Reprint from *The Immune System.*

27. Mosbacher Colloquium

Edited by F. Melchers and K. Rajewsky

Springer-Verlag
Berlin Heidelberg New York

Differentiation of Immunoglobulin Genes

S. Tonegawa and N. Hozumi

Introduction

Recent advances in the technology of nucleic acid biochemistry, in
particular the availability of highly purified, light and heavy chain
mRNAs from mouse plasmacytomas, have enabled several groups of re-
searchers to study immunoglobulin genes at the molecular level. In
this article we will summarize our recent efforts directed to the
following two questions:

What is the genetic origin of the enormous diversity of antibody
molecules?

How is the genetic information encoded in two seemingly separate seg-
ments of DNA, V and C, integrated to generate a contiguous polypep-
tide chain?

The Problem of the Genetic Origin of Antibody Diversity

One of the most intriguing features of the immune system is the vast-
ness of the diversity of its components. On the molecular level this
diversity is seen in the primary structure of antibody molecules. On
the cellular level it is seen in the clones of antibody producing
cells. The two kinds of diversity are connected by the well-established
clonal selection theory (Burnet, 1957; Jerne, 1955). The genetic origin
of such diversity has been one of the central issues in immunology.
The issue is whether or not structural genes for the full repertoire
of antibody molecules are contained as such in energy cell of an or-
ganism. The alternative is that a large proportion of the diversity
is generated by somatic process, such as somatic mutation. The immense
complexity of the vertebrate genome has hampered direct approaches to
this question. Inevitably, arguments for and against these contrasting
views have been based on indirect experimental observations (primarily
of amino acid sequences of immunoglobulin chains) and general principles
of genetics or evolution. (See for example Capra and Kihoe, 1974;
Wigzell, 1973).

Availability of light and heavy chain mRNAs of a purity sufficient
for detailed nucleic acid hybridization studies has opened a way for
direct experimental approaches to this controversial problem.

The Principles of the Approach

Given purified preparations of mRNAs coding for immunoglobulin chains,
several approaches to the present issue can be envisaged. Among these
the ultimate appraoch is probably to isolate, perhaps making use of
nucleotide sequence complementarity, the immunoglobulin genes both

from germ line and plasma cells, and determine whether or not the nucleotide sequences in the corresponding genes from the two sources are different. Such an approach, in spite of the enormous complexity of eucaryotic genomes, now seems quite feasible thanks to the recently developed techniques for cloning recombinant DNA molecules constructed in vitro (Morrow et al., 1974; Thomas et al., 1974). Another approach which we have employed in the past few years is to count the number of immunoglobulin genes by RNA-DNA hybridization. In this approach the experimental steps and logic listed below were followed:

1. Purify mRNAs coding for particular light of heavy chains from mouse myeloma cells.

2. Radioiodinate the purified mRNA to a specific activity higher than 10 million cpm per microgram.

3. Determine the reiteration frequency of the DNA sequences complementary to the mRNA. This is done by annealing under proper conditions the radioiodinated RNA with denatured, sonicated liver DNA (liver is in this case a surrogate of germ line cells), and then determining, as a function of time, the fraction of RNA in the hybrid by measuring the resistance of hybridized RNA to RNase. Under a given set of annealing conditions, the fraction of hybridized RNA (f) is dependent upon the RNA concentration (R_O), the DNA concentration (C_O) and the time of incubation (t). If DNA is present in large excess, f is independent of R_O, and depends only on the product of C and t ($C_O t$). The value of $C_O t$ necessary to achieve 50 % hybridization, $C_O t_{1/2}$, will depend upon the fraction of the DNA which is complementary to the RNA. The reiteration frequency (F) of the gene in question is calculated by the following formula

$$F = F* \frac{C_O t_{1/2}*}{C_O t_{1/2}} \frac{C}{C*}$$

where C designates the genome complexity and * designates a gene used for standardization and whose reiteration frequency F* is known by an independent method (Melli et al., 1971).

4. Carry out competition hybridization experiments by annealing the radioiodinated RNA used in the $C_O t$ curve experiment in the presence of varying amounts of unlabeled mRNA coding for another V region. The purpose of the competition experiments is to determine the degree of base sequence homology between the two mRNAs, and consequently between the two corresponding V genes. The results will tell us whether and how much the radioiodinated mRNA used in the $C_O t$ curve experiment would cross-hybridize with other V genes if they existed as separate germ line genes.

