

THE STRUCTURE OF MAMMALIAN GENES:

- (1) ANTIBODY HEAVY CHAIN VARIABLE REGION GENES: ORGANIZATION,  
DIVERSITY, AND SOMATIC MUTATION.
- (2) STRUCTURE AND TRANSCRIPTION OF THE DNA ENCOMPASSING THE  
ORIGIN OF REPLICATION OF HUMAN MITOCHONDRIAL DNA.

Thesis by

Stephen Thomas Crews

In Partial Fulfillment of the Requirements

For the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

1983

(Submitted September 22, 1982)

This thesis is dedicated to my parents and wife  
for their love and encouragement.



## ACKNOWLEDGEMENTS

I would like to thank the two advisors that I worked with, Leroy Hood and Giuseppe Attardi. Both are outstanding scientists and good teachers with a strong concern for their craft.

I am also grateful to other faculty members that I was able to interact and learn with. These include Norman Davidson, Ray Owen, Eric Davidson, Jim Strauss, Ed Lewis, Judy Campbell, Charles Brokaw, Bill Dreyer, and Jerome Vinograd.

Much of the work in this thesis could not have been accomplished without the collaborative contributions of Elizabeth Springer, Johanna Griffin, Henry Huang, David Lo, Kathryn Calame, Deanna Ojala, Chris Merkel, Bob Gelfand, Jerry Nishiguchi, Karel Grohmann, and Jim Posakony. I have also had consistently rewarding discussions with Ed Ching, Francois Amalric, Mark Davis, Tim Hunkapiller, Mitch Kronenberg, and Joel Buxbaum.

Finally, I would like to thank all of the people who make Caltech an enjoyable place to work. These include Bertha Jones, Maria DeBruyn, Doris Finch, Gloria Engel, Richard and Frank in the shop, and the staffs of the Biology office, secretarial office, Biology stockroom, and graphic arts.

## ABSTRACT

This thesis describes two experimental systems utilized to study mammalian gene structure and expression: (1) antibody heavy chain variable region genes and (2) mitochondrial DNA.

In order to study the organization and structure of antibody genes and the relative germline and somatic contributions towards antibody diversity, we have analyzed the germline genes encoding the murine immune response to phosphorylcholine. Molecular cloning studies were undertaken and conclusively show that there is only one germline  $V_H$  gene segment encoding the immune response to phosphorylcholine. Protein sequencing work on monoclonal antibodies that bind phosphorylcholine reveals many different protein sequences related to one predominant sequence. We are able to conclude that these variant sequences are the result of somatic diversification operating on one germline gene segment. We are further able to show that this diversification is mutational and not recombinational. Finally, somatic mutation is correlated with the class of the antibody; IgG and IgA antibodies undergo somatic mutation, IgM antibodies do not.

We have isolated and sequenced a family of four closely related  $V_H$  gene segments designated V1, V3, V11 and V13. Their function varies: V1 encodes the immune response to phosphorylcholine, V3 is a pseudogene, V11 encodes the immune response to influenza hemagglutinin, and V13 has an unknown function but is not obviously a pseudogene. Structural analysis of recombinant clones containing this family of related  $V_H$  gene segments and other  $V_H$  gene segments reveals several important points about the organization of  $V_H$  gene segments. First, closely related  $V_H$  gene segments can be clustered together within the  $V_H$  gene locus. Second, the spacing distance between adjacent  $V_H$  gene segments is variable; it may be as short as 5 kb and greater than 30 kb. Finally, the average spacing distance between  $V_H$  gene segments is large, at least 23 kb. Assuming a minimum of 200 germline  $V_H$  gene

segments, the size of the  $V_H$  gene locus may be greater than 5 million base pairs.

The human mitochondrial genome is the second system that has been chosen to study gene structure and expression, and to accomplish this, we have applied both DNA and RNA sequencing technologies. We sequenced the DNA encompassing the origin of DNA replication and then localized the origin at the nucleotide level. The human mitochondrial origin of DNA replication shares structural characteristics with other known origins of DNA replication; in particular, the presence of extensive secondary structure in the form of a stem-loop structure. In order to precisely localize mitochondrial transcripts to the DNA, we developed techniques that allowed the isolation and sequencing of the 5'-ends of mitochondrial transcripts. This technology was utilized to precisely localize the 5'-end of the mitochondrial 12S rRNA species 457 nucleotide pairs 5'-to the origin of DNA replication. Analysis of the DNA sequence in this region revealed a phenylalanine tRNA gene whose 3'-end was joined end-to-end with the 5'-end of the 12S rRNA. This analysis first demonstrated the extreme economy of genetic material in mammalian mitochondrial DNA.

## TABLE OF CONTENTS

Acknowledgements . . . . .	iii
Abstract. . . . .	iv
Introduction Part I . . . . .	1
Chapter 1. A Single $V_H$ Gene Segment Encodes the Immune Response to Phosphorylcholine: Somatic Mutation is Correlated With the Class of the Antibody . . . . .	18
Chapter 2. An Immunoglobulin $V_H$ Pseudogene . . . . .	27
Chapter 3. The Chromosomal Organization of Antibody $V_H$ Gene Segments. . . .	37
Conclusion Part I . . . . .	60
Introduction Part II . . . . .	67
Chapter 4. Nucleotide Sequence of a Region of Human Mitochondrial DNA Containing the Precisely Identified Origin of Replication . . . . .	73
Chapter 5. The Sequences of the Small Ribosomal RNA Gene and the Phenylalanine tRNA Gene are Joined End to End in Human Mitochondrial DNA . . . . .	81
Chapter 6. A Small Polyadenylated RNA (7S RNA), Containing a Putative Ribosome Attachment Site, Maps Near the Origin of Human Mitochondrial DNA Replication . . . . .	92
Conclusion Part II. . . . .	106

INTRODUCTION

PART I

## INTRODUCTION

### **Antibody Structure and Synthesis**

The vertebrate immune system protects the animal from pathogenic organisms and substances. One of the most elaborate parts of that system is the antibody response. Antibodies are proteins that are able to recognize and tightly bind molecules. They are generally characterized by their extensive diversity which allows them to bind an almost limitless variety of substances. After a substance is bound by an antibody, its destructive potential is either directly neutralized or else eliminated by interaction with macrophages or complement.

The basic structural unit of the antibody consists of two heavy chains and two light chains. Protein sequence analysis of many different antibody molecules has shown that for each chain the N-terminal 110 amino acids is more variable in amino acid sequence than the rest of the molecule, thus each chain is conveniently divided into a variable and constant region (Wu and Kabat, 1970). Furthermore, within the variable regions there are regions of hypervariability, three regions for each chain, which are separated by framework regions (Wu and Kabat, 1970; Capra and Kehoe, 1974). It is now well established that the variable regions combine to form the antigen binding part of the molecule and X-ray crystallographic studies have shown that the hypervariable regions are the regions within the variable region that generally interact with antigen (Padlan et al., 1973; Amzel et al., 1974). The constant regions carry out the effector functions of the antibody, such as binding complement and also determine the class or type of the chain. There are two types of light chain: kappa and lambda, and five classes of heavy chain,  $\mu$ ,  $\delta$ ,  $\gamma$ ,  $\epsilon$ , and  $\alpha$ . The heavy chain classes determine the class of the antibody, IgM, IgD, IgG, IgE, and IgA.

Antibodies are synthesized by B lymphocytes. Adult mice generate precursor B cells from bone marrow-derived stem cells. These precursor B cells express

antibody as a receptor on their cell surface. When antigen is first recognized by these B cells, they become activated and proliferate. Some of the progeny cells begin to secrete antibody. This antibody constitutes the primary immune response and is characterized by the predominant synthesis of IgM. Another characteristic of the primary response is that the antibody generally has a low affinity for the antigen. There is another class of cells derived from antigen-stimulated B cells. These are memory cells and they are long-lived cells capable of eliciting a rapid, intense response upon antigenic stimulation at a later time. This is called the secondary immune response and is characterized by antibodies that are predominantly IgG. During the secondary response, the affinity of the antibody for the antigen usually increases in a process called affinity maturation (Eisen and Siskind, 1964). It is worth mentioning that cells synthesizing IgG, IgA, IgE or IgD are derived from cells that initially expressed IgM (Davis et al., 1980). This process is referred to as the class switch.

### **Theories of Antibody Diversity**

It has been estimated that mammals can synthesize up to  $10^7$  different antibodies. This explains why an organism's antibodies can bind almost any substance that it comes in contact with. The wide range in antigen-binding specificity can be explained by the amino acid sequence diversity of the antibody variable regions. Numerous theories have been proposed to account for the genetic basis of this diversity. The most prominent are the (1) germline gene theory, (2) mini-gene theory, (3) somatic mutation theory, and (4) somatic recombination theory.

The germline gene theory hypothesizes that there are a large number of variable region genes in the germline DNA (Hood et al., 1976). Expression of these genes in lymphocytes provides the antibody repertoire of the individual. The mini-gene hypothesis states that the variable region genes are assembled from smaller gene segments during development (Kabat et al., 1979). For example, in the germline

DNA, there might be a separate set of gene segments corresponding to each of the heavy chain framework and hypervariable regions. There would be multiple germline DNA gene segments for each mini-gene set. These gene segments could be assembled combinatorially. If each of the four framework regions could be encoded by two possible gene segments and each of the three hypervariable regions could be encoded by 10 possible gene segments, combinatorial association would lead to  $2 \times 10 \times 2 \times 10 \times 2 \times 10 \times 2 = 8 \times 10^3$  different heavy chains. The somatic mutation theory predicts that there are a small number of germline variable region genes (Jerne, 1971; Cohn et al., 1974). As precursor B cells are generated during the development of the immune system these germline genes undergo somatic mutation. Different precursor B cells would undergo different mutational events. In that way, the repertoire would be sufficiently large. There is a variation of the somatic mutation theory called the somatic recombination theory. It postulates that somatic recombination events between variable region genes could produce hybrid genes thus expanding the diversity from a limited number of germline genes (Edelman and Gally, 1967). With the advent of recombinant DNA technology, the availability of monoclonal antibodies, and advances in protein sequencing methodology, these models have become directly testable.

### **Antibody Gene Structure**

Recently, a detailed picture of the structure and expression of antibody genes has emerged. There are three families of antibody genes in mouse, each on a different chromosome. The lambda and kappa gene families encode light chains and the heavy gene family encodes heavy chains. A heavy chain variable region ( $V_H$ ) is encoded by three different mini-gene segments:  $V_H$  (variable) gene segment, D (diversity) gene segment, and  $J_H$  (joining) gene segment (Early et al., 1980). The  $V_H$  gene segment codes for approximately the first 100 amino acids, the D segment is variable in length and codes for the next 1 to 10 amino acids and the  $J_H$  segment



codes for the last 14 to 16 residues. These gene segments are separated from each other in germline DNA. During B cell differentiation, they are joined together, thus creating an assembled variable region gene. There are a large number of germline gene segments, thus fulfilling the prediction of the germline gene hypothesis (Roy Riblet, personal communication). There are probably greater than 200  $V_H$  gene segments, 10-20 D gene segments, and four  $J_H$  gene segments. The  $J_H$  gene segments lie 6 kb 5' to the  $C_\mu$  gene. After the  $V_H$  and D gene segments have been joined to one of the  $J_H$  gene segments, the gene is transcribed and processed into  $\mu$  mRNA. Later, another type of DNA rearrangement called the switch recombination can occur resulting in the switch of the antibody class (Davis et al., 1980). Mice have eight constant ( $C_H$ ) region genes clustered together:  $\mu$ ,  $\delta$ ,  $\gamma_3$ ,  $\gamma_1$ ,  $\gamma_{2b}$ ,  $\gamma_{2a}$ ,  $\alpha$ , and  $\epsilon$  (Shimizu et al., 1982). There is a sequence 5' to each  $C_H$  gene called the Switch (S) sequence where the switch recombination event occurs. The recombination between the S region of the  $C_\mu$  gene and the S region of the  $C_\alpha$  gene results in deletion of the intervening DNA, including the  $C_\mu$  gene, and allows expression of  $\alpha$  chains with the same variable region utilized by the  $\mu$  chains.

The organization of the light chain genes is similar but not as complex as the heavy chain genes. Mice have only one kappa constant region gene ( $C_\kappa$ ). It is preceded by a cluster of five  $J_\kappa$  gene segments; four are functional and one is a pseudogene (Max et al., 1979; Sakano et al., 1979). In addition, there appear to be a large number of  $V_\kappa$  gene segments. Estimates range from 100 to 2000  $V_\kappa$  gene segments (Cory et al., 1981; Zeelon et al., 1981). There are no D gene segments; the variable region is encoded by only a  $V_\kappa$  and a  $J_\kappa$  gene segment. The formation of a functional light chain gene requires that a  $V_\kappa$  gene segment be joined to a  $J_\kappa$  gene segment. A nuclear transcript containing the  $V_\kappa$ ,  $J_\kappa$ , and  $C_\kappa$  gene segments is synthesized and spliced to form the mature kappa mRNA.

The lambda light chain gene family is very small in mice, consisting of two  $V_\lambda$  and three functional  $C_\lambda$  gene segments, each  $C_\lambda$  gene segment having its own  $J_\lambda$  gene

segment (Blomberg et al., 1981). The limited germline gene diversity of lambda chains with respect to kappa chains is consistent with the observation that lambda chains contribute only 5% of the serum antibody in mouse.

### **Origins of Antibody Diversity**

The "Theories of antibody diversity" section listed several ideas postulated to explain antibody diversity. The recently discovered structure of antibody genes and an understanding of how the functional gene is assembled have yielded several answers to the diversity problem. Variable region genes are assembled from mini-genes and there are a large number of germline gene segments. Additionally, there are somatic diversification mechanisms involved in antibody gene formation.

**Germline and combinatorial mechanisms.** Both heavy and light chain gene families contain an abundant number of germline gene segments. The smallest gene family, the lambda family, has only two V gene segments and 3 functional J gene segments. However, the kappa family probably has at least 250 V gene segments and 4 J gene segments. It is assumed that any V gene segment can join to any J gene segment. By combinatorial association of V and J gene segments, at least  $250 \times 4 = 1000$  different light chains can be formed. The heavy chain family can be estimated to consist of 250 V, 10 D and 4 J gene segments. Combinatorial association results in  $250 \times 10 \times 4 = 10000$  different heavy chains. Finally, if any heavy chain can associate with any light chain, then  $1000 \times 10000 = 10^7$  different antibodies are possible. Clearly, the large number of germline gene segments and their combinatorial properties can result in an extensive antibody repertoire.

**Junctional diversity.** The first type of somatically derived diversification mechanism occurs during the joining together of the gene segments that comprise the variable region gene. Initially shown for murine kappa chains, the recombination mechanism that joins the gene segments is imprecise (Max et al., 1979; Sakano et al., 1979; Weigert et al., 1980). By shifting the site of recombination, the same germline

gene segments can have different codons at the recombination site. Codon deletions and insertions are also possible. The only requirement for functional light chain variable region gene formation is that the correct translational reading frame be maintained after the joining of  $V_L$  and  $J_L$ . There are examples where the recombination event has occurred and altered the translational reading frame thereby leading to a nonfunctional gene (Max et al., 1980; Altenburger et al., 1980). The occurrence of these abortive rearrangements may be the cost incurred in order to have junctional diversity.

Heavy chain variable region genes also show the occurrence of junctional diversity (Early et al., 1980). Heavy chain gene formation has an extra level of complexity since there are two recombination events necessary to create a V gene: the V-D joining event and the D-J joining event. Junctional diversity probably occurs at both boundaries. It is formally possible that the same D gene segment could be utilized in separate joining events and be read in all three reading frames since frameshifts created at the V-D junction could be corrected at the D-J junction (Hood et al., 1980). Although the diversity that arises from joining together gene segments is limited to only a small portion of the variable region, it may be an important mechanism in generating antibody diversity since these regions can play a significant role in the binding of antibody to antigen (Hood et al., 1980).

**Somatic mutation.** Several studies have indicated that there might be another mechanism of somatic diversification that could scatter nucleotide changes throughout the variable region gene, not just at the junctions of variable region gene segment joining. The first evidence was provided by an analysis of murine lambda chains. In a large study, complete  $V_\lambda$  protein sequences were obtained from eighteen monoclonal antibodies secreted by myelomas (Weigert et al., 1970). Twelve of them were identical (designated  $\lambda_0$ ) and the other six were identical to  $\lambda_0$  except that they had from one to two amino acid substitutions. Subsequently, it was shown by molecular

cloning that BALB/c mice have only one germline gene that could code for these light chains and its sequence is identical to  $\lambda_0$  (Bernard et al., 1978). The substitutions seen in the variant lambda chains were thought most likely to be the result of somatic mutation of the  $\lambda_0$  gene. However, other explanations were possible: the substitutions may be due to polymorphism within the BALB/c mouse population or the result of transformation and propagation of the antibody-producing cells as a tumor (Ohta, 1980; Seidman et al., 1978).

Further evidence supporting the existence of somatic mutation came from an analysis of the  $V_{\kappa}$  21 group of kappa light chains. Complete protein sequences of these closely related  $V_{\kappa}$  segments revealed 16 different protein sequences (Weigert et al., 1978). Solution hybridization experiments utilizing a cDNA probe of one of these  $V_{\kappa}$  21 mRNAs revealed that the mouse genome contained only 4-6 germline genes that could encode this group of variable regions (Valbuena et al., 1978). It was concluded that some of these light chain sequences must be derived by somatic mutation of germline  $V_{\kappa}$  genes.

### **Somatic Mutation and the Immune Response to Phosphorylcholine**

Since the generality and even the existence of somatic mutation was uncertain, we chose to examine the relative germline and somatic contributions towards antibody diversity of the heavy chain. The experimental system utilized is the immune response to phosphorylcholine. This simple compound, often found on the cell surfaces of bacteria, elicits a response in BALB/c mice that is characterized by antibodies of limited heterogeneity (Claflin et al., 1976). The primary reason for studying this response is that similar to the lambda and  $V_{\kappa}$  21 light chain systems, there is an extensive body of  $V_H$  protein sequence data for heavy chains that bind phosphorylcholine.

Initially, there was complete  $V_H$  region protein sequence data on nine monoclonal antibodies derived from murine myelomas that bind phosphorylcholine

(Figure 1) (Hood et al., 1976). The sequence is divided into the three segments that correspond to the 3 gene segments:  $V_H$ , D, and  $J_H$ . All of the heavy chains use the same  $J_H$  segment,  $J_H1$ . The D segments are variable, both in length and sequence. The  $V_H$  segment which contains the first 101 amino acids of the variable region shows a very limited diversity. Five of the nine sequences have the same  $V_H$  segment sequence (designated T15). The other four are very closely related to the T15 prototype sequence but differ from it by one to eight substitutions. All four variants are different from each other. The key question is whether these variants are the product of different germline genes or whether they represent somatic variants of one germline gene segment which presumably is identical to the T15  $V_H$  segment.

The approach towards solving this problem is clear: isolate and sequence all of the possible germline gene segments that could encode the antibodies that bind phosphorylcholine and compare them to the protein sequences. Chapter 1 of the thesis describes these experiments. We were able to convincingly show that all of the  $V_H$  segments of antibodies that bind phosphorylcholine are encoded by one germline  $V_H$  gene segment, which when translated is identical in sequence to the prototype T15 sequence. The variants must be derived by a somatic mutational mechanism and from analysis of the protein sequences of additional phosphorylcholine-binding IgM and IgG hybridoma antibodies, we were able to make the interesting observation that somatic mutation was correlated with the class of the antibody. Only IgG and IgA undergo somatic mutation, IgM does not.

### **Structure and Organization of Antibody $V_H$ Gene Segments**

The goal of a better understanding of the structure, expression, and evolution of antibody variable region genes has provoked a strong interest in the cloning and sequencing of these genes. There are hundreds of variable region genes, so generally the most practical approach has been to investigate smaller families of closely related genes. As a model system we have analyzed a small family of genes

homologous to the  $V_H$  gene segments used by antibodies that bind phosphorylcholine. Utilizing a cDNA clone homologous to the mRNA encoding this  $V_H$  segment, hybridization to southern blots of mouse sperm DNA cut with restriction enzymes indicates that there are four strongly hybridizing bands, presumably corresponding to four different germline genes. Chapters 1 and 2 of the thesis describe the isolation of recombinant clones containing these four genes and the subsequent DNA sequencing of them. The sequence analysis indicates that these four genes, designated V1, V3, V11, and V13 and collectively referred to as the T15  $V_H$  gene family, function as follows. V1 encodes the  $V_H$  segment of antibodies that bind phosphorylcholine. V3 contains multiple mutations that would preclude its use as a functional gene, hence it is a pseudogene. V11 encodes the  $V_H$  segment of antibodies that bind the influenza hemagglutinin (Walter Gerhard, personal communication), and additional rare somatic variants may be able to bind phosphorylcholine. The gene also encodes the  $V_H$  segment of the antibody from the myeloma M47A, whose binding specificity is unknown. V13 does not correspond to the  $V_H$  segment of any antibody that has been sequenced. However, nothing in its sequence would indicate that it is a pseudogene. The characterization of this small, multigene family in BALB/c mice hopefully will be a starting point for an understanding of the evolution of  $V_H$  gene segments and the functional significance of the nucleotide sequences contained in and around the  $V_H$  gene segment.

Additionally, we have utilized this system to study the arrangement of  $V_H$  genes along the chromosome. Despite a good understanding of the organization of the murine constant region genes, we know relatively little about the arrangement of variable region genes. Genetic studies and analysis of cloned genes indicate that heavy chain variable region genes are linearly arranged along chromosome 12, 5' to the constant region gene locus (Weigert and Riblet, 1978). However, there are several basic questions that remain unanswered. Are closely related  $V_H$  gene

segments clustered as discrete families or are they interspersed among less related genes? Does the arrangement of the gene along the chromosome help determine how frequently it is joined to the D and J gene segments or when, in mouse development, the gene can be joined and thus expressed as antibody? How far apart are V genes spaced and do they share extensive amounts of flanking sequence homology? A key question is whether we can learn by studying their chromosomal organization and sequence how V genes and their diversity have been created and maintained throughout evolution.

Chapter 3 of the thesis describes the isolation and restriction mapping analysis of both lambda and cosmid clones containing genes of the T15  $V_H$  gene family and also less related  $V_H$  gene segments. These genes are grouped into clusters of overlapping clones. The picture we have thus far obtained is that the organization of the  $V_H$  gene locus is complex. We present evidence that closely related  $V_H$  gene segments can be closely linked, however the generality of this observation is still unknown. Analysis of around 500,000 nucleotide pairs of nonoverlapping DNA containing  $V_H$  genes indicates that the spacer distance between  $V_H$  genes is variable; some  $V_H$  genes are only 4 kb apart whereas others may be greater than 30 kb apart. The average spacing distance between  $V_H$  genes is greater than 23 kb. Assuming that there are at least 250  $V_H$  genes, the size of the  $V_H$  gene locus may be greater than 5 million nucleotide pairs.



## REFERENCES

- Altenburger, W., Steinmetz, M. and Zachau, H. G. (1980) Functional and nonfunctional joining in immunoglobulin light chain genes of a mouse myeloma. *Nature* **287**, 603-607.
- Amzel, L. M., Poljak, R., Saul, F., Varga, J. and Richards, F. (1974) The three-dimensional structure of a combining region-ligand complex of immunoglobulin NEW at 3.5 Å resolution. *Proc. Natl. Acad. Sci. USA* **71**, 1427-1430.
- Bernard, O., Hozumi, N. and Tonegawa, S. (1978) Sequences of mouse immunoglobulin light chain genes before and after somatic changes. *Cell* **15**, 1133-1144.
- Blomberg, B., Traunecker, A., Eisen, H. and Tonegawa, S. (1981) Organization of four mouse  $\lambda$  light chain immunoglobulin genes. *Proc. Natl. Acad. Sci. USA* **78**, 3765-3769.
- Capra, J. D. and Kehoe, J. M. (1974) Variable-region sequences of five human immunoglobulin heavy chains of the  $V_{HIII}$  subgroup: definitive identification of four heavy-chain hypervariable regions. *Proc. Natl. Acad. Sci. USA* **71**, 845-849.
- Claflin, J. L. and Rudikoff, S. (1976) Uniformity in a clonal repertoire: a case for a germ-line basis of antibody diversity. *Cold Spring Harbor Symp. Quant. Biol.* **41**, 725-734.
- Cohn, M., Blomberg, B., Geckeler, W., Raschke, W., Riblet, R. and Weigert, M. (1974) First order considerations in analyzing the generation of diversity. In: *The Immune System: Genes, Receptors, Signals*. E. Sercarz et al. (eds.) (New York: Academic Press), pp. 89-117.
- Cory, S., Tyler, B. M. and Adams, J. M. (1981) Sets of immunoglobulin  $V_{\kappa}$  genes homologous to ten cloned  $V_{\kappa}$  sequences: implications for the number of germline  $V_{\kappa}$  genes. *J. Mol. Appl. Genet.* **1**, 103-116.



- Davis, M. M., Calame, K., Early, P. W., Livant, D. L., Joho, R., Weissman, I. L., and Hood, L. (1980) An immunoglobulin heavy-chain gene is formed by at least two recombinational events. *Nature* **283**, 733-738.
- Early, P., Huang, H., Davis, M., Calame, K. and Hood, L. (1980) An immunoglobulin heavy chain variable region is generated from three segments of DNA:  $V_H$ ,  $D$ , and  $J_H$ . *Cell* **19**, 981-992.
- Edelman, G. M. and Gally, J. A. (1967) Somatic recombination of duplicated genes: an hypothesis on the origin of antibody diversity. *Proc. Natl. Acad. Sci. USA* **57**, 353-358.
- Eisen, N. H. and Siskind, G. W. (1964) Variations in affinities of antibodies during the immune response. *Biochemistry* **3**, 996-1008.
- Hood, L., Davis, M., Early, P., Calame, K., Kim, S., Crews, S. and Huang, H. (1980) Two types of DNA rearrangements in immunoglobulin genes. *Cold Spring Harbor Symp. Quant. Biol.* **45**, 887-898.
- Hood, L., Loh, E., Hubert, J., Barstad, P., Eaton, B., Early, P., Fuhrman, J., Johnson, N., Kronenberg, M. and Schilling, J. (1976) The structure and genetics of mouse immunoglobulins: an analysis of NZB myeloma proteins and sets of BALB/c myeloma proteins binding particular haptens. *Cold Spring Harbor Symp. Quant. Biol.* **41**, 817-836.
- Jerne, N. K. (1971) The somatic generation of immune recognition. *Eur. J. Immunol.* **1**, 1-9.
- Kabat, E. A., Wu, T. T. and Bilofsky, H. (1979) Evidence supporting somatic assembly of the DNA segments (minigenes) coding for the framework and complementarity-determining segments of immunoglobulin variable regions. *J. Exp. Med.* **149**, 1299-1313.
- Max, E. E., Seidman, J. G. and Leder, P. (1979) Sequence of five potential recombination sites encoded close to an immunoglobulin  $\kappa$  constant region gene. *Proc. Natl. Acad. Sci. USA* **76**, 3450-3454.

- Max, E. E., Seidman, J. G., Miller, A. H. and Leder, P. (1980) Variation in the cross-over point of kappa immunoglobulin gene V-J recombination: evidence from a cryptic gene. *Cell* **21**, 793-799.
- Ohta, T. (1980) Amino acid diversity of immunoglobulins as a product of molecular evolution. *J. Mol. Evol.* **15**, 29-35.
- Padlan, E. A., Segal, D. M., Spande, T. F., Davies, D. R., Rudikoff, S. and Potter, M. (1973) Structure at 4.5 Å resolution of a phosphorylcholine-binding Fab. *Nature New Biol.* **245**, 165-167.
- Potter, M. (1977) Antigen-binding myeloma proteins of mice. In: *Advances in Immunology*, Vol. 25. (New York: Academic Press) pp. 141-211.
- Sakano, H., Hölz, K., Heinrich, G. and Tonegawa, S. (1979) Sequences at the somatic recombination sites of immunoglobulin light-chain genes. *Nature* **280**, 288-294.
- Seidman, J. G., Leder, A., Nau, M., Norman, B. and Leder, P. (1978) Antibody diversity. *Science* **202**, 11-17.
- Shimizu, A., Takahashi, N., Yaoita, Y., and Honjo, T. (1982) Organization of the constant-region gene family of the mouse immunoglobulin heavy chain. *Cell* **28**, 499-506.
- Valbuena, O., Marcu, K. B., Weigert, M. and Perry, R. P. (1978) Multiplicity of germline genes specifying a group of related mouse  $\kappa$  chains with implications for the generation of immunoglobulin diversity. *Nature* **276**, 780-784.
- Weigert, M. G., Cesari, H. M., Yonkovich, S. J. and Cohn, M. (1970) Variability in the lambda light chain sequences of mouse antibody. *Nature* **228**, 1045-1047.
- Weigert, M., Gattmaitan, L., Loh, E., Schilling, J. and Hood, L. (1978) Rearrangement of genetic information may produce immunoglobulin diversity. *Nature* **276**, 785-790.
- Weigert, M., Perry, R., Kelley, D., Hunkapiller, T., Schilling, J. and Hood, L. (1980) The joining of V and J gene segments creates antibody diversity. *Nature* **283**, 497-499.

- Weigert, M. and Riblet, R. (1978) The genetic control of antibody variable regions in the mouse. Springer Seminars in Immunopathol. **1**, 133-169.
- Wu, T.T. and Kabat, E. A. (1970) An analysis of the sequences of the variable regions of Bence-Jones proteins and myeloma light chains and their implications for antibody complementarity. J. Exp. Med. **132**, 211-250.
- Zeelon, E. P., Bothwell, A. L. M., Kantor, F. and Schechter, I. (1981) An experimental approach to enumerate the genes coding for immunoglobulin variable-regions. Nucl. Acids Res. **9**, 3809-3820.

**Figure 1.** The complete protein sequences of nine  $V_H$  regions of myeloma proteins binding phosphorylcholine. A line indicates that the sequence is identical to T15 at that position. Above the sequence, the sections of the  $V_H$  region derived from the V, D, and J gene segments are shown. Dotted horizontal lines demarcate the positions of the three hypervariable regions.

Fig. 1

[illegible]

## Chapter 1

A Single  $V_H$  Gene Segment Encodes the Immune Response to  
Phosphorylcholine: Somatic Mutation is Correlated with the  
Class of the Antibody

This paper was published in Cell.

# A Single $V_H$ Gene Segment Encodes the Immune Response to Phosphorylcholine: Somatic Mutation Is Correlated with the Class of the Antibody

Stephen Crews,\* Johanna Griffin,\* Henry Huang,\* Kathryn Calame† and Leroy Hood\*

\*Division of Biology  
California Institute of Technology  
Pasadena, California 91125

†Department of Biological Chemistry and  
the Molecular Biology Institute  
University of California  
Los Angeles, California 90024

## Summary

The immune response in BALB/c mice to phosphorylcholine is highly restricted in its heterogeneity. Of the 19 immunoglobulins binding phosphorylcholine for which complete  $V_H$ -segment amino acid sequences have been determined, 10 employ a single sequence, denoted T15 after the prototype  $V_H$  sequence of this group of antibodies. The remaining 9 of these  $V_H$  segments are variants differing by 1 to 8 residues from the T15 sequence. Using a cloned  $V_H$  cDNA probe complementary to the T15 sequence, we isolated from a mouse sperm genomic library clones corresponding to four  $V_H$  gene segments that by DNA sequence analysis are >85% homologous to one another. These four  $V_H$  gene segments have been denoted the T15  $V_H$  gene family. These  $V_H$  gene segments are most, if not all, of the germline  $V_H$  gene segments that could encode the  $V_H$  sequences of antibodies that bind phosphorylcholine. One of these four genes contains the T15- $V_H$ -coding sequence. When the T15-family  $V_H$  gene segments were compared with the complete  $V_H$  protein sequences of 19 hybridoma and myeloma immunoglobulins that bind phosphorylcholine, several striking conclusions could be drawn. First, all of these  $V_H$  regions must have arisen from the germline T15  $V_H$  gene segment. Thus virtually the entire immune response to phosphorylcholine is derived from a single  $V_H$ -coding sequence. Nine of the 19  $V_H$  regions were variants differing from the T15- $V_H$ -coding sequence and, accordingly, must have arisen by a mechanism of somatic diversification. Second, the variants appear to be generated by a somatic mutation mechanism. They cannot be explained by recombination or gene conversion among members of the T15 gene family. Third, somatic mutation is correlated with the class of the antibody. All of the somatic variation is found in the  $V_H$  regions derived from antibodies of the IgA and IgG classes. The IgM molecules express the germline T15  $V_H$  gene segment exclusively.

## Introduction

For the past 80 years immunologists have attempted to understand the mechanisms responsible for anti-

body diversity. We are now in a unique position to analyze the phenomenon of antibody diversity because of the advent of recombinant DNA techniques; the possibility of generating homogeneous antibodies of any desired specificity through the hybridoma technology; and the recent advances in the speed and sensitivity of protein sequence analysis (Hunkapiller and Hood, 1980; Hewick et al., 1981). Thus we can now directly compare the antibody genes and proteins generated during a particular immune response.

The antibody molecule is composed of light (L) and heavy (H) polypeptide chains. Three unlinked families of genes encode the antibody polypeptides—two for light chains,  $\lambda$  and  $\kappa$ , and one for heavy chains. Light chains are encoded by three distinct types of gene segments, variable ( $V_L$ ), joining ( $J_L$ ) and constant ( $C_L$ ) (Brack et al., 1978; Sakano et al., 1979; Seidman et al., 1979), whereas heavy chains are encoded by four different types of gene segments,  $V_H$ , D (diversity),  $J_H$  and  $C_H$  (Early et al., 1980). Two types of DNA rearrangements occur in antibody-producing (B) cells. First, the  $V_L$  and  $J_L$ , and the  $V_H$ , D and  $J_H$  gene segments, are joined to generate  $V_L$  and  $V_H$  genes, respectively (Brack et al., 1978; Sakano et al., 1979, 1980; Seidman et al., 1979; Early et al., 1980). Second, the B cell initially expresses IgM molecules, and later this cell or its progeny switch to the expression of other immunoglobulin classes. This class switch is mediated by a DNA rearrangement in which the  $C_\mu$  gene is replaced by one of the other  $C_H$  genes, such as  $C_\alpha$  (IgA) or  $C_\gamma$  (IgG).

The studies of antibody polypeptides and genes in mice have delineated three sources of antibody diversity. First, it is estimated that the  $\kappa$  and heavy-chain gene families each contain 100–300 germline V and four germline J gene segments (Seidman et al., 1978; Davis et al., 1979; Hood et al., 1980; Sakano et al., 1980). The heavy-chain gene family probably contains 10 or more germline D gene segments (Hood et al., 1980; Sakano et al., 1981). In mice, the  $\lambda$  family appears to have only two or three germline V and J gene segments. Hence, there is extensive *germline* diversity. Second, the gene segments may be joined in a combinatorial manner to amplify V-gene diversity. For example, if any  $V_L$  may be joined to any  $J_L$ , any  $V_H$  to any D, and any D to any  $J_H$  gene segments, then 800  $V_L$  ( $200 V_\kappa \times 4 J_\kappa$ ) and 8000  $V_H$  ( $200 V_H \times 10 D \times 4 J_H$ ) genes may be generated by *combinatorial joining*. At the protein level any light chain probably may join with any heavy chain. Thus the *combinatorial association* of light and heavy chains amplifies antibody diversity still further (for example,  $800 L \times 8000 H = 6.4 \times 10^6$  different antibody molecules). A third category of diversification mechanisms employs two types of somatic alterations of the V-gene sequences. First, junctional diversity arises at the boundaries of the gene segments because their joining together may occur at different points in their base sequences (Max

et al., 1979; Sakano et al., 1979; Early et al., 1980; Hood et al., 1980; Weigert et al., 1980). Thus many different sequences and sequence lengths can be generated at the  $V_L$ - $J_L$ ,  $V_H$ -D and D- $J_H$  boundaries. Second, several studies suggest that somatic mutation may occur throughout the V genes (Weigert et al., 1970, 1978; Weigert and Riblet, 1976; Brack et al., 1978; Bernard et al., 1978; Valbuena et al., 1978).

We have studied the immune response in inbred BALB/c mice to the simple hapten phosphorylcholine with regard to the relative contributions of the mechanisms outlined above (Barstad et al., 1974; Hood et al., 1976; Gearhart et al., 1981). The complete amino acid sequences of 9 myeloma and 10 hybridoma  $V_H$  segments (that portion of the  $V_H$  region encoded by the  $V_H$  gene segment) are given in Figure 1. Two striking patterns emerged from these data. First, 10  $V_H$  segments are identical. Presumably they represent the repeated expression of a single germline  $V_H$  gene segment. This putative germline  $V_H$  sequence is denoted T15, after the initial  $V_H$  region sequenced in this identical set. Second, nine different variants differing by 1 to 8 residues from the T15  $V_H$  sequence are noted. The question we have attempted to answer is which, if any, of these variants are somatically derived and which might be encoded by germline  $V_H$  gene segments.

We report the isolation and sequence analysis of four distinct germline  $V_H$  gene segments belonging to the T15  $V_H$  gene family. One of these  $V_H$  gene segments codes for a  $V_H$  segment identical to T15. A

comparison of the four different germline  $V_H$  gene sequences with 19  $V_H$  regions derived from myeloma and hybridoma immunoglobulins binding phosphorylcholine allows us to draw three fundamental conclusions. First, all of the  $V_H$  regions are derived from the T15 germline  $V_H$  segment. Second, the  $V_H$  variants arose by a somatic mutation mechanism. They cannot be explained by somatic recombination or gene conversion. Third, somatic mutation appears to be correlated with the class of the antibody.

## Results and Discussion

### Strategy

A cloned cDNA probe to the T15  $V_H$  gene sequence was prepared from mRNA isolated from the S107 tumor (S107 and T15 have identical  $V_H$ -segment protein sequences; see Figure 1; Early et al., 1979). This probe would be expected to hybridize to all of the  $V_H$  gene segments that might code for the  $V_H$  segments from phosphorylcholine-binding antibodies (Figure 1). For example, the S107  $V_H$  probe hybridizes strongly with mRNA from the M167 myeloma tumor (Early et al., 1979). The M167  $V_H$  sequence differs from the T15  $V_H$  sequence more than any other T15  $V_H$  variant segments studied (Figure 1). Four major bands hybridized to the S107  $V_H$  probe upon Southern blot analysis of germline (sperm) DNA. Four distinct genes corresponding to the four major hybridization bands were isolated from a sperm genomic library. We believe that we have isolated most, if not all, of the

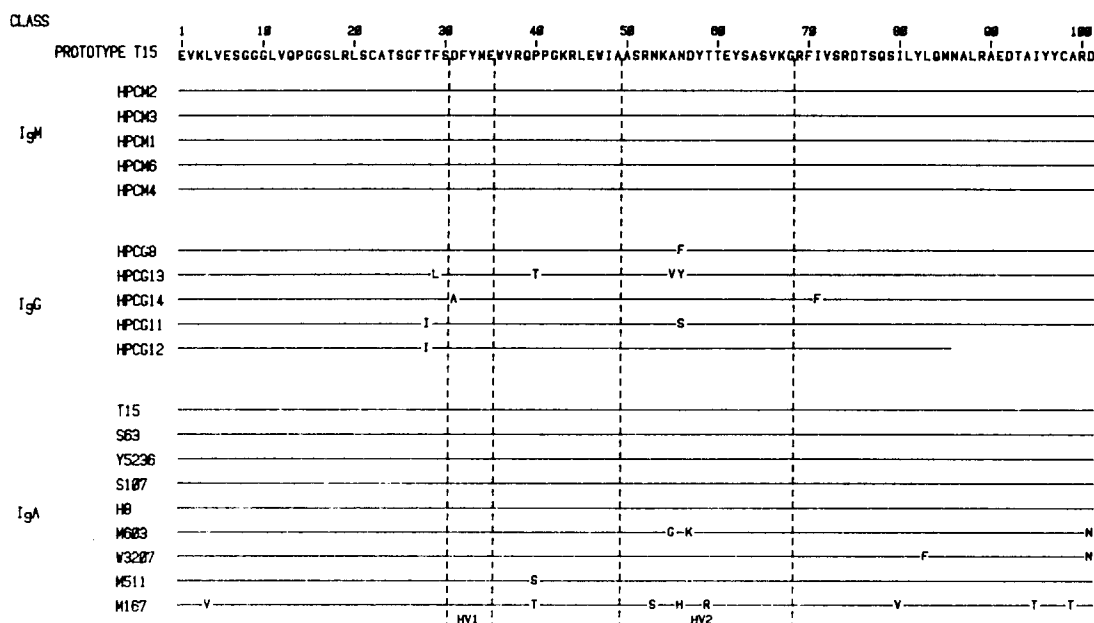


Figure 1. The  $V_H$ -Segment Protein Sequences of Myeloma and Hybridoma Antibodies That Bind Phosphorylcholine

The IgM and IgG antibodies are all derived from hybridomas and the IgA antibodies are the products of myelomas. The protein sequences are compared with the  $V_H$  segment of T15 (the one-letter code for amino acids is used). The first hypervariable region (HV1) and the second hypervariable region (HV2) are demarcated by vertical dashed lines. Adapted from Gearhart et al. (1981).



members of the T15 V<sub>H</sub> gene family. We determined the DNA sequences of the four members of the T15 V<sub>H</sub> gene family and compared these germline sequences with their phenotypic counterparts, the V<sub>H</sub> segments from myeloma and hybridoma immunoglobulins that bind phosphorylcholine.

