A Biological Arms Race:
Site Specific DNA Recombination in Competing Immunofunctional Proteins

Thesis by

Joseph Thomas Meier

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

Biology Division

California Institute of Technology

Pasadena, California

1993

Thesis Defense 17 September, 1992

If you want to know how to do a thing you must first have a complete desire to do that thing. Then go to kindred spirits - others who have wanted to do that thing - and study their ways and means, learn from their successes and failures and add your quota. Thus you may acquire from the experience of the race. And with this technical knowledge you may go forward, expressing through the play of forms the music that is inside you and which is very personal to you.

Robert Henri
The Art Spirit

Anybody who looks back over an experimental development which .... has continued for many years, can hardly fail to notice that it has pursued an exceedingly wobbly course. If the surveyor himself is an experimenter, he will know that the recorded wanderings are fewer and less extensive than those which actually occurred.

Frederick Bartlett
Thinking, An Experimental and Social Study

We can hardly expect a committee to acquiesce in the dethronement of tradition. Only an individual can do that, an individual who is not responsible to the mob. Now that the truly independent man of wealth has disappeared, now that the independence of the academic man is fast disappearing, where are we to find the conditions of partial alienation and irresponsibility needed for the highest creativity?

Garrett Hardin, biologist
Nature and Man's Fate, 1959

ACKNOWLEDGMENTS

The acquisition of a doctorate is considered to be one pinnacle of achievement in formal scholastic training, and properly so. Given this level of importance, I believe it is necessary to acknowledge those persons who have significantly contributed to this accomplishment, however far removed in time.

Completion of this endeavor is accompanied by the movement of Leroy Hood's laboratory from Caltech to the University of Washington. As a "Hoodlum," I was part of an environment few are fortunate enough to experience: a laboratory with enough money, space, and technical expertise to permit concentration on scientific matters rather than political or economic issues. I have gained much from my time in Lee's apprenticeship, and I deeply appreciate the opportunities he provided. Elliot Meyerowitz, Barbara Wold and Pamela Bjorkman, all members of my thesis committee, should be commended for their patience. I particularly wish to mention Barbara's facility at providing me with insightful analyses, both of my "gedanken" experiments and myself, and Pamela's encouragement and commitment to finishing at a time when my own flagged. During my stay here I have had the great fortune to be able to interact with many talented and creative students and post-docs, including Margaret Fahnestock, Steven Clark, Charles Spence, Ruedi Aebersold, and Gerry Siu. However, if I have progressed far enough to truly merit this degree, I believe that the tutelage of Dr. Susanna M. Lewis is primarily responsible. Her professional skill and personal integrity have commanded my respect, especially in light of the arduous course these studies sometime took.

Work on the *Borrelia* project would not have been possible without the support and guidance of another committee member, Dr. Alan G. Barbour. His generosity and personal attention enabled me to learn a great deal about a fascinating organism. I found everyone in the Laboratory of Microbial Structure and Function at the RML interested in assisting me.

Combined with the surrounding Bitterroot Valley, it made an ideal situation for doing research. One other academician requires notice here: Dr. Douglas Phinney, whom I first encountered at the Johnson Space Center's Ion Microprobe Facility. Doug, who later became my adviser/employer when I finished my undergraduate thesis project at the Lunar and Planetary Institute, was the sponsor of an unofficial *salon* which met in a cubbyhole office crammed with electronics, books, and ideas. It was a memorable time, and his friendship and sage advice has continued to be important during my graduate career.

Finally, I would like to thank my family: my parents, Thomas and Nancy, for their constant encouragement (beginning with the nurturing of my childhood love of the written word), and my siblings, Timothy, Ann, Mary, and T. J., for their willingness to help me whenever possible.

ABSTRACT

This thesis is a compilation of inquiries into the molecular biology of two disparate organisms, each using site-specific recombination to generate diversity in and regulate the production of a protein. Coincidentally, each of these proteins functions in the context of the vertebrate immune system: one is the major defensive weapon of a eubacterial pathogen, and the other the *sine qua non* of the system designed to recognize and destroy infectious agents, the antibody. The first section describes a series of experiments designed to explore the molecular basis for antigenic variation in *Borrelia hermsii*, the eubacterial agent responsible for relapsing fever. A serotype 7 *vmp* gene fragment was cloned using mixed sequence oligonucleotide probes derived from the sequencing of CNBr peptides from VMP 7. Use of this fragment in northern and southern blot experiments demonstrated that *B. hermsii* DNA sequences duplicate and rearrange, and that these duplications correlate with differential expression of VMPs (in a pattern remarkably reminiscent of the trypanosomes). These striking results formed the basis for several subsequent studies, which are also discussed. The final section details two separate projects involving V(D)J recombination. Inital effort was directed at producing a non-lymphoid cell line capable of performing V(D)J recombination. Our strategy was based upon the ability of retroviruses to transcriptionally activate genes distant from the site of integration. Due to reports of the cloning of RAG-1 and -2, the project was discontinued, but not before producing one line with an interesting phenotype. Following largely anecdotal reports of a previously unnoticed pattern of base addition during V(D)J recombination, we decided to perform a rigorous examination of the hypothesis, using both experiment and a detailed examination of published data. While we were able to confirm the existence of palindromic, non-templated bases, our results contradicted other reports with regard to the origins and characteristics of