5. Estimate, on the basis of such competition experiments and on the available amino acid sequence data, the number of different V regions whose genes would have cross-hybridized with the particular mRNA used in the $C_O t$ curve experiment. If this number is significantly larger than the experimentally determined reiteration frequency, it constitutes formal evidence for somatic generation of antibody diversity.

Evidence for Somatic Generation of Antibody Diversity

We have applied the approach described in the last section both to mouse κ and λ chains (Tonegawa et al., 1974; Tonegawa, 1976). Partial
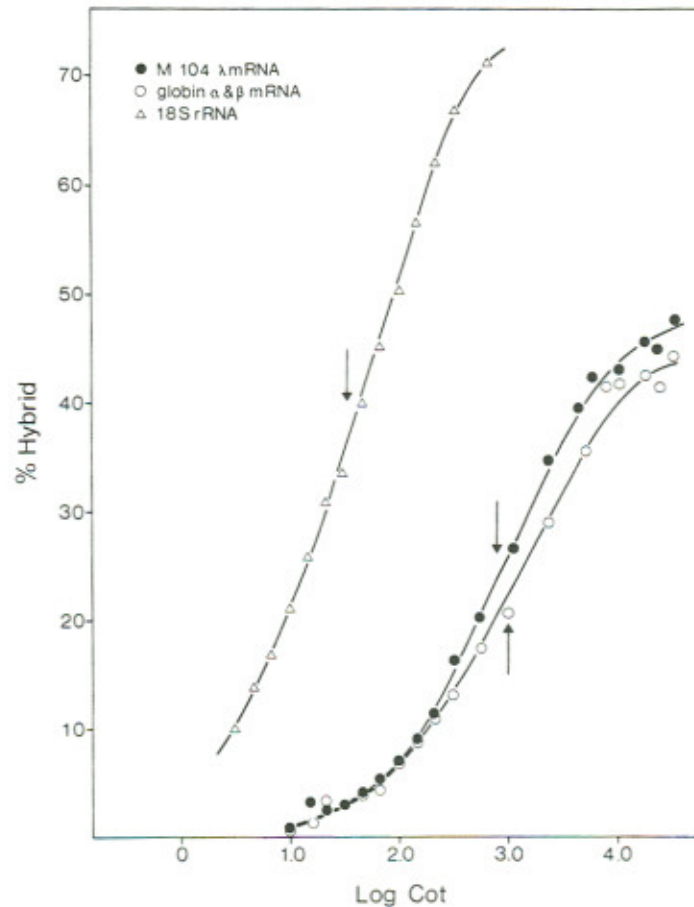
Fig. 1. Hybridization kinetics of mouse RNAs with mouse liver DNA (C_Ot curve).
Arrows: positions of $C_Ot_{1/2}$

application was also made to heavy chains (Bernardini and Tonegawa, 1974; Schuller, 1976). Because they are simple and straight-forward we will describe here the experimental results obtained with λ chains in some detail.