#### Isolation and Characterization of V<sub>H</sub> Gene Segments Belonging to the T15 V<sub>H</sub> Gene Family

The S107 V<sub>H</sub> probe was hybridized to a genomic blot containing Eco RI- or Bam HI-cut sperm DNA. Under the hybridization conditions employed, we expected to detect all bands containing DNA sequences as closely related as any of the variant sequences of the phosphorylcholine-binding antibody V<sub>H</sub> segments (Figure 1). Using either enzyme, four major bands as well as several fainter bands were seen (Figure 2). A similar pattern has been observed by Cory and Adams (1980) with an S107 V<sub>H</sub> probe.

To isolate all of the germline V<sub>H</sub> gene segments that hybridized with the S107 V<sub>H</sub> probe, we used this probe to screen a Charon 4A bacteriophage recombinant DNA library containing BALB/c mouse sperm DNA partially digested with the restriction enzymes Hae III and Alu I (Davis et al., 1980). Twelve different genomic clones were isolated from the sperm library. Another clone,  $\lambda$ V1, was previously isolated from a library containing sperm DNA partially digested with Eco RI (Early et al., 1980). Two additional genomic clones,  $\lambda$ V3207.1 and  $\lambda$ V603.1, that hybridized with the S107 V<sub>H</sub> probe were isolated from Charon 4A genomic libraries containing DNA from the W3207 and M603 myeloma tumors, respectively, partially digested with Eco RI (M. Davis, S. Kim and P. Early, unpublished data). The V<sub>H</sub> gene segments of these two clones appear to be in the germline configuration, because they show no evidence of DNA rearrangements from their organization in sperm DNA. Restriction map analyses in conjunction with DNA blot hybridizations with the S107 V<sub>H</sub> probe were used to localize the V<sub>H</sub> genes within the genomic clones. All together, we have isolated 15 different V<sub>H</sub> genomic clones. Twelve clones have one V<sub>H</sub> gene segment, one clone has two V<sub>H</sub> gene segments, and two clones have three V<sub>H</sub> gene segments. The genomic clones are denoted with the prefix  $\lambda$ , and the corresponding V gene segments are designated with the appropriate number (for example,  $\lambda$ V13 and V13).

To restrict our study to the V<sub>H</sub> gene segments most homologous to the T15 V<sub>H</sub> segment, we performed a genomic blot reconstruction experiment with our clones. Single-copy equivalents of the different clones were digested with a restriction enzyme, run in parallel with similarly digested genomic DNA, blotted and hybridized to the S107 V<sub>H</sub> probe (data not shown). The genomic clones fell into two categories. First, 11 clones showed significant hybridization to the probe. All of these clones had a V<sub>H</sub>-hybridizing restriction

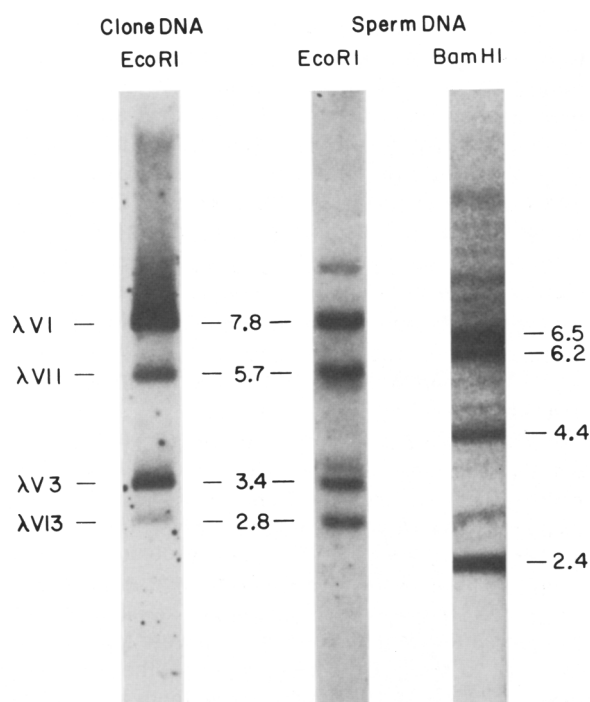


Figure 2. Hybridization of the S107 V<sub>H</sub> Probe to Genomic Blots of Mouse Sperm DNA

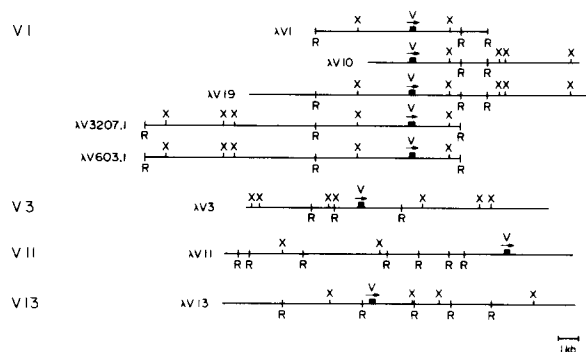
BALB/c mouse sperm DNA was digested by either Eco RI or Bam HI, blotted and hybridized to pS107V1 (Early et al., 1980). Parallel to the Eco RI-cut sperm DNA was Eco RI-cut DNA from  $\lambda$  clones representing the four T15-family V<sub>H</sub> gene segments. The comigration of the V<sub>H</sub>-containing Eco RI fragment of  $\lambda$ V11 with the 5.7 kb Eco RI fragment of sperm DNA may be coincidental. We have not yet located the Eco RI site 3' to the V<sub>H</sub> gene segment in any of our clones for this gene. However, the Bam HI restriction fragment of  $\lambda$ V11 containing the V<sub>H</sub> gene segment comigrates with the 6.2 kb Bam HI fragment of sperm DNA. Both Bam HI sites flanking the V11 gene segment are within the  $\lambda$ V11 insert.

fragment comigrating with a strongly hybridizing band of sperm DNA when cut with appropriate restriction enzymes. Restriction mapping and heteroduplex analysis indicated that these 11 clones contained four different V<sub>H</sub> gene segments. Figure 2 also shows that four different  $\lambda$  clones gave distinct Eco RI fragments corresponding to the four major V<sub>H</sub> bands seen in Eco RI-cleaved sperm DNA. Reconstruction blotting experiments with various amounts of the  $\lambda$  clone DNAs cut with a restriction enzyme and compared with similarly cut sperm DNA indicates that each of the four major bands seen on a genomic blot contains no more than one or two V<sub>H</sub> gene segments. We have therefore isolated most, if not all, of the V<sub>H</sub> gene segments that could encode the V<sub>H</sub> segments of antibodies that bind phosphorylcholine. These V<sub>H</sub> gene segments, as shown by DNA sequencing, exhibit greater than 85% DNA sequence homology to one another and are denoted members of the T15 V<sub>H</sub> gene family. Restriction maps of eight  $\lambda$  clones ( $\lambda$ V1,  $\lambda$ V3207.1,  $\lambda$ V603.1,  $\lambda$ V3,  $\lambda$ V10,  $\lambda$ V11,  $\lambda$ V13,  $\lambda$ V19) representing these

four gene segments are shown in Figure 3. The other three clones overlap the eight shown. Second, the remaining four clones showed very faint bands upon hybridization to the S107 V<sub>H</sub> probe. DNA sequences of V<sub>H</sub> gene segments within these poorly hybridizing clones are approximately 75% homologous to the T15-V<sub>H</sub>-gene-segment DNA sequence (S. Crews, unpublished data). Three of these four clones have more than one V<sub>H</sub> gene segment. These less homologous clones are by our definition not members of the T15 V<sub>H</sub> gene family and will not be considered further.

## The T15 V<sub>H</sub> Gene Family Has Four V<sub>H</sub> Gene Segments

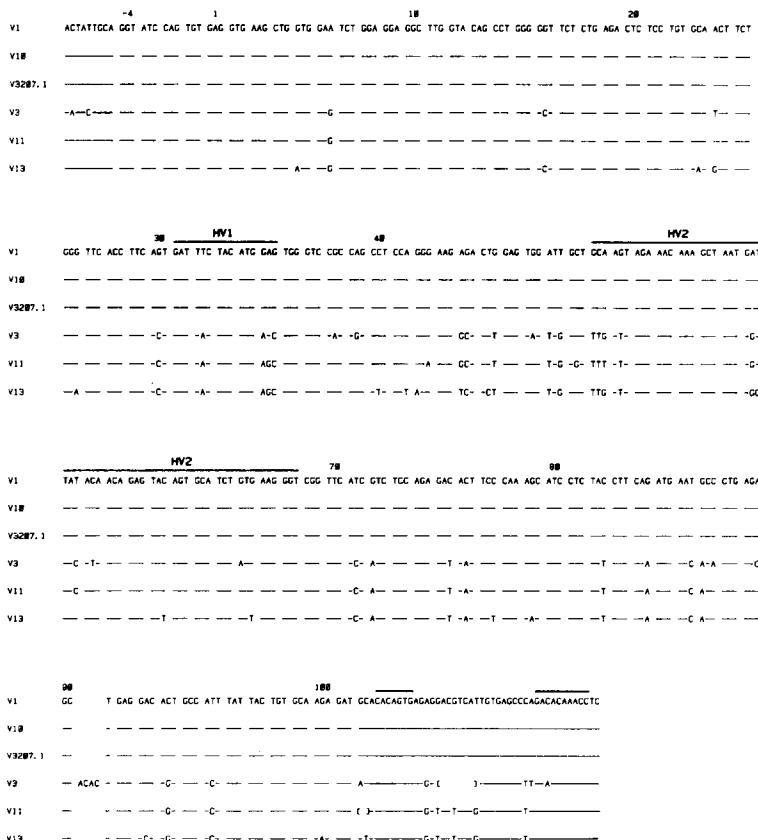
The DNA sequences of the four V<sub>H</sub> gene segments of the T15 V<sub>H</sub> gene family, V1, V3, V11 and V13, are given in Figure 4. The V1 V<sub>H</sub> gene segment was previously sequenced, and when translated was shown to be identical to the T15 V<sub>H</sub> protein sequence (Early et al., 1980). The V<sub>H</sub> gene segments of two clones (λV10 and λV3207.1) that are identical to λV1 in their overlapping flanking regions were found to be identical to V1 in DNA sequence. Another V<sub>H</sub> gene from the clone, λV19, which also is identical to λV1 in its overlapping flanking sequences, has been partially sequenced and is identical to V1 (P. Early, unpub-



**Figure 3. Restriction Maps of the  $\lambda$  Clones Containing V<sub>H</sub> Genes of the T15 V<sub>H</sub> Gene Family**

The Eco RI (R) and Xba I (X) restriction sites are designated by vertical lines. The name of each  $V_H$  gene segment is indicated on the far left, with the corresponding clones listed adjacent to the restriction maps. Arrows: 5'  $\rightarrow$  3' direction of transcription of the V gene (raised bar).

Earlier work showed by heteroduplex analysis that the flanking-sequence homology between  $\lambda$ V1 (ChSpVPC3) and Ch603 $\alpha$ 125, which is a clone containing the rearranged, expressed heavy-chain gene from the M603 myeloma tumor, extended 10.9 kb 5' to the V<sub>H</sub> gene (Davis et al., 1980). Subsequently, we have discovered that in both clones the 6.9 kb Eco RI fragment adjacent 5' to the V<sub>H</sub> gene is a Charon 4A internal Eco RI fragment. Thus the mouse chromosomal DNA represented in both clones extends only to the Eco RI site 5 kb 5' to the V<sub>H</sub> gene.



**Figure 4. DNA-Sequence Comparisons of the T15-Family V<sub>H</sub> Gene Segments**

Coding sequences are indicated by triplets. The triplet reading frame is that of T15. V3 has a 4 nucleotide insertion at position 90, causing a frameshift. The V<sub>H</sub> segment begins at residue 1 and ends at 101. The amino acid residues -1 through -4 are part of the signal peptide that ultimately is cleaved from the heavy chain. Overbars: the two hypervariable regions, as well as the heptamer and decamer sequences believed to be involved in the joining of V-region gene segments.

lished data). The remaining members of the T15 V<sub>H</sub> gene family were 86–96% homologous to one another and to the T15-coding sequence.

**The Variants Related to T15 Arose by Somatic Diversification from the T15 V<sub>H</sub> Gene Segment**

Because we have determined the sequences of the T15 family of V<sub>H</sub> gene segments and many of the V<sub>H</sub> regions from hybridoma and myeloma immunoglobulins that bind phosphorylcholine, we are in a unique position to analyze the genotype (germline V<sub>H</sub> genes) and phenotype (V<sub>H</sub> regions) of the immune response to a simple hapten. Nineteen V<sub>H</sub> regions from myeloma and hybridoma immunoglobulins that bind phosphorylcholine have been completely sequenced (Gearhart et al., 1981). Nine V<sub>H</sub> variants have been observed in the 19 completely sequenced V<sub>H</sub> regions and each variant has been observed only once. We have cloned and sequenced most, if not all, of the germline V<sub>H</sub> genes that could code for these variants, and have found that each variant V<sub>H</sub> protein sequence is different from all of the V<sub>H</sub> gene segments of the T15 family (Figure 5). Thus these variants must have arisen by somatic diversification.

The most striking feature of the comparison of the proteins and gene segments of the T15 family is that all of the V<sub>H</sub> protein sequences appear to be derived from the germline T15 V<sub>H</sub> gene segment (V1). The variants differ by only 1 to 8 residues from the T15 sequence, whereas each variant differs substantially (>16 residues) from the V<sub>H</sub> sequences of V3, V11 and V13. Perhaps the most compelling argument to suggest that these V<sub>H</sub> regions are derived from the T15 V<sub>H</sub> gene segment is that this sequence can be distinguished from each of the other three V<sub>H</sub> gene seg-

ments at 11 positions (codons 30, 32, 48, 51, 57, 71, 72, 76, 85, 91 and 93 in Figure 5) and the variant V<sub>H</sub> sequences share nearly all of these positions with the T15 gene segment. Analysis of the clone, Ch603α125, containing the rearranged, expressed heavy-chain gene of M603 demonstrates that the M603 V<sub>H</sub> segment is derived from the V1 gene segment by virtue of its identity in their 5' flanking sequences (Davis et al., 1980). Similar analyses of clones containing the rearranged, expressed heavy-chain genes of T15, M167 and HPCM2 indicate that they also are derived from the V1 gene segment (K. Calame, S. Kim and M. Davis, unpublished data). Virtually the entire immune response to phosphorylcholine is therefore derived from the T15 V<sub>H</sub> gene segment. Moreover, the T15 V<sub>H</sub> gene undergoes somatic diversification.

**Somatic Diversification Is Correlated with the Class of the Antibody**

The data in Figure 1 demonstrate that the V<sub>H</sub> segments derived from IgM immunoglobulins are all identical to the T15 sequence, while all of the V<sub>H</sub> segments derived from IgG molecules are variants, as are about half the V<sub>H</sub> regions associated with IgA immunoglobulins. Since our data demonstrate that these variants arose by a somatic diversification mechanism, there is a correlation between the somatic diversification of rearranged germline V<sub>H</sub> gene segments and the class switch from IgM to IgG or IgA molecules. Somatic variation in the V<sub>κ</sub> regions derived from these immunoglobulins also is associated with the class switch (Gearhart et al., 1981). Similarly, an analysis of the in vitro response of B cells to secondary immunization with phosphorylcholine suggests that the antibody-

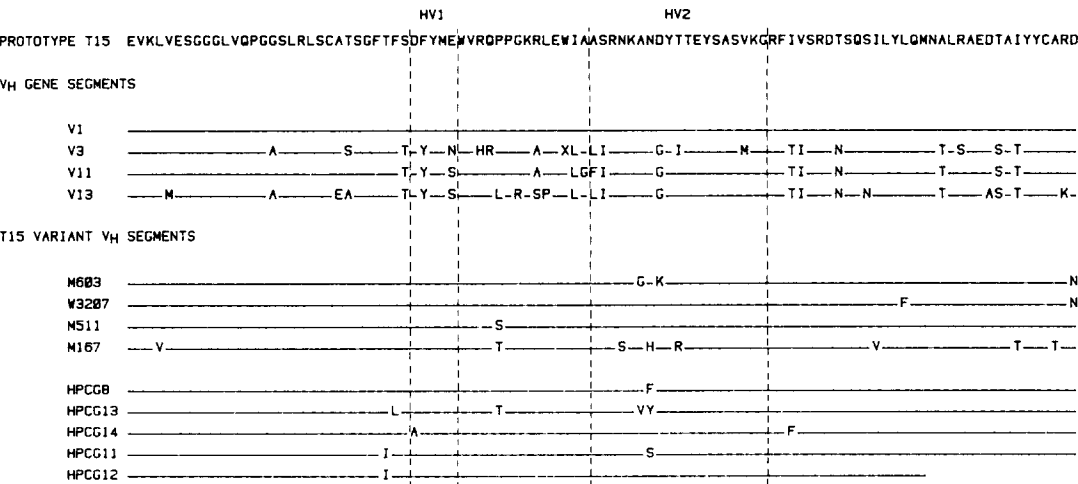


Figure 5. Comparison of the V<sub>H</sub> Gene Segments of the T15 V<sub>H</sub> Gene Family and the T15 Variant V<sub>H</sub> Segments  
The four gene segments of the T15 V<sub>H</sub> gene family have been translated to protein sequences and are here compared with the sequences of the somatic variants (the one-letter code for amino acids is used). The 4 nucleotide insertion in V3 that occurs at position 90 and changes the reading frame has not been included so as to facilitate comparisons.

producing cells secrete IgA and IgG that are more diverse than IgM (Chang and Rittenberg, 1981).

There are two general explanations for the correlation between somatic diversification and class switching. Perhaps the diversification mechanism operates throughout much of the lifetime of the developing B cell. If so, those B cells with longer lives (that is, more cell division cycles) may accumulate more changes (Rodwell and Karush, 1980). By this reasoning, B cells that have undergone the class switch would be older and, accordingly, would have acquired more changes. Alternatively, perhaps the act of class switching itself activates a special diversification mechanism—one that operates in trans, such that  $V_L$  as well as  $V_H$  gene segments diversify in conjunction with the heavy-chain class switch. This special diversification mechanism would be activated during the class switch, operate at high rates and then be turned off. Since the phosphorylcholine response studied here is T-cell-dependent, the T-cell interactions required for class switching may lead to the activation of a somatic diversification mechanism. It is currently impossible to distinguish between these alternative explanations for the association of somatic variation of V gene segments and class switching.

### Constraints on Mechanisms of Somatic Diversification

Somatic alteration of V gene segments may be explained by two kinds of models: somatic recombination or gene conversion (Edelman and Gally, 1967; Seidman et al., 1978), and somatic mutation (Brenner and Milstein, 1966; Baltimore, 1974; Cohn et al., 1974).

The diversification of  $V_H$  gene segments may occur by somatic recombination or gene conversion among the members of the T15  $V_H$  gene family. This hypothesis became more attractive with the observation that related  $V_k$  gene segments share extensive flanking-sequence homologies that could potentially enhance recombination between related genes (Seidman et al., 1978). Indeed, several members of the T15  $V_H$  gene family also share extensive flanking-sequence homologies (K. Calame and S. Crews, unpublished observations). To test the recombination model of antibody diversification, we compared the germline products and the somatic variants at the protein level (Figure 5). Only 1 of 24 positions of somatic variation can be explained by the gene conversion or somatic recombination models (position 95 in M167). We believe that we have isolated most of the  $V_H$  gene segments of the T15  $V_H$  gene family. If somatic recombination were a major mechanism for generating somatic variants in V segments, we should have been able to account for more of the variants by recombination among the germline  $V_H$  gene segments of the T15  $V_H$  gene family. We conclude that most V-segment diver-

sity in the immune response to phosphorylcholine does not arise from gene conversion or somatic recombination.

The other alternative is that the somatic variants arose from some type of mutation process. Two possibilities are that the variants arose from events occurring at the basal somatic mutation rate or that they are the consequence of a V-gene-specific mutation mechanism. Regardless of the mechanism, the mutation rate can be very high. For example, there are 8 amino acid substitutions in the M167  $V_H$  segment. Furthermore, the M603  $V_H$  gene segment has 7 base substitutions, 3 of which are silent changes not leading to amino acid substitutions (Early et al., 1980).

### Selection of Variant B Cells May Occur by Two General Mechanisms

After the variant B cell has arisen in a mouse, it represents one in a population of  $10^9$  lymphocytes. This variant B cell must be selectively amplified to constitute a significant fraction of the B-cell population before there is a reasonable probability that it will be detected with cellular assay systems (myelomas and hybridomas). This selection can occur by one of two general mechanisms. First, variant B cells whose antibody receptors exhibit increased affinity for the antigen may be selectively driven to proliferate more rapidly than their lower-affinity, unmutated counterparts. This process is denoted antigen-driven selection. Indeed, secondary immune responses often lead to the synthesis of antibody molecules with higher affinity than those of the primary immune response. Presumably this affinity maturation is explained by antigen-driven selection. Second, the B-cell immune response appears to be controlled by T cells through an idotype-anti-idotype regulatory network. This process may select B-cell variants if the generation of variants lacking a particular idotype releases the variant B cells from the inhibitory influences of suppressor T cells. Thus B cells may clonally expand upon release from T-cell suppression.

Selection may operate to shift the pattern and distribution of germline as opposed to somatic variation. The distribution of substitutions among the four distinct germline coding sequences comprising the T15  $V_H$  gene family is scattered throughout the V gene segment (Figure 5). In contrast, about 50% of the amino acid substitutions of antibodies binding phosphorylcholine fall within the hypervariable regions, which constitute about 25% of the  $V_H$  segment. However, variant proteins can be expressed that have no substitutions in the hypervariable regions (for example, M511 and HPCG-12 in Figure 5). This pattern is in striking contrast to that seen for the variant mouse  $\lambda_1$  chains, where all of the substitutions fall within the hypervariable regions (Weigert and Riblet, 1976). However, substitutions in the framework as well as the

hypervariable regions can lead to changes in the specificity of the resulting antibody molecule (Scharff et al., 1981).

#### Functional Role of the T15 $V_H$ Gene Family

Clearly the V11 and V13  $V_H$  gene segments do not generally encode  $V_H$  regions binding to phosphorylcholine. We believe that both V11 and V13 may each encode immune responses to antigens different than phosphorylcholine. A  $V_H$  region from the IgA molecule synthesized by the myeloma tumor M47A is identical to the translated protein sequence of V11 (Robinson and Appella, 1977). However, the specificity of this antibody is unknown. A  $V_H$  region encoded by the V13 gene segment has not been observed to date. The V3 gene segment is a pseudogene and cannot be expressed (Huang et al., 1981).

The T15  $V_H$  region only encodes antibodies binding phosphorylcholine when it is associated with certain light chains (Barstad et al., 1978). Hence, even the T15  $V_H$  region in association with other light chains may encode responses distinct from that to phosphorylcholine. Accordingly, the T15  $V_H$  gene family probably encodes a variety of different immune responses.

#### The Immune Response to Phosphorylcholine Employs All of the Mechanisms of Antibody Diversification

For the first time an immune response has been characterized in great detail at the level of germline potential and at the level of somatic expression. This response displays all of the proposed means of generating diversity. The phosphorylcholine response in BALB/c mice employs one germline  $V_H$ , one  $J_H$  ( $J_{H1}$ ) and three to six D gene segments that display diversity arising from combinatorial joining. Similarly, antibodies that bind phosphorylcholine use combinatorial association, in that the same heavy chain is combined with light chains from three different groups (Barstad et al., 1978). The phosphorylcholine response also shows both types of somatic alterations. Of 16 sequences examined, 10 show junctional diversity, presumably arising from alternative recombination points, at the  $V_H$ -D and the D- $J_H$  boundaries. In addition, somatic mutation occurs throughout the  $V_H$  segment (Figure 1). An intriguing question that remains is the quantitation of the relative contribution of the germline, combinatorial and somatic variation mechanisms to the functional repertoire of antibody molecules.

The IgM response may represent a first line of defense against pathogenic organisms by virtue of the expression of germline V, D and J gene segments in all of their combinatorial forms. This extensive combinatorial and germline diversity provides vertebrate organisms with an effective antibody response to most immunogens. Somatic mutation operating on the rearranged  $V_H$  and  $V_L$  gene segments generates variants.

These variants, expressed predominantly as IgA or IgG molecules, may be selectively expanded in the B-cell population. However, this extra level of diversity is imposed only on antibodies derived from antigen-stimulated B cells. Thus, rather than simply expanding the antibody repertoire available for recognizing foreign antigens, somatic mutation may allow the "fine tuning" of the immune response to a particular antigen.

#### Experimental Procedures

##### Genomic Blot Hybridizations

Ten micrograms of BALB/c sperm DNA were cut with the appropriate restriction enzyme and subjected to electrophoresis on a horizontal agarose slab gel. After denaturation and neutralization, the DNA was transferred to nitrocellulose paper by blotting with  $20\times$  SSC (Nagamine et al., 1980). Hybridizations were carried out in a stoppered graduated cylinder rotated in an oven at  $68^\circ\text{C}$ . The hybridization solution consisted of 5 ml of 1 M NaCl, 0.05 M Tris-HCl (pH 7.5), 0.1% NaPPi, 0.1% SDS,  $10\times$  Denhardt's solution, 150  $\mu\text{g}/\text{ml}$  salmon sperm DNA, 50  $\mu\text{g}/\text{ml}$  poly(rC) and 0.75  $\mu\text{g}$  nick-translated  $^{32}\text{P}$ -labeled pS107V1 cDNA probe ( $1.5-2 \times 10^8$  dpm/ $\mu\text{g}$ ). Prehybridization of the filter in hybridization solution without the labeled probe was carried out for 3 hr at  $68^\circ\text{C}$ . The hybridizations with the probe were performed for 15 to 20 hr at  $68^\circ\text{C}$ . Following hybridization, the filters were washed three times in high-salt wash buffer (1 M NaCl, 0.05 M Tris-HCl (pH 7.5), 0.1% NaPPi, 0.1% SDS and  $5\times$  Denhardt's) and twice in low-salt wash buffer ( $1\times$  SSC, 0.1% NaPPi and 0.1% SDS). Each wash was carried out for 30 min at  $68^\circ\text{C}$ . Filters were exposed for 24 hr at  $-70^\circ\text{C}$  with an intensifying screen.

##### Isolation of Recombinant Clones

The Charon 4A recombinant bacteriophage libraries containing BALB/c sperm and M603 myeloma DNA have been previously described (Early et al., 1979; Davis et al., 1980). The Charon 4A recombinant library of genomic DNA from the W3207 myeloma tumor was constructed by S. Kim and M. Davis (unpublished results). These libraries were screened with a nick-translated  $^{32}\text{P}$ -labeled cloned S107  $V_H$  cDNA probe (Early et al., 1980) according to the method of Benton and Davis (1977).

##### DNA Sequencing

Restriction fragments containing the  $V_H$  genes were subcloned into pBR322 or pUK2 (D. Goldberg, B. Seed and R. Parker, unpublished results; Huang et al., 1981). Restriction mapping of the plasmid subclones was facilitated by blot hybridizations with the S107  $V_H$  cDNA sequence cloned into M13mp73. DNA sequencing was according to the method of Maxam and Gilbert (1980). Restriction fragments were labeled at their 5' ends with  $\gamma$ - $^{32}\text{P}$ -ATP and polynucleotide kinase, and at their 3' ends with  $\alpha$ - $^{32}\text{P}$ -dNTPs and DNA polymerase I (Klenow fragment). After the sequencing reactions, the products were fractionated on 40 cm long 25% polyacrylamide/7 M urea gels and 80 cm long 8% or 5% polyacrylamide/7 M urea gels. This allowed us to read well past 300 nucleotides from the end of a fragment.

##### Acknowledgments

The authors are indebted to Mark Davis and Mitch Kronenberg for valuable advice and discussions. We would also like to thank Phil Early and Stuart Kim for gifts of cloned DNAs, and Michael Douglas, Tim Hunkapiller and Bernita Larsh for preparation of the manuscript.

This work was supported by a grant from the NIH. S. C. was supported by fellowships from the NRSA and NIH; K. C. was supported by grants from the California Institute for Cancer Research and the University of California Cancer Research Coordinating Committee.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received March 16, 1981; revised April 27, 1981

## References

- Baltimore, D. (1974). Is terminal deoxynucleotidyl transferase a somatic mutagen in lymphocytes? *Nature* 248, 409-411.
- Barstad, P., Farnsworth, V., Weigert, M., Cohn, M. and Hood, L. (1974). Mouse immunoglobulin heavy chains are coded by multiple germline variable region genes. *Proc. Nat. Acad. Sci. USA* 71, 4096-4100.
- Barstad, P., Hubert, J., Hunkapiller, M., Goetze, A., Schilling, J., Black, B., Eaton, B., Richards, J., Weigert, M. and Hood, L. (1978). Immunoglobulins with hapten-binding activity: structure-function correlations and genetic implications. *Eur. J. Immunol.* 8, 497-503.
- Benton, W. D. and Davis, R. W. (1977). Screening  $\lambda$ gt recombinant clones by hybridization to single plaques in situ. *Science* 196, 180-182.
- Bernard, O., Hozumi, N. and Tonegawa, S. (1978). Sequences of mouse immunoglobulin light chain genes before and after somatic changes. *Cell* 15, 1133-1144.
- Brack, C., Hiram, M., Lenhard-Schuller, R. and Tonegawa, S. (1978). A complete immunoglobulin gene is created by somatic recombination. *Cell* 15, 1-14.
- Brenner, S. and Milstein, C. (1966). Origin of antibody variation. *Nature* 211, 242-243.
- Chang, S. and Rittenberg, M. (1981). Immunologic memory to phosphorylcholine in vitro. I. Asymmetric expression of clonal dominance. *J. Immunol.* 126, 975-980.
- Cohn, M., Blomberg, B., Geckeler, W., Raschke, W., Riblet, R. and Weigert, M. (1974). First order considerations in analyzing the generation of diversity. In *The Immune System: Genes, Receptors, Signals*, E. Sercarz et al., eds. (New York: Academic Press), pp. 89-117.
- Cory, S. and Adams, J. M. (1980). Deletions are associated with somatic rearrangement of immunoglobulin heavy chain genes. *Cell* 19, 37-51.
- Davis, M., Early, P., Calame, K., Livant, D. and Hood, L. (1979). The organization and rearrangement of heavy chain immunoglobulin genes in mice. In *Eukaryotic Gene Regulation, XIV, ICN-UCLA Symposium* (New York: Academic Press), pp. 393-406.
- Davis, M. M., Calame, K., Early, P. W., Livant, D. L., Joho, R., Weissman, I. L. and Hood, L. (1980). An immunoglobulin heavy-chain gene is formed by at least two recombinational events. *Nature* 283, 733-738.
- Early, P. W., Davis, M. M., Kaback, D. B., Davidson, N. and Hood, L. (1979). Immunoglobulin heavy chain gene organization in mice: analysis of a myeloma genomic clone containing variable and  $\alpha$  constant regions. *Proc. Nat. Acad. Sci. USA* 76, 857-861.
- Early, P., Huang, H., Davis, M., Calame, K. and Hood, L. (1980). An immunoglobulin heavy chain variable region is generated from three segments of DNA:  $V_H$ , D and  $J_H$ . *Cell* 19, 981-992.
- Edelman, G. M. and Gally, J. A. (1967). Somatic recombination of duplicated genes: an hypothesis on the origin of antibody diversity. *Proc. Nat. Acad. Sci. USA* 57, 353-358.
- Gearhart, P. J., Johnson, N. D., Douglas, R. and Hood, L. (1981). IgG antibodies to phosphorylcholine exhibit more diversity than their IgM counterparts. *Nature* 291, 29-34.
- Hewick, R. M., Hunkapiller, M. W., Hood, L. E. and Dreyer, W. J. (1981). A gas-liquid solid phase peptide and protein sequencer. *J. Biol. Chem.*, in press.
- Hood, L., Loh, E., Hubert, J., Barstad, P., Eaton, B., Early, P., Fuhrman, J., Johnson, N., Kronenberg, M. and Schilling, J. (1976). The structure and genetics of mouse immunoglobulins: an analysis of NZB myeloma proteins and sets of BALB/c myeloma proteins binding particular haptens. *Cold Spring Harbor Symp. Quant. Biol.* 41, 817-836.
- Hood, L., Davis, M., Early, P., Calame, K., Kim, S., Crews, S. and Huang, H. (1980). Two types of DNA rearrangements in immunoglobulin genes. *Cold Spring Harbor Symp. Quant. Biol.* 45, in press.
- Huang, H., Crews, S. and Hood, L. (1981). An immunoglobulin  $V_H$  pseudogene. *J. Mol. Appl. Genet.*, in press.
- Hunkapiller, M. and Hood, L. (1980). A new protein sequencer with increased sensitivity. *Science* 207, 523-525.
- Max, E. E., Seidman, J. G. and Leder, P. (1979). Sequence of five potential recombination sites encoded close to an immunoglobulin  $\kappa$  constant region gene. *Proc. Nat. Acad. Sci. USA* 76, 3450-3454.
- Maxam, A. M. and Gilbert, W. (1980). Sequencing end-labeled DNA with base-specific chemical cleavages. *Meth. Enzymol.* 65, 499-560.
- Nagamine, Y., Sentenac, A. and Fornageo, R. (1980). Selective blotting of restriction DNA fragments on nitrocellulose membranes at low salt concentrations. *Nucl. Acids Res.* 8, 2453-2460.
- Robinson, E. A. and Appella, E. (1977). Amino acid sequence of the first 217 residues of a mouse heavy chain (MOPC47A) with a domain deletion. *Proc. Nat. Acad. Sci. USA* 74, 2465-2469.
- Rodwell, J. D. and Karush, F. (1980). Restriction in IgM expression. I. The  $V_H$  regions of equine anti-lactose antibodies. *Mol. Immunol.* 17, 1553-1561.
- Sakano, H., Huppi, K., Heinrich, G. and Tonegawa, S. (1979). Sequences at the somatic recombination sites of immunoglobulin light-chain genes. *Nature* 280, 288-294.
- Sakano, H., Maki, R., Kurosawa, Y., Roeder, W. and Tonegawa, S. (1980). Two types of somatic recombination are necessary for the generation of complete immunoglobulin heavy-chain genes. *Nature* 286, 676-683.
- Sakano, H., Kurosawa, Y., Weigert, M. and Tonegawa, S. (1981). Identification and nucleotide sequence of a diversity DNA segment (D) of immunoglobulin heavy-chain genes. *Nature* 290, 562-564.
- Scharff, M., Cook, W., Giusti, A., Kwan, S.-P., Thammana, P., Yelton, D., Zack, D. and Rudikoff, S. (1981). Antigen binding mutants of mouse myeloma cells. In *Immunoglobulin Idiotypes and Their Expression, ICN-UCLA Symposium*, E. Sercarz et al., eds. (New York: Academic Press), in press.
- Seidman, J. G., Leder, A., Nau, M., Norman, B. and Leder, P. (1978). Antibody diversity. *Science* 202, 11-17.
- Seidman, J. G., Max, E. E. and Leder, P. (1979). A  $\kappa$ -immunoglobulin gene is formed by site-specific recombination without further somatic mutation. *Nature* 280, 370-375.
- Valbuena, O., Marcu, K. B., Weigert, M. and Perry, R. P. (1978). Multiplicity of germline genes specifying a group of related mouse  $\kappa$  chains with implications for the generation of immunoglobulin diversity. *Nature* 276, 780-784.
- Weigert, M. and Riblet, R. (1976). Genetic control of antibody variable regions. *Cold Spring Harbor Symp. Quant. Biol.* 41, 837-846.
- Weigert, M. G., Cesari, H. M., Yonkovich, S. J. and Cohn, M. (1970). Variability in the lambda light chain sequences of mouse antibody. *Nature* 228, 1045-1047.
- Weigert, M., Gatmaitan, L., Loh, E., Schilling, J. and Hood, L. (1978). Rearrangement of genetic information may produce immunoglobulin diversity. *Nature* 276, 785-790.
- Weigert, M., Perry, R., Kelley, D., Hunkapiller, T., Schilling, J. and Hood, L. (1980). The joining of V and J gene segments creates antibody diversity. *Nature* 283, 497-499.

## Chapter 2

### An Immunoglobulin V<sub>H</sub> Pseudogene

This paper was published in the Journal of Molecular and Applied Genetics.

# An Immunoglobulin V<sub>H</sub> Pseudogene

Henry Huang, Stephen Crews, and Leroy Hood

*Division of Biology, California Institute of Technology, Pasadena, California 91125, U. S. A.*

---

**Summary:** In the course of studying the members of the T15 group of V<sub>H</sub> gene segments, some of which participate in the immune response to phosphorylcholine in the mouse, we identified a V<sub>H</sub> gene segment that contains three mutations preventing its expression. The mutations are an in-frame stop codon, a 4-base insertion which causes a termination codon to be shifted into the reading frame, and a modification of the recognition elements involved in the joining of V<sub>H</sub> and D gene segments during variable region formation. This pseudogene, which is 88–96% homologous to the other members of the T15 V<sub>H</sub> gene group, is probably of relatively recent origin and will presumably be deleted from the V<sub>H</sub> gene family eventually. We suggest that pseudogenes can only arise in multigene families and that the occurrence of pseudogenes will be a relatively frequent phenomenon in these families. Because the antibody gene families are made up of multiple gene elements, undergo two types of DNA rearrangements during differentiation, and employ several different RNA splicing mechanisms for expression, there are many different ways a particular antibody gene segment may become a pseudogene. **Key Words:** Pseudo-V<sub>H</sub> gene—V<sub>H</sub> gene family—Multigene families.

