The Organization and Diversity of Immunoglobulin Genes*

PHILIP LEDER, TASUKI HONJO, SEYMOUR PACKMAN, DAVID SWAN, MARION NAU, AND BARBARA NORMAN

Laboratory of Molecular Genetics, National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, Maryland 20014

ABSTRACT We have used purified mouse immunoglobulin light chain mRNA and synthetic DNA which is complementary to it to assess the reiteration frequency of gene sequences corresponding to the ζ constant region of the mouse immunoglobulin light chain. These studies indicate that the constant region sequence is represented only two to three times per haploid mouse genome, a finding that rules out a simple stringent germ line mechanism which would require the constant region sequence to be represented hundreds if not thousands of times. Hybridization studies involving 35S-labeled myeloma light chain mRNA yield interesting results which may eventually permit us to distinguish between the remaining somatic mutation and recombinational germ line hypotheses. These results reveal a major component of relatively unique frequency and a minor component with a reiteration frequency of approximately 30 to 50 copies per haploid genome. As discussed, these results do not permit us to distinguish unambiguously between a germ line model and a type of somatic mutation model which permits germ line genes corresponding to each ζ subgroup. The results do, however, clearly rule out the existence of thousands of variable region sequences so closely related to the MOPC-41 V-region as to permit extensive stable cross-hybridization.

Introduction

The functional responsibilities of the immune system include the potential to respond—with great specificity—to a virtually limitless array of antigens. Almost as a minimum, one must postulate that the immune system has the potential to produce a million different antibody molecules—each fairly specific and each subject to some form of regulation which ensures that the correct antibodies are produced in very large amounts in response to specific antigenic stimuli. We have been interested in determining how this vast amount of information is organized in the cellular genome and how we can account, in genetic terms, for the tremendous diversity encoded in the immune system.

Many structural and serologic studies carried out over the last decade have provided an important insight into the way specificity and, thereby, diversity are built into the antibody molecule. A given class of antibody molecules can have absolutely identical amino-acid sequences in their C-terminal halves, but all differ in their N-terminal regions. It is this variable portion of heavy and light chains that forms the antibody combining site and, hence, gives the antibody molecule its unique specificity. The major questions, then, confronting modern immunology are how this diversity arises, how immunoglobulin genes are organized, and how their expression is regulated.

A number of distinguished geneticists and immunologists have provided a variety of very useful genetic models to account for the diversity of antibody molecules (see refs. in 1). These models fall into three general categories which are depicted diagrammatically in Table 1. Our representation of these models is based on the reasonable assumption that about 1000 light chains and 1000 heavy chains could combine to give a million different antibody molecules which would represent a reasonable antibody repertoire (2). The simplest, and in many ways most appealing model, is a stringent germ line hypothesis which holds that each immunoglobulin subunit, light and heavy chain (we shall restrict our arguments to the light chain, inasmuch as analogous arguments hold for the heavy chain) is represented, intact in the genome of every cell of the organism. Thus, 1000 light chain sequences would require 1000 variable region genetic sequences immediately adjacent to the 1000 constant region genetic sequences. In contrast to this straightforward model, the remaining two major hypotheses, the recombinational germ line hypothesis and the somatic mutation hypothesis, have incorporated a critical feature originally suggested by Dreyer and Bennett (3), namely, that variable and constant regions are encoded by two separate genes which are (or whose products are) ultimately joined to one another to form an intact immunoglobulin sequence. These models differ from the stringent germ line hypothesis in that they require only one or very few copies of the constant region gene sequence. They differ from one another in that the recombinational germ line hypothesis holds that immunoglobulin diversity has arisen through evolutionary processes so that the variable regions are represented individually in the germ line and thus in every cell of the organism. The somatic mutation hypothesis, in contrast, holds that diversity has arisen—at least in part—during somatic differentiation; and, therefore, only very few copies of the variable region sequences are required in the germ line since these will differentiate during somatic development.

The three models differ clearly from one another in the number of constant and variable gene sequences which they require. The stringent germ line hypothesis requires many copies of the constant region sequences, the remaining hypotheses require very few. The recombinational germ line hypothesis and the somatic mutation hypothesis differ from one another in that the former requires many more copies cor-

Abbreviations: cDNA, complementary DNA; Ct (or Cx) is defined as the concentration of DNA (or RNA) in annealing reaction × the time of incubation and is expressed as moles × sec/liter; C-region, constant region of immunoglobulin; V-region, variable region of immunoglobulin.