In the λ chain studies, two mRNAs coding for two different V_λ regions were prepared to a purity of 90 % or higher. The sources of the mRNAs were the Balb/s plasmacytomas MOPC 104E and HOPC 2020. The two V regions differ only in two amino acids (see Fig. 3). Each of the two pairs of amino acids is related by a single base substitution in the corresponding triplet codons. The C_Ot curve of M 104E mRNA is illustrated in Figure 1. A mixture of mouse α and β globin mRNAs as well as 18S ribosomal RNA were used as kinetic standards. The gene reiteration frequencies of these RNAs have previously been determined at less than 3 and at 200-250 respectively (Harrison et al., 1972; Melli et al., 1971). The reiteration frequency of M 104E λ chain gene is calculated to be 2 to 3 from the results represented in Figure 1.
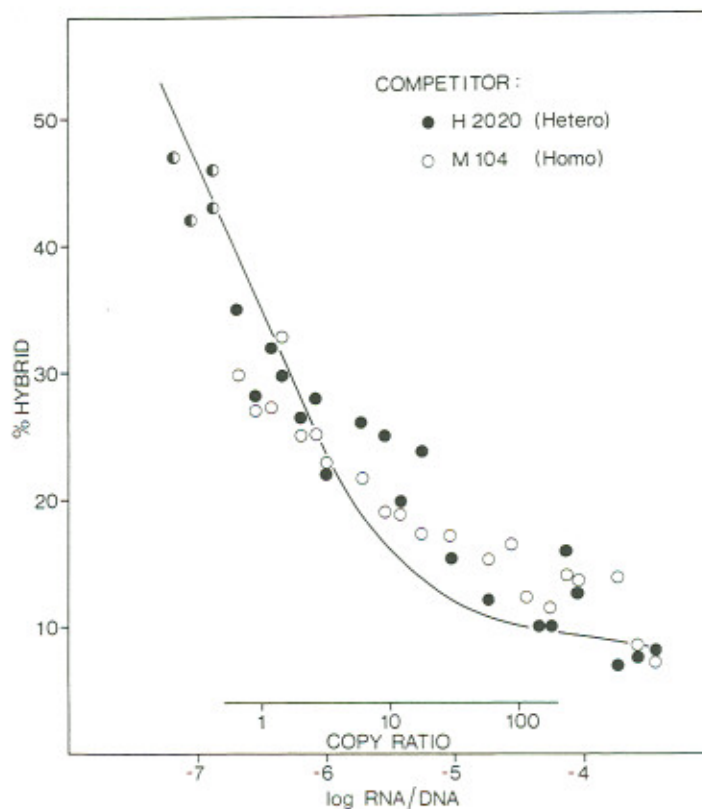
Fig. 2. Competition hybridization of λ mRNAs. Theoretical curve derived by computer program and represents homologous competition of hypothetical RNA mixture containing 85 % unique, major component and 15 % impurities, assumed to consist of 1500 different equimolar species

In the competition hybridization experiments illustrated in Figure 2, a fixed amount of radioactive M 104E mRNA and an excess of denatured liver DNA were annealed to a high C_0t value (C_0t = 10,000) in the presence of varying amounts of unlabeled H 2020 mRNA (heterologous competition) or M 104E mRNA (homologous competition). Radioactive M 104E mRNA in the hybrid is plotted as a function of the ratio of total RNA to DNA. The fact that the two competition curves are indistinguishable indicates that there is extensive nucleotide sequence homology between the two mRNAs and hence between the two V_λ genes. Is such an extensive sequence homology unique to this particular pair of λ chains? That this is not the case is strongly suggested in the amino acid sequence data of myeloma λ chains summarized in Figure 3. As discussed elsewhere (Tonegawa and Steinberg, 1976), a simple statistical calculation applied to the data in Figure 3 suggests that there must be at least 20 to 30 different V_λ regions in the pool of Balb/c mouse myeloma proteins. Each of these V_λ regions is expected to be as homologous to M 104E in sequence as is H 2020. Thus if there were a germ line V gene for each of these sequences, M 104E mRNA would cross-hybridize with every one of them. Yet the number of germ line V_λ genes was determined to be no more than 2 to 3 by the C_0t curve experiment.

SUMMARY OF AMINO ACID REPLACEMENTS IN V_λ

(M. Weigert et al.)