---

A pseudogene is defined as a DNA sequence that displays significant homology to a functional gene, but has mutations preventing its expression (1). Pseudogenes have been found in several multigene families, including the 5S rRNA genes (2), the  $\alpha$ -globin genes (1, 3–5) and the  $\beta$ -globin genes (6–10). As do other multigene families, the antibody gene families present a variety of alternative pathways by which a gene segment may become a pseudogene. The immunoglobulin gene families are encoded by multiple gene segments that undergo DNA rearrangements during the differentiation of individual antibody-producing cells. The mechanisms for DNA rearrangements, reviewed briefly in the

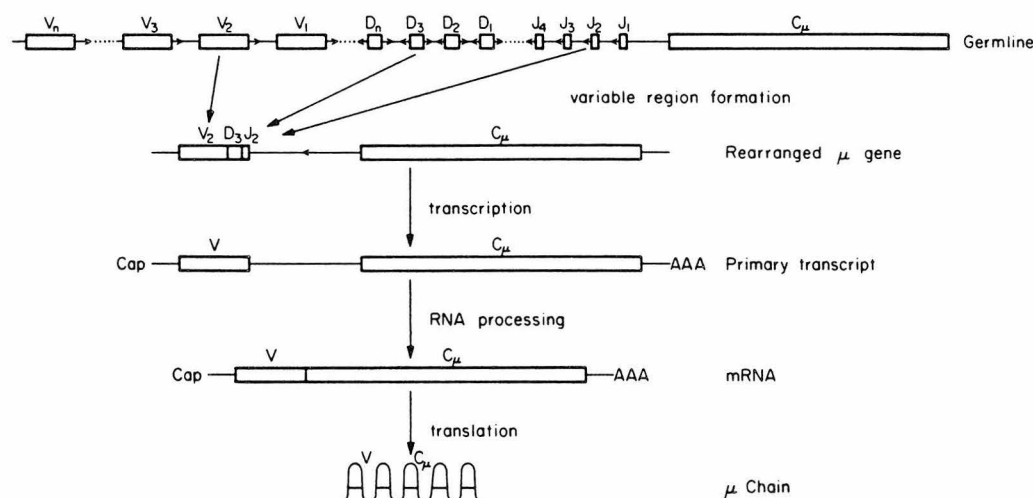
following paragraphs, provide novel pathways whereby immunoglobulin pseudogenes may arise.

Three separate multigene families, lambda ( $\lambda$ ), kappa ( $\kappa$ ), and heavy (H), encode the immunoglobulin light and heavy chains. Each polypeptide chain is encoded by several distinct gene segments (11): variable (V<sub>L</sub>), joining (J<sub>L</sub>), and constant (C<sub>L</sub>) for light chains (12–17), and V<sub>H</sub>, diversity (D), J<sub>H</sub>, and C<sub>H</sub> for heavy chains (18–20)(Fig. 1). These gene segments are separated by intervening DNA sequences in the germline and are rearranged with respect to one another during the differentiation of antibody-producing cells. Light chain genes are formed by the joining of V<sub>L</sub> and J<sub>L</sub>, and heavy chain genes are formed by the joining of V<sub>H</sub>, D, and J<sub>H</sub> gene segments. There are two highly conserved sequences, 7 and 10 nucleotides in length, which are

---

Received February 11, 1981; accepted May 11, 1981.  
Address correspondence and reprint requests to Dr. Hood.





**FIG. 1.** The expression of an immunoglobulin heavy chain  $\mu$  gene. The conserved, joining signals flanking  $V_H, D$ , and  $J_H$  gene segments are indicated by arrows. The introns in the  $C_\mu$  gene segment are omitted for simplicity. The precise location of  $D$  gene segments has not been determined.

located on the 3' side of all germline  $V$  gene segments and 5' to all germline  $J$  gene segments studied to date. The location and relative invariance of these conserved sequences in the  $\lambda, \kappa$ , and heavy gene families suggest that they play a fundamental role in the antibody DNA rearrangements. These conserved sequences are postulated to be recognition signals for enzymes which mediate  $V_L-J_L$  joining (12, 15, 16) or  $V_H-D-J_H$  joining (19, 20; P. Early and L. Hood, in preparation). The conserved sequences are separated from each other by either 11 bases (one DNA helical turn) as in  $J_\lambda, V_\kappa$ , and  $D$  gene segments, or 20–23 nucleotides (two DNA helical turns) as in  $V_H, J_H, J_\kappa$ , and  $V_\lambda$  gene segments. In all cases a gene segment flanked by recognition signals separated by one turn is joined to a gene segment flanked by signals separated by two turns (19). This has been denoted the "one-turn to two-turn" joining rule (19).

The joined  $V_L-J_L$  and  $V_H-D-J_H$  gene segments encode the variable regions or antigen-binding domains of light and heavy chains, respectively. The joined  $V$  gene is separated from the  $C$  gene segment by intervening DNA sequences. This rearranged configuration is transcriptionally active, and the primary RNA transcript contains the coding regions as well as the intervening sequences. There is evidence suggesting that signals at either end of the intervening sequences are involved in RNA proc-

essing (21, 22). RNA processing removes the intervening sequences to generate the mature mRNA containing uninterrupted coding sequences for the variable and constant regions. In the heavy chain family, transcripts can be processed in one of two ways to give rise to alternate forms of mRNA which encode membrane-bound or secreted polypeptides (23–25). Alternate RNA processing pathways also may give rise to alternative mRNAs encoding  $\mu$  or  $\delta$  chains (26). Translation of the mRNA results in a nascent polypeptide which undergoes folding, disulfide bonding, and other post-translational modifications such as removal of the leader peptide and glycosylation, to give rise to the functionally active polypeptide.

The  $V_H$  gene, initially associated with the  $C_\mu$  gene segment, may then subsequently be associated with any of a number of other  $C_H$  gene segments. This phenomenon is called  $C_H$  switching, and involves DNA rearrangements distinct from  $V_H-D-J_H$  joining (reviewed in Ref. 27).

The complex series of events that leads to immunoglobulin gene expression suggests that immunoglobulin pseudogenes can result from many different kinds of mutations. Two pseudogene segments have been reported to date. First, there is a mouse  $J_\kappa$  gene segment,  $J_{\kappa 3}$ , which lacks a proper RNA splicing consensus sequence to its 3' side (15, 16). Since the protein sequence encoded by  $J_{\kappa 3}$  has not

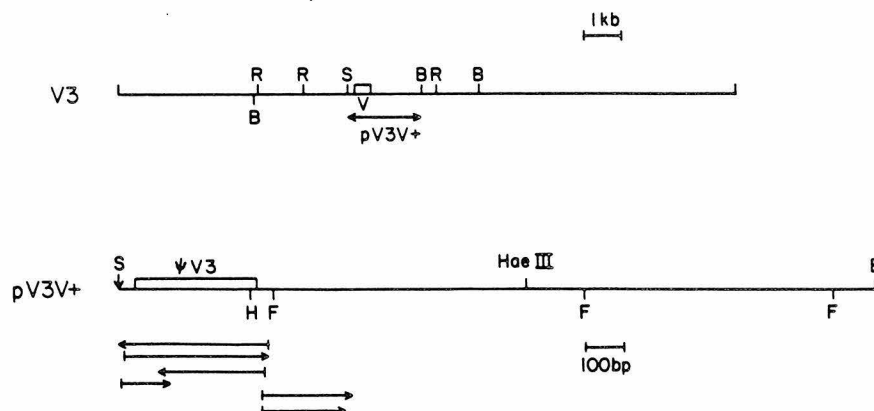
been observed in any of the 35 or more  $J_k$  segments sequenced to date (28), it probably cannot participate in the removal of the intervening sequence between the V and C coding regions. In contrast, many examples of  $J_k$ 1-,  $J_k$ 2-,  $J_k$ 4-, and  $J_k$ 5-encoded protein sequences are known (28). Furthermore, the human  $J_k$  gene segment cluster does not contain a  $J_k$ 3-like sequence, suggesting that the mouse pseudogene segment  $J_k$ 3 was formed by a recent duplication (29). Second, a human  $V_k$  pseudogene was recently reported (30). This pseudogene has base substitutions as well as insertions and deletions which would result in amino acid replacements at highly conserved residues. The coding region also contains frameshifts and in-phase termination codons. In addition, in the 3' flanking region the 10-base  $V_k$  joining signal appears to be abnormal.

In our laboratory we are analyzing all of the  $V_H$  gene segments that are closely related to the T15  $V_H$  gene segment. Some of the members of this group participate in the immune response to the simple hapten phosphorylcholine. In the course of cloning and sequencing all of these gene segments (S. Crews et al., in press), we determined that one of the five members of this small multigene family is a pseudogene. We can clearly delineate three mutational events that inactivate the gene segment. Moreover, because we know the sequences of the other germline  $V_H$  gene segments in the T15 group, we can speculate about the evolutionary history of this pseudogene.

## MATERIALS AND METHODS

Restriction enzymes were obtained from New England Biolabs and were used according to the manufacturer's instructions. Bacterial alkaline phosphatase was obtained from Bethesda Research Laboratories, Inc., and used at 65° C and a concentration of 6 u/pmol 5' end. The V3 clone (Fig. 2) was isolated from a Charon 4A recombinant phage library (31) of BALB/c mouse sperm DNA partially digested with *Hae*III and *Alu*I (32). The library was screened with a <sup>32</sup>P-labeled T15-like  $V_H$  region cDNA subcloned in pBR322 (33) as a probe. The probe is derived from the S107 myeloma tumor whose immunoglobulin  $V_H$  region is identical to that of T15. The genomic library was screened by the Benton-Davis filter screening procedure (34). Southern blot analysis was performed as described by Southern (35). The 1.9 kilobase (kb) *Sau*3AI-*Bam*HI fragment containing the V gene segment (Fig. 2) was ligated with *Bam*HI-digested and bacterial alkaline phosphatase-treated plasmid pUK2 and grown in HB101 bacterial host. The pUK2 plasmid is a derivative of pBR322 with the *Hae*II 235-*Hae*II 2351 region deleted (36), and a synthetic *Eco*RI-*Bam*HI-*Eco*RI linker inserted into the *Eco*RI site (D. Goldberg, B. Seed, and R. Parker, unpublished). This subclone was named pV3V<sup>+</sup>.

The nucleotide sequence of the V gene segment was determined from the pV3V<sup>+</sup> subclone by base-



**FIG. 2.** The  $\lambda$  and plasmid clones containing  $\psi$ V3. The V3 clone is a Charon 4A clone in which the  $\psi$ V3 gene segment was mapped to be within the 1.9 kb *Sau*3AI(S)-*Bam*HI(B) fragment. This fragment was subcloned into the pUK2 plasmid and named pV3V<sup>+</sup>. The 5' end-labeled fragments from the *Sau*3AI(S), *Hha*I(H), and *Fnu*4HI(F) sites were sequenced as indicated by the arrows. The *Eco*RI sites in V3 are indicated by R. The single *Hae*III site in the pV3V<sup>+</sup> insert is also indicated.

specific chemical degradation of 5' end-labeled fragments (37, 38). The sequencing strategy is shown in Fig. 2.

## RESULTS AND DISCUSSION

### The Isolation of a Pseudogene Denoted $\Psi V3$

Southern blot analyses showed that the 3.4 kb *Eco*RI fragment of V3 that contains the V gene segment hybridizes strongly with the T15-like probe, and it comigrates with one of four bands in *Eco*RI-digested sperm or embryo DNA that strongly hybridizes to the same probe (39; S. Crews et al., in press). Thus by the criterion of strong hybridization, the V3 clone contains a V<sub>H</sub> gene segment that is a member of the T15 group. The V<sub>H</sub> gene segment in this clone is denoted  $\Psi$ V3 for reasons that will become apparent subsequently.

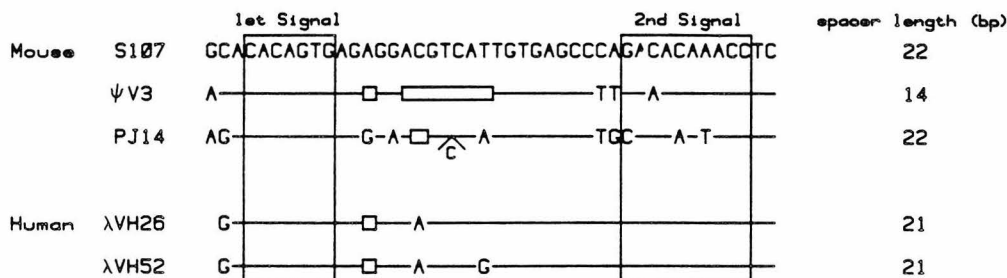
The DNA sequence of  $\Psi V3$  encodes the last four codons of the leader peptide (40, 41) and amino acid residues 1–101 of the heavy chain V region, similar to other  $V_H$  gene segments (19, 20, 42). A comparison of the nucleotide sequence of  $\Psi V3$  with the T15 germline  $V_H$  gene segment, the prototype for the T15 group, is shown in Fig. 3. The  $\Psi V3$  and T15  $V_H$  gene segments are 88% (35 differences/303 bases) homologous in the coding region. The  $\Psi V3$  is 96% (13/303) homologous and 90% (29/303) homologous to two additional members of the T15 group, V11 and V13, respectively (S. Crews et al., in press). This homology led us to assign  $\Psi V3$  as a member of the T15  $V_H$  gene group.

### The $\Psi$ V3 Gene Segment is a Pseudogene

There are three aberrant features in the  $\Psi V3$  sequence that lead us to conclude  $\Psi V3$  must be a pseudogene (Fig. 3). First, the codon for the highly conserved tryptophan residue (59 occurrences/63

**FIG. 3.** Comparison of the coding region of  $\psi$ V3 and S107 (T15)  $V_H$  gene segments. Sequence identity is indicated by horizontal dashes. The data for S107, which is identical to T15, are from Ref. 19. The translated protein sequence also is given. Note the termination codons starting at bases 151 and 286 in  $\psi$ V3. The cysteine 98, which is disulfide bonded to cysteine 22 in the S107 heavy chain, is to the carboxy-terminal side of the termination codons. Thus the disulfide bridge, a fundamental part of all antibody domains, cannot be formed. A gap in the S107 DNA sequence indicates the position of the 4-base insertion in  $\psi$ V3.

31	G I Q C E V K L V E S G G L V Q P G A S L R L S C A S S G F T F T D	31
105	GST ATC CAG TGT GAG GTG AAG CTG GAG TCT GGA GGC TTG GTA CAG CCT GGG GCT TCT CTG AGA CTC TCC TGT GCA TCT TCT GGG TTC ACC TTC ACT GAT	105
105	---	105
31	G I Q C E V K L V E S G G L V Q P G G S L R L S C A T S G F T F S D	31
66	Y Y M N W V H R P P G K A L E X L A L I R N K A N G Y I T E Y S A S M	66
210	TAC TAC ATG AAC TGG GTC CAC CGG CCT CCA GGG AAG GCA CTT GAG TAG TTG GCT TTG ATT AGA AAC AAA GCT AAT GGT TAC ATA ACA GAG TAC AGT GCA TCT ATG	210
210	-T- --- -G-G --- -G- -A- --- -G- A-T --- GCA -G- --- -G- A-T -C- --- -G- --- -G- ---	210
66	F Y M E W V R Q P P G K R L E W I A A S R N K A N D Y T T E Y S A S V	66
102	K G R F T I S R D N S Q S I L Y L Q M N T L S A H X G Q C H L L C K R	102
319	AAG GGT CGG TTC ACC ATC TCC AGA GAT AAT TCC CAA AGC ATC CTC TAT CTT CAA ATG AAC ACA CTG AGC GCA CAC TGA GGA CAG TGC CAC TTA TTA CTG TGC AAG AGA T	319
316	---	316
101	K G R F I V S R D T S Q S I L Y L Q M N A L R A E D T A I Y Y C A R D	101



**FIG. 4.** Comparison of joining signals and their spacers 3' to mouse and human  $V_H$  gene segments. Sequence identity is indicated by a horizontal line; the 1-base insertion in PJ14 is indicated by the arrowhead; deletions are indicated by boxed regions. The joining signals are boxed and the spacer lengths (in bases) are shown to the right of each sequence. Data are taken from the following references: T15 (19), PJ14 (20), human  $V_H$  (42).

sequences)(28) at base 151 has mutated to a termination codon (TAG). Second, a 4-base insertion, CACA or ACAC, after base 280 or 281, respectively, brings a termination codon (TGA) normally out of frame into the translational reading frame. Finally, instead of the normal 20–23 nucleotide separation between the 7- and 10-base joining signals, the  $\Psi V3$  gene segment has a deletion of 8 bases, resulting in a spacer of only 14 nucleotides (Fig. 4). Since the  $\Psi V3$  gene segment has an abnormal 14-base spacer, it probably cannot participate in  $V_H$ -D joining. This conclusion is based on the fact that the joining signals and the precise spacer relationships have been conserved over the more than 500 million years of vertebrate evolution—suggesting that enormous precision is required in the lengths of these elements for appropriate  $V_L$ - $J_L$  or  $V_H$ -D- $J_H$  joining. The  $\Psi V3$  gene segment probably exists in the BALB/c mouse germline since these three mutations are independent; it is unlikely that they all arose as cloning artifacts.

There is one conceivable scheme by which the  $\Psi V3$  gene segment could have a function. If recombination or gene conversion were an important mechanism for producing somatic variation among V gene segments (43, 44), then the  $\Psi V3$  gene segment could recombine with or convert other germline  $V_H$  gene segments to generate novel V gene segments, as long as the defective information was not included in the recombined V gene segment.

#### The $\Psi V3$ Gene Segment Diverged Recently from Another Member of the T15 $V_H$ Gene Family

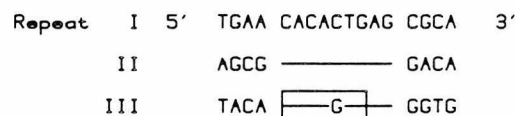
Since the  $\Psi V3$  gene segment is 96% homologous to a second member of the T15 gene family, V11

(S. Crews et al., in press), it appears that these two genes diverged from one another relatively recently. We believe that the  $\Psi V3$  pseudogene is a nonfunctional  $V_H$  gene segment of relatively recent origin now in the process of free accumulation of random mutations unchecked by natural selection. We propose that one of the three inactivating mutations appeared first and permitted the free accumulation of the other inactivating mutations. It is impossible to determine which of the three inactivating mutations occurred first.

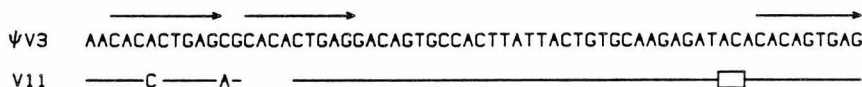
#### Origin of the Insertion

Besides shifting the translational reading frame, the 4-base insertion in the  $\Psi V3$  gene segment creates a sequence CACACTGAG which is a perfect direct repeat of the sequence 3–11 nucleotides 5' to the insertion (Fig. 5). Remarkably, a portion of this repeated sequence, CACACTG, is almost identical to one of the  $V_H$ -D- $J_H$  joining signals, the 7-base CACAGTG which is 33 bases 3' to repeat II. Thus the insertion also created three almost-perfect direct repeats.

The direct repeats I and II are similar to the short direct repeats generated by the duplication of a host sequence during the insertion of transposable genetic elements such as insertion sequences (IS) and



**FIG. 5.** Three direct repeats found in the  $\psi V3$  gene segment. The sequence of each repeat is shown together with 4 bases of flanking 5' and 3' sequences. The first  $V_H$ -D- $J_H$  joining signal is boxed.

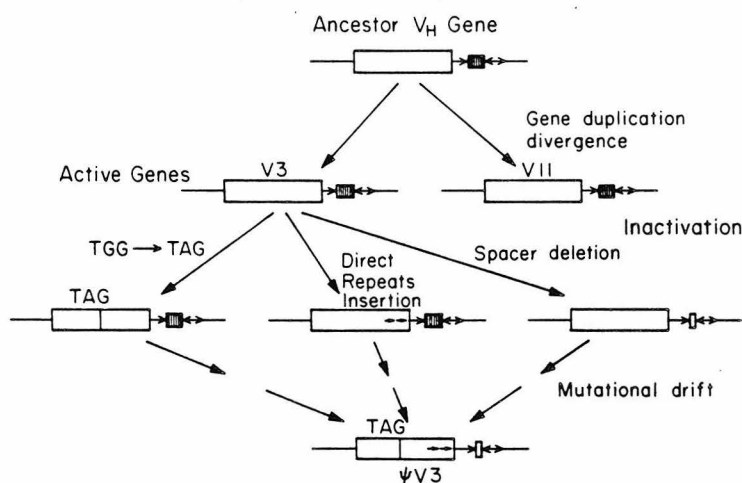


**FIG. 6.** Comparison of the  $\psi V3$  and V11 sequences in the region of the  $\psi V3$  repeats. The repeats are indicated by the arrows above the  $\psi V3$  sequence. Homology is indicated by a horizontal line. Gaps were introduced into the V11 sequence to maximize homology with  $\psi V3$ . The first gap corresponds to the insertion in  $\psi V3$ .

transposons in bacteria (reviewed in Ref. 45), TY1 in yeast (46), *copia* in *Drosophila* (47), and a number of retroviruses in mammals (reviewed in Ref. 48). It is possible that some transposable genetic element could have generated the direct repeats upon its insertion, and subsequent imprecise excision could have resulted in the direct repeats separated by two nucleotides. We find this model unappealing since imprecise excision of a transposable element would result in a portion of the element being left in the host DNA. The  $\psi V3$  sequence between repeats II and III is identical to that of V11 except for a 2-base deletion in V11 (Fig. 6), precluding insertion of a mobile element in this region. Examination of the V11 gene segment indicates that it requires only one base change to generate repeat I, and that the insertion alone generates repeat II. Furthermore, more than half of the repeats I and II, CTGAG, is already present in V11. Thus the putative transposable element probably did not insert between repeats I and II, and simple mutational events rather than transposable elements seem sufficient to explain the direct repeats (Fig. 7). At present it is not clear if the insertion resulted from errors during recombination, repair, or replication.

### Pseudogenes Usually Occur in Multigene Families

Genes which are members of a multigene family may be less exposed to natural selection than genes which are alleles at a single locus (49). The extent of shielding from natural selection depends upon the number of genes in the family. It also depends upon the degree of overlap of the function of the genes within the multigene family. For example, the decrease in activity or the total loss of activity of a single  $V_H$  gene segment that is important in producing a specific hapten-binding immunoglobulin may be compensated for by other  $V_H$  gene segments that perform a very similar function. In contrast, if an allele at a single locus is defective, the functional impairment may be quite significant. Thus it can be seen intuitively that since natural selection does not operate intensely upon their removal, defective genes in multigene families would persist over a longer evolutionary time, and such pseudogenes would be present in more individuals in the species than would mutant genes of a single locus system (50, 51). Thus mutant genes of single locus systems, which have short evolutionary life-



**FIG. 7.** Postulated evolutionary descent of the  $\psi V3$  gene segment. The  $\psi V3$  and V11 gene segments are closely related and probably resulted from duplication of an ancestral gene with subsequent mutational drift and gene inactivation leading to  $\psi V3$ . The coding region is indicated by the open box; flanking sequences are indicated by horizontal lines. The conserved  $V_H$ -D-J $_H$  joining signals are drawn as single- and double-headed arrows, respectively, and the spacer between them is the hatched box. Direct repeats in  $\psi V3$  are shown as single-headed arrows pointing in the same direction. The TGG  $\rightarrow$  TAG mutation is indicated by the vertical line labeled TAG.



TABLE 1. Possible sources of immunoglobulin pseudogenes

	Defect in	References
DNA level		
V-(D)-J joining:	Signals, spacer	30, This paper
C <sub>H</sub> switch:	Signals	27
RNA level		
Transcription:	Promoter signals	52
	Signals for poly(A) addition	24, 53, 54
Processing:	Signals for splicing out intervening sequences	15, 16
Signals for alternate pathways:	Membrane versus secreted	23-25
	μ versus δ	26
Protein level		
Translational:	Ribosome binding	55-57
	Initiation	55-57
	Termination	30, This paper
Post-translational:	Folding (e.g., disulfide bonding)	This paper
	Modification (e.g., glycosylation)	58
Membrane passage:	Leader peptide	19, 41
Membrane anchorage:	Hydrophobic tail	58

Steps in immunoglobulin gene expression where defects may be found in contemporary immunoglobulin pseudogenes. References are given where the defect has been observed, otherwise the reference refers to the step involved.

spans, occur at low frequencies in the species, whereas pseudogenes, which have much longer evolutionary lifespans, may occur throughout a species. The longer evolutionary lifespan is reflected in the multiple mutational events associated with most of the pseudogenes studied to date. In contrast, mutant genes typically contain only one or a few mutations, reflecting their short evolutionary lifespans. We predict that most pseudogenes will be a DNA sequence in a *multigene family* that displays significant homology to a functional gene in the family, but has mutations preventing its expression.

### Immunoglobulin Genes May Become Pseudogenes in Many Different Ways

The immunoglobulin multigene families are large, and gene members have extensive functional overlap. Thus it would not be surprising to find these families harboring many pseudogenes. Indeed, two pseudogenes for V gene segments have been found in some 20 examples analyzed to date. Defects in immunoglobulin genes may operate at a number of levels of gene expression (Table 1). Moreover, since the immunoglobulin families utilize DNA rear-

rangements and a variety of RNA splicing mechanisms to express their genes, defects in these mechanisms may provide a class of pseudogenes which will not be found in most multigene families (Table 1). In conclusion, we predict that there will be many immunoglobulin pseudogenes displaying defects at a variety of steps of immunoglobulin gene expression.

### ACKNOWLEDGMENTS

This work was supported by a grant from the National Institutes of Health. We thank Dr. Johanna Griffin for critically reviewing this manuscript.

### REFERENCES

1. Proudfoot NJ, Maniatis T: The structure of a human  $\alpha$ -globin pseudogene and its relationship to a  $\alpha$ -globin gene duplication. *Cell* 21:537-544, 1980
2. Jacq C, Miller JR, Brownlee GG: A pseudogene structure in 5S DNA of *Xenopus laevis*. *Cell* 12:109-120, 1977
3. Nishioka Y, Leder A, Leder P: An unusual alpha globin-like gene that has cleanly lost both globin intervening sequences. *Proc Natl Acad Sci USA* 77:2806-2809, 1980

4. Vanin EF, Goldberg GI, Tucker PW, Smithies O: A mouse alpha globin-related pseudogene ( $\Psi\alpha 30.5$ ) lacking intervening sequences. *Nature* 286:222–226, 1980
5. Lauer J, Shen JC-K, Maniatis T: The chromosomal arrangement of human  $\alpha$ -like globin genes: Sequence homology and  $\alpha$ -globin gene deletions. *Cell* 20:119–130, 1980
6. Cleary ML, Haynes JR, Schon EA, Lingrel JB: Identification by nucleotide sequence analysis of a goat pseudoglobin gene. *Nucleic Acids Res* 8:4791–4802, 1980
7. Fritsch EF, Lawn RM, Maniatis T: Molecular cloning and characterization of the human  $\beta$ -like globin gene cluster. *Cell* 19:959–972, 1980
8. Hardison RC, Butler ET III, Lacy E, Maniatis T: The structure and transcription of four linked rabbit  $\beta$ -like globin genes. *Cell* 18:1285–1297, 1979
9. Jahn CL, Hutchinson CA III, Phillips SJ, Weaver S, Haigwood NL, Voliva CF, Edgell MH: DNA sequence organization of the  $\beta$ -globin complex in the BALB/c mouse. *Cell* 21:159–168, 1980
10. Lacy E, Maniatis T: The nucleotide sequence of a rabbit  $\beta$ -globin pseudogene. *Cell* 21:545–553, 1980
11. Dreyer WJ, Bennett JC: The molecular basis of antibody formation: A paradox. *Proc Natl Acad Sci USA* 54:864–868, 1965
12. Bernard O, Hozumi N, Tonegawa S: Sequences of mouse immunoglobulin light chain genes before and after somatic changes. *Cell* 15:1133–1144, 1978
13. Brack C, Tonegawa S: Variable and constant parts of the immunoglobulin light chain gene of a mouse myeloma cell are 1250 nontranslated bases apart. *Proc Natl Acad Sci USA* 74:5652–5656, 1977
14. Brack C, Hiram M, Lenhard-Schuller R, Tonegawa S: A complete immunoglobulin gene is created by somatic recombination. *Cell* 15:1–14, 1978
15. Max EE, Seidman JG, Leder P: Sequence of five potential recombination sites encoded close to an immunoglobulin  $\kappa$  constant region gene. *Proc Natl Acad Sci USA* 76:3450–3454, 1979
16. Sakano H, Huppi K, Heinrich G, Tonegawa S: Sequences at the somatic recombination sites of immunoglobulin light-chain genes. *Nature* 280:288–294, 1979
17. Seidman JG, Max EE, Leder P: A  $\kappa$ -immunoglobulin gene is formed by site-specific recombination without further somatic mutation. *Nature* 280:370–375, 1979
18. Schilling J, Clevinger B, Davie JM, Hood L: Amino acid sequence of homogeneous antibodies to dextran and DNA rearrangements in heavy chain V region gene segments. *Nature* 283:35–40, 1980
19. Early P, Huang H, Davis M, Calame K, Hood L: An immunoglobulin heavy chain variable region gene is generated from three segments of DNA:  $V_H$ , D and  $J_H$ . *Cell* 19:981–992, 1980
20. Sakano H, Maki R, Kurosawa Y, Roeder W, Tonegawa S: The two types of somatic recombination necessary for generation of complete immunoglobulin heavy chain genes. *Nature* 286:676–683, 1980
21. Breathnach R, Benoist C, O'Hare K, Gannon F, Chambon P: Ovalbumin gene: Evidence for a leader sequence in mRNA and DNA sequences at the exon-intron boundaries. *Proc Natl Acad Sci USA* 75:4853–4857, 1978
22. Lerner M, Boyl J, Mount S, Wolin S, Steitz J: Are snRNP involved in splicing? *Nature* 283:220–224, 1980
23. Alt FW, Bothwell ALM, Knapp M, Siden E, Mather E, Koshland M, Baltimore D: Synthesis of secreted and membrane-bound immunoglobulin mu heavy chains is directed by mRNAs that differ at their 3' ends. *Cell* 20:293–302, 1980
24. Early P, Rogers J, Davis M, Calame K, Bond M, Wall R, Hood L: Two mRNAs can be produced from a single immunoglobulin  $\mu$  gene by alternative RNA processing pathways. *Cell* 20:213–320, 1980
25. Rogers J, Early P, Carter C, Calame K, Bond M, Hood L, Wall R: Two mRNAs with different 3' ends encode membrane-bound and secreted forms of immunoglobulin  $\mu$  chain. *Cell* 20:303–312, 1980
26. Moore KW, Rogers J, Hunkapiller T, Early P, Nottenburg C, Weissman I, Bazin H, Wall R, Hood LE: The expression of immunoglobulin D may employ both DNA rearrangements and RNA splicing mechanisms. *Proc Natl Acad Sci USA* 78:1800–1804, 1981
27. Davis MM, Kim SK, Hood L: Immunoglobulin class switching: Developmentally regulated DNA rearrangements during differentiation. *Cell* 22:1–2, 1980
28. Kabat EA, Wu TT, Bilofsky H *Sequences of Immunoglobulin Chains*, Washington D.C., U.S. Department of Health, Education, and Welfare Public Health Service, National Institutes of Health, 1979
29. Hieter PA, Max EE, Seidman JG, Maizel JV Jr, Leder P: Cloned human and mouse kappa immunoglobulin constant and J region genes conserve homology in functional segments. *Cell* 22:197–207, 1980
30. Bentley DL, Rabbitts TH: Human immunoglobulin variable region genes—DNA sequences of two  $V_\kappa$  genes and a pseudogene. *Nature* 288:730–733, 1980
31. Maniatis T, Hardison RC, Lacy E, Lauer J, O'Connell C, Quon D, Sim GK, Efstratiadis A: The isolation of structural genes from libraries of eukaryotic DNA. *Cell* 15:687–701, 1978
32. Davis MM, Calame K, Early PW, Livant DL, Joho R, Weissman IL, Hood L: An immunoglobulin heavy chain gene is formed by at least two recombinational events. *Nature* 283:733–738, 1980
33. Early PW, Davis MM, Kaback DB, Davidson N, Hood L: Immunoglobulin heavy chain gene organization in mice: Analysis of a myeloma genomic clone containing variable and  $\alpha$  constant regions. *Proc Natl Acad Sci USA* 76:857–861, 1979
34. Benton WD, Davis RW: Screening  $\lambda$ gt recombinant clones by hybridization to single plaques *in situ*. *Science* 196:180–182, 1977
35. Southern EM: Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J Mol Biol* 98:503–517, 1975
36. Sutcliffe JG: Complete nucleotide sequence of the *Escherichia coli* plasmid pBR322. *Cold Spring Harbor Symp Quant Biol* 43:77–90, 1978
37. Maxam AM, Gilbert W: Sequencing end-labeled DNA with base-specific chemical cleavages. In: *Methods in Enzymology*, Vol. 65, ed. by L Grossman and K Moldave, New York, Academic Press, 1980, pp 499–560
38. Smith DR, Calvo JM: Nucleotide sequence of the *E. coli*

- gene coding for dihydrofolate reductase. *Nucleic Acids Res* 8:2255-2274, 1980
39. Cory S, Adams JM: Deletions are associated with somatic rearrangement of immunoglobulin heavy chain genes. *Cell* 19:37-51, 1980
  40. Blöbel G, Dobberstein B: Transfer of proteins across membranes. I. Presence of proteolytically processed and unprocessed nascent immunoglobulin light chains on membrane-bound ribosomes of murine myeloma. *J Cell Biol* 67:835-851, 1975
  41. Milstein C, Brownlee GG, Harrison TM, Mathews MB: A possible precursor of immunoglobulin light chains. *Nature* 239:117-120, 1972
  42. Matthysens G, Rabbitts TH: Structure and multiplicity of genes for the human immunoglobulin heavy chain variable region. *Proc Natl Acad Sci USA* 77:6561-6565, 1980
  43. Edelman GM, Gally JA: Somatic recombination of duplicated genes: An hypothesis on the origin of antibody diversity. *Proc Natl Acad Sci USA* 57:353-358, 1967
  44. Seidman JG, Leder A, Nau M, Norman B, Leder P: Antibody diversity. *Science* 202:11-17, 1978
  45. Calos MP, Miller JH: Transposable elements. *Cell* 20:579-595, 1980
  46. Farabaugh PJ, Fink GR: Insertion of the eukaryotic transposable element Ty1 creates a 5-base pair duplication. *Nature* 286:352-356, 1980
  47. Dunsmuir P, Brorstein WJ Jr, Simon MA, Rubin GM: Insertion of the *Drosophila* transposable element *copia* generates a 5-base pair duplication. *Cell* 21:575-579, 1980
  48. Temin HM: Origin of retroviruses from cellular movable genetic elements. *Cell* 21:599-600, 1980
  49. Hood L, Campbell JH, Elgin SCR: The organization, expression, and evolution of antibody genes and other multigene families. *Annu Rev Genet* 9:305-353, 1975
  50. Nei M: *Molecular Population Genetics and Evolution*. Amsterdam, North-Holland, 1975
  51. Hood JM, Huang HV, Hood L: A computer simulation of evolutionary forces controlling the size of a multigene family. *J Mol Evol* 15:181-196, 1980
  52. Efstratiadis A, Posakony JW, Maniatis T, Lawn RM, O'Connell C, Spritz RA, DeRiel JK, Forget BG, Weissman SM, Slightom JL, Blechl AE, Smithies O, Baralle FE, Shoulders CC, Proudfoot NJ: The structure and evolution of the human  $\beta$ -globin gene family. *Cell* 21:653-668, 1980
  53. Proudfoot NJ, Brownlee GG: 3' non-coding region sequences in eukaryotic messenger RNA. *Nature* 263:211-214, 1976
  54. Konkel DA, Tilghman SM, Leder P: The sequence of the chromosomal mouse  $\beta$ -globin major gene: Homologies in capping, splicing, and poly(A) sites. *Cell* 15:1125-1132, 1978
  55. Baralle FE, Brownlee GG: AUG is the only recognizable signal sequence in the 5' non-coding regions of eukaryotic mRNA. *Nature* 274:84-87, 1978
  56. Hagenbuchle O, Santer M, Steitz JA, Mans RJ: Conservation of the primary structure at the 3' end of 18S rRNA from eukaryotic cells. *Cell* 13:551-563, 1978
  57. Kozak M: How do eukaryotic ribosomes select initiation regions in messenger RNA? *Cell* 15:1109-1123, 1978
  58. Kehry M, Ewald S, Douglas R, Sibley C, Raschke W, Fambrough D, Hood L: The immunoglobulin  $\mu$  chains of membrane-bound and secreted IgM molecules differ in their C-terminal segments. *Cell* 21:393-406, 1980



## Chapter 3

### The Chromosomal Organization of Antibody V<sub>H</sub> Gene Segments

## **The Chromosomal Organization of Antibody V<sub>H</sub> Gene Segments**

Stephen Crews, Elizabeth Springer, and Leroy Hood

### **INTRODUCTION**

Most mammalian genes are members of multigene families, that is, they share structural homology and often a similar function with other genes. The functional significance of genes being multicopy can differ. Large gene families for rRNA genes have evolved to meet cellular requirements for large amounts of their products. The genes in these families are characterized by homogeneity in sequence (Brown et al., 1971). Other multigene families have members that individually perform different functions. Antibody variable (V) region genes are an example of a multigene family of this latter category. In this case different variable region genes encode antibodies with different antigenic specificities. Unlike the rRNA gene families, which are characterized by sequence homogeneity, the antibody variable region genes are characterized by their extensive diversity (Kabat et al., 1979). This diversity has evolved to meet the demand for diversity in the immune response, so that the host can neutralize any foreign substance or microbe.

The homogeneity of gene sequence, characteristic of the rRNA gene families is probably maintained by unequal crossing over and gene correction events (Smith, 1976). The large variable region gene repertoire that exists is due to extensive gene duplication that most reasonably occurred by unequal crossing over. However, genetic events that promote homogeneity such as gene correction events, would reduce the diversity of the variable gene repertoire. It is of interest to see if this fundamental difference between multigene families that are homogeneous in sequence and those that are diverse in sequence is reflected in the organization of these genes on the chromosome.

There may be even greater significance to the arrangement of V genes and V gene families along the chromosome. It is possible that the position of a V gene along

the chromosome may determine how frequently it can be joined to the other gene segments (D and J) and make a functional antibody gene or perhaps even when it can be expressed during development. Alternatively, a V gene may be expressed in a similar fashion regardless of its location within the  $V_H$  locus. There is suggestive evidence that certain  $V_H$  genes may be expressed before others during mouse development (Klinman, 1980). Precursor B cells bearing antibody binding to certain antigens arise at specific times during mouse development. For example, B cells expressing antibodies that bind trinitrophenol appear at birth, whereas B cells responsive to phosphorylcholine appear 5-7 days later. One interpretation of these results is that the V gene segments utilized in these responses are activated to undergo the variable gene formation DNA rearrangement at different times in development. However, there has been no work done at the chromosomal level to confirm this hypothesis and there are alternative explanations.

In order to study the organization of V genes, we have begun to study by molecular cloning the mouse heavy chain variable region gene locus. There is probably a minimum of several hundred  $V_H$  gene segments. Since the task of physically trying to link them to each other would be formidable, we have tried to study the organization of a small family of closely related  $V_H$  gene segments. This set of gene segments, referred to as the T15  $V_H$  gene family, contains four germline gene segments. All four gene segments have previously been isolated and sequenced (Crews et al., 1981; Huang et al., 1981).