* This paper was presented at the Annual Meeting of the National Academy of Sciences in Washington, D.C. on April 22-24, 1974. It was presented in A Symposium on Organization and Transcription of the Eukaryotic Genome.
Table 1. Predicted frequencies of light chain constant and variable gene sequences in cloned immunocyte DNA

<table>
<thead>
<tr>
<th>Stringent germ line hypothesis</th>
<th>Theoretical gene frequency</th>
<th>Expected frequency as determined by hybridization with MOPC-41 light chain mRNA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DNA source</strong></td>
<td><strong>Variable</strong></td>
<td><strong>Constant</strong></td>
</tr>
<tr>
<td>MOPC-41 $V_i$ $C_k$ $V_k$ $C_k$ $V_k$ $C_k$ ...</td>
<td>~1000</td>
<td>~1000</td>
</tr>
<tr>
<td>Other $V_i$ $C_k$ $V_k$ $C_k$ $V_k$ $C_k$ ...</td>
<td>~1000</td>
<td>~1000</td>
</tr>
<tr>
<td>Recombinational germ line hypothesis</td>
<td><strong>DNA source</strong></td>
<td><strong>Variable</strong></td>
</tr>
<tr>
<td>MOPC-41 $V_i$ $V_2$ $V_3$ $V_4$ ... $V_{10}-10^9$ $C_k$</td>
<td>~1000</td>
<td>&lt;10</td>
</tr>
<tr>
<td>Other $V_i$ $V_2$ $V_3$ $V_4$ ... $V_{10}-10^9$ $C_k$</td>
<td>~1000</td>
<td>&lt;10</td>
</tr>
<tr>
<td>Somatic mutation hypothesis</td>
<td><strong>DNA source</strong></td>
<td><strong>Variable</strong></td>
</tr>
<tr>
<td>MOPC-41 $V_i$ $V_2$ $V_3$ $V_4$ ... $V_{10}-10^9$ $C_k$</td>
<td>&lt;100</td>
<td>&lt;10</td>
</tr>
<tr>
<td>Other $V_i$ $V_2$ $V_3$ $V_4$ ... $V_{10}-10^9$ $C_k$</td>
<td>&lt;100</td>
<td>&lt;10</td>
</tr>
</tbody>
</table>

* According to the somatic mutation hypothesis, the gene sequence corresponding to the variable region of the MOPC-41 subgroup in a MOPC-41 immunocyte should differ from the comparable sequence in other clones. Therefore, hybrids of the MOPC-41 variable region mRNA and the comparable DNA regions derived from other clones should be mismatched and incomplete were this hypothesis correct.

**Table 2. Diagrammatic representation of MOPC-41 mRNA and its complementary [3H]cDNA**

<table>
<thead>
<tr>
<th>Light chain</th>
<th>Variable : Constant : 1300 Bases</th>
</tr>
</thead>
<tbody>
<tr>
<td>3' : U* : Pref</td>
<td>325 : 325 : 1500 : 200</td>
</tr>
</tbody>
</table>

**5' :** Putative untranslated sequence.

The purified mRNA thus obtained served as an efficient template for the synthesis of complementary MOPC-41 DNA (cDNA) (10). As we have pointed out previously, by annealing an oligo(dT) primer to the 3' poly(A) end of the message, we were able to phase the RNA-dependent DNA polymerase and assure that copying occurs from the 3' to the 5' end. In the case of the light chain message, the cDNA thus synthesized was approximately half the length of the immunoglobulin message. Careful chain length determinations carried out using authentic restriction enzyme fragments of phage λ indicated that the MOPC-41 cDNA had a chain length of approximately 630 nucleotides. Since reverse transcription is from the 3' end, the cDNA probe would correspond to the C-terminal portion or constant region-coding portion of the immunoglobulin light chain mRNA (see diagram, Table 2). We will discuss the evidence supporting this assertion below.