		NO. OF λ SEQUENCED	MIN. NO. O BASE CHA
SER SER — ILE GLY ALA — TYR HIS	I	12, M 104 etc.	0
ASN	II	1, S 176	1
LEU	III	1, RPC 20	1
VAL	IV	1, Y 5606	1
CYS	V	1, Y 5444	1
THR GLY	VI	1, H 2020	2
ASN ASN ARG	VII	1, S 178	4

```
0        25      50      75      100
         26 30   48 55           92 95
N-TERMINAL                              C-TERMINAL
```
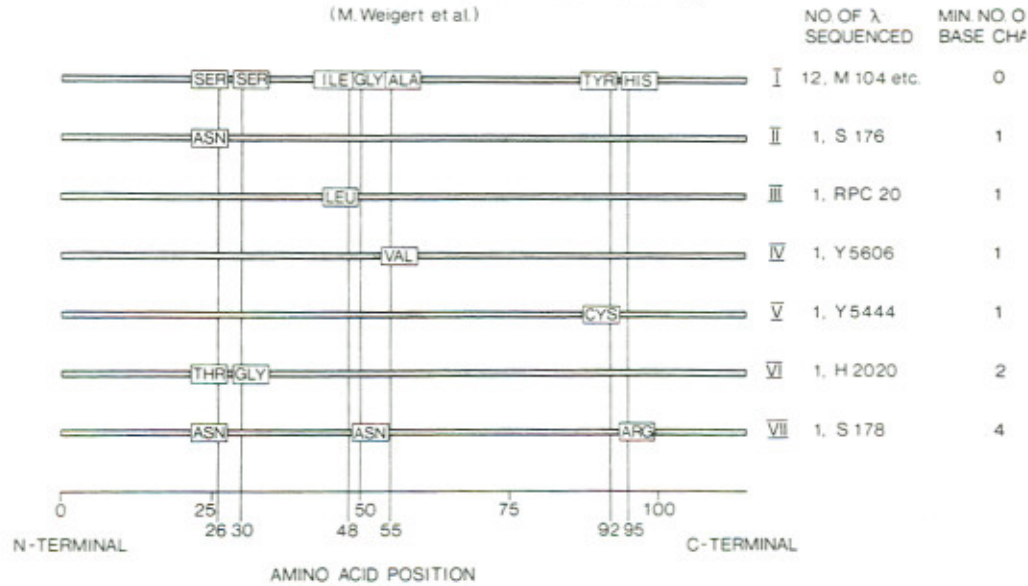
AMINO ACID POSITION

Fig. 3. Schematic representation of differences among λ chains. (After Weigert et al., 1970)

Sequence variability is considerably greater in mouse κ chains than in λ chains. This makes it more difficult to apply the kind of approa described above to κ chains. Nevertheless, we could define a group of κ chains, comprising over 10 % of all κ chains for which sequence information was available, within which there was over 80 % amino acid homology. For this group of κ chains we have shown that the numbers of germ line V genes are no more than 1 to 2 (Tonegawa et al., 1974).

Hence, the hybridization evidence indicates that the numbers of germ line V genes are too small to account for the observed diversity in antibody molecules.

The Problem of Integration of V and C Gene Information

Uniqueness (one gene per haploid genome) of C genes has been suspecte from normal Mendelian segregation of allotypic markers (Milstein and Munro, 1972). The hybridization studies just described confirm this notion. The same hybridization studies demonstrate that a group of closely related V regions are somatically generated from a few, possibly a single, germ line V gene(s). They do not, however, give us any reliable estimate of total number of germ line V genes. Given the enormous diversity of V regions, however, the existence of multiple germ line V genes is almost inescapable. This raises a second intriguing problem, one to which Dreyer and Bennett addressed themselves more than a decade ago (Dreyer and Bennett, 1965). How is the informa tion in the V and C genes integrated? Since V and C gene sequences exist in a single mRNA molecule as a contiguous stretch (Milstein et al, 1974), such integration must take place either directly at the
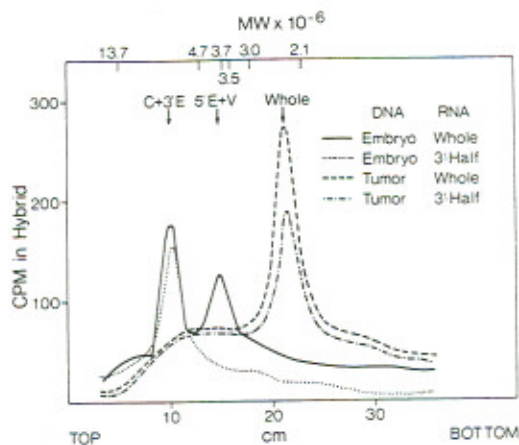
SPECIFICITY OF RESTRICTION ENZYMES
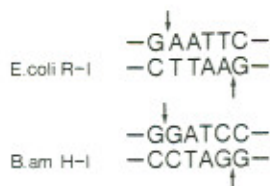
Fig. 4. Base sequence specificities of 2 re-
striction endonuclease

```
              ↓
            —GAATTC—
E.coli R–I  —CTTAAG—
              ↑

              ↓
            —GGATCC—
B.am H–I    —CCTAGG—
              ↑
```


Fig. 5. Hybridization patterns of
Bam H–I digests of mouse DNA with
MOPC 321 κ chain mRNA and its
fragments

DNA level or during transcription, or after transcription by joining
of two RNA molecules.

We will describe experiments which provide evidence for joining of V
and C sequences at the DNA level.