In this paper, we show the isolation and analysis of overlapping lambda and cosmid clones containing these  $V_H$  gene segments. We find that two of these four gene segments are closely linked but each of the other two has not been physically linked to any other  $V_H$  gene segment. We conclude that closely related  $V_H$  gene segments can be clustered, but the spacing distance between genes is variable and can be quite large, greater than 30 kb. We have also isolated and mapped a large set of

cosmid clones that contain other  $V_H$  genes and further substantiate the previous conclusion that  $V_H$  gene spacing distance can be variable and the average spacing distance is large, greater than 23 kb. Assuming at least 200  $V_H$  gene segments, the size of the  $V_H$  locus may be greater than 5 million nucleotide pairs.

## MATERIALS AND METHODS

### Preparation of the S107 $V_H$ Gene Probe

The radioactive probe utilized for all of the clone library screens and blot hybridizations was the pS107  $V_H$  cDNA clone previously described by Early et al. (1980). The entire plasmid subclone was used for hybridization with phage clones. Hybridizations with cosmid clones were performed with the purified fragment cut from pBR322 with Pst I. The DNA probe was labeled by nick translation to a specific activity of approximately  $1 \times 10^8$  dpm/ $\mu$ g.

### Isolation of Recombinant Clones

The isolation of clones from the bacteriophage lambda clone libraries was as previously described (Crews et al., 1981). The cosmid library screened was constructed by Michael Steinmetz and consists of BALB/c sperm DNA cloned into the cosmid vector, pT15 (Steinmetz et al., 1982). We screened about 2-300,000 clones containing up to three genome equivalents of mouse DNA. The hybridizations were done with the pS107  $V_H$  cDNA probe at a concentration of 10 ng/ml at 65°C in a hybridization solution consisting of 1M NaCl, 0.05 M Tris-HCl (pH 7.5), 0.1% NaPPi, 0.1% SDS, 10x Denhardt's solution, 50  $\mu$ g/ml poly (rC), 1  $\mu$ g/ml Hae III cut, denatured pBR322. Following hybridization for 16 hours, the filters were washed six times in a high salt wash buffer consisting of 3x SSC, 0.5% SDS and 10x Denhardt's at 65°C for 15 minutes each. This was followed by three washes at 65°C for 15 minutes each with 1x SSC, 0.1% SDS. The filters were exposed by autoradiography for two days with a screen intensifier. Positive colonies were then rescreened in a similar fashion and

mini-cosmid preparations and large scale preparations of cosmid DNA were made (Birnboim and Doly, 1979; Davis et al., 1980).

### **Restriction Mapping and Blot Hybridizations**

Cosmid clones were mapped by analysis of double digests and two-dimensional mapping using low-melting temperature agarose (Parker and Seed, 1980). Hybridizations of the pS107 V<sub>H</sub> cDNA probe and other probes to nitrocellulose blots of various clone DNAs were carried out under conditions similar to those used for screening the cosmid clones.

## **RESULTS AND DISCUSSION**

### **Chromosomal Arrangement of the T15 Family V<sub>H</sub> Gene Segments**

As a model system to study the chromosomal organization of V<sub>H</sub> genes, we chose to study the genes that are closely related to the S107 V<sub>H</sub> gene segment. As previously shown, hybridization under moderately stringent conditions with a cloned S107 V<sub>H</sub> cDNA probe to a blot of mouse sperm DNA cut with Eco RI shows four intense bands upon autoradiography (Cory and Adams, 1980; Crews et al., 1981). The four genes, designated the T15 V<sub>H</sub> gene family, corresponding to these bands have previously been isolated from libraries of BALB/c mouse sperm DNA cloned into a bacteriophage lambda vector (Crews et al., 1981). These clones were mapped with the restriction enzymes Eco RI, Bam HI, and Xba I and the genes sequenced (Crews et al., 1981; S. Crews, unpublished data).

Examination of the restriction maps of these clones in conjunction with heteroduplex data indicated that clones containing the V1 and V3 genes overlapped (Figure 1, a). The two genes were thus physically linked on the mouse chromosome at a distance of 16 kb. The transcriptional orientation was in the same direction with V3 being the 5'-most gene. The restriction maps of the clones containing the V13 gene did not show any obvious overlap with clones containing any of the other three genes.

The same was also true for clones containing the V11 gene. These observations were further confirmed by the lack of hybridization of radioactively labeled restriction fragments at the ends of the clones to Southern blots of the other clones cut with various restriction enzymes.

The observation that the V1 and V3 genes were closely linked on the mouse chromosome provided a good reason to think that all four genes would be closely linked since the V11 and V13 genes are more homologous in nucleotide sequence to V3 than V1 is. We decided to use the pS107 V<sub>H</sub> cDNA probe to screen a cosmid library containing BALB/c mouse sperm DNA. The size of the cosmid clone inserts is usually greater than 33 kb and could significantly expand the amount of cloned DNA flanking the genes that we had previously isolated. We screened 2 to 3 genome's equivalent of cosmid clones and obtained 16 different clones.

These clones were restriction mapped and compared to the other lambda clones. We mapped the cosmid clones using the enzymes Nae I, Nar I, Nru I, Xho I, Sma I, Sal I, Cla I, Kpn I, Hpa I, and Eco RI. The first seven enzymes cut infrequently in mouse DNA due to the presence in their restriction enzyme recognition sequence of the dinucleotide 5'-CpG-3', which is rare in mammalian DNA (Swartz et al., 1962). The enzymes, Kpn I and Hpa I, even though they do not contain the 5'-CpG-3' dinucleotide in their restriction enzyme recognition sequence, cut relatively infrequently in mouse DNA, thus making them useful. Positioning of the sites cut by these enzymes easily allows the mapping of frequently cutting enzymes such as Eco RI. Analysis of the restriction maps and hybridization studies indicate that we obtained one cosmid clone containing each gene of the T15 V<sub>H</sub> gene family and 13 other cosmid clones containing genes less related to the pS107 V<sub>H</sub> cDNA probe.

Two clones that we have analyzed, the cosmid cV19 and the lambda clone,  $\lambda$ V11, appear to contain the V1 and V11 genes, respectively, by comparison of the restriction sites surrounding the genes and by hybridization experiments. However,

they diverge from the other overlapping clones for these genes at one end of the clone. We believe that these clones contain these two genes but are the result of a cloning artifact whereby during the library construction there was a tandem ligation of DNA not contiguous in the mouse genome. The  $V_H$  genes on these clones are being sequenced and other experiments are being performed to confirm this interpretation. cV19 has not been utilized in the analysis described here and only the homologous part of  $\lambda$ V11 is shown in Figure 1c.

As can be seen in Figure 1, the cosmid clones greatly extended the amount of cloned DNA surrounding the T15 family  $V_H$  genes. Nevertheless, comparison of the restriction maps of the clones indicated that the clones containing the V11 gene segment were not physically linked to the clones containing the other  $V_H$  gene segments, and the same was true for the clones containing the V13 gene segment. Presently, the T15  $V_H$  gene family has been divided into three gene clusters by molecular cloning. The first contains the V1 and V3 genes and encompasses 45 kb of DNA. The second cluster contains the V13 gene and has 45 kb of DNA, while the V11 gene constitutes the third cluster and contains 33 kb of DNA.

An important question is whether there are other, weakly related  $V_H$  genes contained on the clones just described that would not be detected with the pS107  $V_H$  probe under our conditions of hybridization. The experiment shown in Figure 2 indicates that this is unlikely. The figure shows an autoradiogram of a Southern blot of Eco RI digested DNA of all of the cosmid clones and several lambda clones that were hybridized to the pS107  $V_H$  probe. Several of the  $V_H$  genes residing on these clones have been sequenced and thus can be directly compared to the nucleotide sequence of the pS107  $V_H$  cDNA probe. Two of the clones that are included are M2 and M49 which have  $V_H$  gene segments 59% and 56% homologous to the pS107  $V_H$  cDNA probe, respectively (Early et al., 1982). The autoradiogram clearly shows that the probe hybridizes well with both genes. Analysis of published  $V_H$  protein and

nucleotide sequences indicates that these two genes are as non-homologous to the S107  $V_H$  probe as any known  $V_H$  segment is to it. Additionally, a MOPC 21  $V_H$  cDNA probe that is only 71% homologous to the S107  $V_H$  probe hybridizes to the same set of restriction fragments; no other fragments hybridize with the MOPC 21  $V_H$  probe (Bothwell et al., 1981; E. Kraig, personal communication). We conclude that we can detect all  $V_H$  genes on our clones using any  $V_H$  probe and that there are no other  $V_H$  genes residing on the clones containing gene segments of the T15  $V_H$  gene family.

There are several conclusions that can be made about the chromosomal organization of the T15  $V_H$  gene family. First, we have provided evidence that closely related  $V_H$  genes can be clustered since V1 and V3 have been physically linked at a distance of 16 kb apart. Second, the V11 and V13 genes are further than this from other  $V_H$  genes; thus the spacing between  $V_H$  genes would appear to be variable. The third point is that the average spacing distance is very large. We have cloned 120 kb of DNA encompassing the gene segments of the T15  $V_H$  gene family and there are only four genes. Thus, the average spacing distance between the genes in this family is greater than 20 kb.

#### **Other $V_H$ Gene Segments Show a Similar Organization as the T15 Family $V_H$ Gene Segments**

In order to see if the organization of the T15 family of  $V_H$  gene segments was typical of the  $V_H$  locus, we have mapped another 13 cosmid clones that contain  $V_H$  genes. The restriction maps of these clones are shown in Figure 3. Analysis of the restriction enzyme cleavage patterns for each clone shows that the 13 clones can be grouped into nine nonoverlapping clusters of DNA, although small 1-2 kb overlaps between some of these clones might not be detected. The number of cosmids obtained for a given gene segment ranges from one to four. Hybridization of the pS107  $V_H$  cDNA probe to Southern blots of Eco RI-cut cosmid DNAs indicates that there are 17 different gene segments residing in these nine clusters of clones. If the



clones are cut with Bam HI or Xba I, blotted, and hybridized with the same probe, generally the same number of fragments hybridize to the probe. The homology of the gene segments in this collection of cosmids to the pS107  $V_H$  cDNA sequence is unknown and awaits sequence analysis. However, with clone blots, all of these gene segments clearly hybridize less efficiently to the probe than members of the T15  $V_H$  gene family; thus they are less than 85% homologous to the pS107  $V_H$  sequence.

Several general observations can be made about the gene organization revealed by these clones. The spacing of  $V_H$  genes is variable; some are less than 5 kb apart whereas others are probably greater than 30 kb apart. As was seen with the T15  $V_H$  gene family, the average spacing distance is quite large, greater than 20 kb.

### **Overview of $V_H$ Gene Organization**

Several features of  $V_H$  gene organization have been revealed in this work. The first is that closely related  $V_H$  gene segments can be clustered. We have shown that two members of the T15  $V_H$  gene family, V1 and V3, are physically linked on the chromosome at a distance of 16 kb. The clustering of related  $V_H$  gene segments has also been seen with other  $V_H$  gene segments. We have also sequenced two gene segments, one of which is the germline gene segment utilized in the immune response to 2,1 levan (S. Crews, unpublished observations). These two gene segments are 90% homologous in their nucleotide sequence and have been physically linked 15 kb from each other. Others have shown the occurrence of clustered gene segments that are closely related in gene families that encode the immune response to NP and two other genes of the levan  $V_H$  gene family (Bothwell et al., 1981; Kemp et al., 1982). However, the generality of this observation remains uncertain as there are no reports of large series of  $V_H$  gene segments contiguously linked. For example, we have obtained clones for the V11 and V13 gene segments of the T15  $V_H$  gene family that comprise 33 and 45 kb of DNA, respectively, yet have been unable to physically link these gene segments to any other. Chromosomal walking experiments utilizing probes

at the ends of each gene cluster to screen for new overlapping clones are in progress to further understand the chromosomal organization of this gene family. In summary, present evidence indicates that closely related  $V_H$  gene segments tend to be clustered together within the  $V_H$  locus but only a small portion of the entire locus has been investigated and the generality of this concept remains to be seen.

The second major point brought out in this work is that the spacing between  $V_H$  gene segments can be variable. Although true for the T15  $V_H$  family of genes, this point is best illustrated with the  $V_H$  gene segments present on the nine other clusters of  $V_H$  gene segments described in this paper. There are examples where there is only one  $V_H$  gene segment in a cosmid cluster (1,3,5,6,7, and 8). In many of these cases, there is at least 30 kb of DNA flanking the  $V_H$  gene segment without another  $V_H$  gene segment present. Although the existence of any of these clusters within the  $V_H$  locus is unknown, it is reasonable to assume that most of them reside there. There are other examples where  $V_H$  gene segments are quite close together. For instance, there are four gene segments contained in cluster 11 which is 33 kb in size. Three of the  $V_H$  gene segments are within 15 kb. There are also two  $V_H$  gene segments within 5 kb in cluster 9. Additionally, we have isolated and mapped two different lambda clones, with inserts less than 18 kb, that contain three  $V_H$  gene segments each (S. Crews, unpublished observations). As can be seen in clusters 2, 4, 9, 10, and 11, intermediate distances, such as those around 15 kb are also observed. Although the work presented here indicates considerable variability in the spacing distance of  $V_H$  gene segments, it will be interesting to see the overall chromosomal location of these clusters. It is possible that the average spacing distance between  $V_H$  gene segments may vary throughout the  $V_H$  locus. Such an organization has been observed for the murine heavy chain constant region gene locus where generally, the spacing distance between adjacent genes becomes smaller as the locus proceeds in a 5' to 3' direction (Shimizu et al., 1982).

The most striking observation to emerge from this work is that the average spacing distance between  $V_H$  gene segments is very large. Figure 4 summarizes the data accumulated from our study of 12  $V_H$  gene clusters. These clusters contain 21 different  $V_H$  gene segments and 492 kb of nonoverlapping DNA. This yields a value of 23 kb of DNA between  $V_H$  gene segments on the average. Since we are analyzing only a subset of  $V_H$  gene segments and have not completely linked together these gene segments, the estimate should be considered preliminary. There are two questions that should be addressed. The first is whether the clones we have isolated and analyzed are biased towards their  $V_H$  gene segment spacing and the second is whether the subset of gene segments that we have analyzed is representative of all  $V_H$  gene segments. The first question is difficult to answer. However we have isolated clones that contain from one to four  $V_H$  gene segments so it would seem that there is no obvious bias against certain clones with regard to how many gene segments they contain. The second question is also difficult to answer, but analyses of a number of other  $V_H$  gene segment families have generally not shown a tight clustering of  $V_H$  gene segments (Bothwell et al., 1981; Kemp et al., 1982; Givol et al., 1981). The genes have either been unlinked to other genes or else linked at distances of at least 15 kb. Clearly, the generality of these observations depends on further structural analysis of the  $V_H$  gene locus, but these data have so far provided the most extensive look at the organization of  $V_H$  gene segments.

There are probably a minimum of 200 germline  $V_H$  gene segments (R. Riblet, personal communication). Utilizing our value of 23 kb of DNA separating  $V_H$  gene segments, we estimate that the size of the  $V_H$  gene locus could be almost 5 million nucleotide pairs in length. It has been estimated that in mice 2 million nucleotide pairs is equivalent to one centimorgan of DNA in genetic map units (V. McKusick, personal communication). Thus, the size of the  $V_H$  locus may be at least 2.5 centimorgans. Interestingly, genetic mapping of idiotypes (which generally correlates

with the germline  $V_H$  gene segment) indicate that the 5'-most  $V_H$  gene segment is 9.8 centimorgans from the constant region gene locus and the 3'-most  $V_H$  gene segment is 0.5-1 centimorgan away from the constant region gene locus (Weigert and Riblet, 1978). Thus, the very large size of the  $V_H$  gene locus that we have calculated is not inconsistent with the large size of the  $V_H$  locus seen genetically.

The view that we have obtained of the  $V_H$  gene locus is that it is very large with the  $V_H$  gene segments scattered throughout it at intervals of great distance. The significance of the large distance between  $V_H$  gene segments is unclear and the nature of the spacer DNA is also unknown. However, this organization is different from that seen in the rRNA multigene families which are characterized by short spacers of similar length (Federoff, 1979). It is thought that the homogeneity in sequence characteristic of the rRNA gene families is maintained by unequal crossing over and gene correction events. Antibody V gene segments are characterized by diversity in their sequence and the occurrence of gene correction events could be detrimental towards the maintenance of the germline V gene repertoire. It is possible that the large distance between  $V_H$  gene segments may tend to reduce the frequency of gene correction events. Further structural analysis may give more insight into the significance of V gene arrangement. Unfortunately, various hypotheses will be difficult to test experimentally.

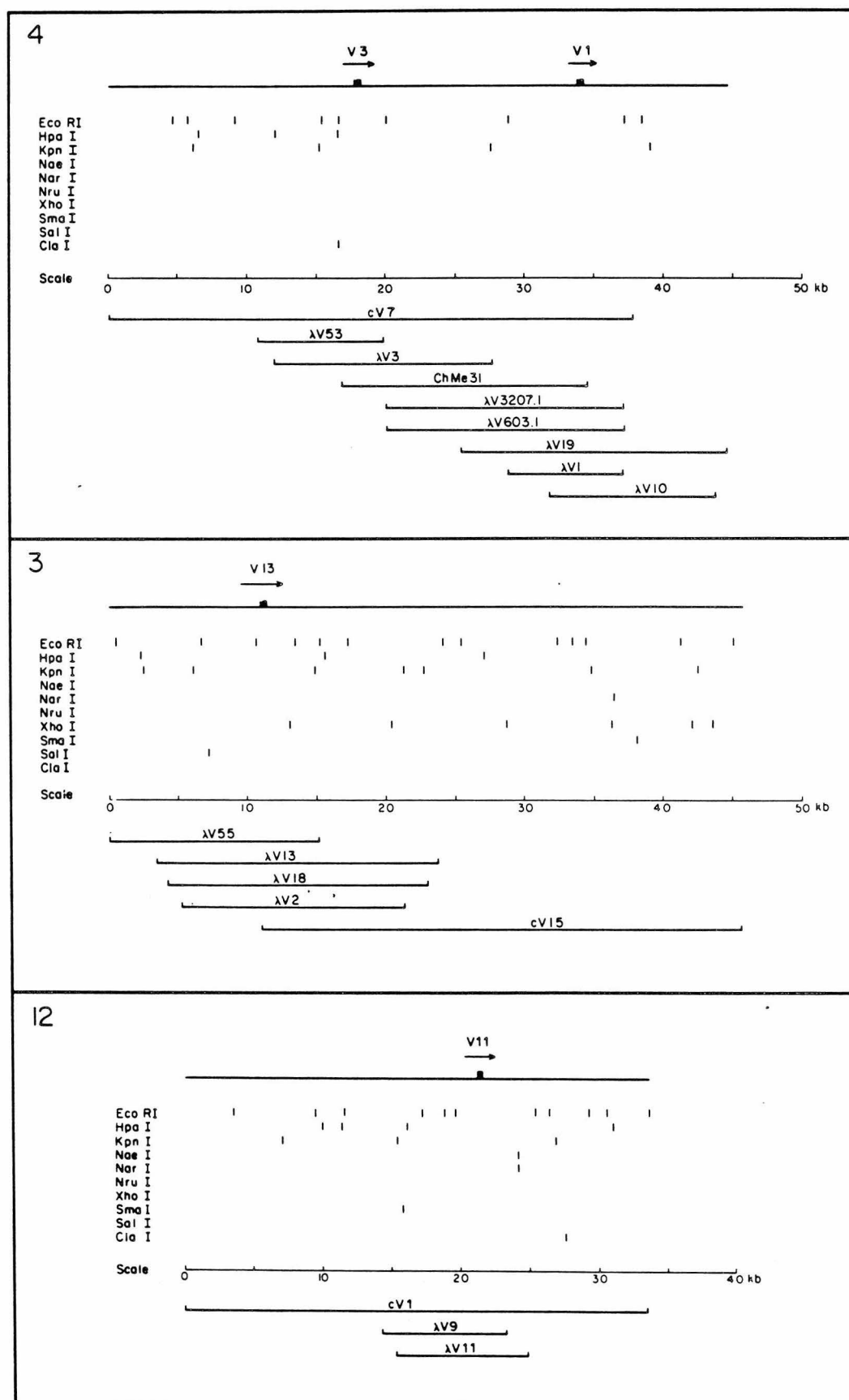
## REFERENCES

- Backman, K. (1980) A cautionary note on the use of certain restriction endonucleases with methylated substrates. *Gene* **11**, 169-171.
- Birnboim, H. C. and Doly, J. (1979) A rapid alkaline extraction procedure for screening recombinant plasmid DNA. *Nucl. Acids Res.* **7**, 1513-1523.
- Bothwell, A. L. M., Paskind, M., Reth, M., Imanishi-Kari, T., Rajewsky, K. and Baltimore, D. (1981) Heavy chain variable region contribution to NP<sup>b</sup> family of antibodies: somatic mutation evident in a  $\gamma$ 2a variable region. *Cell* **24**, 625-637.
- Brown, D. D., Wensink, P. C. and Jordan, E. (1971) A comparison of the rDNAs of *Xenopus laevis* and *Xenopus mulleri*: the evolution of tandem genes. *J. Mol. Biol.* **63**, 57-73.
- Cory, S. and Adams, J. M. (1980) Deletions are associated with somatic rearrangement of immunoglobulin heavy chain genes. *Cell* **19**, 37-51.
- Crews, S., Griffin, J., Huang, H., Calame, K. and Hood, L. (1981) A single V<sub>H</sub> gene segment encodes the immune response to phosphorylcholine: somatic mutation is correlated with the class of the antibody. *Cell* **25**, 59-66.
- Davis, R. W., Botstein, D. and Roth, J. R. (1980) *Advanced Bacterial Genetics*. (Cold Spring Harbor, New York: Cold Spring Harbor Laboratory).
- Early, P., Huang, H., Davis, M., Calame, K. and Hood, L. (1980) An immunoglobulin heavy chain variable region is generated from three segments of DNA: V<sub>H</sub>, D, and J<sub>H</sub>. *Cell* **19**, 981-992.
- Early, P., Nottenburg, C., Weissman, I. and Hood, L. (1982) Immunoglobulin gene rearrangements in normal mouse B cells. *Mol. Cell. Biol.*, in press.
- Federoff, N. V. (1979) On spacers. *Cell* **16**, 697-710.
- Huang, H., Crews, S. and Hood, L. (1981) An immunoglobulin V<sub>H</sub> pseudogene. *J. Mol. Appl. Genet.* **1**, 93-101.

- Givol, D., Zakut, R., Effron, K., Rechavi, G., Ram, D. and Cohen, J. B. (1981) Diversity of germline immunoglobulin  $V_H$  genes. *Nature* **292**, 426-430.
- Kabat, E. A., Wu, T. T. and Bilofsky, H. (1979) Sequences of immunoglobulin chains. U.S. Dept. Health, Education, and Welfare, NIH Pub. No. 80-2008.
- Kemp, D. J., Cory, S. and Adams, J. M. (1979) Cloned pairs of variable region genes for immunoglobulin heavy chains isolated from a clone library of the entire mouse genome. *Proc. Natl. Acad. Sci. USA* **76**, 4627-4631.
- Kemp, D. J., Tyler, B., Bernard, O., Gough, N., Gerondakis, S., Adams, J. M. and Cory, S. (1982) Organization of genes and spacers within the mouse immunoglobulin  $V_H$  locus. *J. Mol. Appl. Genet.* **1**, 245-261.
- Klinman, N. (1980) B-cell maturation and repertoire expression. In: *Immunoglobulin Genes and B Cell Differentiation*. Battisto/Knight (Eds.) (New York: Elsevier North Holland, Inc.) pp. 193-205.
- Parker, R. C. and Seed, B. (1980) Two-dimensional agarose gel electrophoresis "SeaPlaque" agarose dimension. In: *Methods in Enzymology*, Vol. 65 (New York: Academic Press) pp. 358-371.
- Shimizu, A., Takahashi, N., Yaoita, Y. and Honjo, T. (1982) Organization of the constant-region gene family of the mouse immunoglobulin heavy chain. *Cell* **28**, 499-506.
- Smith, G. P. (1976) Evolution of repeated DNA sequences by unequal crossovers. *Science* **191**, 528-535.
- Steinmetz, M., Winoto, A., Minard, K. and Hood, L. (1982) Clusters of genes encoding mouse transplantation antigens. *Cell* **28**, 489-498.
- Swartz, M. N., Trautner, T. A. and Kornberg, A. (1962) Enzymatic synthesis of deoxyribonucleic acid. XI. Further studies on nearest neighbor base sequences in deoxyribonucleic acids. *J. Biol. Chem.* **237**, 1961-1967.
- Weigert, M. and Riblet, R. (1978) The genetic control of antibody variable regions in the mouse. *Springer Seminars in Immunopathol.* **1**, 133-169.

**Figure 1.** The restriction enzyme maps for the three gene clusters containing the four T15 family  $V_H$  gene segments are shown: (a) V1-V3 cluster, (b) V13 cluster, and (c) V11 cluster. In the upper left hand corner of each cluster is a number that identifies the cluster in Figure 4. The upper line represents the chromosomal DNA arrangement as defined by the overlapping clones. The filled-in box on the line represents a  $V_H$  gene segment, the arrow indicates the 5' to 3' transcriptional orientation of the gene segment, and the name of the gene segment is written above the arrow. The restriction map of the cluster is shown with the scale below it. The bacteria harboring all of these recombinant clones were  $dam^+$  and therefore *Cla* I may not cut at all of the ATCGAT sites within these recombinant clones (Backman, 1980; M. Garfinkel, personal communication). While the *Cla* I restriction map reflects the unmethylated sites of these clones, it may not completely reflect the map of the mouse genomic DNA. The names and position of each overlapping clone shown are at the bottom.

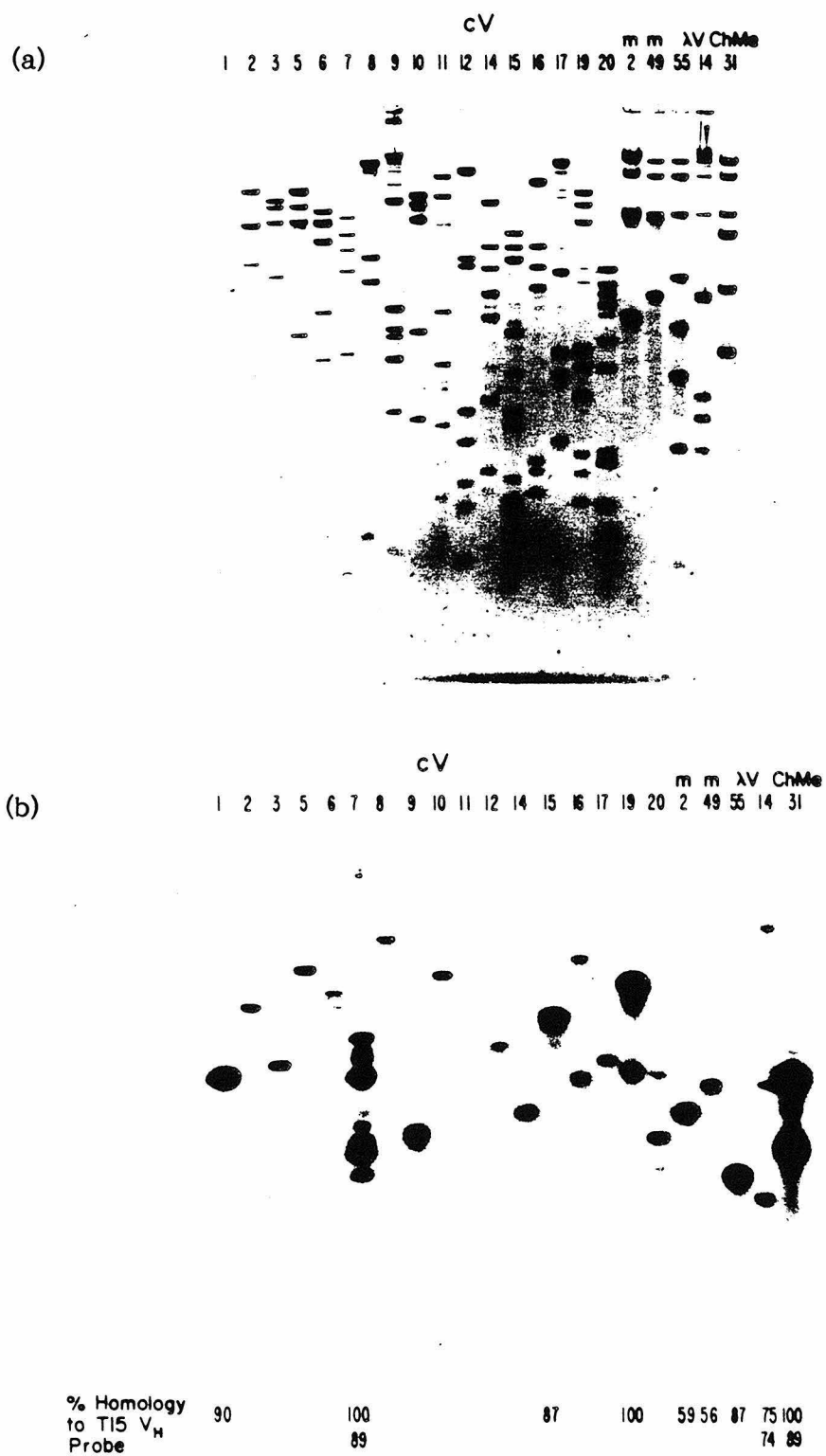
Fig. 1





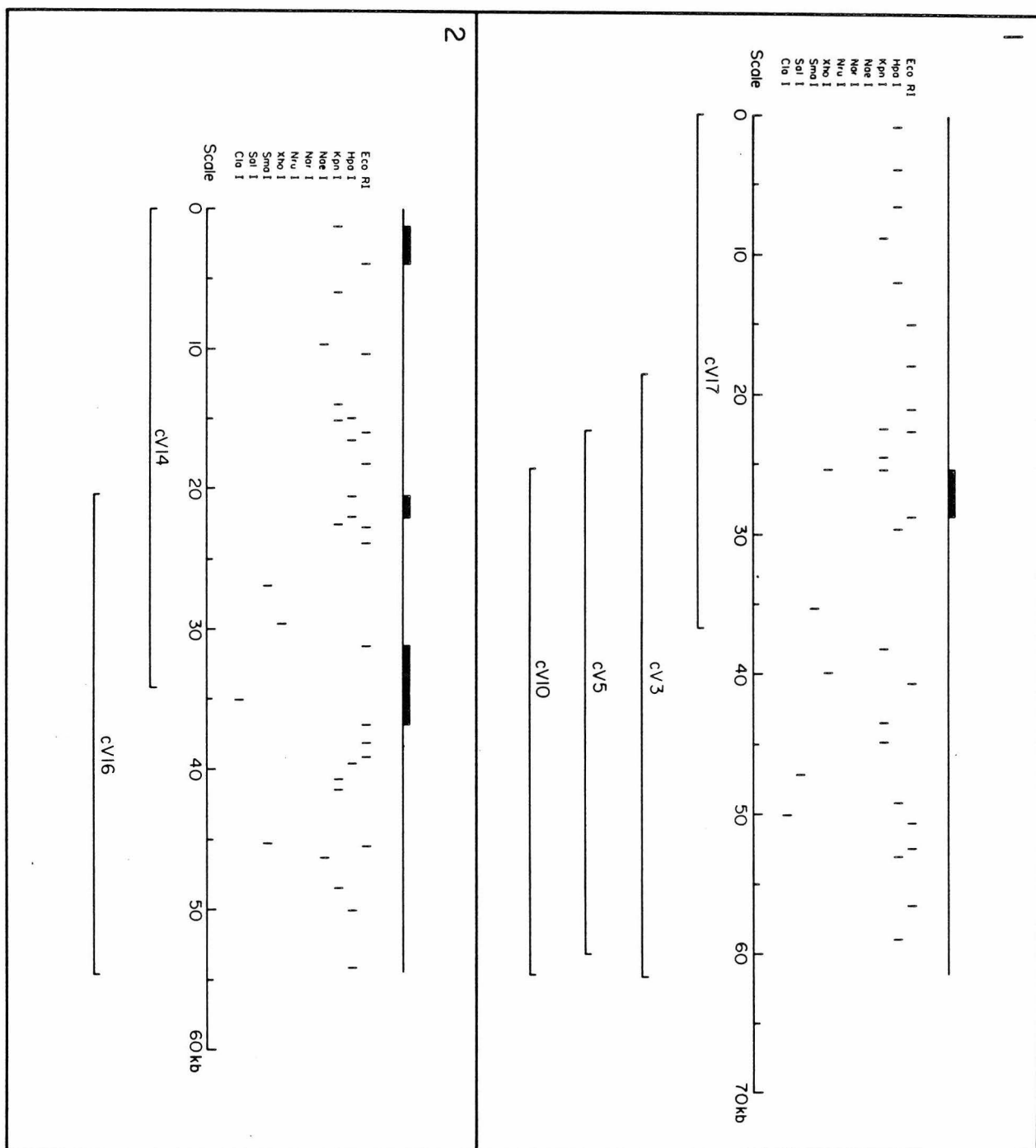
**Figure 2.** (a) Photograph of ethidium bromide stained gel of cloned DNAs cut with the restriction enzyme Eco RI. cV1-3, 5-12, 14-17, and 19-20 are the cosmid clones described in this paper. M2, M49, and  $\lambda$ V14 are lambda clones containing sequenced  $V_H$  gene segments from gene families other than the T15 family (Early et al., 1982; S. Crews, unpublished data)  $\lambda$ V55 is a lambda clone containing the V13 gene segment and ChMe31 is a clone containing the V1 and V3 gene segments (Kemp et al., 1979). (b) Autoradiogram of a Southern blot of the gel shown in "(a)" hybridized to the pS107  $V_H$  probe. cV7 contains two genes, V1 and V3, each lying on one Eco RI fragment; the extra bands of hybridization seen are due to additional restriction fragments derived from a deletion product of the intact clone. Below the autoradiogram, the % homology to the pS107  $V_H$  probe is shown for those genes that have been sequenced.

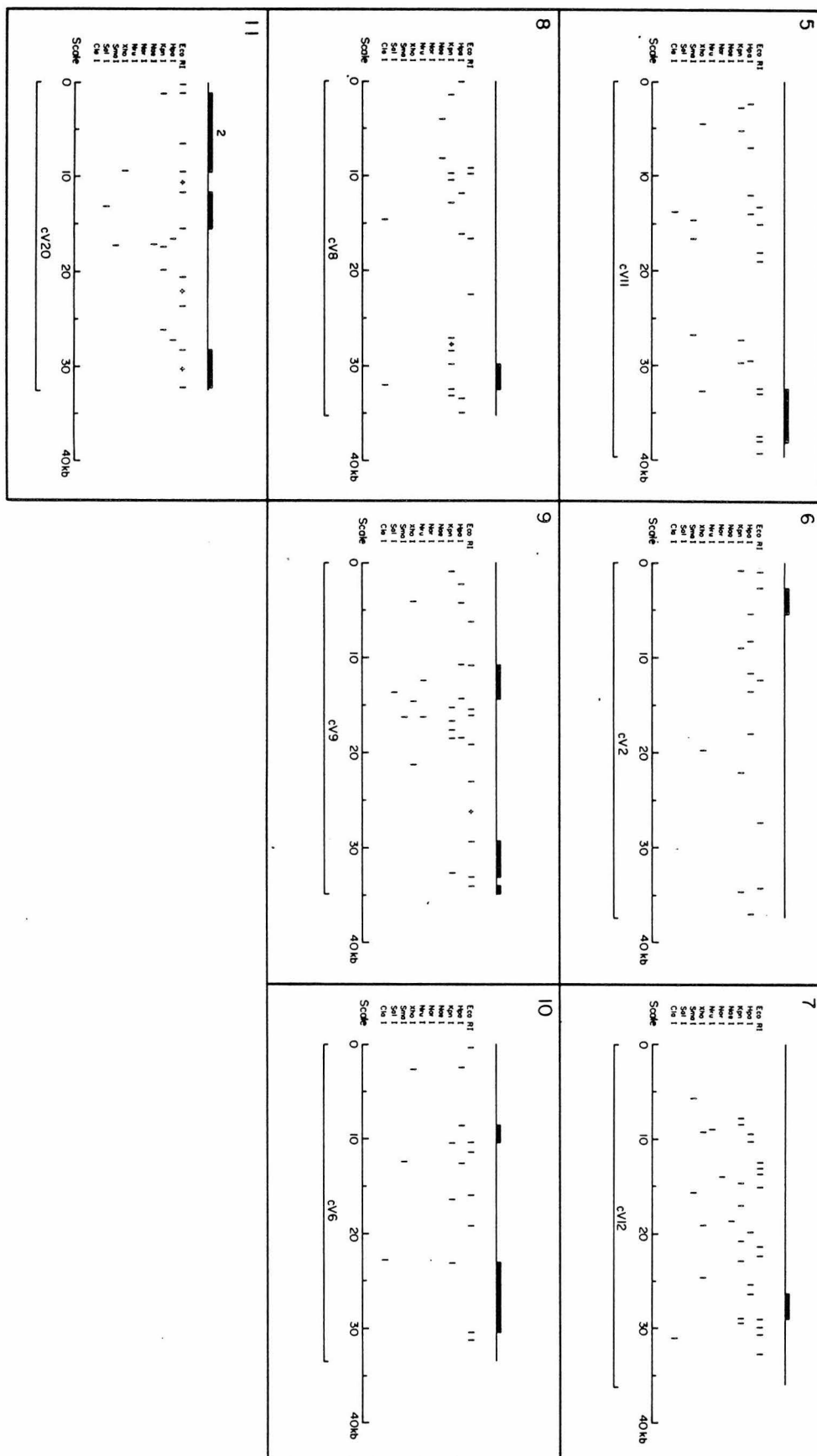
54



**Figure 3.** The restriction enzyme maps for the nine gene clusters that contain gene segments that are not members of the T15  $V_H$  gene family are shown. The 5' to 3' direction of transcription is unknown and the left to right orientation is arbitrary. The description of this Figure is the same as Figure 1. A cross (+) in the restriction map (e.g., see cluster 8-Kpn map) indicates that there is a restriction enzyme cutting site somewhere between the two flanking sites but its exact position is unknown. The "2" above the  $V_H$  gene segment in cluster 11 indicates that each of the two Eco RI fragments shown below the filled-in box contains a  $V_H$  gene segment.













Fig. 3





**Figure 4.** Summary description of the clusters of  $V_H$  genes. Each cluster is identified with respect to Figures 1 and 3. The organization shown in column 2 is drawn to scale with each cluster centered at the 50 kb position mark on the scale. The length of each cluster is the total amount of nonoverlapping DNA. The last column is the number of overlapping cosmid clones per cluster. Clusters 3, 4, and 12 also have a number of overlapping lambda clones. Below the columns the data are summed. At the bottom is shown a calculation of the average amount of non-overlapping DNA per  $V_H$  gene. This is obtained by dividing column 3 by column 2.

Clusters of  $V_H$  Genes Homologous to the T15  $V_H$  Gene

Cluster	Organization	Length kb	Overlapping Cosmid Clones
	0 50 100kb		
1		62	4
2		55	2
3 (V13)		46	1
4 (V1, V3)		45	1
5		40	1
6		38	1
7		36	1
8		35	1
9		35	1
10		34	1
11		33	1
12 (V11)		33	1
21 $V_H$ genes		492	16

$$\text{DNA} / V_H \text{ Gene} = \frac{492}{21} = 23 \text{ kb}$$

CONCLUSION

PART I



## CONCLUSION

The first three chapters of this thesis described experiments that addressed the question of the genetic basis of antibody diversity. We were able to conclusively show that somatic mutation of antibody  $V_H$  genes plays a significant role in the generation of antibody diversity and began also to understand the organization of  $V_H$  gene segments along the chromosome. In this section, I would like to summarize our current understanding of somatic mutation, V gene organization, and the future outlook for these subjects.