An additional test of the homogeneity of the immunoglobulin light chain mRNA

The availability of the [3H]cDNA probe makes it possible to analyze the purity of the immunoglobulin mRNA preparation using the technique of hybridization kinetic analysis (25, 26). Under conditions in which the mRNA concentration is in excess, the C1 value at which one-half the cDNA probe is hybridized (the C1/2) is proportional to the relative concentration of mRNA complementary to it in the reaction mixture. Thus, we can use purified reticuloocyte globin mRNA and its complementary probe as a standard for an mRNA sequence of 1200 bases in length (α plus β chain) and view the hybridization of a globin cDNA probe approximately 570 nucleotides in length (27). The C1/2 value thus obtained is $3.7 \times 10^{-4}$. Thus, the C1/2 value for the MOPC-41 mRNA which is slightly longer, 1300 nucleotides, would be $4 \times 10^{-4}$ as shown in Fig. 1. The experimentally obtained value is $4.2 \times 10^{-4}$, which is in very close agreement to the expected theo-
retical value. Also shown is the thermal elution profile of the hybrid formed between MOPC-41 mRNA and its [3H]cDNA. The sharp melting profile with a Tm of 87°C indicates that the hybrid formed is congruently base-paired with little or no mismatching.

Cross-hybridization between MOPC-41 cDNA and other \( \kappa \) light chain mRNAs

We have argued above from the known mechanism of the RNA-dependent DNA polymerase and the structure of the mRNA that the [3H]cDNA probe corresponds, as a minimum, to a major portion of the constant region of the light chain mRNA (see Table 2). Inasmuch as the mRNA is 1300 bases long, contains a 200 base poly(A) sequence, and requires about 650 bases to encode the mature form of the light chain, we are left to account for a 450 nucleotide sequence. We must assign an additional 250 bases to encode the putative light chain precursor (9). At least 150 of these bases are located at the 5'-terminal portion of the translated sequence (D. McKean, T. Honjo, M. Nau, D. Swan, and P. Leder, unpublished results). There still remain 200 nucleotides to be accounted for and this sequence, presumably untranslated, can occupy either the 3' or the 5' end of the molecule. In either case, the cDNA probe which is transcribed from the middle of the poly(A) region at the 3' end of the molecule will encode, as a minimum, about 70% of the constant region sequence and, as a maximum, will be complementary to all of it as well as a small portion of the variable region sequence. Consistent with this expectation, the MOPC-41 [3H]cDNA probe cross-hybridizes readily to mRNA derived from a myeloma tumor which produces a \( \kappa \) chain of a different subgroup (Table 3). As shown, the hybridization results in protection of approximately 70% of the cDNA probe. The remaining, unprotected portion of the probe might correspond to a region overlapping the variable or an untranslated sequence.

Genetic representation of the constant region sequence. Two genes encoding one polypeptide chain?

We have already indicated that the most stringent germ line hypothesis would require hundreds if not thousands of copies of the C-region sequence. Nevertheless, a number of genetic analyses (28-35) which have examined the linkage, allotype, and crossover of heavy and light chain C-region genes have buttressed the original arguments of Dreyer and Bennett (3) suggesting that only a few copies of the C-region gene are required in the genome. Implicit in this argument is the notion that (at least) two genes encode one polypeptide chain.

A direct test of the stringent germ line hypothesis, therefore, can be performed using the MOPC-41 cDNA as a hybridization probe to assess the reiteration frequency of the constant region gene sequence. MOPC-41 [3H]cDNA was hybridized to a vast excess of unlabeled total DNA derived from MOPC-41 tumor (Fig. 2). A similar analysis using mouse globin cDNA is shown for comparison. The result lends little room for ambiguity. The C4t/4, value for MOPC-41 cDNA is 1130 as compared to a C4t/4 value of 3000 for a “unique copy” mouse DNA. This corresponds to a reiteration frequency of 2.7 copies per haploid genome. The sharp thermal elution profile (Fig. 2, inset) indicates the closely matched nature of the hybrid formed. Similar analyses were carried out using MOPC-41 [3H]cDNA and total DNA from a \( \lambda \) chain-producing tumor, RPC-20, and from normal mouse liver. These results gave reiteration frequencies of 3.0 and 3.2 per haploid genome, respectively, and, taken together, indicate that only two or three copies of the C-region gene are present per haploid genome. These are represented to the same extent regardless of whether the cells are actively producing \( \kappa \) chains.