Evidence for Somatic Joining of V and C Genes

Bacterial restriction endonucleases recognize and cleave specific se-
quences of base-pairs within a DNA duplex. The specificity of two such
enzymes, *Escherichia coli* R–I (Eco R–I) and *Bacillus amyloliquefaciens* H–I
(Bam H–I) is shown in Figure 4. We have purified these enzymes in a
large quantity and digested high molecular weight mouse DNA from various
sources. The resulting DNA fragments were fractionated according to
size in a preparative agarose gel electrophoresis. The gel was cut into
slices, and DNA extracted from each slice was assayed for V and C gene
sequences by hybridization with radioiodinated whole light chain mRNA
or its 3'-end-containing fragments. The assay is based on the fact that
the sequences corresponding to V and C genes are on the 5'-end and 3'-
end halves of the mRNA molecules respectively. The details and justi-
fication of such assay procedures are described elsewhere (Hozumi and
Tonegawa, submitted).

Figure 5 illustrates such an experiment carried out for a κ chain.
DNA from 12-day-old Balb/c mouse embryos or from MOPC 321 tumors was

digested with the Bam H-I enzyme. Assay was carried out with MOPC 321 κ-mRNA. Results obtained from two separate gels, one of embryo DNA and the other of tumor DNA, are superimposed. Since both electrophoresis and hybridization were performed under identical conditions, profiles from the two gels are comparable. With the embryo DNA, two DNA components of 6.0 and 3.9 million MW hybridized with the whole κ-mRNA, whereas only the former hybridized with the 3'-end half. Hence the $C_κ$ gene is in the 6.0 million component and the $V_κ$ gene should be in the 3.9 million component. The fact that the extent of hybridization in the larger component is nearly identical with the two RNA probes supports this interpretation. Since the two genes are in separate DNA fragments, they are probably some distance away from each other in the embryo genome. The possibility that the enzyme cleaved contiguously arranged V and C genes near the boundary is not entirely eliminated. Two candidates for such cleavage sites are represented by the amino acids in the V region at positions 93-95 and 97-98. However, the probability that either of these amino acids provides the exact nucleotide sequence required is low.

The pattern of hybridization is completely different in the DNA from the homologous tumor. Both RNA probes hybridized with a new DNA component of 2.4 million MW. There is no indication that either of these RNA probes hybridizes with other DNA components above the general background level. These results strongly suggest that both V and C gene sequences are contained in the 2.4 million MW component of this tumor DNA. The whole RNA hybridizes with this component nearly twice as well as does the 3'-end half, thereby supporting this notion. Hence the $V_κ$ and $C_κ$ genes, which are most likely some distance away from each other in the embryo genome, are brought together in the plasma cells presumably to form a contiguous nucleotide stretch. Such rearrangement of immunoglobulin genes takes place in both the homologous chromosomes. An alternative explanation, namely that fortuitous mutations by loss and gain of Bam H-I sites generated the observed pattern difference, is not impossible. It is, however, extremely unlikely because of the multiplicity of events which must be assumed. Furthermore such pattern changes are not unique to this particular combination of enyzme and DNA. With *E. coli* R-I enzyme, and embryo and H 2020 DNAs we have obtained results which lead us to a similar conclusion (Hozumu and Tonegawa, manuscript in preparation).

Models for V - C Gene Joining

Various models have been suggested for integration of V and C information at the DNA level. Some of these models are schematically illustrated in Figure 6. The "copy-insertion" model assumes that a specific V gene is duplicated and the copy is inserted at a site adjacent to the C gene (Dreyer et al., 1967). The "lateral duplication" model envisages multiple V genes which are arranged in parallel, each adjacent to the C gene (Smithies, 1970). According to either of these two models, lymphocytes expressing a particular V gene should retain embryonic context as far as this V gene is concerned. Our results are clearly incompatible with this prediction. The "excision insertion" model suggests that a specific V gene is excised to form an episome-like structure, which in turn is integrated adjacent to the C gene (Gally and Edelman, 1970). The "deletion" model originally presented for the globin gene system can be applied to the immunoglobulin genes. According to this model DNA intervening a specific V gene and the C gene lopps out and is lost upon subsequent cell multiplication (Kabat, 1972). The latter two models are both consistent with the present results.

A copy insertion

B lateral gene
duplication.