In Chapter 1, we showed that the immune response to phosphorylcholine in BALB/c mice is encoded by one germline gene segment. Previously, it had been demonstrated by Hood and his colleagues that there were a large number of antibodies that possessed  $V_H$  segment protein sequences that were variants of this germline gene segment (Gearhart et al., 1981). It was also clear that these variants were not the result of polymorphism within the BALB/c mouse population and it seems unlikely, although cannot be disproved, that they are not the result of a tumor artifact (Gearhart et al., 1981). Thus, we concluded that the variants were derived by somatic diversification. Other scientists investigating the mechanisms of diversity utilized by antibodies involved in other immune responses have come to similar conclusions, namely that somatic diversification plays a prominent role in the generation of antibody diversity (Bernard et al., 1978; Bothwell et al., 1981; Pech et al., 1981; Valbuena et al., 1978). Below, I would like to briefly list some of the features of somatic diversification.

The mechanism of somatic diversification appears to be mutational and not recombinational. This conclusion was derived from the data presented in Chapter 1. The  $V_H$  segment variants of antibodies that bind phosphorylcholine contained substitutions that could not be explained by recombination between the V1 gene segment and the V3, V11, and V13 gene segments. These were the most homologous

in sequence to the V1 gene segment and thus most likely to undergo this recombination event. It seems reasonable that somatic recombination between nonallelic V genes may occur; however, mutation probably accounts for most somatic diversification.

It is clear that somatic mutation can be extensive. Although many  $V_H$  segment variants have only one or a small number of substitutions, others have undergone many mutations. We may be seeing only those antibodies that have undergone relatively few mutations, since many mutations in the V gene segment will result in a nonfunctional antibody and escape our analysis that only selects functional antibodies. The most extreme example is the phosphorylcholine-binding antibody from the myeloma, M167. Its  $V_H$  segment has undergone eight amino acid substitutions. Recent analysis of the DNA sequence of the expressed gene has shown that there are a total of 44 nucleotide substitutions that are scattered throughout the gene and its immediate flanking sequences (Kim et al., 1981).

Analysis of the types of mutation that have been seen in the V region variants does not reveal any striking feature. There are no specific sites or sequences that mutate exclusively and there is no strong tendency for any base (A, C, G, or T) to mutate to any other. Mutations can occur in both the  $V_H$  gene segment and the  $J_H$  gene segment. It is unclear if somatic mutation occurs in the D segment since we do not know if all of the germline D gene segments have been isolated and sequenced. Within the  $V_H$  gene segment, substitutions have been seen in all three framework and both hypervariable regions although there is a clustering of mutations in the second hypervariable region. In Chapter 1, we discussed several models for how certain somatic variants might be selected from the population of antibody-producing cells during the immune response. These include the selection of antibody-producing cells based on their possessing an antibody with a relatively high affinity for antigen, or regulation by lymphocytes through idiotype- anti-idiotype networks.

The most important conclusion that was observed regarding somatic mutation of  $V_H$  genes was that it was correlated with the class of the antibody. Only IgG and IgA antibodies undergo somatic mutation; IgM antibodies do not. Although this observation is consistent with data from antibodies of other immune responses, the most extensive data are derived from those that bind phosphorylcholine. Analysis of the complete  $V_H$  segment of five IgG antibodies indicates that all have undergone somatic mutation. Four out of nine IgA antibodies had undergone somatic mutation whereas five had not. Most interestingly, none of the five complete IgM segments had undergone somatic mutation. More extensive data derived from N-terminal sequences support these results (Hood et al., unpublished data). There have been 29  $V_H$  segments from IgG and IgA antibodies sequenced over their first 40 amino acid residues and 12 had undergone somatic mutation. Analysis of the N-terminal sequences of 13 IgM  $V_H$  segments indicated that none had undergone somatic mutation.

Protein sequence analysis of the light chains of phosphorylcholine-binding antibodies combined with molecular cloning studies indicates that the light chain genes undergo somatic mutation also (Selsing and Storb, 1981; Gershenfeld et al., 1981). Similar to  $V_H$  genes the occurrence of somatic mutation is correlated with the class of the antibody. Only the light chains from IgG and IgA antibodies undergo somatic mutation; those from IgM antibodies do not.

The existence of somatic mutation of antibody variable region genes has been conclusively shown in response to phosphorylcholine and other antigens. Somatic mutation also appears to contribute significantly to antibody diversity. The humoral immune response can be divided into two stages. The primary response occurs when the organism encounters a foreign substance and is characterized by the production of IgM antibodies. These antibodies are encoded by unmutated, germline antibody genes. Since the immune system can respond to a large variety of substances, this represents

an enormous germline and combinatorial repertoire. The secondary immune response occurs when the immune system encounters antigen at times subsequent to the first encounter. These antibodies are generally encoded by IgG antibodies and are often the product of somatic mutation. The fact that somatic mutation seems to occur or be expressed only after the organism has been exposed to the foreign substance implies that it does not increase the available repertoire capable of responding to a substance but may act to fine-tune the response.

The above view of the role of somatic mutation in the immune response is speculative. There are a number of uncertainties that remain to be investigated. It is important to know when in B cell development somatic mutation occurs and also in which cells. For instance, it is possible that IgM producing cells undergo somatic mutation but there is selection at some level that removes the variants from the B cell population that we assay. Another issue that is important to understand is the mechanism of somatic mutation. Is there a special V gene specific mutational mechanism or are the V gene somatic variants that have been observed the result of the same type of somatic mutation that affects all genes? It should be pointed out that there exists little information regarding "ordinary" somatic mutation that presumably affects all mammalian genes. Before we ascribe certain features of V gene somatic mutation as being unique to V genes, more should be known about normal somatic mutation.

The first step that must be taken to answer these questions is the isolation of pure B cells at various stages of differentiation, both before and throughout the course of a specific immune response. If stage-specific B cells can be purified or cloned, then questions regarding somatic mutation and other molecular events can be approached. It obviously is desirable to work with normal, untransformed cells if possible. Ideally, B cells could be isolated and stimulated to undergo somatic mutation. Utilizing recombinant DNA technology and gene transfer techniques, as

well as classical techniques in protein chemistry, we could begin to understand the character and enzymology of the mutational process.

We are obtaining a clear view of the structure of the antibody gene locus. The utilization of molecular cloning and DNA sequencing allows us to isolate and study at the nucleotide level entire antibody gene families. Already, the constant region gene loci have been mapped and sequenced and their organization revealed (Shimizu et al., 1982). The structure of variable region genes has been slower both because there are a large number of gene segments and, as shown in Chapter 3 of this thesis, the size of the locus is immense, as reflected in the large spacer distance between variable region gene segments.

Better understanding of the structure and function of the variable region gene locus can come from continued efforts to clone and examine the arrangement of variable region genes. Evolutionary comparisons may be useful. Unfortunately, since we are interested in the large-scale chromosomal organization of the antibody gene locus, genetic and molecular manipulations by the genetic engineer will be difficult to do and questions such as the effect of variable gene position on the expression of that gene as antibody will be very difficult to answer. Naturally, the effect of the gene arrangement on the evolution of the variable gene locus cannot be directly tested but can only be a subject of speculation. It is to be hoped that structural work combined with a molecular analysis of the development of variable gene expression will shed some light on the effect of variable region gene organization on the function of the gene.

## REFERENCES

- Bernard, O., Hozumi, N. and Tonegawa, S. (1978) Sequences of mouse immunoglobulin light chain genes before and after somatic changes. *Cell* **15**, 1133-1144.
- Bothwell, A. L. M., Paskind, M., Reth, M., Imanishi-Kari, T., Rajewsky, K., and Baltimore, D. (1981) Heavy chain variable region contribution to NP<sup>b</sup> family of antibodies: somatic mutation evident in a  $\gamma 2a$  variable region. *Cell* **24**, 625-637.
- Gearhart, P. J., Johnson, N. D., Douglas, R. and Hood, L. (1981) IgG antibodies to phosphorylcholine exhibit more diversity than their IgM counterparts. *Nature* **291**, 29-34.
- Gershenfeld, H. K., Tsukamoto, A., Weissman, I. and Joho, R. (1981) Somatic diversification is required to generate the  $V_{\kappa}$  genes of MOPC 511 and MOPC 167 myeloma proteins. *Proc. Natl. Acad. Sci. USA* **78**, 7674-7678.
- Kim, S., Davis, M., Sinn, E., Patten, P. and Hood, L. (1981) Antibody diversity: somatic hypermutation of rearranged  $V_H$  genes. *Cell* **27**, 573-581.
- Pech, M., H6chtel, J., Schnell, H. and Zachau, H. G. (1981) Differences between germline and rearranged immunoglobulin  $V_{\kappa}$  coding sequences suggest a localized mutation mechanism. *Nature* **291**, 668-670.
- Selsing, E. and Storb, U. (1981) Somatic mutation of immunoglobulin light-chain variable-region genes. *Cell* **25**, 47-58.
- Shimizu, A., Takahashi, N., Yaoita, Y. and Honjo, T. (1982) Organization of the constant-region gene family of the mouse immunoglobulin heavy chain. *Cell* **28**, 499-506.
- Valbuena, O., Marcu, K. B., Weigert, M. and Perry, R. P. (1978) Multiplicity of germline genes specifying a group of related mouse  $\kappa$  chains with implications for the generation of immunoglobulin diversity. *Nature* **276**, 780-784.

INTRODUCTION

PART II

## INTRODUCTION

Mitochondria are the subcellular site for a number of metabolic processes in eukaryotic cells. They contain the biosynthetic machinery for the isoleucine-valine pathways, the tricarboxylic acid cycle and the nitrogen degradative pathway. However, another important function is the production of ATP by the mechanisms of electron transport and oxidative phosphorylation. The ATP synthesizing components, which include the cytochrome  $bc_1$  complex, cytochrome c, cytochrome oxidase and the F1 ATPase, are found within the mitochondrial inner membrane. The biogenesis of these proteins is particularly interesting; they are the products of two genomes: the nuclear and the mitochondrial. The most interesting feature of mitochondria is that they contain their own DNA (Nass and Nass, 1963; Luck and Reich, 1964) genome and have their own transcriptional and translational apparatus. The genome itself is small and the bulk of the mitochondrial proteins are encoded by nuclear genes and synthesized on cytoplasmic ribosomes.

Mammalian mitochondrial ribosomes synthesize up to 25 different proteins. The function and characterization of these proteins is incomplete but there is good evidence that three subunits of cytochrome oxidase, one subunit of the cytochrome  $bc_1$  complex, and one subunit of the F1 ATPase are encoded by the mitochondrial genome and synthesized on mammalian ribosomes (Attardi et al., 1981). By analogy with studies on mitochondria of lower eukaryotes, the mammalian mitochondrial genome may encode one of the proteins of mitochondrial ribosomes, one or two additional components of the F1 ATPase and perhaps proteins involved in the processing of mitochondrial RNA (Tzagoloff et al., 1979). Certainly, one of the foremost challenges in mitochondrial biology is a characterization and functional understanding of all of the mitochondrial gene products.

The human mitochondrial genome is a closed circular DNA molecule about 16,500 nucleotide pairs in length. It contains the genes for the 12S and 16S RNA



species found in mitochondrial ribosomes and all of the genes for the tRNAs utilized by the mitochondrial protein synthesizing apparatus. There are 18 relatively stable RNA species containing poly(A) that are transcribed from the mitochondrial genome (Amalric et al., 1978). Many of these are mRNA coding for the mitochondrially translated polypeptides previously mentioned. It is known that both strands of the mitochondrial DNA are completely transcribed and the bulk of the L-strand transcripts are metabolically unstable with respect to the H-strand transcripts (Murphy et al., 1975; Aloni and Attardi, 1971). Hybridization studies have shown that there are a significant number of tRNA genes on both strands; however, the two rRNA species and all but one of the abundant, poly(A) containing RNAs hybridize to the H-strand (Wu et al., 1974; Angerer et al., 1976; Amalric et al., 1978). Hybridization experiments revealed an overall view of the arrangement of these transcripts and genes along the mitochondrial genome (Ojala et al., 1980). Additionally, electron microscopic observations and hybridization experiments allowed the crude localization of the origin of DNA replication (Robberson et al., 1974; Ojala and Attardi, 1978).

Chapter 4 shows how using DNA sequencing techniques, we were able to sequence the DNA encompassing the origin of replication and precisely determine the site where DNA replication begins. With the advent of DNA and RNA sequencing techniques, it became obvious that we not only could sequence the mitochondrial genome but we could locate the position of the transcription products at the nucleotide level. Utilizing the mitochondrial DNA sequence and the corresponding RNA sequences we would then be able to look in detail at the structure of mitochondrial genes, the expression of these genes and the evolution of the genome. The DNA sequence analysis carried out by us and more extensively by Sanger's laboratory was straightforward, but the sequence analysis of the mitochondrial transcripts required the development of techniques that allowed the purification in

sufficient quantities of individual mitochondrial RNA species.

Chapter 5 demonstrates how we achieved this goal, first with the two rRNA species. The procedure involved micrococcal nuclease treatment of isolated mitochondria to eliminate cytoplasmic RNA contamination, extraction of mitochondrial RNA, and then isolation of the mitochondrial RNA species from preparative agarose gels. The purified rRNA species were then end-labeled with [ $\gamma$ - $^{32}$ P] ATP and polynucleotide kinase, purified on preparative agarose gels, and the 5'-end of the RNAs were sequenced by partial enzymatic reactions. The technique allowed an accurate sequence of the ends of the two rRNAs and allowed us to position them precisely on the mitochondrial genome. The most interesting result from this analysis was that the 5'-end of the 12S rRNA is directly juxtaposed to the 3'-end of the phenylalanine tRNA gene without any intervening nucleotides. This observation first demonstrated the extreme economy in gene spacing found in mammalian mitochondrial DNA. In Chapter 6, this technique was further utilized to locate the position of one of the polyadenylated RNA species. This RNA species (designated as 7S RNA), is the only stable polyadenylated RNA species coded by the L-strand. We were able to show that it begins 270 nucleotide pairs 5' to the origin of replication and the transcript ends directly at the origin. Combined with the work in Chapters 4 and 5, we have obtained a complete and detailed picture of the genes surrounding and including the origin. Utilizing the same sequencing technology for other polyadenylated species of mitochondrial RNAs, the ends of all of the abundant RNAs could be sequenced and their positions on the genome precisely located.

## REFERENCES

- Aloni, Y. and Attardi, G. (1971) Symmetrical in vivo transcription of mitochondrial DNA in HeLa cells. *Proc. Natl. Acad. Sci. USA* **68**, 1757-1761.
- Amalric, F., Merkel, C., Gelfand, R. and Attardi, G. (1978) Fractionation of mitochondrial RNA from HeLa cells by high-resolution electrophoresis under strongly denaturing conditions. *J. Mol. Biol.* **118**, 1-25.
- Angerer, L., Davidson, N., Murphy, W., Lynch, D. and Attardi, G. (1976) An electron microscope study of the relative positions of the 4S and ribosomal RNA genes in HeLa cell mitochondrial DNA. *Cell* **9**, 81-90.
- Attardi, G., Cantatore, P., Ching, E., Crews, S., Gelfand, R., Merkel, C., Montoya, J. and Ojala, D. (1981) Organization and expression of genetic information in human mitochondrial DNA. In: *International Cell Biology 1980-1981*. H. G. Schweiger (ed.) (Berlin/Heidelberg: Springer-Verlag) pp. 225-238.
- Luck, D. J. L. and Reich, E. (1964) DNA in mitochondria of *Neurospora crassa*. *Proc. Natl. Acad. Sci. USA* **52**, 931-938.
- Murphy, W. I., Attardi, B., Tu, C. and Attardi, G. (1975) Evidence for complete symmetrical transcription in vivo of mitochondrial DNA in HeLa cells. *J. Mol. Biol.* **99**, 809-814.
- Nass, M. M. K. and Nass, S. (1963) Intramitochondrial fibers with DNA characteristics. I. Fixation and staining reactions. *J. Cell Biol.* **19**, 593-611.
- Ojala, D. and Attardi, G. (1978) Precise localization of the origin of replication in a physical map of HeLa cell mitochondrial DNA and isolation of a small fragment that contains it. *J. Mol. Biol.* **122**, 301-319.
- Ojala, D., Merkel, C., Gelfand, R. and Attardi, G. (1980) The tRNA genes punctuate the reading of genetic information in human mitochondrial DNA. *Cell* **22**, 393-403.

- Robberson, D. L., Clayton, D. A. and Morrow, J. F. (1974) Cleavage of replicating forms of mitochondrial DNA by Eco RI endonuclease. *Proc. Natl. Acad. Sci. USA* **71**, 4447-4451.
- Tzagoloff, A., Macino, G. and Sebald, W. (1979) Mitochondrial genes and translation products. *Ann. Rev. Biochem.* **48**, 419-441.
- Wu, M., Davidson, N., Attardi, G. and Aloni, Y. (1972) Expression of the mitochondrial genome in HeLa cells. XIV. The relative positions of the 4S RNA genes and of the ribosomal RNA genes in mitochondnrial DNA. *J. Mol. Biol.* **71**, 81-93.

## Chapter 4

Nucleotide Sequence of a Region of Human Mitochondrial DNA  
Containing the Precisely Identified Origin of Replication

This paper was published in Nature.

# Nucleotide sequence of a region of human mitochondrial DNA containing the precisely identified origin of replication

Stephen Crews, Deanna Ojala, James Posakony, Jerry Nishiguchi & Giuseppe Attardi

Division of Biology, California Institute of Technology, Pasadena, California 91125

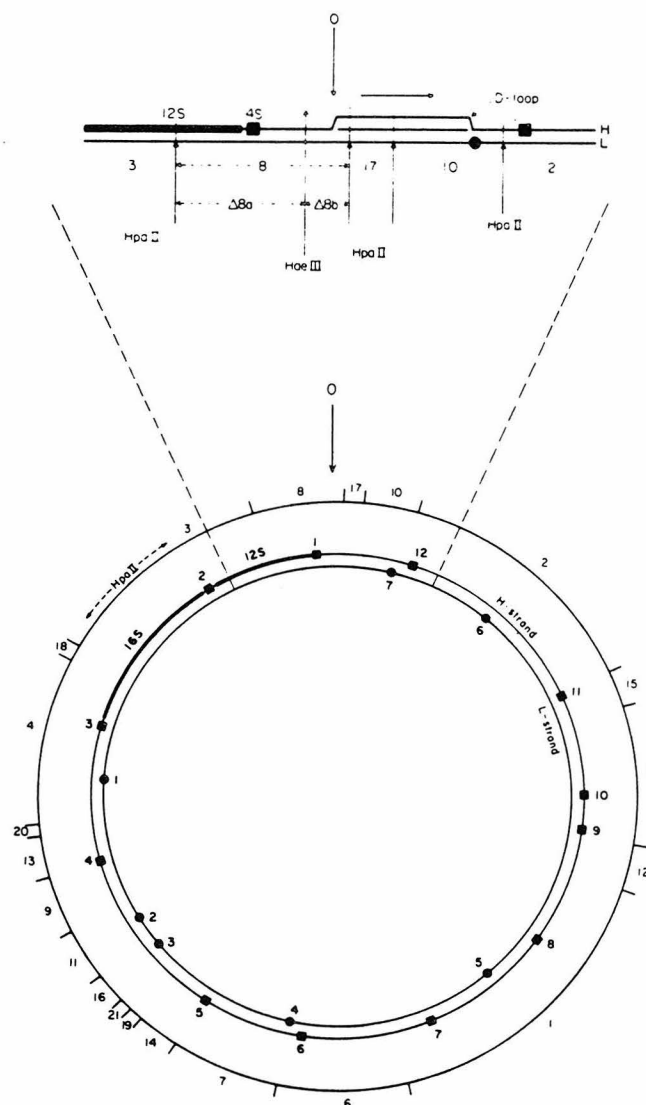
*A fragment of HeLa cell mitochondrial DNA containing the origin of replication has been sequenced. The precise position of the origin in this sequence has been identified by determining the nucleotide order in the 5'-end proximal portion of the heavy strand initiation fragment (7S DNA), and by aligning the two sequences.*

MITOCHONDRIAL DNA synthesis in animal cells has been shown to have very distinctive features. In mouse L cells, where this process has been studied in most detail, DNA replication starts at a site located near the 5'-end of the 12S rRNA coding sequences with the displacement synthesis of a daughter H-strand, which grows in the direction away from the rRNA genes. When the nascent H-strand is at least two-thirds complete, synthesis of the complementary L-strand DNA begins at a point approximately 67% of the genome length from the H-strand origin, proceeding in the opposite direction, and ultimately replication is completed<sup>1,2</sup>. The same type of asymmetrical replication of mit-DNA has been observed in other animal cell systems such as HeLa cells, chick embryo, chick liver, rat liver, *Xenopus* cells and sea urchin oocytes<sup>3</sup>. However, the degree of asynchrony in the replication of the two strands varies considerably between different systems and also in the same system. Furthermore, in most systems investigated, including rat liver and rat hepatoma cells<sup>4</sup>, sea urchin oocytes<sup>5</sup>, *Drosophila melanogaster*<sup>6</sup> and HeLa cells (D. Carré and G. A., unpublished observations), a varying, though generally small, proportion of mit-DNA molecules has been observed to undergo a synchronous unidirectional replication of the two strands, resulting in totally duplex replication forms. The significance and possible regulation of this variability in the degree of asynchrony between displacement synthesis and complement synthesis in mit-DNA replication are unknown.

Equally unique to mit-DNA of most animal cells investigated so far is the presence, in a substantial portion of the molecules, of a displacement loop (D-loop) near the origin of replication. This loop results from the synthesis of a short segment of H-strand DNA which displaces the parental H-strand<sup>7</sup>. These single-strand segments of DNA (referred to as 7S DNA because of their sedimentation coefficient in sucrose gradients) have recently been shown to exist as discrete size classes ranging between 500 and 700 nucleotides depending on the cell type. In mouse L cells, most of the 7S DNA molecules have a common 5'-end with a variability in length at the 3'-end<sup>8</sup>, whereas, in human cells, the variability has been reported to be at the 5'-end<sup>9</sup> or at both ends<sup>8</sup>. In L cells there also seems to be a microheterogeneity at the 5'-end of each predominant size class<sup>8</sup>. Recent studies have shown that the 7S DNA is unstable, with a half life of about 1 h in L cells<sup>10</sup> and 2.8 h in human KB cells (D. Bogenhagen and D. Clayton, personal communication). Therefore, if 7S DNA is a primer for DNA replication, only a small fraction of the 7S DNA molecules can serve this role, as mit-DNA is known to be stable. The true function(s) of 7S DNA is still unknown. It may reflect a need for the cell to be able to recognise the origin or to continuously expose the single-stranded parental H-strand for initiation of transcription.

Conceivably also, the D-loop may be involved in the recently discovered interaction between the mitochondrial inner membrane and the origin of mitochondrial DNA replication<sup>11</sup>.

This article describes a sequence analysis of the region surrounding the origin of mit-DNA replication in HeLa cells. In



**Fig. 1** Physical and genetic maps of HeLa cell mit-DNA. The restriction endonuclease map determined with the enzyme *HpaII*<sup>12</sup> has been aligned with the *EcoRI* and *HindIII* maps<sup>14</sup> and with the map of the positions of the complementary sequences for the 12S and 16S rRNA on the H-strand and for 4S RNAs on the H(■) and L strands (●)<sup>15</sup>. In the upper part, a region of the *HpaII* physical map of HeLa cell mit-DNA has been expanded to show the precise positions of the origin of replication (vertical arrow (marked O)) and of the D-loop. The rightward arrow indicates the direction of H-strand synthesis. H: heavy strand; L: light strand. The arrows marked *HpaII* indicate *HpaII* cleavage sites; the arrow marked *HaeIII* indicates an *HaeIII* site in *HpaII* fragment 8 (modified from D.O. and G.A.<sup>13</sup>).





**Fig. 2** Autoradiograph of the separated strands of 5'-end labelled fragment  $\Delta 8b^{Hae}$ . The restriction fragment  $\Delta 8b^{Hae}$  was obtained by cleaving purified *HpaII* fragment 8 with *HaeIII* and electrophoresis of the digestion products on a 1.5% agarose gel in 0.032 M Tris, 0.02 M sodium acetate, 0.002 M EDTA, pH 7.4, for 4 h at 4 V  $cm^{-1}$ . The fragment was eluted by diffusion at 37 °C in 0.001 M Tris-HCl, pH 7.4, 0.001 M EDTA. In the left track, fragment  $\Delta 8b^{Hae}$  was labelled, without prior dephosphorylation, with [ $\gamma$ - $^{32}P$ ]ATP and T4 polynucleotide kinase (PL Biochemicals) using the exchange reaction. In the right track,  $\Delta 8b^{Hae}$  was dephosphorylated with 4  $\mu$ g of bacterial alkaline phosphatase (Worthington BAPF) in 100  $\mu$ l of 0.01 M Tris-HCl, pH 8.0, at 37 °C for 1 h (freed of divalent cations by treatment with BioRad Chelex 100)<sup>19</sup>. The phosphatase was removed by extraction with phenol-chloroform-isoamyl alcohol (50:50:1), or inactivated by the addition of nitrilotriacetic acid to 0.005 M at 25 °C<sup>20</sup>. The dephosphorylated 5'-ends were 'kinased' by the procedure described by Maxam and Gilbert<sup>16</sup>, which included a heat denaturation step in the presence of 0.001 M spermidine. The labelled fragments were denatured

in 0.001 M Tris-HCl, pH 7.4, 0.001 M EDTA for 5 min at 90 °C, quickly cooled in an ice bath and, after addition of a Ficoll-bromophenol blue mixture, electrophoresed on a 6% polyacrylamide (1:30 bisacrylamide) gel in 0.05 M Tris-borate, pH 8.3, 0.001 M EDTA. The gel was run at 4 °C for 15 h at 6 V  $cm^{-1}$ . The two separated strands were excised from the gel and eluted by diffusion at 37 °C in 0.5 M ammonium acetate, 0.1% SDS,  $10^{-4}$  M EDTA. In later experiments, *HpaII* fragment 8 was cleaved with *HaeIII*, and the products were directly kinased and subjected to strand separation as above.

these cells, by electron microscopic mapping of the D-loop with respect to certain restriction enzyme cutting sites, the origin of mit-DNA replication had been previously located in one of the three *HpaII* fragments, 8, 10 or 17 (ref. 12). More recently, by using 7S DNA as a template for fragment-primed DNA synthesis by *Escherichia coli* DNA polymerase I and by carrying out a polyacrylamide gel electrophoretic analysis of the *in vitro* products after restriction enzyme cleavage, it has been possible to localise the position of the 5'-end of the 7S DNA in *HpaII* fragment 8, at about 80 nucleotides from the *HpaII* cutting site between fragments 8 and 17 (ref. 13). It has been further shown that *HpaII* fragment 8 can be subcleaved by *HaeIII* into two smaller fragments designated as  $\Delta 8a^{Hae}$  (~560 base pairs) and  $\Delta 8b^{Hae}$  (~200 base pairs), and that the origin is contained in  $\Delta 8b^{Hae}$  (ref. 13). This information, summarised in Fig. 1, has thus provided the foundation for the sequencing work described here.

### Sequence analysis of the region of mit-DNA containing the origin of replication

The approach followed here to sequence the region of mit-DNA around the origin of replication involved isolating the restriction fragment  $\Delta 8b^{Hae}$ , labelling the 5'-ends with [ $\gamma$ - $^{32}P$ ]ATP and T4 polynucleotide kinase, separating the strands and sequencing both strands by the technique of Maxam and Gilbert<sup>16</sup>. As the fragment  $\Delta 8b^{Hae}$  was known to be around 200 nucleotides long, we hoped that by ordering more than 100 nucleotides from each end, we could sequence the entire fragment with a substantial sequence overlap. Fragment  $\Delta 8b^{Hae}$  was isolated by gel electrophoresis, eluted and subjected to two kinds of kinase reactions: (1) the exchange reaction<sup>17,18</sup>, which does not require prior dephosphorylation of the 5'-terminus of the DNA, and (2) removal of the 5'-phosphate with bacterial alkaline phosphatase and then treatment with polynucleotide kinase<sup>16</sup>. Figure 2 shows the results of the kinase reactions on strand separation on a polyacrylamide gel. Significantly, in the DNA labelled by the exchange reaction, the faster migrating band contained only 15% as much radioactivity as the slower moving band, whereas

DNA labelled after dephosphorylation revealed a faster migrating band labelled to the extent of about 65% of the slower band. As the incorporation into the slower migrating band was similar in the material labelled by the two procedures, in the present work, the DNA was routinely dephosphorylated before treatment with polynucleotide kinase.

After elution of the material in the two labelled DNA bands, aliquots of each sample were hybridised to an excess of either the heavy or the light strand of HeLa cell mit-DNA, incubated with *S<sub>1</sub>* nuclease, and the acid precipitability measured. The DNA from the faster migrating band hybridised to the extent of about 56% of the input c.p.m. to the H-strand and of 2% to the L-strand, whereas about 42% of the input c.p.m. in the material from the slower migrating band hybridised to the L-strand and 0.6% to the H-strand. The lack of complete hybridisation to mit-DNA of the two separated  $\Delta 8b$  strands is presumably due to the occurrence of some breaking at the ends of the hybrids, with the *S<sub>1</sub>* nuclease cleaving off a portion (40–50%) of the 5'-labelled nucleotides. On the basis of these data, we have designated the faster migrating DNA strand,  $\Delta 8b^{Hae}(H)$ , and the slower migrating DNA strand,  $\Delta 8b^{Hae}(L)$ . It is not clear why the dephosphorylated and kinased  $\Delta 8b^{Hae}(H)$  contained a lower amount of radioactivity than  $\Delta 8b^{Hae}(L)$  (varying, in different experiments between 75% and 31%). This difference is probably due to incomplete dephosphorylation, or to renaturation and subsequent lower incorporation during kinase treatment at the flush end created by the *HaeIII* enzyme relative to the staggered 5'-end produced by *HpaII*, or to some other feature, such as secondary structure.

Sequence analysis of the labelled DNA strands was carried out by the Maxam and Gilbert procedure<sup>16</sup>, using the G > A, A > C, C + T and C reactions. The sequence of nucleotides 2 to 25 was determined by fractionating the cleavage products on a 25% polyacrylamide–7 M urea gel; nucleotides 26 to 148 for  $\Delta 8b^{Hae}(L)$  and nucleotides 25 to 123 for  $\Delta 8b^{Hae}(H)$  were identified by fractionating the reaction products on 10% polyacrylamide–7 M urea gels (Fig. 3). The total length of the fragment was found to be 221 nucleotides, and thus it was possible to obtain an approximately 50 nucleotide overlap.

**Table 1** 5'-terminal nucleotide analysis of  $\Delta 8b^{Hae}(L)$ ,  $\Delta 8b^{Hae}(H)$  and 7S DNA

	% of total $^{32}P$ c.p.m.*							
	dAMP	dCMP	dGMP	dTMP	AMP	CMP	GMP	UMP
$\Delta 8b^{Hae}(L)$	0.15	98.6	0.61	0.68				
$\Delta 8b^{Hae}(H)$	0.43	91.8	5.2	2.6				
7S DNA								
After isolation								
Expt 1	0.6	1.0	3.3	94.2	0.1	ND	ND	ND
Expt 2	1.0	1.4	4.4	93.1	ND	ND	ND	ND
In situ	2	5	3	89	ND	1	1	ND

5'-end-labelled  $\Delta 8b^{Hae}(L)$  and  $\Delta 8b^{Hae}(H)$  were dissolved in 15  $\mu$ l 50 mM ammonium acetate, pH 5.4, and digested with 1  $\mu$ g *P<sub>1</sub>* endonuclease from *Penicillium citrinum* for 1 h at 37 °C<sup>21</sup>. After addition of 10 nmol each of 5'-dAMP, dCMP, dGMP and dTMP, the mixture was applied to a PEI cellulose thin layer strip. After washing the strip in methanol, one-dimensional thin layer chromatography was carried out using 1 M acetic acid for 2 cm and then 1 M acetic acid–3 M LiCl(9:1) for 16 cm<sup>22</sup>. 7S DNA, with 15  $\mu$ g of yeast tRNA carrier, was dissolved in 15  $\mu$ l 50 mM ammonium acetate, pH 5.4. The mixture was digested with 2.5  $\mu$ g of *P<sub>1</sub>* at 37 °C for 3 h. After addition of marker 5'-dNMPs, the digest was spotted onto a PEI cellulose plate and run in the first dimension as above. The plate was washed successively with Tris-methanol and methanol, and then run in a second dimension for 16 cm using 6 g  $Na_2B_4O_7 \cdot 10H_2O$ , 3 g  $H_3BO_3$  and 25 ml ethylene glycol in 70 ml water<sup>22</sup>. All thin layer plates were examined by autoradiography sensitive enough to detect 1% of the input radioactivity (3 d.p.m. in a 1-mm<sup>2</sup> circle), and then the marker spots were visualised with a short-wave UV lamp and the spots cut out. They were eluted quantitatively with 1 ml of 0.02 M Tris-HCl, pH 7.4, 0.7 M  $MgCl_2$  and counted in 10 ml of a Triton-xylene based scintillation fluid. ND, not detectable.

\* The values refer to the radioactivity which migrated from the origin. The % of the input radioactivity recovered from the origin was, respectively, 0.5 and 4.9% for  $\Delta 8b^{Hae}(L)$  and  $\Delta 8b^{Hae}(H)$ , not detectable and 1.2% for isolated 7S DNA (experiments 1 and 2, respectively) and 2.0% for the *in situ* labelled 7S DNA.

To complete the sequence, the end-labelled separated strands were digested exhaustively with endonuclease P<sub>1</sub>, which generates 5'-mononucleotides<sup>21</sup>, and the products were separated by one-dimensional chromatography on thin layer polyethylenimine strips. The results shown in Table 1 indicate that the 5'-terminal nucleotide of both the H and L strands is dCMP, as expected for the *Hpa*II and *Hae*III cutting sites. The complete sequence of the fragment  $\Delta 8b^{Hae}$  is shown in Fig. 4.

### Isolation and characterisation of 7S DNA

To precisely localise the origin of mitochondrial DNA replication in the  $\Delta 8b^{Hae}$  fragment, the approach followed was to sequence the 5'-end portion of the 7S DNA and to align this sequence with that of  $\Delta 8b^{Hae}$ . Previous reports of 7S DNA size heterogeneity in human, mouse and other animal cells and also of an apparent 7S DNA 5'-end microheterogeneity in mouse L cells required that HeLa cell 7S DNA be first characterised in its size and 5'-termini.

Isolation of 7S DNA was carried out by the procedure of Kasamatsu *et al.*<sup>7</sup>, in which the 7S DNA molecules are released from closed circular mit-DNA by brief heat treatment and then separated by sucrose gradient centrifugation. The material in the 7S DNA peak was further purified from RNA contaminants by RNase digestion or nitrocellulose chromatography<sup>23</sup>, and then isolated by polyacrylamide or agarose gel electrophoresis in denaturing conditions.

Figure 5 illustrates the purification procedures used. Slot 2 shows the pattern obtained by electrophoresing through a 4% polyacrylamide-7M urea gel a sample of sucrose gradient-isolated 7S DNA labelled *in vitro* at the 5'-end with [ $\gamma$ - $^{32}\text{P}$ ]ATP and T4 polynucleotide kinase. A prominent band (indicated as 1) is seen migrating at a position corresponding to a length of about 680 nucleotides (estimated by comparison with marker HeLa cell *Hpa*II fragments). A fainter band (indicated as 2), corresponding to a size of about 630 nucleotides, can also be seen. The further purification achieved by nitrocellulose chromatography is shown in slot 3, where the discrete and heterogeneous components visible in the lower part of the gel and presumably arising from RNA contaminants have been eliminated. Slot 4 shows the electrophoretic pattern obtained with 7S DNA prepared, using the isolation procedure described above, from cells labelled *in vivo* with  $^{32}\text{P}$ -orthophosphate. The pattern observed is very similar to that of the *in vitro* labelled material. Here, in addition to the two above mentioned bands, there is also a third faint band migrating at a position corresponding to a size of about 590 nucleotides. These observations on *in vivo* labelled material indicate that the appearance

of the minor 7S DNA components is not due to artefacts introduced by the *in vitro* labelling conditions.

Comparison of the electrophoretic patterns of sucrose gradient-purified 7S DNA with those published by Brown *et al.*<sup>9</sup> and Gillum and Clayton<sup>8</sup> suggests that the major and the minor bands observed here correspond to the three major forms of 7S DNA detected by these authors in several human cell types. However, the relative intensities of the bands observed by the cited authors and those shown here are strikingly different. Instead of three components in approximately equal amounts, our material shows a preponderance of the largest, 680 nucleotide long, component. To exclude the possibility that this difference resulted from losses or degradation occurring during purification of the 7S DNA, the *in situ* labelling technique used by Gillum and Clayton<sup>8</sup> was followed. In this procedure, the 7S DNA is labelled without prior separation from closed circular mit-DNA. In Fig. 5, slot 5, where the results of this experiment are presented, the electrophoretic pattern seen is very similar to that of the *in vitro* labelled samples with the 680 nucleotide long component again being strongly predominant. Densitometric analysis of these autoradiograms indicate that this component represents 70–80% of the total 7S DNA in preparations labelled *in situ*, and more than 95% in preparations labelled after sucrose gradient isolation, assuming equal efficiency of *in vitro* labelling of the various components.

