleotides in fragments over 12,000 nucleotides away from the 3′ poly(A) ("messenger" end), while double-stranded regions are found between 3000 and 8000 nucleotides from this end.

That HnRNA may be the precursor of mRNA was suggested by a number of investigators soon after its discovery (Scherrer *et al.*, 1963; Penman *et al.*, 1963). Several models of genetic regulation have been proposed based on this concept (Scherrer and Marcaud, 1968; Georgiev, 1969, 1974; Darnell *et al.*, 1973). Although differing in many details, these authors all share the view that HnRNA's consist of a noninformative region, which presumably has some regulatory function, and an informative segment representing mRNA. In agreement with the view that messenger RNA's are formed from the posttranscriptional modification of higher molecular weight precursor RNA, it has been found that rRNA's are processed in a nonconservative fashion from 45 S RNA in which approximately 50% of the molecule is lost in the formation of 28 S and 18 S products (Maden, 1971; Choi and Saunders, 1974). Similarly, tRNA is processed from a higher molecular weight pre-tRNA (Burdon and Clason, 1969).

Several lines of experimental evidence exist implicating HnRNA as the precursor of mRNA (for review, see Darnell, 1975). For instance, sequences transcribed from integrated DNA of tumor virus and sequences present in hemoglobin and immunoglobulin have been found in HnRNA (Lindberg and Darnell, 1970; Melli and Pemberton, 1972; Imaizumi *et al.*, 1973; Williamson *et al.*, 1973). More recently, Herman *et al.* (1976) prepared cDNA to mRNA of HeLa cells and found that the bulk of the cDNA hybridized to isolated HeLa cell HnRNA. If HnRNA indeed contains mRNA as a part of its structure, the question then arises as to how it may participate in the regulation of mRNA formation. Although precise knowledge is not available, it is conceivable that the nonmessage part of HnRNA is needed either for transcriptional or posttranscriptional control of mRNA formation. In the former case, the nonmessage part may be transcribed from sites of recognition similar to those found in bacteria. The presence of possible recognition sites in eukaryotic cells has been proposed in several models to explain gene regulation in higher eukaryotes (Britten and Davidson, 1969; Georgiev, 1969, 1974; Darnell *et al.*, 1973). In contrast to the bacterial system, however, transcription of a large portion of nonmessage sequence into HnRNA raises the possibility of posttranscriptional regulation in addition to or in place of transcriptional control. Models for this type of regulation have been proposed (Scherrer and Marcaud, 1968; Darnell *et al.*, 1973) in which regulatory proteins would interact with HnRNA either to activate or suppress its processing into mRNA (Darnell, 1975).

IV. CORRELATION OF TRANSCRIPTION WITH SEQUENCE ARRANGEMENT PATTERNS

In correlating the information from recent studies of sequence arrangement and transcription of eukaryotes, many new questions arise. For instance, there is no satisfactory explanation for the discrepancy between the small amount of messenger RNA (1–2%) found and the large excess of repeat-contiguous single copy DNA (~40%) of putative structural genes. Even more puzzling is the function of long period interspersed single copy DNA which constitute an additional 40% of the total DNA and for which no specific function has been formulated.

Part of this problem stems from our lack of knowledge of the total complexities of messenger and heterogeneous nuclear RNA that are expressed in different stages of development and in different cell types. Additive studies such as those performed in *Drosophila* (Bishop *et al.*, 1975) on the complexities of mRNA should be extended to mammalian HnRNA and mRNA as well to obtain independent determination on the approximate number of structural genes in different species. On the other end of the spectrum are the difficulties encountered in interpreting the results obtained from sequence arrangement studies. Since DNA renaturation techniques are used in these studies, the results heavily depend on the fidelity of duplex reassociation. This process is best described as an "association" because, when dealing with repetitive sequences, the two strands forming the duplex may not be identical to the original paired strands, and mismatching due to base substitution would greatly affect the structures seen. Therefore, it is difficult to be certain that all of the nonassociated sequences adjacent to repetitive segments are representative of unique copy DNA and not of repetitive components with large accumulations of base substitutions. It is conceivable that, just as intermediate repetitive sequences may have evolved from simple sequences disguised by mismatching, further base substitution can cause such sequences to reassociate as single copy DNA (Walker, 1971).

V. FRACTIONATION OF NATIVE DOUBLE-STRANDED DNA

A different approach in the study of the eukaryotic genome makes use of fractionation of total native DNA through the use of cesium salt density gradient fractionation. Progress was first made with this technique in the early 1960's when minor components differing in buoyant densities from the bulk of the DNA were observed in the mouse, guinea pig, calf, and crab when centrifuged in neutral cesium chloride density gradients (Kit, 1961,