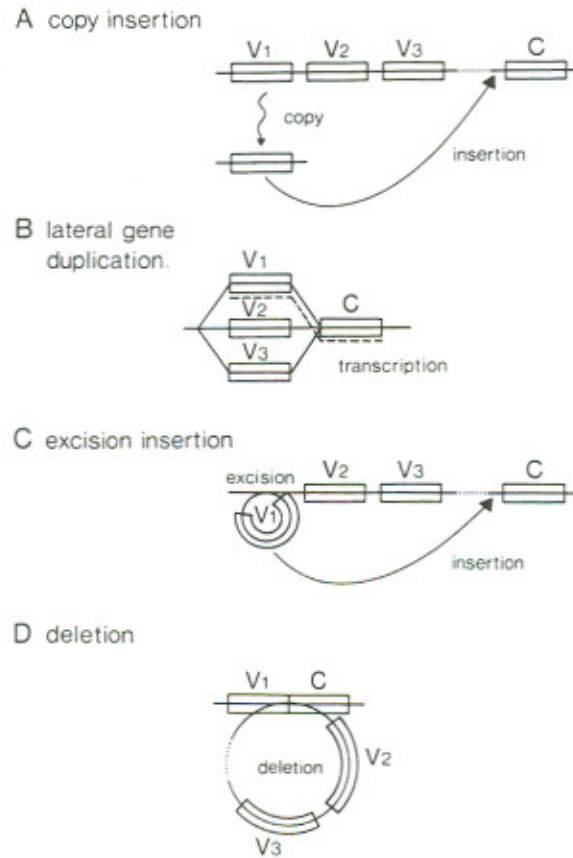C excision insertion

D deletion

Fig. 6. Models for V-C gene joining

One may note, however, that the excision insertion model requires at least two recombinations, whereas the deletion model requires only one. Further experiments are required to distinguish between the two models.

Concluding Remarks

The hybridization experiments described in this article demonstrated that both content and context of immunoglobulin genes are altered during differentiation of lymphocytes. We know, however, very little about the molecular mechanisms underlying these phenomena. The gene isolation experiments now underway in several laboratories are expected to help in understanding these mechanisms as well as other problems associated with immunoglobulin genes.

Are alterations in genetic content and context restricted to the immune system? As mentioned above, the "deletion" model was originally proposed for the globin system. We feel that this mechanism will turn out to be important in the differentiation of other eucaryotic systems as well.

References

Bernardini, A., Tonegawa, S.: FEBS Lett. 40, 73 (1974)

Burnet, F.M.: Australian J. Sci. 20, 67 (1957)

Capra, J.D., Kehoe, J.M.: Scand. J. Immunol. 3, 1 (1974)

Cohn, M., Blomberg, B., Geckeler, W., Raschke, W., Riblet, R., Weigert, M.: In: The Immunse System: Genes, Receptors, Signals. Sercarz, E.E. Williamson, A.R., Fox, C.R. (eds.). New York, London: Academic Press 1974, p. 89

Dreyer, W.J., Bennett, J.C.: Proc. Nat. Acad. Sci. 54, 864 (1965)

Dreyer, W.J., Gray, W.R., Hood, L.: Cold Spring Harbor Symp. Quant. Biol. 32, 353 (1967)

Gally, J.A., Edelman, G.M.: Nature (Lond.) 227, 341 (1970)

Harrison, R.P., Hell, A., Birnie, G.D., Paul, J.: Nature (Lond.) 239, 219 (1972)

Jerne, N.K.: Proc. Nat. Acad. Sci. 41, 849 (1955)

Kabat, D.: Science 175, 134 (1972)

Melli, M., Whitfield, C., Rao, K.V., Richardson, M., Bishop, J.O.: Nature (New Biol.) 231, 8 (1971)

Milstein, C., Brownlee, G.G., Cartwright, E.M., Jarvis, J.M., Proudfoot N.J.: Nature (Lond.) 252, 354 (1974)

Milstein, C., Munro, A.J.: Ann. Rev. Microbiol. 24, 335 (1972)

Morrow, J.F., Cohen, S.N., Chang, A.C.Y., Boyer, H.W., Goodman, H.M., Helling, R.B.: Proc. Nat. Acad. Sci. 71, 1743 (1974)

Schuller, R.: M.Sc. thesis, Univ. Alberta (1976)