Naturally this result is not compatible with the stringent germ line hypothesis which requires many copies of the C-region gene. It is, however, entirely consistent with one of the cardinal notions of the Dreyer-Bennett model which requires only a few copies of the constant region gene. Given the considerable data on the differences among mouse \( \kappa \) chain V-region amino-acid sequences, it seems clear that a minimum of 25 V-region sequences (and possibly many more) must be encoded in the germ line genome. These two conclusions force us to ask how these many V-region sequences can be joined to the few C-region sequences which we observed. The possibility that the polypeptide products of these genes are joined is ruled out by the fact that V- and C-regions are encoded in a single mRNA molecule (1). The remaining stages at which this joining could occur would be at the level of mRNA or at

![Fig. 1. Hybridization of purified mouse globin and MOPC-41 light chain mRNAs and their respective [3H]cDNAs. The hybridization reaction was carried out in 0.6 M NaCl-0.2 mM EDTA-15 mM Tris-HCl, pH 7.2 at 75°. mRNA is present at 10- to 100-fold excess over cDNA. Aliquots were taken at time intervals for the assay of the hybrid formed by S1 nuclease digestion (27). Thermal elution profiles (inset) were determined by hydroxyapatite chromatography (12)].

<table>
<thead>
<tr>
<th>Source of light chain mRNA</th>
<th>Maximum extent of hybrid formation, %</th>
<th>Tm, °C</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOPC-41</td>
<td>93</td>
<td>85</td>
</tr>
<tr>
<td>MPC-11</td>
<td>71</td>
<td>84.5</td>
</tr>
</tbody>
</table>

Hybridization reactions and assays were carried out as indicated in the legend to Fig. 2. Thermal denaturation was assayed by hydroxyapatite elution as described previously (12).
The cDNA and cellular DNA were prepared as described (10, 12), with the former having a specific activity of $4.6 \times 10^{9}$ dpm/µg, and a base length of 630 as determined by alkaline sucrose gradient centrifugation. The hybridization procedure and $S$ nuclease assays were as described (38), and were performed at DNA concentrations and salt concentrations appropriate for each portion of the kinetic curve (see below). The $C_T$ values on the abscissa are those that would obtain at 0.18 M Na$^+$ (26). $C_T/v$, for MOPC-41 cDNA × MOPC-41 DNA equals 2100. The filled symbols represent MOPC-41 cDNA × MOPC-41 DNA: (a) 0.99 mg/ml of DNA, 0.18 M Na$^+$; (b) 9.9 mg/ml of DNA, 0.54 M Na$^+$; (c) 0.9 mg/ml of DNA, 1.06 M Na$^+$. The open symbols (O) represent globin cDNA × MOPC-41 DNA. Thermal elution profiles (inset) were determined by diluting the hybrid to 0.24 M Na$^+$, heating at each temperature for 8–10 min, and determining resistance to $S$ nuclease (1). In the inset, the filled symbols (●) represent MOPC-41 cDNA × MOPC-41 DNA; the open symbols (○) represent MOPC-41 cDNA × mouse spleen DNA.

cross the level of the gene itself. While there is no evidence favoring either of these possibilities, efficiency argues that a single joining event occurring at the level of the gene would be more likely than thousands of joining events occurring at the level of the message. Several interesting models for genetic recombination or re-arrangement of these multiple genes have been proposed (3, 36, 37).

**Somatic mutation or evolution as a source of diversity**

The studies noted above focus on the genetic representation of the C-region sequence and, therefore, do not address the question of origin of diversity in the V-region sequences. Nevertheless, a number of studies addressing this question have been carried out using $[1^{125}]$I-mRNA derived from myeloma tumors (11, 20, 22, 23). The interpretation of data obtained from such studies is necessarily limited by uncertainties regarding the purity of the hybridization probe and/or the portion of the hybridization curve representing the V- and C-regions. Inasmuch as the MOPC-41 mRNA preparation we have used appears by several criteria and the unique nature of the constant region sequence has been determined, we have carried out preliminary studies using intact $[1^{125}]$I-labeled mRNA (19).