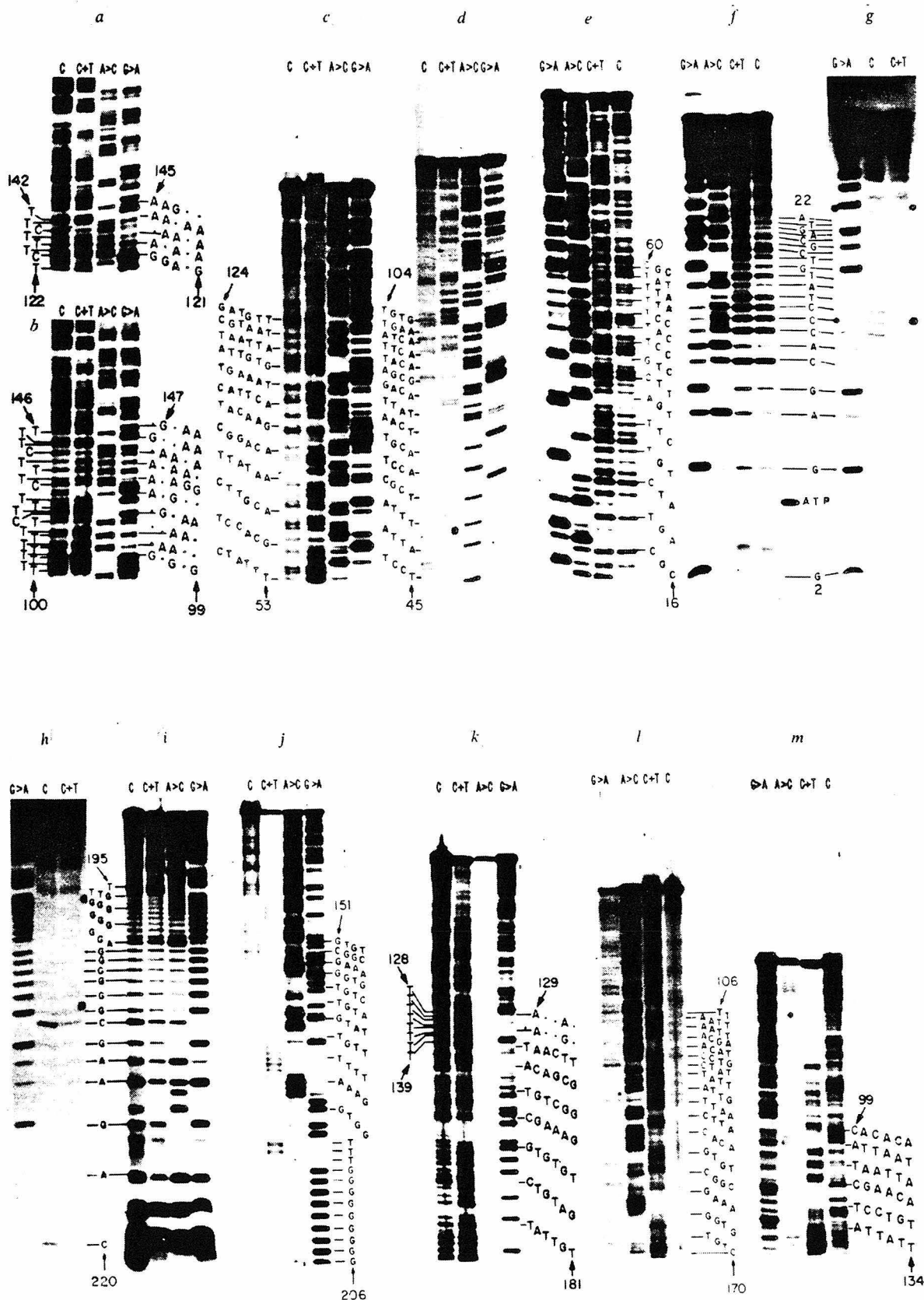
The fact that the relative intensities of the bands produced by the two labelling methods were comparable did not rule out the possibility that the difference in mobility between the 7S DNA components might be due to some factor other than heterogeneity in length. However, treatment with RNase (Fig. 5, slot 4) or pronase (in the presence of SDS) before electrophoresis failed to alter the observed number, relative abundance, or electrophoretic mobility of these components, thereby excluding any possible role of a covalently linked RNA primer or of DNA-associated protein in generating this phenomenon. In addition, when the *in situ* labelled component 2 and the *in vitro* labelled component 1 were isolated from the urea gel, heat denatured and re-run through a second urea gel in the same conditions (Fig. 5, slots 6 and 7, respectively), or through an agarose-CH<sub>3</sub>HgOH gel<sup>26</sup> (not shown), they maintained the same relative electrophoretic mobility. The above results support the conclusion that the multiple bands arise from discrete length differences.

To investigate the possibility of 5'-end microheterogeneity (which would not have been detected by the above electrophoretic analysis), sucrose-gradient-isolated 7S DNA was labelled at the 5'-end with [ $\gamma$ - $^{32}$ P]ATP and T4 polynucleotide



**Fig. 4** Nucleotide sequence of the fragment  $\Delta 8b^{Hae}$  and of the 5'-terminal segment of 7S DNA.





**Fig. 3** Representative autoradiographs of sequencing gels of  $\Delta 8b^{Hae}$  (L) (a-g) and  $\Delta 8b^{Hae}$  (H) (h-m). The DNA strands, labelled at the 5'-end with  $[\gamma-^{32}P]ATP$  and polynucleotide kinase after alkaline phosphatase treatment (see text), were sequenced using the chemical methods of Maxam and Gilbert. The sequence reactions used and the times of reaction for the base modification step were G > A (10 min), A > C (10 min), G + T (15 min) and C (15 min). The degradation products were electrophoresed at 1,000 V for a, 15.5; b, 13; m, 12; c, 1, 8.5; d, k, 7.25; f-i, 7 and e, j, 4 h on either 10% polyacrylamide (1:20 bisacrylamide)-7 M urea gels (a-e, j-m) or 25% polyacrylamide (1:30 bisacrylamide)-7 M urea gels (f-i) ( $40 \times 30 \times 0.15$  cm) in 0.05 M Tris-borate, pH 8.3, 0.001 M EDTA, that had been pre-electrophoresed for 2 and 3 h, respectively, at 900 V. The gels were wrapped with Saran wrap and exposed at  $-70^{\circ}C$  using pre-fogged Kodak X-Omat XR-5 film with a DuPont Cronex Lightning-Plus intensifying screen.

kinase and subjected to digestion with the *Hpa*II restriction enzyme. (*Hpa*II is known to cleave single-stranded DNA<sup>27</sup>, and previous experiments in this laboratory had indicated that the *Hpa*II site present in the single stranded portion of the D-loop (which is equivalent in sequences to the 7S DNA) is cut at a low frequency in standard digestion conditions<sup>13</sup>.) The digest was analysed on a 7.5% polyacrylamide–7M urea gel; the results obtained (not shown) indicated that a small percentage of the 7S DNA molecules had been cut at a position corresponding to a distance of ~86 nucleotides from the labelled 5'-end (as estimated from the migration of the small fragment thus produced relative to that of denatured  $\Phi$ X174 RF DNA *Hae*III fragments)<sup>28</sup>. As the resolving power of the gel electrophoresis used would have shown a difference of 1 or 2 bases in length, the above results indicate that the 5'-end of the main 7S DNA component occurs at a unique position within the mitochondrial genome. To obtain further evidence on this question, a 5'-end base analysis was carried out. 7S DNA was dephosphorylated at the 5'-end by alkaline phosphatase either *in situ* or *in vitro* after sucrose gradient isolation, labelled with [ $\gamma$ -<sup>32</sup>P]ATP and polynucleotide kinase, purified by nitrocellulose chromatography and electrophoresed through a 4% acrylamide–7M urea gel. The main component was eluted from the gel and subjected to digestion with endonuclease P<sub>1</sub>. The products were analysed by two-dimensional chromatography on polyethylenimine thin layer plates, and the results are given in Table 1. In the *in vitro* labelled sample, more than 90% of the counts recovered are represented by dTMP, confirming the conclusion that the 5'-end of the predominant HeLa cell 7S DNA component is unique. Furthermore, no radioactivity, or only traces of radioactivity, were detected in ribonucleotides. The results obtained for the *in situ* labelled sample are very similar, although a slightly larger proportion of radioactivity seems to be associated with deoxyribonucleotides other than dTMP; this may reflect the lower degree of purification of *in situ* labelled 7S DNA due to the omission of the sucrose gradient purification step.

## Sequence analysis of 7S DNA

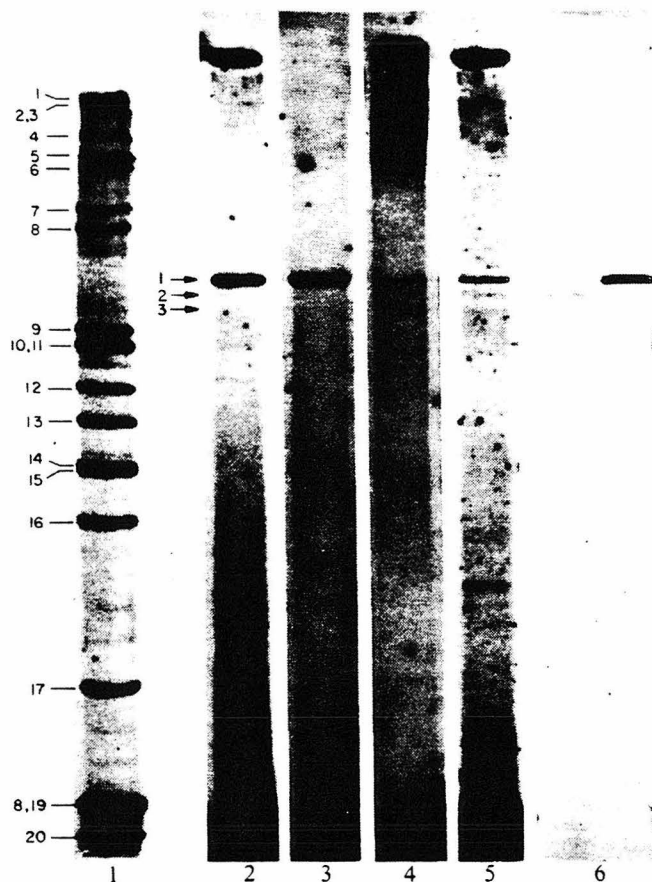
The main 7S DNA component was subjected to alkaline phosphatase treatment and the kinase reaction either *in situ* or after sucrose gradient purification, passed through a nitrocellulose column and then run on a polyacrylamide–urea gel. After elution of the 7S DNA from the gel, it was submitted to the Maxam–Gilbert sequencing reactions.

Figure 6 shows a sequencing gel of 5'-end labelled, sucrose gradient-purified 7S DNA. A sequence of around 25 nucleotides could be determined unambiguously. This sequence is identical to the portion of the sequence of  $\Delta 8b^{Hae}$  (H) from residue 87 to residue 63. This alignment allowed the positioning of the 5'-end of 7S DNA to correspond with residue 87 of  $\Delta 8b^{Hae}$ . Sequencing experiments using *in situ* labelled 7S DNA gave identical results<sup>28</sup>.

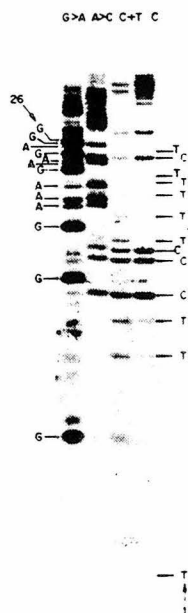
## Discussion

In the experiments reported here, we have precisely located the origin of HeLa cell mit-DNA synthesis in the *Hpa*II map of this DNA and examined the sequence surrounding it. The location of the 5'-end of the 7S DNA, determined by sequence analysis to be at 87 nucleotides from the *Hpa*II cutting site between fragments 8 and 17, is in excellent agreement with the 7S DNA priming experiments that indicated the origin to be at approximately 80 nucleotides from the *Hpa*II site, and with the experiments in which 5'-end labelled 7S DNA was cut with *Hpa*II giving a labelled DNA fragment about 86 nucleotides long. The size of  $\Delta 8b^{Hae}$  is 221 nucleotide pairs, which is about 10% longer than the value that was earlier derived from its electrophoretic mobility in agarose gels relative to that of *Hpa*II fragments.

There are several interesting features of the  $\Delta 8b^{Hae}$  sequence (Fig. 4), although their significance is unclear. Beginning at nucleotide 215 and proceeding clockwise (the clockwise direction is here defined as the direction of H-strand synthesis), there



**Fig. 5** Autoradiographs, after electrophoresis through 4% polyacrylamide–7 M urea gels, of HeLa cell 7S DNA labelled *in vitro* and purified in different ways. All samples, in 0.001 M Tris-HCl, pH 7.4, 0.001 M EDTA, were adjusted to 6 M urea, heated at 100 °C for 1 min and loaded on a 4% polyacrylamide (1:30 bisacrylamide)–7 M urea slab gel. Electrophoresis was carried out at 5 V cm<sup>-1</sup> for 15 h. *Slot 1*: *Hpa*II digest of HeLa cell mit-DNA labelled *in vitro* with [ $\alpha$ -<sup>32</sup>P]dNTP and *E. coli* DNA polymerase I<sup>12</sup>. *Slots 2 and 3*: 7S DNA was released from mit-DNA by brief heating, isolated by sucrose gradient centrifugation<sup>7</sup>, and labelled *in vitro* at the 5'-end as follows: A sample of ~0.2 pmol of 7S DNA (estimated by assuming a proportion of about 10% D-loop containing molecules in HeLa cell mit-DNA preparations<sup>24</sup>, and a complete recovery of the isolated 7S DNA) was dephosphorylated and 5'-end labelled as described in Fig. 2 legend. After ethanol precipitation, the pellet was dissolved in 0.001 M Tris-HCl, pH 7.4, 0.001 M EDTA. A portion of the sample was analysed directly (slot 2) and a portion was further purified by nitrocellulose (NC) chromatography. For this the sample (50  $\mu$ l) was adjusted to 0.5 M KCl and passed through a 1.5  $\times$  0.3 cm nitrocellulose column (Hercules, prepared as described by Boezi and Armstrong<sup>25</sup>), equilibrated with 0.5 M KCl, 0.01 M Tris-HCl, pH 7.4. The column was washed with equilibration buffer to remove all the unbound radioactivity, then eluted with 0.001 M Tris-HCl, pH 7.4, 0.001 M EDTA (slot 3). *Slot 4*: 7S DNA from  $1.5 \times 10^8$  cells labelled *in vivo* with <sup>32</sup>P-orthophosphate<sup>12</sup> was isolated by sucrose gradient centrifugation and run on gel after treatment with pancreatic RNase (40  $\mu$ g ml<sup>-1</sup>, 10 min at 25 °C). *Slot 5*: 7S DNA labelled *in situ* following the procedure of Gillum and Clayton<sup>8</sup>: a sample of 10  $\mu$ g mit-DNA in 100  $\mu$ l 0.01 M Tris-HCl, pH 8.5, 0.01 M NaCl, was incubated with 0.05  $\mu$ g calf intestine alkaline phosphatase (Boehringer Mannheim) for 1 h at 57 °C, and then labelled with [ $\gamma$ -<sup>32</sup>P]ATP and T4 polynucleotide kinase. *Slots 6 and 7*: 7S DNA components 2 and 1, which had been labelled at the 5'-end *in situ* or after sucrose gradient isolation, respectively, then purified by nitrocellulose chromatography, and run through a 4% polyacrylamide–7 M urea gel, were eluted and rerun in the same conditions.



**Fig. 6** Representative autoradiograph of 7S DNA sequencing gel. Sucrose gradient isolated 7S DNA was dephosphorylated and 5'-end labelled, as detailed in Fig. 2 legend, purified by nitrocellulose chromatography and then run through a 4% polyacrylamide-7 M urea gel. The main component (1) was eluted and subjected to the sequencing reactions G>A, A>C, C>T and C. Due to the low amount of radioactivity in the 7S DNA preparations, the base modification reactions were carried out for a time 10 times longer than used for the  $\Delta 8b^{Hae}$  strands (legend to Fig. 3). This extended reaction time resulted in a decreased mobility of the products in the A>C reaction. The cleavage products were electrophoresed on a 25% polyacrylamide-7 M urea gel for 6.3 h.

is a stretch of 17 G·C base pairs in a row with an A·T pair interruption at position 207. Furthermore, there is a strong purine bias on the H-strand. Nearby, from position 189 to 181, there is a stretch of 9 A·T base pairs in succession. There is another A·T base pair stretch running from position 137 to 128, and, contained within it, an 8 nucleotide true palindrome from position 137 to 130. Closer to the origin is a 13 A·T base pair stretch which contains two true palindromes (113-104 and 116-105) and a sequence with a twofold axis of symmetry (115-104). Repeated sequences are relatively rare; the sequence 5' TTGTTAT 3' at nucleotides 140-134 is repeated 3' AACAAATA 5' at positions 182-176, and the sequence 5' GTGGAAA 3' is found at nucleotides 166-160 and at 193-187. The sequence 5' TATGATG 3' 3' ATACTAC 5', often found to be associated with promoter sites<sup>29</sup>, is located at nucleotides 178 to 172 and might possibly be associated with an L-strand promoter site.

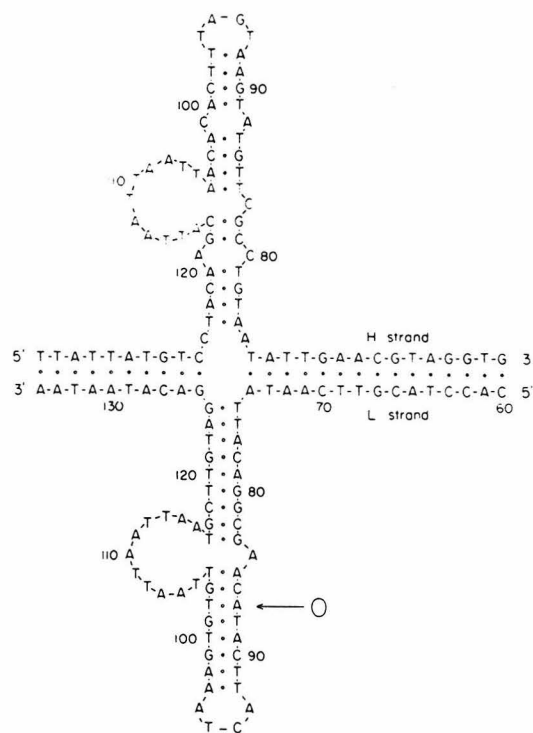
The origin itself is preceded, going in a clockwise direction, by a 29-nucleotide long A·T rich region that extends 3 nucleotides past the origin, is followed by a 6-nucleotide G·C rich region and then by a 10-nucleotide long A·T rich region. A·T rich sequences have previously been noted near other origins of DNA replications; however, we do not detect any sequence characteristics in common with the other origins of replication and primer RNA-DNA junctions that have been accurately localised (ColE1<sup>30</sup> RNA-DNA junction,  $\Phi$ X174 viral strand origin<sup>31</sup>, fd<sup>32</sup> and G4 complementary strand<sup>33</sup> origins) nor with the L-cell 7S DNA sequence<sup>8</sup>.

Computer analysis of the sequence data indicates a theoretically stable hairpin structure that could form when the DNA double helix is unwound, as occurs during DNA synthesis. As shown in Fig. 7, this looped structure contains the origin of H-strand replication and an interior A·T rich loop, 12 nucleotides long. As calculated using the rules of Tinoco *et al.*<sup>35,36</sup> and Gralla and Crothers<sup>37</sup>, the  $\Delta G$  of the L-strand hairpin structure is approximately  $-13 \text{ kcal mol}^{-1}$  and that of the H-strand hairpin is approximately  $-5 \text{ kcal mol}^{-1}$ , with the difference being due to two A·C pair mismatches in the H-strand stem. Although this secondary structure is probably thermodynamically stable at 25 °C in a moderate ionic strength buffer at neutral pH, it is not known whether it is ever formed *in vivo*. However, the presence of the origin in the hairpin at only two nucleotides from the A·T rich interior loop is suggestive of a functional role for this

structure. Secondary structure loops have been strongly implicated in the initiation of phage fd<sup>32</sup> and G4<sup>33</sup> complementary strand synthesis, and may be involved in the initiation and regulation of DNA replication in the mammalian papovaviruses SV40<sup>38,39</sup> and BKV<sup>40</sup>.

We have not detected any ribonucleotides covalently attached to the 5'-end of the 7S DNA that might represent an RNA primer or residues of it. Clayton *et al.*<sup>8</sup> have presented evidence for the presence of an RNA primer in mouse L cell 7S DNA. The reason for this difference is unknown. In HeLa cell 7S DNA, we have also observed a greater homogeneity in size than previously reported in 7S DNA from other human cell types and other organisms. Nor have we detected the 5'-end microheterogeneity described in these systems. Our sequence analysis has not revealed any repetition of sequence or other structural features that might point to the existence of multiple initiation points, thereby explaining the existence of the multiple size classes. Instead, we think that the discrete size classes and the relative amounts of 7S DNA contained in each of them, the 5'-end microheterogeneity and the presence of RNA primer residues may reflect various rates or pathways of 7S DNA processing which could be dependent on cell type and/or physiological conditions.

This work was supported by NIH grants GM-11726 and ST32GM-07616. We thank A. Chomyn for the computer analysis of the DNA sequence, A. Maxam and W. Gilbert for advice on sequence determination, N. Davidson for valuable discussions and A. Drew for technical assistance.



**Fig. 7** Possible secondary structure of the region of fragment  $\Delta 8b^{Hae}$  surrounding the origin of replication. The arrow (O) points to the nucleotide in the L-strand which is complementary to the 5'-end nucleotide of 7S DNA. The  $\Delta G$  values of the L-strand and of the H-strand portions of the structure were calculated on the basis of the thermodynamic data for Watson-Crick nearest-neighbour sequences, loops, mismatches, and G·U pairs, as compiled for ribonucleic acids (see text). The free energies of base-pair addition were reduced by about 15% whenever the base-pair doublet included a G·C pair, in order to correct for the greater stability of the rG·rC pair as compared to dG·dC<sup>34</sup>. In view of the many assumptions underlying these calculations, the  $\Delta G$  values reported in the text have to be considered only as rough indicators of stability.

Received 13 November; accepted 1 December 1978.

1. Robberson, D. L., Kasamatsu, H. & Vinograd, J. *Proc. natn. Acad. Sci. U.S.A.* **69**, 737-741 (1972).
2. Robberson, D. L. & Clayton, D. A. *Proc. natn. Acad. Sci. U.S.A.* **69**, 3810-3814 (1972).
3. Kasamatsu, H. & Vinograd, J. A. *Rev. Biochem.* **43**, 695-719 (1974).
4. Wolstenholme, D. R., Koike, K. & Cochran-Fouts, P. J. *Cell Biol.* **56**, 230-245 (1973).
5. Matsumoto, L., Kasamatsu, H., Piko, L. & Vinograd, J. *J. Cell Biol.* **63**, 146-159 (1974).
6. Rubenstein, J. L. R., Brutlag, D. & Clayton, D. A. *Cell* **12**, 471-482 (1977).
7. Kasamatsu, H., Robberson, D. L. & Vinograd, J. *Proc. natn. Acad. Sci. U.S.A.* **68**, 2252-2257 (1971).
8. Gillum, A. M. & Clayton, D. A. *Proc. natn. Acad. Sci. U.S.A.* **75**, 677-681 (1978).
9. Brown, W. M., Shine, J. & Goodman, H. M. *Proc. natn. Acad. Sci. U.S.A.* **75**, 735-739 (1978).
10. Bogenhagen, D. & Clayton, D. A. *J. molec. Biol.* **119**, 49-68 (1978).
11. Albring, M., Griffith, J. & Attardi, G. *Proc. natn. Acad. Sci. U.S.A.* **74**, 1348-1352 (1977).
12. Ojala, D. & Attardi, G. *Plasmid* **1**, 78-105 (1977).
13. Ojala, D. & Attardi, G. *J. molec. Biol.* **122**, 301-319 (1978).
14. Brown, W. M. & Vinograd, J. *Proc. natn. Acad. Sci. U.S.A.* **71**, 4617-4621 (1974).
15. Angerer, L., Davidson, N., Murphy, W., Lynch, D. & Attardi, G. *Cell* **9**, 81-90 (1976).
16. Maxam, A. M. & Gilbert, W. *Proc. natn. Acad. Sci. U.S.A.* **74**, 560-564 (1977).
17. Van de Sande, J. H., Kleppe, K. & Khorana, H. G. *Biochemistry* **12**, 5050-5055 (1973).
18. Berkner, K. L. & Folk, W. R. *J. biol. Chem.* **252**, 3176-3184 (1977).
19. Efstratiadis, A., Kafatos, F. C. & Maniatis, T. *Cell* **10**, 571-585 (1977).
20. Simoncsits, A., Brownlee, G. G., Brown, R. S., Rubin, J. R. & Guillely, H. *Nature* **269**, 833-836 (1977).
21. Fujimoto, M., Kuninaka, A. & Yoshino, H. *Agr. biol. Chem.* **38**, 1555-1561 (1974).
22. Randerath, K. & Randerath, E. *Analyt. Biochem.* **13**, 575-579 (1965).
23. Miyamoto, C. & Denhardt, D. T. *J. molec. Biol.* **116**, 681-707 (1977).
24. Flory, J. P. thesis, Calif. Inst. Tech. (1976).
25. Boezi, J. A. & Armstrong, R. L. *Meih. Enzym.* **12**, A, 684-686 (1967).
26. Bailey, J. M. & Davidson, N. *Analyt. Biochem.* **70**, 75-85 (1977).
27. Godson, G. N. & Roberts, R. J. *Virology* **73**, 561-567 (1976).
28. Attardi, G., Crews, S. T., Nishiguchi, J., Ojala, D. K. & Posakony, J. W. *Cold Spring Harb. Symp. quant. Biol.* **43** (in the press).
29. Pribnow, D. *Proc. natn. Acad. Sci. U.S.A.* **72**, 784-788 (1975).
30. Tomizawa, J.-I., Ohmori, H. & Bird, R. E. *Proc. natn. Acad. Sci. U.S.A.* **74**, 1865-1869 (1977).
31. Langeveld, S. A. *et al. Nature* **271**, 417-420 (1978).
32. Geider, K., Beck, E. & Schaller, H. *Proc. natn. Acad. Sci. U.S.A.* **75**, 645-649 (1978).
33. Giddes, J. C., Barrell, B. G. & Godson, G. N. *Proc. natn. Acad. Sci. U.S.A.* **75**, 1081-1085 (1978).
34. Kallenbach, N. R. *J. molec. Biol.* **37**, 445-466 (1968).
35. Tinoco, I., Jr *et al. Nature new Biol.* **246**, 40-41 (1973).
36. Borer, P. N., Dengler, B., Tinoco, L., Jr. & Uhlenbeck, D. C. *J. molec. Biol.* **86**, 843-853 (1974).
37. Gralla, J. & Crothers, D. M. *J. molec. Biol.* **73**, 497-511 (1973).
38. Subramanian, K. N., Dhar, R. & Weissman, S. M. *J. biol. Chem.* **252**, 355-367 (1977).
39. Tijan, R. *Cell* **13**, 165-179 (1978).
40. Dhar, R., Lai, C.-J. & Khoury, G. *Cell* **13**, 345-358 (1978).

## Chapter 5

The Sequences of the Small Ribosomal RNA Gene and the Phenylalanine  
tRNA Gene are Joined End to End in Human Mitochondrial DNA

This paper was published in Cell.



# The Sequences of the Small Ribosomal RNA Gene and the Phenylalanine tRNA Gene Are Joined End to End in Human Mitochondrial DNA

Stephen Crews and Giuseppe Attardi

Division of Biology  
California Institute of Technology  
Pasadena, California 91125

## Summary

The 5' end proximal regions of the two HeLa cell mitochondrial rRNAs (16S and 12S) have been sequenced by partial enzymatic digestions of the 5' end <sup>32</sup>P-labeled RNAs followed by electrophoretic fractionation of the products on polyacrylamide/urea gels. Likewise, a 600 nucleotide mitochondrial DNA (mit-DNA) fragment, previously designated  $\Delta 8a^{Hae}$ , that contains the 5' end of the 12S rRNA, has been sequenced by the method of Maxam and Gilbert. The first 71 nucleotides of the 12S rRNA and the DNA coding sequence have been aligned and found to be colinear. This observation extends to the 5' end proximal segment of the 12S rRNA gene the conclusion of earlier experiments, indicating the absence of intervening sequences in the body of the small rRNA gene. A comparison of the 12S rRNA coding sequence determined here (286 nucleotides) with that of an *E. coli* 16S rRNA gene has revealed significant homologies. Previous electron microscopic analysis of hybrids between the heavy (H) strand of mit-DNA and ferritin-labeled mitochondrial 4S RNAs had shown the presence of a 4S RNA gene near the 5' end of the 12S rRNA coding sequence. In the present work, a search of the DNA sequence for a cloverleaf structure has indeed revealed the occurrence of a tRNA<sup>Phe</sup> gene. The unexpected finding, however, has been that the 3' end of this gene is contiguous to the 5' end of the 12S rRNA coding sequence without any intervening nucleotides.

## Introduction

Electron microscopic mapping experiments (Wu et al., 1972; Angerer et al., 1976) had previously shown that the coding sequences for the two mitochondrial ribosomal RNA (rRNA) species in HeLa cell mitochondrial DNA (mit-DNA) are located very close to each other on the heavy (H) strand, with a spacer approximately 160 nucleotides in length separating them. Furthermore, experiments of hybridization using ferritin-labeled 4S RNA had revealed the presence of one 4S RNA site in the H strand spacer between the 12S and 16S rRNA coding sequences, one 4S site immediately adjacent to the other end of the 12S RNA sequence and one 4S site adjacent to the other end of the 16S RNA sequence (Wu et al., 1972; Angerer et al., 1976; Figure 1). The same gene organization of the rRNA region was later observed in *Xenopus laevis* mit-DNA

(Ramirez and Dawid, 1978), and, as concerns the relative positions of the two rRNA genes, also in *Drosophila* (Klukas and Dawid, 1976) and mouse cell mit-DNA (Battey and Clayton, 1978), suggesting a high degree of conservation among animal cells.

From the alignment of the positions of the rRNA genes and of the origin of replication with respect to restriction maps of HeLa cell mit-DNA (Figure 1) and from the direction of H strand synthesis (Brown and Vinograd, 1974), it was inferred that the transcription of the rRNA genes proceeds from the smaller to the larger one (Ojala and Attardi, 1977), in agreement with what has been observed in all rDNA analyzed thus far (Lewin, 1976). Furthermore, the 5' end of the 12S rRNA gene was located within Hpa II fragment 8 of HeLa cell mit-DNA. In more recent mapping experiments using the S1 protection technique (Berk and Sharp, 1977, 1978), it has been possible to localize this position more precisely in fragment 8 at approximately 270 nucleotides from the Hpa II site between fragments 3 and 8 (Figure 1) (Attardi et al., 1979; Ojala and Attardi, 1980).

The above-mentioned studies have not provided any further information concerning the precise location, relative to the 12S rRNA gene, of the 4S RNA gene that the EM analysis had placed very close to the 5' end of the small rRNA gene. The nature of the DNA sequences located on the 5' side of the 12S rRNA gene, and in particular the arrangement on the H strand of the 12S rRNA and 4S RNA coding sequences, was of particular interest to us in connection with questions concerning the mechanisms of synthesis and processing of the two RNAs and the location of their promoter sequence(s). This paper describes a sequencing analysis of a region of HeLa cell mit-DNA corresponding to a 5' end proximal segment of the 12S rRNA gene and to the adjacent DNA stretch. In particular, the subfragment  $\Delta 8a^{Hae}$ , produced by cleavage with Hae III of Hpa II fragment 8 (Ojala and Attardi, 1978), has been sequenced and aligned with the 5' end proximal sequence of approximately 70 nucleotides of 12S rRNA: this alignment has allowed the positioning of the 5' end of 12S rRNA at 286 nucleotides from the Hpa II site between fragments 3 and 8. An analysis of the DNA sequence flanking, on the 5' end site, the 12S rRNA gene has revealed the presence of a tRNA<sup>Phe</sup> gene. Unexpectedly, this tRNA gene is immediately adjacent, without any intervening nucleotides, to the 5' end of the 12S rRNA gene.

## Results

### Purification and 5' End Labeling of Mitochondrial rRNA

The high degree of purity of RNA required for sequencing work made special demands on the procedure followed for the isolation of mitochondrial rRNA.

Generally, HeLa cell mitochondria are heavily contaminated with endoplasmic reticulum and chromatin, thus yielding RNA that contains large amounts of cytoplasmic rRNA and mRNA and nuclear DNA. To circumvent this problem, the mitochondrial fraction, after EDTA treatment (to help break apart cytoplasmic, membrane-bound ribosomes; Attardi, Cravioto and Attardi, 1969) and before SDS lysis, was incubated in an isotonic solution containing micrococcal nuclease. At high concentrations, this nuclease degrades non-mitochondrial nucleic acids into small fragments and leaves the intramitochondrial RNA intact. Although other ribonucleases and deoxyribonucleases can be used (Amalric et al., 1978), micrococcal nuclease is preferable, since it can easily be inactivated by incubation with chelating agents. The gel pattern obtained by running 12S and 16S rRNAs purified by this procedure through an agarose methylmercuric hydroxide slab gel was virtually identical to that obtained with total oligo(dT)-cellulose unbound mitochondrial RNA from cells labeled with  $^{32}\text{P}$ -orthophosphate in the presence of actinomycin D or camptothecin, both inhibitors of nuclear RNA synthesis (Amalric et al., 1978). The 12S and 16S rRNA species were estimated to be more than 80% pure (see below), and there was no evidence of intramitochondrial RNA degradation during the micrococcal nuclease treatment. Presumably, any damaged mitochondria that take up the nuclease have their RNA completely degraded.

The purified RNAs were labeled at their 5' ends with  $\gamma$ - $^{32}\text{P}$ -ATP and polynucleotide kinase after dephosphorylation with bacterial alkaline phosphatase at 65°C. The labeled RNAs were run on an agarose/methylmercuric hydroxide gel, bands corresponding to intact 12S and 16S rRNAs were observed and the RNA in these bands was eluted for the enzymatic sequencing reactions. Human 5S rRNA was also labeled in vitro at its 5' end, and likewise showed a band corresponding to intact 5S rRNA (gel not shown). This 5S rRNA, previously sequenced by Forget and Weissman (1967) using the method of Sanger, Brownlee and Barrell (1965), served as a control for the in vitro labeling and sequencing technology used for the mitochondrial rRNAs.

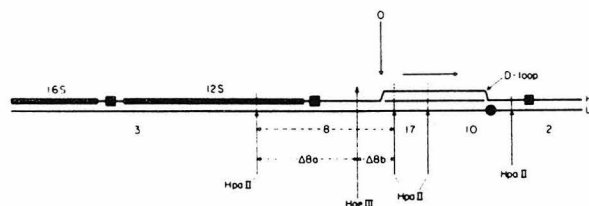


Figure 1. Region of the HeLa Cell mit-DNA Physical Map Showing the Precise Positions of the Origin of DNA Replication (Marked O) and of the rRNA Genes

The rightward arrow above the D loop indicates the direction of H strand synthesis. 4S rRNA coding sequences on the L strand (●) and the H strand (○) are shown together with the Hpa II restriction enzyme cutting sites and the Hae III site found in Hpa II fragment 8.

### Sequence Analysis of Mitochondrial rRNA

The in vitro 5' end-labeled 5S, 12S and 16S rRNAs were first characterized in their 5' termini by a complete P1 digestion and fractionation on PEI-cellulose TLC plates. Besides yielding sequence information, this analysis was also expected to provide an approximate measure of purity of the RNAs and a test of the validity of the in vitro labeling technique used. The results obtained are shown in Table 1, and it can be seen that the 5' end of each of the three RNA species was quite homogeneous with respect to base composition. Furthermore, the 5' terminal nucleotide of the human 5S rRNA had previously been demonstrated to be G (Forget and Weissman, 1967; Hatlen, Amaldi and Attardi, 1969), and two-dimensional fingerprints by high-voltage electrophoresis of in vivo  $^{32}\text{P}$ -labeled mitochondrial 12S and 16S rRNAs, completely digested by a mixture of ribonucleases A, T1, U2 and T2, had indicated that pA and pG were, respectively, the 5' termini of the two rRNAs (S. Crews and K. Grohmann, unpublished observations). Thus the results obtained with the in vitro labeled RNAs indicating that the predominant labeled 5' terminal nucleotide is G, A and G for the 5S, 12S and 16S rRNAs, respectively, not only demonstrate the purity of the RNAs but also strongly suggest that the in vitro labeling technique has faithfully labeled the true 5' end.

Sequencing of the 5' ends of mitochondrial rRNAs was performed utilizing the method of Donis-Keller, Maxam and Gilbert (1977) and Simoncsits et al. (1977). This method uses base-specific enzymatic digestions of 5' end-labeled RNA followed by the fractionation of the digestion products on polyacrylamide/urea sequencing gels. Figures 2 and 3 show autoradiograms of sequencing gels of the 12S and 16S rRNAs. The 12S rRNA gel patterns (Figure 2) showed a shadowing or repetition of each band at the next position on the ladder. This phenomenon and the relative intensities of the bands were reproducible and apparently not sensitive to the conditions of the in vitro labeling. Our interpretation of the 12S rRNA sequencing gels is that there exist two species of 12S rRNA that differ in molecular weight by one nucleotide at the 5' end, and that the more abundant species is one nucleotide shorter. In any case, the sequence is unambiguous and is shown in Figure 4. The 16S rRNA gel pattern (Figure 3) did not show evidence of the phenomenon observed for 12S rRNA, pointing to the

Table 1. 5' End Terminal Nucleotide Analysis of in Vitro Labeled RNAs

RNA Species	Percentage of Total $^{32}\text{P}$ Radioactivity			
	AMP	CMP	GMP	UMP
Cytoplasmic 5S rRNA	2.9	1.4	94.6	1.1
Mitochondrial 12S rRNA	92.1	1.8	1.6	4.5
Mitochondrial 16S rRNA	5.7	2.0	88.9	3.4

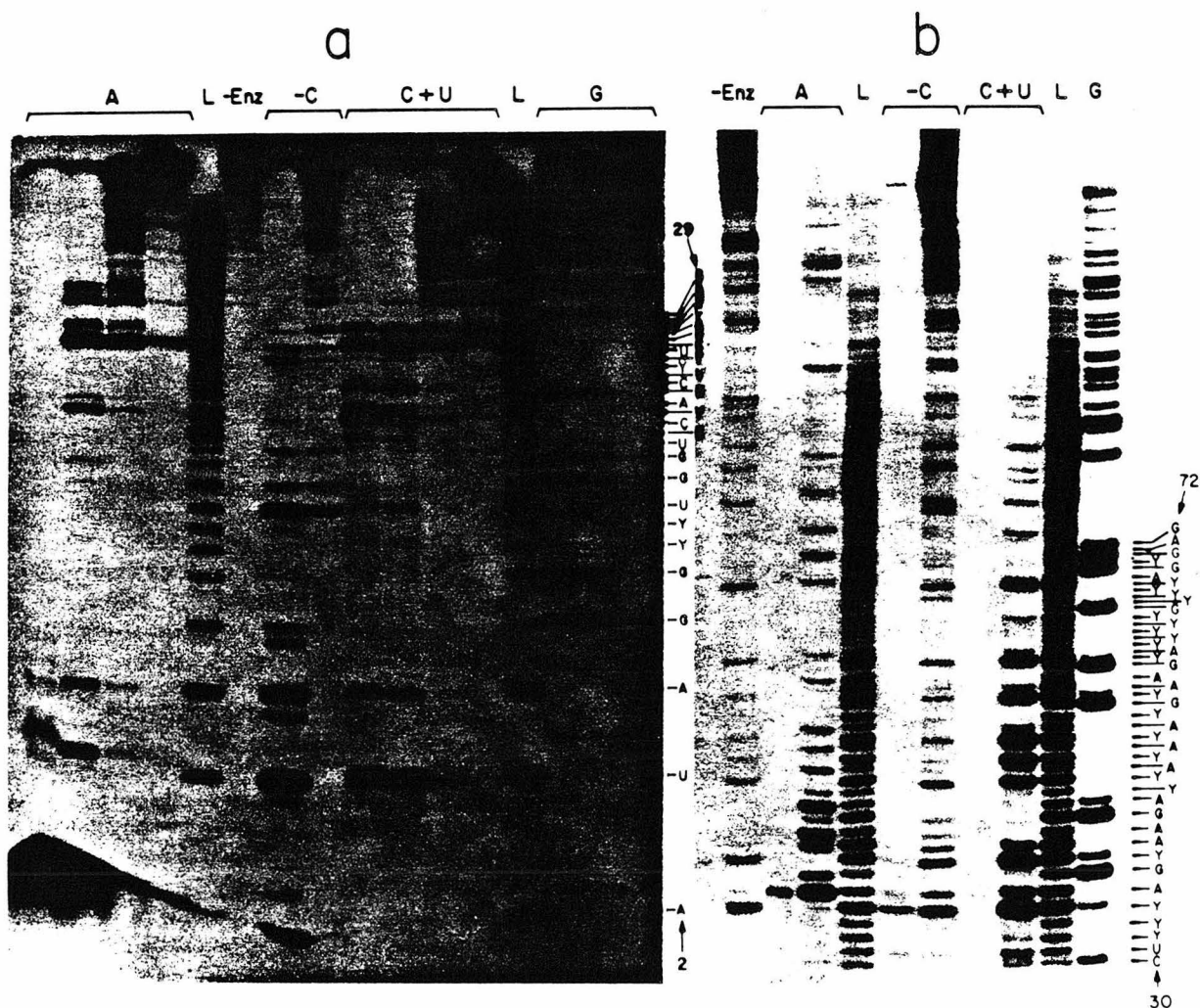


Figure 2. RNA Sequencing Gels of Mitochondrial 12S rRNA

(a) Autoradiogram of 25% sequencing gel (showing nucleotides 1–30) of partial enzymatic digestions of 5' end-labeled 12S rRNA. Shown at the top of the gel are the cleavage specificities: RNAase U2 (A), RNAase Phy I (–C), RNAase A (CU), RNAase T1 (G), formamide ladder (L) and RNA incubated without enzyme (–Enz).