Smithies, O.: Cold Spring Harbor Symp. Quant. Biol. 38, 725 (1970)

Thomas, M., Cameron, J.R., Davies, R.W.: Proc. Nat. Acad. Sci. 71, 4579 (1974)

Tonegawa, S.: Proc. Nat. Acad. Sci. 73, 203 (1976)

Tonegawa, S., Steinberg, C.: In: The Generation of Antibody Diversity. Cunningham, A.J. (ed.). London, New York: Academic Press 1976, pp. 175-182

Tonegawa, S., Steinberg, C., Dube, S., Bernardini, A.: Proc. Nat. Acad. Sci. 71, 4027 (1974)

Wigzell, H.: Scand. J. Immunol. 2, 199 (1973)

Discussion

Dr. Hood: I think the two last papers are very exciting in the sense that (1) they give us an overview of nucleic acid chemistry with a clear picture of the messenger for antibody molecules that has resulted from it, (2) the results reported by Tonegawa suggest rather strongly that the joining mechanism of V and C regions of antibody molecules that one has talked about for the past ten years is indeed a legitimate mechanism that occurs at the level of DNA and as such is almost certainly going to be a fundamental mechanism for differentiating the immune system and perhaps even other complicated kinds of system with a large information-handling problem, and (3) the hybridization data reported both by Drs. Milstein and Tonegawa has suggested to them very strongly that each subgroup of κ chains on one hand, or λ chains on the other, contains gene numbers that are lower than are consistent with amino acid sequences that are available today. Let me again parenthetically add that, in my own mind, I still have great reservations about this latter point in the sense that we do not really understand the limits of the error bars. In fact I was amused to see Dr. Tonegawa use hemoglobin as his standard when in the case of human hemoglobins one can point out that some years ago Eric Davidson did some hybridization studies and said the human hemoglobulins were consistent with single cot characteristics and it is only more recently that a very careful cot analysis of the α chain of human hemoglobulins suggests that there are at least four closely related genes and this is consistent with the sequence data from which we know that at least in certain individuals there are probably three different closely related genes. So I think the reservations that we should keep in mind are how big are the error bars, that is, is it possible still to have 5-10 genes per subgroup. Now, if that is a possibility, and I just raise it for your consideration, the next really important question one has to consider is, what is the total repertoire of antibody molecules. Clearly, if it is 10^7 or 10^8, there are too many antibody molecules to be accounted for by a germ-line model. If there are 10^5 or 10^6 and in addition other mechanism such as combinatorial association of V_H and V_L, then perhaps a germ-line contribution would still end up by being freely a major contribution to antibody genes. But with that brief summary, I would like to open these two fine papers for questions from the audience.

Dr. Rajewsky: I should like to ask the speakers to answer the Chairman's challenge. Cesar Milstein was clearly saying that he wanted to make a point on the basis of his data, namely that he concludes that the number of germ-line genes is lower than would be required for the expression of antibody variability. Dr. Tonegawa made the same point. The Chairman was saying that he was not convinced. Wouldn't you like to comment on this?

Dr. Milstein: In this story of number of genes and diversity there have been all sorts of ups and downs and the fashion generally accepted has dramatically changed in a short number of years. There have always been a number of stubborn people who kept saying "I am not convinced," who eventually changed the fashion by some new experiments. I think it is very healthy that there are some stubborn people who will look for small flaws which do occur in these experiments. The experiments point in one direction but they are not absolutely cut and dried. If one wants to look for uncertainties, one could say, we are not really absolutely sure, and, in fact, he is right in raising the point. We are not absolutely sure and we could certainly destroy each individual experiment. I think, what to me is particularly convincing (and I would like to perhaps address now the Chairman and explain to him why I am convinced, or *fairly* convinced, anyway) is the accumulation of data. There are several cases and not one which is well done, there are several which are well done and that is what convinces me.

Dr. Hood: Yes, I have to agree that the bulk of the nucleic acid hybridization data certainly points to a smaller number of antibody genes; the issue, however, is open whether you need to have somatic mutation as an important mechanism. However, one can look at other experiments, too, such as the very beautiful ones reported by Dr. Eichmann which suggest that rather closely related genes the products of which all bind with a particular kind of antigen are pseudo-alleles, that is, separate gene products that can be mapped into the genome. I think this kind of data raises very intriguing and important questions. It is not as elegant nor as satisfying as the nucleic acid data will eventually become, because we are one step removed from the direct source of information, the genes themselves, but to me this argues in a very compelling way that the final number of antibody V genes, whether or not somatic mutation is important, is going to be rather large.