As we have indicated previously, the remaining hypotheses, the recombinational germ line and the somatic mutation, differ from one another in that the former requires many, possibly thousands, of variable region gene sequences whereas the latter requires few, perhaps a hundred or less. Whereas, these requirements are quite clear, there are consequences in terms of a hybridization experiment which require qualification. Referring to Table 1 we may ask what results might be expected from hybridization analyses were either model correct? The differences in amino-acid sequences between variable regions among the $k$ subgroups are so great that we might not expect stable hybrid formation between sequences derived from different subgroups. Variable sequences derived from the same subgroup in which over 80% of the amino-acid sequences are identical, in contrast, would be expected to form stable cross-hybrids (12). Were each subgroup to include approximately 50 different variable sequences (a not unreasonable expectation) and were the germ line hypothesis correct, we would expect to observe a reiteration frequency of about 50 or fewer copies for the variable region. Were the somatic mutation hypothesis correct and the subgroup represented by a single germ line gene, we would expect $C_T/v$ values consistent with a unique representation of this gene. Furthermore, the somatic mutation hypothesis stipulates that the genetic makeup of the cloned immunocytes would differ from one another due to their having diverged by a somatic mutation. We might then expect that hybrids formed between mRNA from one clone and DNA from another would be less stable and hybridization less extensive were the somatic mutation hypothesis correct. Our initial hybridization kinetic analyses of $[1^{125}]$I-mRNA obtained from MOPC-41 in the presence of vast excesses of total cellular homologous MOPC-41 DNA are shown in Fig. 3. A similar analysis using $[1^{125}]$I-labeled mouse globin mRNA is shown for comparison. A major portion of the $[1^{125}]$I-mRNA hybridizes with a relatively unique $C_T/v$, (3200). In contrast to the observation of others who used less well characterized RNA preparations, there is no very highly reiterated material (11, 20). The $C_T$ curve is not obviously biphasic. Nevertheless, approximately 25% of the $[1^{125}]$I-mRNA does hybridize over a $C_T$ range of between 2 and 800. This material can be assigned a reiteration frequency of 30 to 50 copies per haploid genome, though it is difficult to feel secure about this number. An important difference between
this mRNA and that derived from globin is that the latter does not contain these moderately reiterated sequences. The extent of hybridization, the amount of hybrid formed when the reaction is complete, is approximately 70%.

We have similarly analyzed DNA derived from a myeloma of a different subgroup, MOPC-265 (Fig. 4). The results are quite similar: a portion of the material is moderately reiterated, the major component is relatively unique and the extent of hybridization is about 66%. Indeed, the thermal elution profiles of hybrids formed between MOPC-41 [188I]mRNA and MOPC-41 and MOPC-265 DNAs are indistinguishable (Fig. 5).

Such results are entirely—but not exclusively—consistent with the recombinational germ line hypothesis. A portion of the mRNA is moderately reiterated, consistent with a number of genes that would be expected to constitute a single sub-

group. The extent of hybridization does not differ significantly between two cloned myeloma tumors, suggesting that sequences are represented equivalently in each cloned line. Still there are very strong reservations to such an uncritical interpretation of the data. It is clear that the extent of hybridization can only be estimated within a wide margin of error (10–15%) and a significant difference might easily go undetected. Furthermore, there is no direct evidence to suggest that the moderately reiterated RNA which we observe in fact corresponds to the variable region. Preparation of V-region specific probe and careful competition experiments using mRNAs derived from various subgroup tumors are necessary to establish this point.

Conclusions

Studies, in which an apparently homogeneous mRNA and its complementary synthetic DNA are used, have allowed us to quantitate the number of gene sequences corresponding to the constant region of the immunoglobulin light chain. These results indicate that the constant region sequence is represented approximately three times per haploid mouse genome and is represented equally in a variety of tumors and organs. Such a result is inconsistent with a stringent germ line hypothesis which requires many, possibly thousands of, copies of the constant region gene sequence. Taken together with other serologic and immunohemical data, this result lends considerable support to the notion that immunoglobulin subunit chains are encoded by at least two gene sequences. This conclusion raises the further requirement of postulating an unprecedented biochemical mechanism for joining constant and variable sequences either at the level of the mRNA or at the level of the gene itself.
Fig. 5. Thermal denaturation of hybrids formed between MOPC-41 \(^{3}H\)mRNA and DNA from MOPC-41 and MOPC-265 tumors. Hybrids were formed in 1.04 M Na\(^+\) as described in the legend to Fig. 3 and diluted 4-fold at the initial temperature of the melt curve. The reaction mixture was kept in a glass test tube in a Haake circulating water bath, and kept at each temperature for 8–10 min prior to sampling. At each temperature, aliquots were withdrawn, diluted 50-fold into 0.3 M NaCl–30 mM Na-citrate and assayed for resistance to 60 μg/ml of RNase A, as described in the legend to Fig. 4. The ordinate represents the % of the total cpm made single-stranded during the melting.

Thus far, hybridization studies in which the entire light chain mRNA is used have not permitted us to distinguish between the recombinational germ line and somatic mutation hypotheses. They do indicate, however, that we should be able to make this important distinction when our technology permits us to focus on the variable region sequence.

We are very grateful to Dr. Michael Potter for supplying the original myeloma tumors used in these studies. We are also grateful for the expert editorial assistance of Ms. Catherine Kunkle in the preparation of this manuscript.