(b) Autoradiogram of 10% sequencing gel (nucleotides 26–71) of the same digests as in (a). The numbering designation refers to the positions of the nucleotides in the longer, less abundant RNA species.

occurrence of one single species; the first 32 nucleotides of the sequence determined are shown in Figure 3. The results obtained by us on the 12S mitochondrial rRNA (ultimately confirmed by sequencing the complementary section of the mit-DNA) and the 5S rRNA (data not shown) have confirmed the essential features of the RNA sequencing method previously described by others (Donis-Keller et al., 1977; Simoncits et al., 1977). Thus RNAase T1 cuts regularly and exclusively after Gs, RNAase U2 cuts regularly and exclusively after As and RNAase A cleaves only after pyrimidines but fails to cleave efficiently after some pyrimidines, particularly when present in pyrimidine stretches; Phy I can effectively distinguish between C and U. The results of the last two reactions are, in general, the most difficult to interpret, and in many cases in the present work we could not identify with

certainly the pyrimidine involved. Since our aim was to align the RNA sequence with the corresponding DNA sequence (which generally can be derived more accurately), however, an absolute identification of the pyrimidines by this method was unnecessary, and we did not attempt to make an assignment unless the sequence was absolutely unambiguous. Control samples (incubated without enzyme) showed preferential cleavage between a pyrimidine and an A.

#### DNA Sequence of the Fragment $\Delta 8a^{\text{Hae}}$

Hpa II fragment 8 was cleaved with Hae III and the 600 nucleotide pair fragment, designated  $\Delta 8a^{\text{Hae}}$  (Ojala and Attardi, 1978), was purified. Two restriction enzymes, Alu I and Mbo I, cut  $\Delta 8a^{\text{Hae}}$  into five fragments and two fragments, respectively. These fragments, as detailed in the sequencing strategy shown



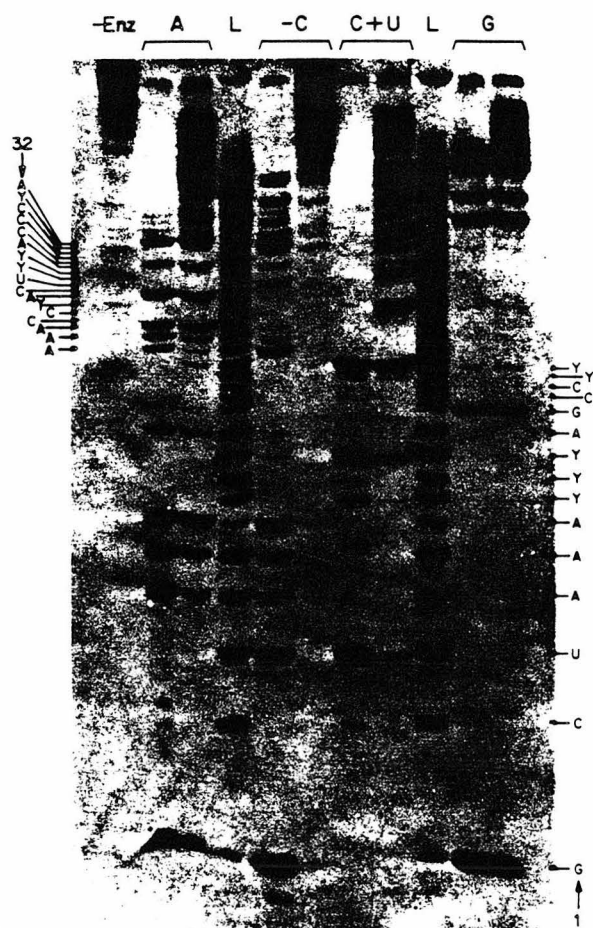


Figure 3. RNA Sequencing Gel of Mitochondrial 16S rRNA  
Autoradiogram of 25% sequencing gel (showing nucleotides 1–32) of partial enzymatic digests of 5' end-labeled 16S rRNA. Gel track designations are as shown in Figure 2.

in Figure 5, were used to sequence the entire  $\Delta 8a$  fragment. The sequence of  $\Delta 8a^{\text{Hae}}$  and the 12S rRNA sequence aligned with it are shown in Figure 6. For the entire stretch of approximately 70 nucleotides which has been sequenced, the 12S rRNA is colinear with the DNA sequence. The 5' end of the RNA corresponds to a residue in the H strand at 286 nucleotides from the Hpa II site between Hpa II fragments 3 and 8. It should be noted that the 5' end lies in a stretch of three As, indicating that the 5' terminus of the minor 12S rRNA species is an A.

#### Identification of the tRNA<sup>Phe</sup> Gene

As described earlier and diagrammed in Figure 1, electron microscopic studies of hybrids formed between mit-DNA and mitochondrial 4S RNAs labeled with ferritin had revealed that near the 5' end of the 12S rRNA coding sequence on the H strand there is a 4S RNA site. It was unknown whether this was a tRNA or some other co-sedimenting low molecular

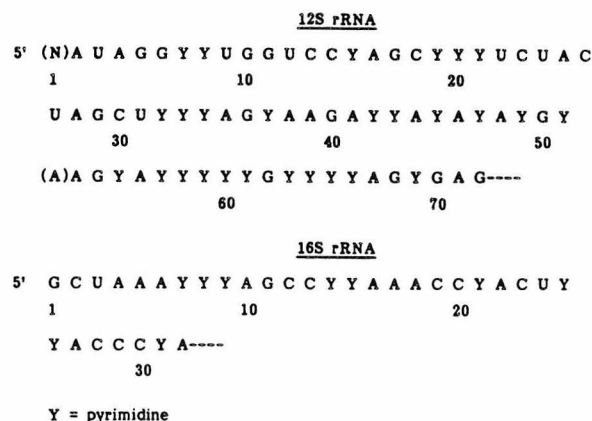


Figure 4. Nucleotide Sequence of the 5' End Regions of Mitochondrial 12S and 16S rRNAs

(Top) Sequence of mitochondrial 12S rRNA derived from Figure 3. (N) pertains to the minor rRNA species that is one nucleotide longer than the major species. (Bottom) Sequence of mitochondrial 16S rRNA derived from Figure 2. (Y) pyrimidine.

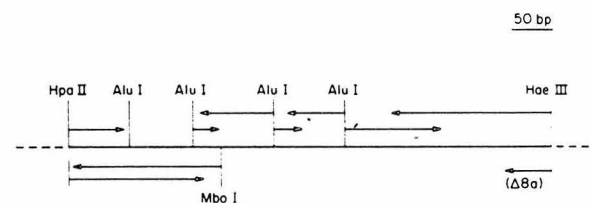
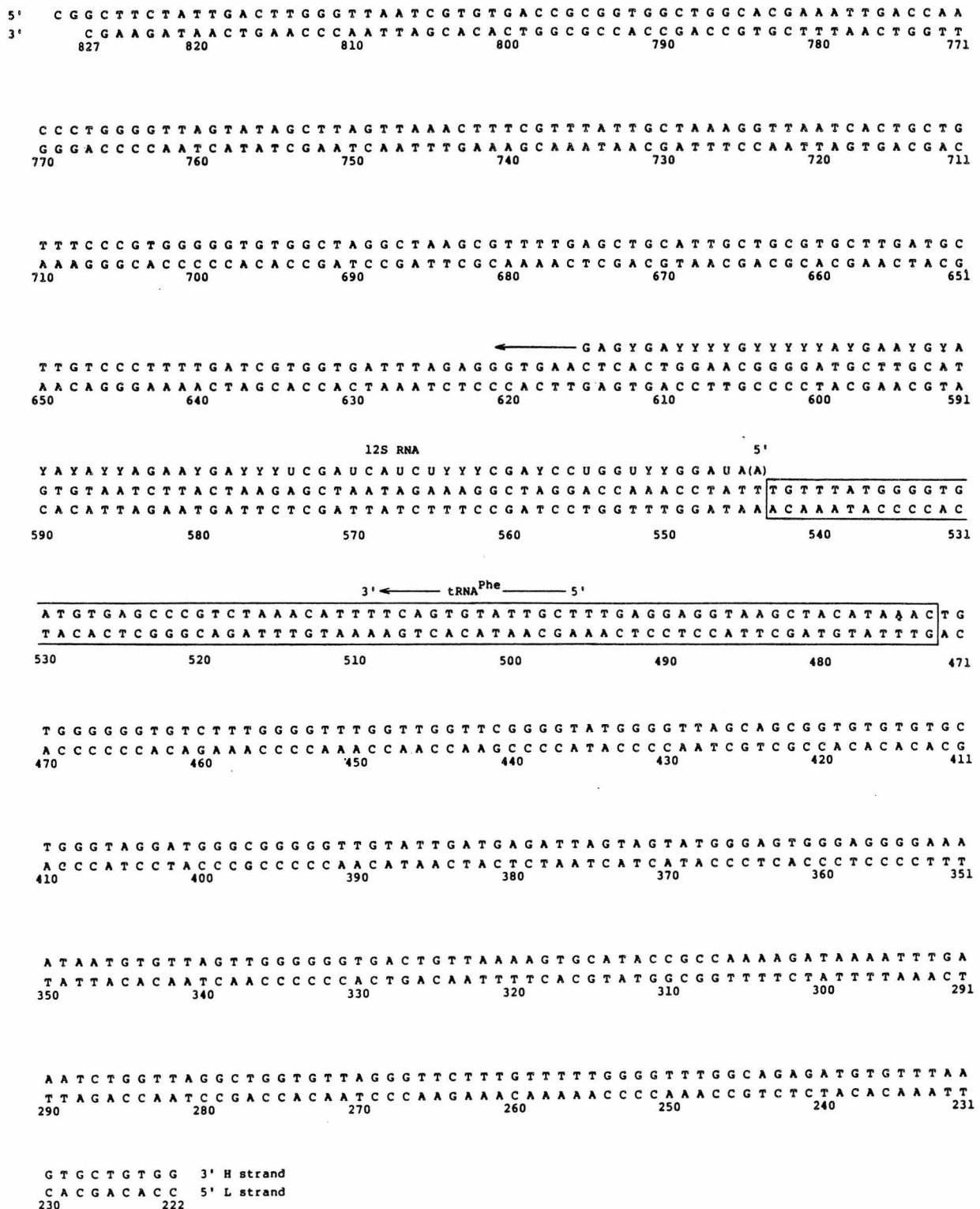


Figure 5. DNA Sequencing Strategy and Restriction Map of Fragment  $\Delta 8a^{\text{Hae}}$

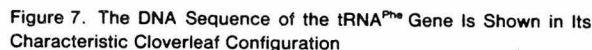
The four Alu I and one Mbo I fragment derived from fragment  $\Delta 8a^{\text{Hae}}$  that were sequenced are shown, with an arrow indicating the extent to which each of the separated single strands was read. Fragment  $\Delta 8a^{\text{Hae}}$  was also labeled at its ends, its strands were separated and the L strand was sequenced for a short distance from its 5' end. The specificity of the strands of the individual fragments was established by sequencing through restriction sites, by determining the homology to the 12S rRNA sequence or by comparison to the sequence of the end of the L strand of  $\Delta 8a^{\text{Hae}}$ .

weight RNA (perhaps a mitochondrial equivalent to the 5S rRNA, so far undiscovered in animal cell mitochondria). Scanning the DNA sequence for a cloverleaf structure revealed the tRNA shown in Figure 7, which has an anticodon corresponding to phenylalanine. Although this tRNA has an overall structure very similar to that of other procaryotic and eucaryotic tRNAs, it possesses several distinctive features which will be discussed below (see Discussion). As shown in Figure 6, the 5' end of the tRNA coding sequence corresponds to residue 473 on the H strand, and the 3' end of the tRNA gene to residue 543. Thus the tRNA<sup>Phe</sup> gene sequence is immediately adjacent to the 5' end of the 12S rRNA.

The sequence from residue 245 to residue 309 (65 nucleotides) can also be folded in a cloverleaf structure; however, this structure shows significant deviations from a standard tRNA pattern.

Figure 6. DNA Sequence of Fragment  $\Delta 8a^{Hae}$ 

The complete sequence of  $\Delta 8a^{Hae}$  is shown together with the 12SrRNA aligned with it. The boxed region represents the tRNA<sup>Phe</sup> sequence. The numbering designation starts from the 5' end nucleotide of the L strand of Hpa II fragment 8 (Crews et al., 1979). The C residue in the 12S rRNA sequence at position 569 does not correspond to the T found in the DNA sequence. This discrepancy may be due to an anomaly in either the RNA or DNA sequencing technique.



There is good evidence to support the conclusion that in the present work the true 5' terminal segments of the two mitochondrial rRNAs have been sequenced and, in the case of the 12S rRNA, aligned with the DNA sequence. In the first place, the data presented here indicate that the 5' end of the 12S rRNA lies at 286 nucleotides from the Hpa II site between Hpa II fragments 3 and 8. This is in excellent agreement with the results of previous experiments (involving RNA-DNA hybridization between labeled fragment 8 and 12S rRNA, S1 nuclease treatment of the hybrids and electrophoretic analysis of the hybridized DNA sequences), which indicated that the 5' end of the 12S rRNA is approximately 270 nucleotides from the above-mentioned Hpa II site (Ojala and Attardi, 1980). Furthermore, the observation that the 5' terminal nucleotide of in vitro 5' end-labeled 12S and 16S rRNAs is A and G, respectively, agrees perfectly with the data obtained with in vivo labeled RNA subjected to complete nuclease digestion and fractionation by two-dimensional high voltage electrophoresis. The micro-

A comparison of the DNA sequence corresponding to the 12S rRNA which has been determined in the present work (286 nucleotides) with that corresponding to an *E. coli* 16S rRNA (Brosius et al., 1978) has revealed significant homologies (Figure 8). Thus starting at 44 nucleotides from the 5' end of the 12S rRNA and proceeding towards the 3' end there is a stretch of 11 nucleotides homologous to the 11 nucleotide stretch at positions 51-61 in *E. coli* 16S rRNA. Though statistically not very significant ( $P < 0.10$ ), this homology is probably meaningful because of the similar positions in the two molecules of the nucleotide stretches involved. Proceeding further towards the 3' end, there is in 12S rRNA a 26 nucleotide stretch at positions 150-175 from the 5' end, which shows a 21/26 homology ( $P < 0.0015$ ) to the stretch at positions 333 to 358 in the 16S rRNA, and, still further, a 29 nucleotide stretch at positions 234-262 from the 5' end, which exhibits a 23/29 homology ( $P < 0.0005$ ) to nucleotides 508-536 in 16S rRNA. No such regions of homology have been found between the *E. coli* 16S rRNA and the portion of Hpa II fragment 8 not coding for the 12S rRNA (543 nucleotide pairs). In view of the lack of sequence data on the small rRNAs from other sources, it would not be justifiable to derive evolutionary implications from the significant homologies detected here between the nucleotide sequences of the human and *E. coli* small rRNA species. It seems probable, however, that such homologies, as well as others that may exist in other portions

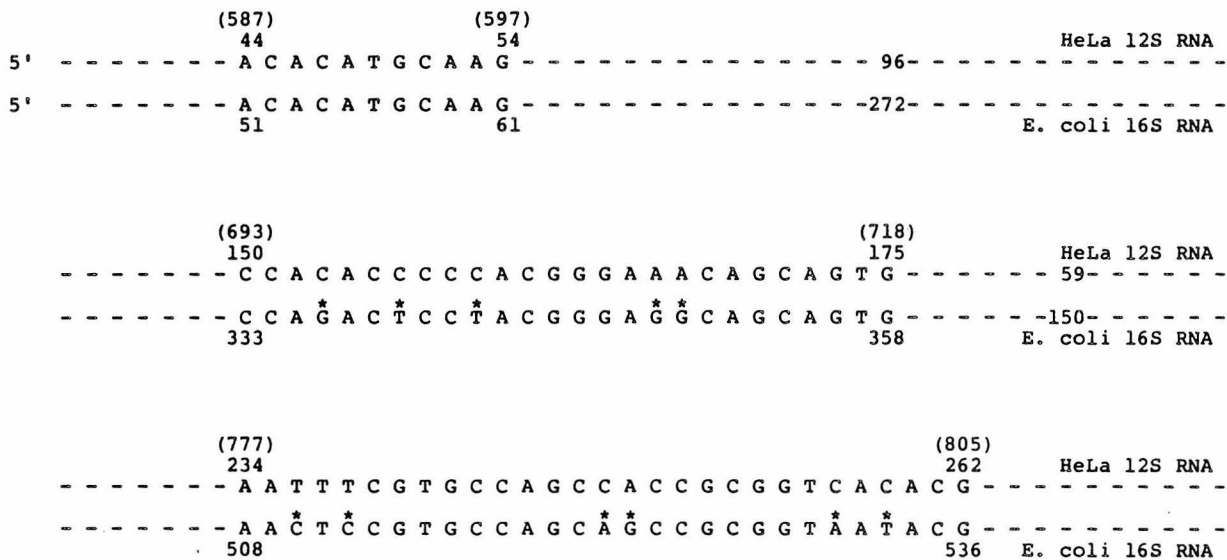


Figure 8. Homology of Segments of Human 12S rRNA and E. coli 16S rRNA

The sequences shown are those of the corresponding segments of the noncoding DNA strands. The numbers not in brackets refer to the nucleotide distance from the 5' terminus of each RNA. The numbers in parentheses above the 12S rRNA sequence correspond to the designation used in Figure 6. The asterisks indicate the nonhomologous nucleotides.

of the small rRNA molecule and in the large rRNA species, underlie the functional similarities existing between mitochondrial and bacterial ribosomes, such as those concerning the sensitivity to antibiotics and the mechanism of initiation of protein synthesis (Borst and Grivell, 1971; Bianchetti et al., 1977).

The immediate juxtaposition of the tRNA<sup>Phe</sup> and 12S rRNA coding sequences in Hpa.II fragment 8 raises interesting questions concerning the mechanisms of synthesis and processing of mitochondrial rRNA species in HeLa cells. In all the rRNA systems analyzed so far, an rRNA precursor has been identified which contains extra sequences at the 5' end of the small rRNA component (Ginsburg and Steitz, 1975; Fedoroff, 1979; DeBoer, Gilbert and Nomura, 1979; Gilbert, de Boer and Nomura, 1979). If this rule applies also to the HeLa cell mitochondrial rDNA transcription unit, all or a portion of the tRNA<sup>Phe</sup> sequence must be a part of the mitochondrial rRNA precursor. It is conceivable that the tRNA<sup>Phe</sup> is processed out of the same precursor by a mechanism involving a precise endonucleolytic cleavage between the two sequences. Alternatively, the mature tRNA<sup>Phe</sup> could derive from an independent transcript.

The significance of the immediate contiguity of the small rRNA gene and the tRNA<sup>Phe</sup> gene in human mitochondrial DNA is uncertain. It may represent an extreme form of tight packing of genetic information, and it will be interesting to see how general this situation is in other portions of the mitochondrial genome. On the other hand, it is tempting to postulate that the tRNA sequence may perform some essential recognition function in connection with the synthesis or processing of

mitochondrial rRNA (Attardi et al., 1979).

A striking feature of yeast mitochondrial tRNA and structural genes sequenced thus far is that the gene coding sequences are surrounded by stretches of AT-rich nucleotides (>95%). In contrast, no AT-rich stretches are seen adjoining the human mitochondrial 12S rRNA and tRNA<sup>Phe</sup> genes. Instead, adjacent to the tRNA<sup>Phe</sup> gene, reading along the H strand, is a long, highly GT-rich region. The sequence 5' GGGGT 3' is repeated 9 times in Δ8a, including once in the body of the tRNA<sup>Phe</sup> gene and twice in the body of the 12S rRNA gene. Two large repeats (nucleotides 448–463 and 532–547) show 15/16 and 14/16 homology to nucleotides 246–265 and 10/12 homology to each other. The stretch from 532 to 547 encompasses the 5' end of the 12S rRNA gene and the 3' end of the tRNA<sup>Phe</sup> gene, whereas the stretch from 448 to 463 is only nine nucleotides from the 5' end of the tRNA<sup>Phe</sup> gene. Furthermore, one of the "GGGGT" repeats (nucleotides 466–470) lies between the 5' end of the tRNA<sup>Phe</sup> gene and the 448–463 repeat. Functional studies of the DNA region on the 5' side of the 12S rRNA gene will be of interest to see whether it contains multiple promoters, as have been discovered for E. coli rRNA cistrons (Young and Steitz, 1979; de Boer et al., 1979; Gilbert et al., 1979) and possibly Xenopus rRNA cistrons (Boseley et al., 1979), and whether such an organization might allow high rates of mitochondrial rRNA transcription.

The tRNA sequence situated in a position immediately adjacent to the 5' end of the 12S rRNA sequence could be folded in a typical cloverleaf structure and was identified as to its amino acid specificity. Exami-

nation of the DNA sequence indicates that there is no intervening sequence in the body of the tRNA gene and no -CCA at the 3' end of the coded sequence; similar observations have been made for the yeast mitochondrial tRNA genes studied thus far (Li and Tzagoloff, 1979; Bos et al., 1979; Martin et al., 1979). The primary sequence of the putative human mitochondrial tRNA<sup>Phe</sup> reveals a substantial agreement with the pattern considered to be "invariant" in all nonmitochondrial tRNAs (Gauss, Gruter and Sprinzl, 1979); there are, however, some striking deviations. Thus the D loop contains the typical A at residue 14 [which may interact with the U at position 8 as in other tRNAs (Quigley and Rich, 1976)], but has a C at position 15 instead of the "invariant" purine and lacks the sequence GG found so far in all tRNAs active in protein synthesis [the *Neurospora crassa* mitochondrial initiator methionine tRNA being an exception (Heckman et al., 1978)]. The "invariant" nucleotide found at the end of the D loop and the three "invariant" nucleotides in the anticodon loop are present, as are the constant pyrimidines (C in the present cases) located in position 11 (base-paired to the "invariant" purine) and at the end of the variable loop. Presumably, the latter C is not able to interact with the C15 residue (by "transpairing" between the two bases) as in "typical" tRNAs (Kim et al., 1974; Robertus et al., 1974). The most interesting structural deviations occur in the T $\psi$ C loop. The loop has a short stem, but probably has adequate stability because of the three GC base pairs forming the stem. The highly conserved T $\psi$ C sequence in the T $\psi$ C loop is unrecognizable in the human mitochondrial tRNA<sup>Phe</sup> sequence. In this tRNA, as in the *N. crassa* mitochondrial initiator methionine tRNA, which has a less extensive modification in the T $\psi$ C loop (Heckman et al., 1978), it is unclear whether the D loop and the T $\psi$ C loop interact with each other, as is the case in all other functional tRNAs, and, if so, what the details of this interaction would be. Other human mitochondrial tRNAs identified from DNA sequencing analysis likewise show profound deviations from the standard tRNA primary sequence and appear similar in length and overall structure to the tRNA<sup>Phe</sup> sequence identified here (B. Barrell and F. Sanger, personal communication). In addition, the yeast and *Neurospora crassa* mitochondrial tRNAs (Heckman et al., 1978; Heckman, Alzner-Deweerd and RajBhandary, 1979) analyzed thus far depart in one or more features from the standard pattern, although in general less than the human tRNAs.

A comparison of the sequences of all tRNAs<sup>Phe</sup> analyzed thus far (Martin et al., 1978; Gauss et al., 1979) reveals that the human mitochondrial tRNA<sup>Phe</sup> has considerably less homology with all of other tRNAs<sup>Phe</sup> (39–54%) than they generally show to each other (54–96%). The degree of homology is approximately the same between the human mitochondrial tRNA<sup>Phe</sup>, on the one hand, and procaryotic (39–54%)

or eucaryotic (43–40%) tRNAs<sup>Phe</sup>, on the other hand. There are a number of structural features common to all of the tRNAs<sup>Phe</sup> that are also found in the human mitochondrial tRNA<sup>Phe</sup>, including the 5' terminal nucleotide (G) and the 3' terminal CA (before the -CCA stem), the residue in position 9, the first three nucleotide pairs of the D stem and the last nucleotide pair of the anticodon stem. There is a considerable similarity between the yeast and human mitochondrial tRNAs<sup>Phe</sup> in the sequence of the anticodon stem, which consists entirely or almost entirely of AU or UU pairs.

#### Experimental Procedures

##### Materials

Micrococcal nuclease was purchased from Sigma, Bacterial alkaline phosphatase from Bethesda Research Laboratories and polynucleotide kinase from either P-L Biochemicals or Boehringer-Mannheim. RNAases used for sequencing were from the following sources: A (Sigma), T1 (Sigma), U2 (Sankyo), Phy I (P-L) and P1 (Calbiochem). All restriction enzymes were from New England Biolabs.

##### Isolation of Mitochondrial rRNA

$2.0 \times 10^9$  HeLa cells, at a concentration of  $9.0 \times 10^5$  cells per ml, were labeled for 2.5 hr with  $^{32}\text{P}$ -orthophosphate (35  $\mu\text{Ci/ml}$ ) in modified Eagle's medium containing  $10^{-4}$  M phosphate and supplemented with 5% dialyzed serum. A postnuclear supernatant was prepared as previously described (Attardi et al., 1969). EDTA was added to a concentration of 0.04 M and the mitochondria were pelleted. The pellet was dissolved gently in 250 mM sucrose, 10 mM Tris-HCl (pH 8.0), 10 mM  $\text{MgCl}_2$  (1 ml per  $1.5 \times 10^8$  cells) and allowed to sit at 20°C for 1 min.  $\text{CaCl}_2$  was added to a final concentration of 3 mM and micrococcal nuclease to 250 U/ml. The mixture was incubated at 20°C for 15 min and then EGTA was added to 6 mM. After 1 min at 20°C, cold 250 mM sucrose, 10 mM Tris-HCl (pH 6.7), 10 mM EDTA was added, the mitochondria were pelleted once, resuspended in 1 ml/ $1.5 \times 10^8$  cells of 150 mM NaCl, 10 mM Tris-HCl (pH 7.4), 1 mM EDTA, incubated for 2 min with 100  $\mu\text{g/ml}$  pronase at 20°C and then lysed with 1% SDS. After incubation at 20°C for 30 min, the RNA was extracted with phenol:chloroform:isoamyl alcohol (50:50:1). The RNA was passed over oligo(dT)-cellulose once and the unbound fraction was run on a preparative agarose (1.65%)/methylmercuric hydroxide slab gel (Amalric et al., 1978). The bands corresponding to the 12S and 16S rRNA species were identified by autoradiography of the wet gel or by ethidium bromide staining, after prior removal of the methylmercuric hydroxide from the gel by reaction with 250 ml of 20 mM dithiothreitol, and cut out of the gel. The RNA species were eluted at 37°C for 15 hr in 1 mM Tris-HCl (pH 7.4), 1 mM EDTA, 1% SDS, and then ethanol-precipitated after the addition of sodium acetate to 0.3 M. Recovery was approximately 70–75%.

##### Isolation of Human 5S rRNA

RNA was extracted from cytoplasmic free polysomes isolated from HeLa cells as previously described (Murphy and Attardi, 1973). After sucrose gradient fractionation to remove 18S and 28S rRNAs, the 4–5S peak was run on a 10% polyacrylamide (30:1 bisacrylamide)/7 M urea gel and the 5S rRNA band was excised, eluted and ethanol-precipitated.

##### 5' End Labeling of RNA

Each purified RNA was dissolved in 10  $\mu\text{l}$  of 10 mM Tris-HCl (pH 8.0) and incubated for 15 min at 65°C with 0.25 units of bacterial alkaline phosphatase. Nitrilotriacetic acid was then added to 5 mM and the mixture incubated at 20°C for 30 min. After addition of 0.5 nmole of  $\gamma$ - $^{32}\text{P}$ -ATP in 10  $\mu\text{l}$  water, 0.1 vol of 10 $\times$  kinase buffer [ $1 \times = 50$  mM Tris-HCl (pH 8.0), 10 mM  $\text{MgCl}_2$ , 5 mM DTT] and 2 units of polynucleotide kinase, the reaction was carried out for 5 min at 37°C.



The enzyme reaction was stopped by the addition of 0.5 M ammonium acetate, carrier tRNA was added and the mixture was ethanol-precipitated and run on an agarose (1.65%)/methylmercuric hydroxide gel. After removal of the methylmercuric hydroxide and autoradiography of the gel, the band corresponding to the end-labeled RNA was excised and eluted from the gel.

#### RNA Sequencing

RNA sequencing was performed by base-specific cleavage of 5' end-labeled RNA followed by electrophoresis on sequencing gels. The reaction conditions described by Donis-Keller et al. (1977) were utilized for the RNAase T1 (G-specific), RNAase A (CU-specific) and RNAase U2 (A-specific) reactions. The Phy I RNAase (AUG-specific) reaction and the hot formamide degradation for the formation of the ladder were performed as described by Simoncsits et al. (1977). Electrophoresis was carried out on thin (0.5 mm) 25% polyacrylamide (30:1 bisacrylamide)/7 M urea gels for nucleotides 1–30 and on 10% polyacrylamide (20:1 bisacrylamide)/7 M urea gels for nucleotides 26–72.

The 5' terminal nucleotide was determined by exhaustive digestion of the in vitro labeled RNA with nuclease P1 (Fujimoto, Kuninaka and Yoshino, 1974) and fractionation of the products on PEI-cellulose TLC plates. After one-dimensional development with 1 M LiCl (Wurst, Vournakis and Maxam, 1978) and autoradiography, the nucleotide composition was determined as described by Crews et al. (1979).

#### DNA Sequencing

Hpa II fragment 8 of human mit-DNA (Ojala and Attardi, 1977) was purified and then cleaved with Hae III in combination with either Alu I or Mbo I as required, and the resulting fragments were fractionated on a 5% polyacrylamide (30:1 bisacrylamide) gel in 50 mM Tris-borate, 5 mM MgCl<sub>2</sub> (Maniatis, Jeffrey and van de Sande, 1975). The purified fragments desired for sequencing were eluted from the gel, dephosphorylated at 65°C with alkaline phosphatase and kinased with  $\gamma$ -<sup>32</sup>P-ATP. The strands of the end-labeled fragments were separated on polyacrylamide gels run in the cold as described previously (Crews et al., 1979). The DNA fragments were sequenced by the method of Maxam and Gilbert (1977) and the reaction products were fractionated on thin 25% (nucleotides 2–30), 10% (25–90) and 8% (75–200) polyacrylamide (20:1 bisacrylamide)/7 M urea gels.

#### Acknowledgments

These investigations were supported by two grants from the NIH. We are grateful to M. Albring for his suggestion to use micrococcal nuclease and to C. Merkel for his help in the first experiments. We are also indebted to J. Posakony, who carried out the computer analysis of the  $\Delta 8a^{100}$  fragment and the sequence comparison with the E. coli 16S rRNA and with whom we have had valuable discussions. The help of R. Murphy and A. Chomyn in writing the computer programs and the technical assistance of A. Drew are gratefully acknowledged.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received October 15, 1979; revised December 17, 1979

#### References

- Amalric, F., Merkel, C., Gelfand, R. and Attardi, G. (1978). Fractionation of mitochondrial RNA from HeLa cells by high-resolution electrophoresis under strongly denaturing conditions. *J. Mol. Biol.* 118, 1–25.
- Angerer, L., Davidson, N., Murphy, W., Lynch, D. and Attardi, G. (1976). An electron microscope study of the relative positions of the 4S and ribosomal RNA genes in HeLa cell mitochondrial DNA. *Cell* 9, 81–90.
- Attardi, B., Cravioto, B. and Attardi, G. (1969). Membrane-bound

- ribosomes in HeLa cells. I. Their proportion to total cell ribosomes and their association with messenger RNA. *J. Mol. Biol.* 44, 47–70.
- Attardi, G., Cantatore, P., Ching, E., Crews, S., Gelfand, R., Merkel, C. and Ojala, D. (1979). The organization of the genes in the human mitochondrial genome and their mode of transcription. In *Extrachromosomal DNA, ICN-UCLA Symposia on Molecular and Cellular Biology*, 15, D. Cummings, P. Borst, I. Dawid, S. Weissman and C. F. Fox, eds. (New York: Academic Press), pp. 443–469.
- Batley, J. and Clayton, D. A. (1978). The transcription map of mouse mitochondrial DNA. *Cell* 14, 143–156.
- Berk, A. J. and Sharp, P. A. (1977). Sizing and mapping of early adenovirus mRNAs by gel electrophoresis of S1 endonuclease-digested hybrids. *Cell* 12, 721–732.
- Berk, A. J. and Sharp, P. A. (1978). Spliced early mRNAs of simian virus 40. *Proc. Nat. Acad. Sci. USA* 75, 1274–1278.
- Bianchetti, R., Lucchini, G., Corsti, P. and Tortora, P. (1977). Dependence of mitochondrial protein synthesis initiation on formylation of the initiator methionyl-tRNA. *J. Biol. Chem.* 252, 2519–2523.
- Borst, P. and Grivell, L. A. (1971). Mitochondrial ribosomes. *FEBS Letters* 13, 73–88.
- Bos, J. L., Osinga, K. A., Van der Horst, G. and Borst, P. (1979). Nucleotide sequence of the mitochondrial structural genes for cysteine-tRNA and histidine-tRNA. *Nucl. Acids Res.* 6, 3255–3266.
- Boseley, P., Moss, T., Mächler, M., Portmann, R. and Birnstiel, M. (1979). Sequence organization of the spacer DNA in a ribosomal gene unit of *X. laevis*. *Cell* 17, 19–31.
- Brosius, J., Palmer, M. L., Kennedy, P. J. and Noller, H. F. (1978). Complete nucleotide sequence of a 16S ribosomal RNA gene from *Escherichia coli*. *Proc. Nat. Acad. Sci. USA* 75, 4801–4805.
- Brown, W. M. and Vinograd, J. (1974). Restriction endonuclease cleavage maps of animal mitochondrial DNAs. *Proc. Nat. Acad. Sci. USA* 71, 4617–4621.
- Crews, S., Ojala, D., Posakony, J., Nishiguchi, J. and Attardi, G. (1979). Nucleotide sequence of a region of human mitochondrial DNA containing the precisely identified origin of replication. *Nature* 277, 192–198.
- de Boer, H. A., Gilbert, S. F. and Nomura, M. (1979). DNA sequences of promoter regions for rRNA operons *rrnE* and *rrnA* in *E. coli*. *Cell* 17, 201–209.
- Donis-Keller, H., Maxam, A. M. and Gilbert, W. (1977). Mapping adenines, guanines, and pyrimidines in RNA. *Nucl. Acids Res.* 4, 2527–2538.
- Fedoroff, N. V. (1979). On spacers. *Cell* 16, 697–710.
- Forget, B. G. and Weissman, S. M. (1967). Nucleotide sequence of KB cell 5S RNA. *Science* 158, 1695–1699.
- Fujimoto, M., Kuninaka, A. and Yoshino, H. (1974). Substrate specificity of nuclease P1. *Agr. Biol. Chem.* 38, 1555–1561.
- Gauss, D. H., Gruter, F. and Sprinzl, M. (1979). Compilation of tRNA sequences. *Nucl. Acids Res.* 6, r1–r19.
- Gilbert, S. F., de Boer, H. A. and Nomura, M. (1979). Identification of initiation sites for the in vitro transcription of rRNA operons *rrnE* and *rrnA* in *E. coli*. *Cell* 17, 211–224.
- Ginsburg, D. and Steitz, J. A. (1975). The 30S ribosomal precursor RNA from *Escherichia coli*. A primary transcript containing 23S, 16S and 5S sequences. *J. Biol. Chem.* 250, 5647–5654.
- Hahn, U., Lazarus, C. M., Lünsdorf, H. and Kuntzel, H. (1979). Split gene for mitochondrial 24S ribosomal RNA of *Neurospora crassa*. *Cell* 17, 191–200.
- Hatlen, L. E., Amaldi, F. and Attardi, G. (1969). Oligonucleotide pattern after pancreatic ribonuclease digestion and the 3' and 5' termini of 5S ribonucleic acid from HeLa cells. *Biochemistry* 8, 4989–5005.
- Heckman, J. E. and RajBhandary, U. L. (1979). Organization of tRNA and rRNA genes in *N. crassa* mitochondria: intervening sequence in the large rRNA gene and strand distribution of the RNA genes. *Cell* 17, 583–595.

- Heckman, J. E., Alzner-Deweerd, B. and RajBhandary, U. L. (1979). Interesting and unusual features in the sequence of *Neurospora crassa* mitochondrial tyrosine transfer RNA. *Proc. Nat. Acad. Sci. USA* 76, 717-721.
- Heckman, J. E., Hecker, L. I., Schwartzbach, S. D., Barnett, W. E., Baumstark, B. and RajBhandary, U. L. (1978). Structure and function of initiator methionine tRNA from the mitochondria of *Neurospora crassa*. *Cell* 13, 83-95.
- Kim, S. H., Suddath, F. L., Quigley, G. J., McPherson, A., Sussman, J. S., Wang, A. H. J., Seeman, N. C. and Rich, A. (1974). Three-dimensional tertiary structure of yeast phenylalanine transfer RNA. *Science* 185, 435-440.
- Klukas, C. K. and Dawid, I. B. (1976). Characterization and mapping of mitochondrial ribosomal RNA and mitochondrial DNA in *Drosophila melanogaster*. *Cell* 9, 615-625.
- Lewin, B. (1976). Order and spacing of ribosomal RNA genes. *Nature* 260, 574-576.
- Li, M. and Tzagoloff, A. (1979). Assembly of the mitochondrial membrane system: sequences of yeast mitochondrial valine and an unusual threonine tRNA gene. *Cell* 18, 47-53.
- Maniatis, T., Jeffrey, A. and van de Sande, H. (1975). Chain length determination of small double- and single-stranded DNA molecules by polyacrylamide gel electrophoresis. *Biochemistry* 14, 3787-3794.
- Mannella, C. A., Collins, R. A., Green, M. R. and Lambowitz, A. M. (1979). Defective splicing of mitochondrial rRNA in cytochrome-deficient nuclear mutants of *Neurospora crassa*. *Proc. Nat. Acad. Sci. USA* 76, 2635-2639.
- Martin, R. P., Sibley, A. P., Schneller, J. M., Keith, G., Stahl, A. J. C. and Dirheimer, G. (1978). Primary structure of yeast mitochondrial DNA-coded phenylalanine tRNA. *Nucl. Acids Res.* 5, 4579-4592.
- Martin, N. C., Miller, D. L., Donelson, J. E., Siquardson, C., Hartley, J. L., Moynihan, P. S. and Pham, H. D. (1979). Identification and sequencing of yeast mitochondrial tRNA genes. In *Extrachromosomal DNA, ICN-UCLA Symposia on Molecular and Cellular Biology*, 15, D. Cummings, P. Borst, I. Dawid, S. Weissman and C. F. Fox, eds. (New York: Academic Press), pp. 357-375.
- Maxam, A. M. and Gilbert, W. (1977). A new method for sequencing DNA. *Proc. Nat. Acad. Sci. USA* 74, 560-564.
- Murphy, W. and Attardi, G. (1973). Stability of cytoplasmic messenger RNA in HeLa cells. *Proc. Nat. Acad. Sci. USA* 70, 115-119.
- Ojala, D. and Attardi, G. (1977). A detailed physical map of HeLa cell mitochondrial DNA and its alignment with the positions of known genetic markers. *Plasmid* 1, 78-105.
- Ojala, D. and Attardi, G. (1978). Precise localization of the origin of replication in a physical map of HeLa cell mitochondrial DNA and isolation of a small fragment that contains it. *J. Mol. Biol.* 122, 301-319.
- Ojala, D. and Attardi, G. (1980). Fine mapping of the ribosomal RNA genes of HeLa cell mitochondrial DNA. *J. Mol. Biol.*, in press.
- Quigley, G. J. and Rich, A. (1976). Structural domains of transfer RNA molecules. *Science* 194, 796-806.
- Ramirez, J. L. and Dawid, I. B. (1978). Mapping of mitochondrial DNA in *Xenopus laevis* and *X. borealis*: the positions of ribosomal genes and D-loops. *J. Mol. Biol.* 119, 133-146.
- Robertus, J. D., Ladner, J. E., Finch, J. T., Rhodes, D., Brown, R. S., Clark, B. F. C. and Klug, A. (1974). Structure of yeast phenylalanine tRNA at 3 Å resolution. *Nature* 250, 546-551.
- Sanger, F., Brownlee, G. G. and Barrell, B. G. (1965). A two-dimensional fractionation procedure for radioactive nucleotides. *J. Mol. Biol.* 13, 373-398.
- Simoncsits, A., Brownlee, G. G., Brown, R. S., Rubin, J. R. and Guillely, H. (1977). New rapid gel sequencing method for RNA. *Nature* 269, 833-836.
- Wu, M., Davidson, N., Attardi, G. and Aloni, Y. (1972). Expression of the mitochondrial genome in HeLa cells. XIV. The relative positions of the 4S RNA genes and of the ribosomal RNA genes in mitochondrial DNA. *J. Mol. Biol.* 71, 81-93.
- Wurst, R. M., Vournakis, J. N. and Maxam, A. M. (1978). Structure mapping of 5'-<sup>32</sup>P-labeled RNA with S1 nuclease. *Biochemistry* 17, 4493-4499.
- Young, R. A. and Steitz, J. A. (1979). Tandem promoters direct *E. coli* ribosomal RNA synthesis. *Cell* 17, 225-234.