Dr. Tonegawa: Well, first of all one technical point, whether one can distinguish one to two copies from say five copies, and I completely agree with the Chairman that this would be very difficult. Now when you start talking about distinguishing 20 to 30 genes from one to five copies I feel quite comfortable. Dr. Hood said that we were using hemoglobin messenger as a standard, in fact I thought I explained that we did use other RNA which is known to be unique as a standard like $E.$ $coli$ RNA made in $vitro$. The data is such that the cot-curve of reiteration frequency of lambda messenger RNA cannot correspond to more than a few genes. My second point is that the total number of germ-line V genes and the relative number of germ-line V genes as compared to the size of the V-region repertoire in an adult animal are two completely separate issues. The issue which we addressed ourselves to is the second one. We still do not know what the total number of V genes is.

Dr. Starlinger: Regarding your translocation mechanisms, if I understood you correctly, you discuss either a direct translocation, that means excision and translocation or, as an alternative, you discuss the removal of intervening material by a Campbell-type mechanism. For the latter I think you could have a test because, if I understood correctly, people are discussing that in the course of the maturation of a specific clone, you have first a specific V gene attached to a μ chain and, then later, you have the same V gene attached to a γ chain. Now, if the excision-type mechanism were correct, then you would expect that the C region for the μ chain should be removed upon putting the V gene to the γ chain and this you could test.

Dr. Tonegawa: I don't want to generalize at this time that this V-C gene joining in the light chain is a similar situation to that existing in the so-called μ-γ- or μ-α-switch. I like the unifying model but we don't know it, and we should be careful about it. There are ways of testing this model but it has not been done.

Dr. Hood: Let me add a postscript to that - I will be very surprised if either of the models you think of are most likely correct because of what we have seen with heavy chains. The pertinent observation there, although not completely verified, it that, on rare occasions, a single cell can make two different classes of heavy chains which apparently have precisely the same V gene by idiotypic analysis. That is the important reservation. If those V-region sequences from both heavy chains are identical, then it means that the mechanism for heavy chain VC joining must be one in which a single V is placed adjacent to many Cs that a single cell can simultaneously secrete both of these gene products. However, the critical analysis that is in progress now, is whether or not those molecules of different class that are idiotypically identical have V regions that are identical.

Dr. Poljak: I would like to ask a general question as a naive member of the audience here and this is, we have heard different proposals about the number of V genes in the germ line, and about somatic mutations giving rise to the diversification. I think positions have in later years tended to come together; also we do know that the maximum number of structural genes that we can have in a vertebrate genome is only about 40,000. Now I would like to ask you as a proponent of the germ-line theory how many genes do you need, and then I would like to ask the people who propose somatic mutations how many genes do they need? and see if the number is not in fact fairly close.

Dr. Hood: Well, there are two points. I suspect there are people who might argue with the 40,000 gene number in either direction. My own view is, I can't answer that question because we have no idea about the total repertoire. Obviously, in a sense, if one believes in a straight germ-line theory you need the square root of the number of genes that are required in the repertoire if you can put all light and heavy chains together in this combinatorial fashion. I would point out that a thousand light and a thousand heavy chain genes can generate 10^6 different specificities to antibody molecules and, in my view, that is more than enough. The whole issue of how big the repertoire is and notwithstanding some recent estimates that are very high, I think it is entirely unsettled and extremely intriguing and important. Suppose the number is 10^5, then I think there would be no problem whatsoever with the number of germ-line genes. There will be people that will jump up and say that can't possibly be right, but I don't think there is compelling data one way or the other.

Dr. Milstein: I said that actually. I said, if you ask me how many genes there are, the first thing I will ask you is, in which species and of which chain. One of the main points I am trying to make is that the number is going to vary. There are other variables which are, in my mind, very important, which are the ones that are going to diminish the load of the number of genes because of the facility of production of mutants which will be useable at the time the animal needs it, which is when it is born. I will say that, as a general rule of thumb, the number of genes is roughly of the order of the number of subgroups which are defined by protein chemistry data. Does that answer your question?