Chapter 6

A Small Polyadenylated RNA (7S RNA), Containing a Putative Ribosome  
Attachment Site, Maps Near the Origin of Human Mitochondrial DNA Replication

This paper was published in the Journal of Molecular Biology.



Reprinted from *J. Mol. Biol.* (1981) **150**, 303–314

**A Small Polyadenylated RNA (7 S RNA), Containing a  
Putative Ribosome Attachment Site, Maps Near the  
Origin of Human Mitochondrial DNA Replication**

DEANNA OJALA, STEPHEN CREWS, JULIO MONTOYA  
ROBERT GELFAND AND GIUSEPPE ATTARDI

### A Small Polyadenylated RNA (7 S RNA), Containing a Putative Ribosome Attachment Site, Maps Near the Origin of Human Mitochondrial DNA Replication

The light (L) strand sequence of HeLa cell mitochondrial DNA which codes for a small polyadenylated RNA (7 S RNA) has been precisely localized, by mapping and sequencing studies, in a region of the genome which immediately precedes the origin of replication in the direction of L-strand transcription, extending from 219 nucleotides to within 20 nucleotides from this origin. A 5'-end sequencing analysis of 7 S RNA has revealed the presence of two major species differing in size by one nucleotide at this end, and possibly three minor species. The 7 S RNA appears to contain, near its 3'-end, a potential reading frame for a polypeptide 23 or 24 amino acids long. Furthermore, in the 5' non-coding region, an 11-nucleotide long sequence has been identified which is complementary to a sequence near the 3'-end of the 12 S ribosomal RNA, and possibly represents a ribosome attachment site.

A particularly intriguing aspect of the transcription process in HeLa cell mitochondrial DNA is its complete symmetry (Aloni & Attardi, 1971). Both the heavy (H) and light (L) strands of this DNA are transcribed over their entire length (Murphy *et al.*, 1975). This symmetry of transcription has to be contrasted with the markedly asymmetrical distribution of informational content in the two strands. The H-strand codes for most of the relatively stable mtDNA<sup>†</sup> transcripts thus far identified, including the two high molecular weight ribosomal RNA species, 16 S and 12 S rRNA, at least 14 transfer RNA species (as shown by the mtDNA sequence (Barrell *et al.*, 1980)) and the majority of the poly(A)-containing RNA species (Amalric *et al.*, 1978; Ojala *et al.*, 1980a): ten among these, on account of their relative abundance, association with partially purified polysomes and a close correlation with significant reading frames in mtDNA, are presumably specific messenger RNAs for mtDNA coded polypeptides (Attardi *et al.*, 1980a; Montoya *et al.*, 1981).

Most of the transcripts of the L-strand have a much shorter half-life than the H-strand transcripts, and do not accumulate to any significant extent (Aloni & Attardi, 1971). Only a few tRNA species (Lynch & Attardi, 1976; Angerer *et al.*, 1976) and a small polyadenylated RNA species (7 S RNA) (Ojala & Attardi, 1974a) have been recognized as relatively stable and abundant L-strand transcripts.

The 7 S RNA (so designated on account of its sedimentation constant in the native state (Ojala & Attardi, 1974a); corresponding to RNA 18 in the classification

<sup>†</sup> Abbreviation used: mtDNA, mitochondrial DNA.

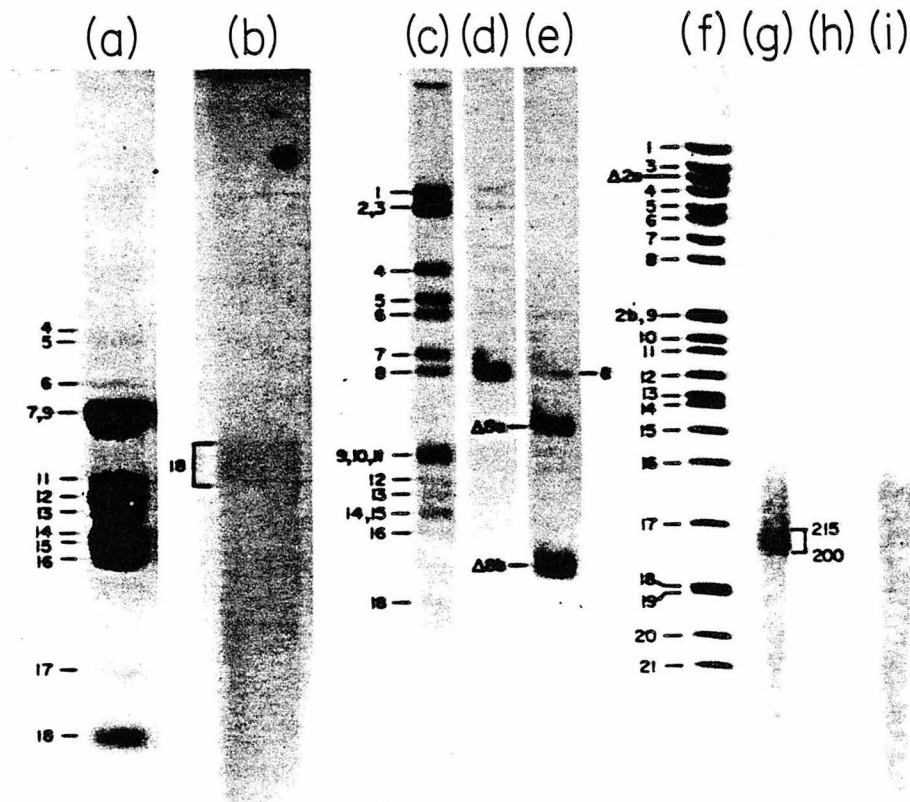


FIG. 1. Isolation, 5'-end labeling and mapping of mitochondrial 7 S RNA. Lane (a), autoradiogram of the oligo(dT)-bound fraction of mitochondrial RNA isolated by the micrococcal nuclease procedure (Crews & Attardi, 1980) from HeLa cells labeled for 2.5 h with [ $^{32}$ P]orthophosphate and fractionated by electrophoresis through a 1.4% agarose/CH<sub>3</sub>HgOH slab gel. Lane (b), rerun, through a 5% polyacrylamide/5 M-urea gel, of 7 S RNA (RNA 18), after elution from the gel track shown in lane (a) and 5'-end labeling with [ $\gamma$ - $^{32}$ P]ATP and polynucleotide kinase (Crews & Attardi, 1980). In the final  $^{32}$ P-labeled 7 S RNA preparation, the contribution of radioactivity from labeling *in vivo* was negligible as compared to the labeling *in vitro*. Lanes (c) to (e), mapping of 7 S RNA by the DNA transfer technique. 7 S RNA was labeled *in vivo* and hybridized with a *Hpa*II plus *Bam*HI mtDNA digest transferred to nitrocellulose paper (Southern, 1975) (lane (d)), or labeled *in vitro* and hybridized with a *Hae*III digest of *Hpa*II fragment 8 transferred to nitrocellulose paper and hybridized with an end-labeled and denatured *Hpa*II mtDNA digest. Lanes (f) to (i), sizing of 7 S RNA by the *S*<sub>1</sub> protection technique (Berk & Sharp, 1977, 1978). The L-strand of *Hpa*II fragment 8 was hybridized with 7 S RNA for 4 h at 66°C in the presence of 0.4 M-NaCl and treated with *S*<sub>1</sub> nuclease, and the protected hybrids were then analyzed by electrophoresis through a 5% polyacrylamide gel in Tris/borate/Mg<sup>2+</sup> buffer as described previously (Ojala *et al.*, 1980a) (lane (g)). The products of control reactions in which the RNA or DNA was omitted are shown in lanes (h) and (i), respectively. Lane (f), size markers provided by an end-labeled *Hpa*II plus *Bam*HI mtDNA digest.

by Amalric *et al.* (1978)) has long been of interest because of its distinctive properties. It is the most abundant poly(A)-containing RNA coded for by mtDNA (Gelfand & Attardi, 1981); its half-life is comparable to that of the putative H-strand coded mRNAs (Attardi *et al.*, 1980b) but, in comparison with these, a smaller proportion of it is found to be associated with polysomes (Ojala & Attardi, 1974b; Amalric *et al.*, 1978). In order to obtain information useful for the understanding of the nature of this L-strand transcript, in the present work, its coding sequence has been precisely localized in HeLa cell mtDNA by mapping and sequencing methods.

As a first step towards localizing the sequence coding for 7 S RNA within the HeLa cell mtDNA *Hpa*II physical map (Ojala & Attardi, 1977), hybridization experiments by the DNA transfer technique (Southern, 1975) were carried out. *In vivo*  $^{32}$ P-labeled 7 S RNA, extracted by the micrococcal nuclease procedure (Crews & Attardi, 1980) and separated by electrophoresis through a 1.4% (w/v) agarose/CH<sub>3</sub>HgOH slab gel (Fig. 1, lane (a)), was incubated with a *Hpa*II plus *Bam*HI digest of HeLa cell mtDNA transferred to a nitrocellulose filter. As shown in Figure 1, lane (d), the RNA hybridized only with *Hpa*II fragment 8, suggesting that its coding sequence is included in this fragment. Further experiments of hybridization of 7 S RNA *in vitro* with the two subfragments of *Hpa*II-8 ( $\Delta$ 8a and  $\Delta$ 8b) produced by *Hae*III cleavage (Ojala & Attardi, 1978) indicated a somewhat stronger hybridization with  $\Delta$ 8b than with  $\Delta$ 8a (Fig. 1, lane (e)): from the position of the *Hae*III site within *Hpa*II-8 and the size of the non-poly(A) portion of the RNZ (estimated to be ~210 nucleotides from its sedimentation rate under

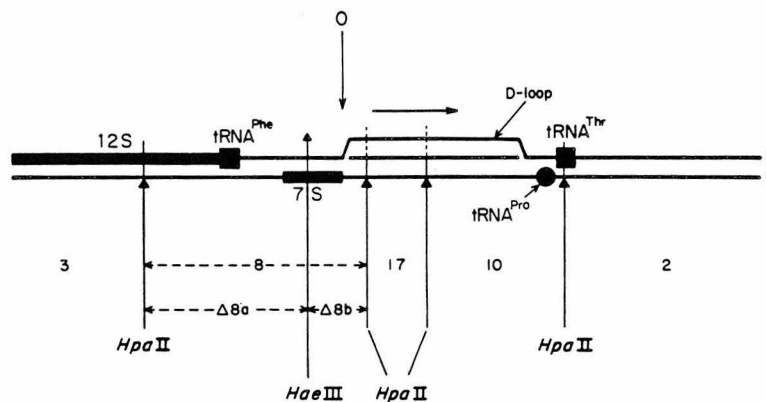


FIG. 2. Region of the HeLa cell mtDNA physical map illustrating the gene organization in the region around the origin of DNA replication. The precise positions of the 12 S rRNA gene on the H-strand and of the tRNA genes on the H-strand (■) and L-strand (●), as derived from DNA and RNA sequence data (Crews & Attardi, 1980; Anderson *et al.*, 1981), the positions of the origin of replication (○) and of the D-loop, as determined by mapping and DNA sequencing analysis (Ojala & Attardi, 1978; Crews *et al.*, 1979), and that of the 7 S RNA coding sequence on the L-strand, as determined in the present work, are shown. (The 12 S rRNA, tRNA and 7 S RNA genes are represented as their anti-sense sequences.) The rightward arrow above the D-loop indicates the direction of H-strand synthesis as well as the direction of transcription of the 7 S RNA coding sequence. The vertical arrows indicate *Hpa*II and *Hae*III cleavage sites.

denaturing conditions (Ojala & Attardi, 1974a)), it was inferred that the coding stretch is located within the half of fragment 8 containing the origin of DNA replication (Fig. 2).

In order to test the colinearity of 7 S RNA and its coding sequence, and in order to obtain another estimate of the size of this RNA, a mapping experiment utilizing the  $S_1$  protection technique (Berk & Sharp, 1977, 1978) was performed. As shown in Figure 1 (lane (g)), a broad band, corresponding to hybrids of 200 to 215 nucleotide pairs in length, was observed against a diffuse background of heterogeneous material. No indication of the presence of intervening sequences was thus found. The size estimate for 7 S RNA obtained here (200 to 215 nucleotides) is in good agreement with the previous determination; the broadness of the band formed by the hybrids may reflect a size heterogeneity of the RNA (see below). The heterogeneous background material is presumably due to the formation of RNA-RNA hybrids, resulting from the presence in the original 7 S RNA preparation of H and L-strand coded complementary RNA sequences: this heterogeneous background was also visible in the “-DNA” control lane (Fig. 1, lane (i)).

In order to localize more precisely the nucleotide stretch coding for 7 S RNA in *Hpa*II fragment 8, a 5'-end proximal segment of this RNA was subjected to sequence analysis. 7 S RNA, isolated as described as above, was labeled at its 5'-end with [ $\gamma$ - $^{32}$ P]ATP and polynucleotide kinase, after treatment with bacterial alkaline phosphatase (Crews & Attardi, 1980). The labeled RNA was further purified by electrophoresis on a 5% (w/v) polyacrylamide/5 M-urea gel in Tris/borate/EDTA buffer (Maniatis *et al.*, 1975) (Fig. 1, lane (b)). The fact that, in the latter gel, the band corresponding to the 5'-end-labeled RNA is broader than the 7 S RNA band in the agarose/ $\text{CH}_3\text{HgOH}$  gel (Fig. 1, lane (a)) presumably is not due to degradative phenomena occurring during the labeling *in vitro*, since a similarly broad band was observed when RNA *in vivo* was run under the same electrophoretic conditions (not shown). The 5'-end-labeled RNA was first characterized as to its 5'-terminal nucleotide by exhaustive digestion with nuclease  $P_1$  and fractionation of the products on PEI-cellulose thin-layer plates, as previously described (Crews & Attardi, 1980). The distribution of total  $^{32}\text{P}$  radioactivity among the four mononucleotides was: AMP, 90.0%; UMP, 4.7%; GMP, 4.1%; CMP, 1.2%. These observations indicated the substantial purity of the 7 S RNA preparation.

For sequence analysis, the 5'-end labeled 7 S RNA was subjected to base-specific partial enzymatic hydrolysis according to the procedures of Donis-Keller *et al.* (1977) and Simoncsits *et al.* (1977), followed by analysis of the digestion products on 25% (w/v) polyacrylamide/7 M-urea gels, as previously described (Crews & Attardi, 1980). The autoradiograms of the sequencing gels (Fig. 3) revealed a complex pattern in which two or, in some cases, three oligonucleotides were observed at the same ladder positions. Similar observations have been made on other mitochondrial RNA species subjected to 5'-end sequence analysis (Crews & Attardi, 1980; Montoya *et al.*, 1981), and have been accounted for by the presence, in the sequencing reactions, of two RNA species differing in size by one nucleotide at the 5'-end. The presence of multiple species differing by one, two, three, etc. nucleotides at the 5'-end produces, in the sequencing gels, a shadowing effect in which a specifically terminated oligonucleotide is found in adjacent ladder positions

## LETTERS TO THE EDITOR

307

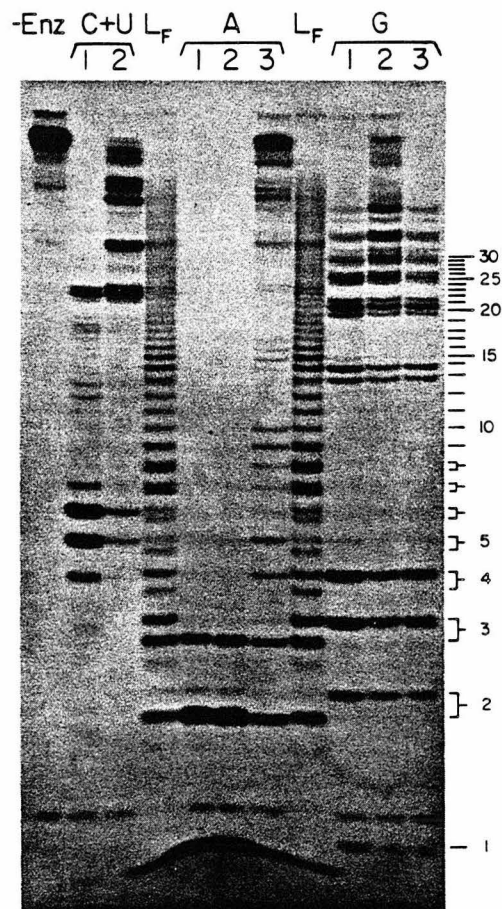


FIG. 3. 5'-end sequencing gel of mitochondrial 7 S RNA. Autoradiogram of a 25% polyacrylamide/7 M-urea gel of partial enzymatic digests of 5'-end-labeled 7 S RNA. Shown at the top of the gel are the cleavage specificities: RNAase A (C+U), RNAase U<sub>2</sub> (A), RNAase T<sub>1</sub> (G), formamide ladder (L<sub>F</sub>) and RNA incubated without enzyme (-Enz). The numbers indicate the ladder positions (see the text for details).

(the number of adjacent positions reflecting the number of species present). In the present case, a comparison with the previously determined DNA sequence of *Hpa*II fragment 8 (Crews *et al.*, 1979; Crews & Attardi, 1980) revealed only one potential coding sequence which could account for the pattern observed in the RNA sequencing gels (Fig. 4), and this sequence is indeed in the general region indicated by the mapping data (Fig. 2). This coding sequence assignment assumes the presence in the 7 S RNA preparation of two major species which differ in size at the 5'-end by one nucleotide (species 2 and 3 in Table 1). An analysis of the gels (Fig. 3) shows an excellent agreement of the ladder positions of the nucleotides of the two postulated major species with the main bands; furthermore, at many of the positions where there is coincidence in migration of identically terminated



FIG. 4. Alignment of the nucleotide sequence of the 5'-end region of 7S RNA with the DNA sequence of *HpaII* fragment 8 (Crews *et al.*, 1979; Crews & Attardi, 1980). The initiator and terminator codons of the putative reading frame of 7S RNA are underlined, and the amino acids corresponding to the individual codons are indicated as one-letter symbols. The underlined sequence at position 238 to 228 is complementary to a sequence at the 3'-end of 12S rRNA. The sequence of the 5'-terminal segment of 7S DNA and its alignment with the *HpaII*-8 sequence (Crews *et al.*, 1979) is also shown. O, origin of mtDNA replication. The hyphens have been omitted for clarity.

oligonucleotides derived from the two species (for example, G<sub>21</sub>, G<sub>26</sub>, A<sub>2</sub>, etc.), the bands show a considerably higher intensity. In the region corresponding to the smaller oligonucleotides, there are some oligonucleotides which do not appear to derive from the two major species (Fig. 3). Their identity is fully compatible with their pertaining to three minor species differing by single nucleotides at the 5'-end from the two major species (species 1, 4 and 5 in Table 1). Although this interpretation is quite plausible and in agreement with the pattern of 5'-end microheterogeneity previously observed in mitochondrial RNAs, the identification of these minor species has to be considered tentative.

TABLE I  
Interpretation of the 7 S RNA 5'-end sequencing gel

	Ladder position number†																													
Molecular species‡	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1?	A	A	A	A	<u>G</u>	A	<u>Y</u>	A	A	A	A	Y	Y	Y	<u>G</u>															
2	A	A	A	G	A	Y	A	A	A	A	Y	Y	Y	G	A	A	A	Y	Y	Y	G	G	Y	Y	A	G	G	Y	Y	G
3	A	A	G	A	Y	A	A	A	A	Y	Y	Y	G	A	A	A	A	Y	Y	Y	G	Y	Y	A	G	G	—	Y	G	
4?	A	G	A	<u>Y</u>	A	A	A	A	—	(Y)	Y	( <u>G</u> )	—	A	A	<u>Y</u>														
5?	<u>G</u>	A	<u>Y</u>	A	A	A	A	(Y)	—	(Y)	—	—	—	A	<u>Y</u>															

† Ladder position in the RNA sequencing gel (Fig. 3), as indicated in that Figure by a tick or bracket.

‡ Molecular species of the 7 S RNA differing in size by 1, 2, 3 or 4 nucleotides at the 5'-end; the identification of species 1, 4 and 5 is tentative (see the text for details).

Y indicates a pyrimidine, a dash (—), an unidentified base, and the parentheses, a tentative identification. In the postulated sequence of each of the minor species (1, 4 and 5) only the underlined nucleotides have been directly recognized as distinct from the nucleotides of the 2 major species, and these sequences have been extended up to the last nucleotide specifically attributable to them.



at position 215 and an A interruption at position 208. A comparison of the HeLa cell mtDNA sequence in the region coding for 7 S RNA (Crews *et al.*, 1979; Crews & Attardi, 1980) with the corresponding region of human placenta mtDNA (Eperon *et al.*, 1980) reveals differences at two locations, i.e. reading on the H-strand, there are, in HeLa mtDNA, three more G nucleotides in the block at positions 214 to 199, and there is a C instead of a T at position 159.

A striking observation is that the 11-nucleotide stretch at positions 238 to 228 has a perfect base complementarity to a sequence very close to the 3'-end of the 12 S rRNA (Fig. 5(a)). In particular, this complementary sequence in the 12 S RNA is located in part in the stem portion of the hairpin structure which can be formed by the 3'-end proximal region of this RNA (Eperon *et al.*, 1980). The latter structure is probably thermodynamically stable at 25°C in a moderate ionic strength environment and at a neutral pH (a  $\Delta G$  value of  $\sim -7$  kcal mol<sup>-1</sup> can be calculated using the rules of Tinoco *et al.* (1973), Gralla & Crothers (1973) and Borer *et al.* (1974)). A similar step-loop structure can be drawn for the small ribosomal subunit RNA from several prokaryotic, eukaryotic cytoplasm and organelle sources (Yuan *et al.*, 1979; Van Etten *et al.*, 1980; Dubin & Baer, 1980): the evolutionary persistence of this stem-loop structure near the 3'-end of the small ribosomal subunit RNA, in spite of the primary sequence divergence, suggests the importance of this secondary structure *per se* in some function related to protein synthesis. In *Escherichia coli* 16 S rRNA, the polypyrimidine sequence C-C-U-C-C, which is considered to be involved in ribosome binding to mRNA (Shine & Dalgarno, 1974; Steitz & Jakes, 1975), is located in part in the stem-loop structure mentioned above. Melting of this

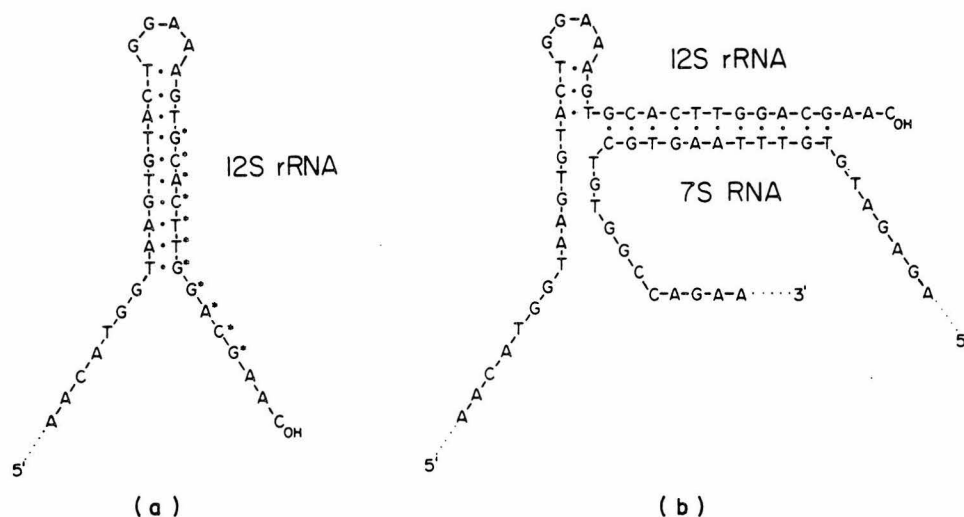


FIG. 5. Postulated pairing between the 3'-end proximal region of 12 S rRNA and 7 S RNA. (a) The 3'-end proximal sequence of human 12 S rRNA, identified by comparison with the 3'-end proximal sequence of the small mitochondrial rRNA from hamster cells (Dubin & Baer, 1980), is shown folded in a hairpin structure (modified from Eperon *et al.*, 1980). The asterisks indicate the nucleotides complementary to the 11-nucleotide stretch in 7 S RNA. (b) The hairpin structure shown in (a) is opened to allow full base-pairing of the 11-nucleotide sequence with the complementary sequence in 7 S RNA.

structure is thus required to expose completely the sequence C-C-U-C-C and thus allow its base-pairing with the complementary sequence in the mRNAs. In human mitochondrial 12 S rRNA, a full base-pairing with the 11-nucleotide complementary sequence found in the 5' non-coding segment of 7 S RNA would likewise require the melting of the hairpin structure near the 3'-end (Fig. 5(b)). It has been suggested that the exchange of base-pairs involved in the transition from *intermolecular* base-pairing (between ribosome and mRNA) of the Shine-Dalgarno sequence to *intramolecular* base-pairing (within the 16 S rRNA) would facilitate the dissociation of the ribosome from the messenger by lowering the energy required to disrupt the mRNA-rRNA interaction (Steitz & Jakes, 1975). The observation that the 12 S RNA sequence complementary to the 11-nucleotide stretch of 7 S RNA is involved in the formation of the hairpin structure near the 3'-end of 12 S rRNA, in a manner which is strikingly similar to the relationship of the Shine-Dalgarno sequence with the stem-loop structure near the 3'-end of *E. coli* 16 S rRNA, is consistent with the idea that the 11-nucleotide stretch of 7 S RNA possibly participates in ribosome binding *in vivo*.

It seems possible that at least a fraction of 7 S RNA functions in HeLa cells as mRNA for a mitochondrially synthesized small polypeptide. The occurrence of a portion of this RNA in a partially purified polysome fraction (Ojala & Attardi, 1974b; Amalric *et al.*, 1978), its abundance, its relative metabolic stability, which is comparable to that of well-characterized mitochondrial mRNAs (Gelfand & Attardi, 1981; Attardi *et al.*, 1980b), and the presence in its sequence of a potential reading frame are all observations which are consistent with the above-mentioned possibility. Among the products *in vivo* of mitochondrial protein synthesis in HeLa cells there is indeed a small polypeptide of ~3500 molecular weight (Ching, 1979; Attardi & Ching, 1979), a size which is fairly close to that expected for a polypeptide encoded in the reading frame found in 7 S RNA (~2900  $M_r$ ). The presence of a possible ribosome attachment site in the 5' non-coding stretch of 7 S RNA would also be consistent with the functional identification of this RNA as a mRNA. The distance between this presumptive ribosome attachment site and the reading frame of 7 S RNA (49 nucleotides) is indeed considerably longer than that which occurs between the stretch complementary to the Shine-Dalgarno sequence and the reading frame in the *E. coli* mRNAs (4 to 10 nucleotides (Steitz & Jakes, 1975)); however, distances up to 38 nucleotides have been reported in eukaryotic mRNAs between the sequence complementary to the 3'-proximal stretch of eukaryotic 18 S rRNA, which had been suggested to have a role in ribosome attachment to some mRNAs (Hagenbüchle *et al.*, 1978), and the reading frame†.

If 7 S RNA indeed functions as mRNA, the existence in its sequence of a long 5' non-coding stretch containing a putative ribosome attachment site would set this RNA apart from all the other HeLa cell mitochondrial mRNAs so far identified. These mRNAs, which are all H-strand coded, either start directly with the initiator codon or have only a few nucleotides preceding this codon (Ojala *et al.*, 1980b; Attardi *et al.*, 1980a; Montoya *et al.*, 1981). It is possible that this structural

† It should be mentioned that the view that complementarity between mRNA and 18 S rRNA plays a role in initiation of translation in eukaryotic cells has been disputed by Kozak (1978).

difference reflects the requirement for a more efficient translation of 7 S RNA as compared to the other mRNAs. Further work is needed to clarify the physiological significance of 7 S RNA. It would be particularly important to know whether the location of the 7 S RNA sequence in the region near the origin of replication, which also contains a main initiation site for L-strand transcription (Cantatore & Attardi, 1980), has any bearing on its function. It is possible that this RNA or a precursor of it may have a role in connection with the initiation of mtDNA replication (as a primer, for example), or with the complete transcription of the L-strand.

These investigations were supported by two grants from the National Institutes of Health (GM-11726 and T32 GM-07616). We are very grateful to Drs F. Sanger and B. Barrell for communicating to us mtDNA sequence data prior to publication. The technical assistance of Ms A. Drew is gratefully acknowledged.

Division of Biology  
California Institute of Technology  
Pasadena, Calif. 91125, U.S.A.

D. OJALA  
S. CREWS  
J. MONTOYA  
R. GELFAND  
G. ATTARDI

Received 7 April 1981

#### REFERENCES

- Aloni, Y. & Attardi, G. (1971). *Proc. Nat. Acad. Sci., U.S.A.* **68**, 1757-1761.
- Amalric, F., Merkel, C., Gelfand, R. & Attardi, G. (1978). *J. Mol. Biol.* **118**, 1-25.
- Anderson, S., Bankier, A. T., Barrell, B. G., de Bruijn, M. H. L., Coulson, A. R., Drouin, J., Eperon, I. C., Nierlich, D. P., Roe, B. A., Sanger, F., Schreier, P. H., Smith, A. J. H., Staden, R. & Young, I. G. (1981). *Nature (London)*, **290**, 457-465.
- Angerer, L., Davidson, N., Murphy, W., Lynch, D. & Attardi, G. (1976). *Cell*, **9**, 81-90.
- Attardi, G. & Ching, E. (1979). In *Methods in Enzymology (Biomembranes)*, vol. 56, pp. 66-79, Academic Press, New York.
- Attardi, G., Cantatore, P., Ching, E., Crews, S., Gelfand, R., Merkel, C., Montoya, J. & Ojala, D. (1980a). In *The Organization and Expression of the Mitochondrial Genome* (Kroon, A. M. & Saccone, C., eds), pp. 103-119, North-Holland, Amsterdam.
- Attardi, G., Cantatore, P., Ching, E., Crews, S., Gelfand, R., Merkel, C., Montoya, J. & Ojala, D. (1980b). In *International Cell Biology 1980-1981* (Schweiger, H. G., ed.), pp. 225-238, Springer-Verlag, Berlin, Heidelberg and New York.
- Barrell, B. G., Anderson, S., Bankier, A. T., de Bruijn, M. H. L., Chen, E., Coulson, A. R., Drouin, J., Eperon, I. C., Nierlich, D. P., Roe, B. A., Sanger, F., Schreier, P. H., Smith, A. J. H., Staden, R. & Young, I. G. (1980). *Proc. Nat. Acad. Sci., U.S.A.* **77**, 3164-3166.
- Berk, A. J. & Sharp, P. A. (1977). *Cell*, **12**, 721-732.
- Berk, A. J. & Sharp, P. A. (1978). *Proc. Nat. Acad. Sci., U.S.A.* **75**, 1274-1278.
- Borer, P. N., Dengler, B., Tinoco, I. Jr & Uhlenbeck, D. C. (1974). *J. Mol. Biol.* **86**, 843-853.
- Cantatore, P. & Attardi, G. (1980). *Nucl. Acids Res.* **8**, 2605-2624.
- Ching, E. P. (1979). Ph.D. thesis, California Institute of Technology, Pasadena.
- Crews, S. & Attardi, G. (1980). *Cell*, **19**, 775-784.
- Crews, S., Ojala, D., Posakony, J., Nishiguchi, J. & Attardi, G. (1979). *Nature (London)*, **277**, 192-198.
- Donis-Keller, H., Maxam, A. M. & Gilbert, W. (1977). *Nucl. Acids Res.* **4**, 2527-2538.

- Dubin, D. T. & Baer, R. J. (1980). In *The Organization and Expression of the Mitochondrial Genome* (Kroon, A. M. & Saccone, C., eds), pp. 231-340, North-Holland, Amsterdam.
- Eperon, I. C., Anderson, S. & Nierlich, D. P. (1980). *Nature (London)*, **286**, 460-467.
- Gelfand, R. & Attardi, G. (1981). *Mol. Cell. Biol.*, in the press.
- Gralla, J. & Crothers, D. M. (1973). *J. Mol. Biol.* **73**, 497-511.
- Hagenbüchle, O., Santer, M., Steitz, J. A. & Mans, R. (1978). *Cell*, **13**, 551-563.
- Kozak, M. (1978). *Cell*, **15**, 1109-1123.
- Lynch, D. C. & Attardi, G. (1976). *J. Mol. Biol.* **102**, 125-141.
- Maniatis, T., Jeffrey, A. & van de Sande, H. (1975). *Biochemistry*, **14**, 3787-3794.
- Montoya, J., Ojala, D. & Attardi, G. (1981). *Nature (London)*, **290**, 465-470.
- Murphy, W. I., Attardi, B., Tu, C. & Attardi, G. (1975). *J. Mol. Biol.* **99**, 809-814.
- Ojala, D. & Attardi, G. (1974a). *J. Mol. Biol.* **88**, 205-219.
- Ojala, D. & Attardi, G. (1974b). *J. Mol. Biol.* **82**, 151-174.
- Ojala, D. & Attardi, G. (1977). *Plasmid*, **1**, 78-105.
- Ojala, D. & Attardi, G. (1978). *J. Mol. Biol.* **122**, 301-319.
- Ojala, D., Merkel, C., Gelfand, R. & Attardi, G. (1980a). *Cell*, **22**, 393-403.
- Ojala, D., Montoya, J. & Attardi, G. (1980b). *Nature (London)*, **287**, 79-82.
- Shine, J. & Dalgarno, L. (1974). *Proc. Nat. Acad. Sci., U.S.A.* **71**, 1342-1346.
- Simonsits, A., Brownlee, G. G., Brown, R. S., Rubin, J. R. & Guillely, H. (1977). *Nature (London)*, **269**, 833-836.
- Southern, E. M. (1975). *J. Mol. Biol.* **98**, 503-517.
- Steitz, J. A. & Jakes, K. (1975). *Proc. Nat. Acad. Sci., U.S.A.* **72**, 4734-4738.
- Tinoco, I. Jr, Borer, P. N., Dengler, B., Levine, M. D., Uhlenbeck, O. C., Crothers, D. M. & Gralla, J. (1973). *Nature New Biol.* **246**, 40-41.
- Van Etten, R. A., Walberg, M. W. & Clayton, D. A. (1980). *Cell*, **22**, 157-170.
- Yuan, R. C., Steitz, J. A., Moore, P. B. & Crothers, D. M. (1979). *Nucl. Acids Res.* **7**, 2399-2418.

*Edited by S. Brenner*

CONCLUSION

PART II

## CONCLUSION

In the second part of this thesis, we showed how techniques in RNA and DNA sequencing could be successfully applied to the study of the human mitochondrial genome. Besides defining the origin of DNA replication at the nucleotide sequence level, we were able to show that the 3'-end of the mitochondrial phenylalanine tRNA gene was joined end-to-end to the 5'-end of the 12S mitochondrial rRNA gene. This dramatically exhibited the extreme economy of gene organization of the human mitochondrial genome.

These techniques and other gene mapping techniques were utilized to map or identify most of the genes and stable transcription products on the genome (Barrell et al., 1979; Ojala et al., 1980a, 1980b; Anderson et al., 1981; Montoya et al., 1981; Ojala et al., 1981). The complete sequencing of the mitochondrial genome along with the mapping of the termini of the stable poly(A)-containing RNAs has allowed analysis of the sequence of mitochondrial genes. Perhaps, most striking is again the extreme economy that we first saw in that the tRNA genes are often joined end-to-end with the rRNA and mRNA genes. Additionally, this type of molecular analysis has provided insight into the processing of the mitochondrial transcripts. Evolutionary comparisons with the sequenced genome of other organisms such as mouse, cow, and yeast should be interesting. The sequence information will also be necessary for understanding the function, expression, and regulation of mitochondrial genes. Lacking useful genetic mutants, these studies may rely on in vitro or gene transfer systems that will provide a challenge for molecular biologists to develop.

## REFERENCES

- Anderson, S., Bankier, A. T., Barrell, B. G., De Bruijn, M. H. L., Coulson, A. R., Drouin, J., Eperon, I. C., Nierlich, D. P., Roe, B. A., Sanger, F., Schreier, P. H., Smith, A. J. H., Staden, R. and Young, I. G. (1981) Sequence and organization of the human mitochondrial genome. *Nature* **290**, 457-465.
- Barrell, B. G., Bankier, A. T. and Drouin, J. (1979) A different genetic code in human mitochondria. *Nature* **282**, 189-194.
- Montoya, J., Ojala, D. and Attardi, G. (1981) Distinctive features of the 5'-terminal sequences of the human mitochondrial mRNAs. *Nature* **290**, 465-470.
- Ojala, D., Merkel, C., Gelfand, R. and Attardi, G. (1980a) The tRNA genes punctuate the reading of genetic information in human mitochondrial DNA. *Cell* **22**, 393-403.
- Ojala, D., Montoya, J. and Attardi, G. (1980b) The putative mRNA for subunit II of cytochrome c oxidase in human mitochondria starts directly at the translation initiator codon. *Nature* **287**, 79-82.
- Ojala, D., Montoya, J. and Attardi, G. (1981) tRNA punctuation model of RNA processing in human mitochondria. *Nature* **290**, 470-474.