these inserts. Some surprises arose, most notably in the influence primary DNA sequence has on the spectrum of product molecules; this adds a new dimension to a process previously thought to be well understood. This work represents the most thorough study of P nucleotide addition to date.

TABLE OF CONTENTS

CHAPTER ONE

Antigenic Variation in *Borrelia hermsii*, Agent of Relapsing Fever in Humans

INTRODUCTION

The power and flexibilty of DNA recombination as a method of gene regulation were first made apparent to me during a medical microbiology survey course given at UC San Diego Medical School by Dr. John Holland. Although the nature of the class precluded detailed examination of any one system, the medical orientation of the course served to underline the four major strategies used by parasites to evade the vertebrate immune response:[1,2]

1) colonization of sites within the organism which are protected from attack (i.e., intra-cellular or immune-priviledged areas);

2) antigen mimicry, where host proteins are adsorbed, or foreign proteins bearing host antigens are produced by the parasite;

3) modulation of immune response (suppression) due to blocking of regulatory proteins;

4) antigenic variation.

The variety of known and inferred mechanisms used by parasites employing the last of these strategies piqued my interest, both esthetically and as potential thesis topics (all known examples of antigenic variation involved rearrangement of the organism's genetic material, a process which captivated my imagination). Of the different mechanisms described, those involving antigenic variation in bacteria seemed to be the most amenable to a straightforward analysis; in paricular, two systems held a particular fascination: the DNA inversion of the Cin/Gin/Hin system[3] (a biological binary switch), and the multiple switches of *Borrelia hermsii*, the agent of relapsing fever.[4]

My interest in the recombinational mechanisms of these organisms led me to the laboratory of Dr. Melvin Simon (then at UC San Diego). I began work on the *Salmonella* Hin project with the understanding that my main goal was in the examination of the molecular basis of antigenic variation in *Borrelia hermsii*. Unfortunately, the major barrier to progress was the absence of any indication that the various *B. hermsii* serotypes had been cloned, or could even be grown reliably

in vitro. At this point, a serendipitous meeting created an opportunity for my participation in this study: a remark overheard at an Agouron Institute beach party for a former graduate student of Dr. Simon's (Michael Silverman) resulted in an introduction to the Rocky Mountain Laboratories (an NIH facility). Coincidentally, Dr. Alan Barbour had recently succeeded not only in producing cloned *Borrelia* serotypes *in vitro*,[5] but also a set of non-crossreactive monoclonal antibodies for each of the 3 cloned serotypes.[6] Discussions with him resulted in a fruitful 2 year stay at the Laboratory of Microbial Structure and Function, Rocky Mountain Laboratory, Hamilton, Montana. The results are detailed in ***Cell*, 41**, 403-409, (1985).

CHAPTER TWO

Antigenic Variation Is Associated with DNA Rearrangements in a Relapsing Fever *Borrelia*. (1985). *Cell* **41**, *403 - 409*.

Cell, Vol. 41, 403–409, June 1985, Copyright © 1985 by MIT 0092–8674/85/060403–07 $02.00/0

Antigenic Variation Is Associated with DNA Rearrangements in a Relapsing Fever Borrelia

Joseph T. Meier,*† Melvin I. Simon,† and Alan G. Barbour*‡
* Department of Health and Human Services, Public Health Service, National Institutes of Health, National Institute of Allergy and Infectious Diseases, and Laboratory of Microbial Structure and Function and Laboratory of Pathobiology
Rocky Mountain Laboratories
Hamilton, Montana 59840
† Division of Biology
California Institute of Technology
Pasadena, California 91125

Summary

Borrelia hermsii, an agent of relapsing fever, undergoes antigenic variation in its host. Surface-exposed proteins with differing primary structures determine the serotype of each organism. Using amino acid sequence data from two of these variable proteins, we synthesized two mixed-sequence oligonucleotides and then used the oligonucleotides to probe mRNA and DNA of three isogenic serotypes of B. hermsii. In Northern blots the probes were specific for the mRNA of the homologous serotype. Southern blots revealed two classes of hybridizing fragments: those common to the three serotypes and those specific for a particular serotype. A serotype-specific DNA fragment, which had hybridized to both oligonucleotide probes, was cloned. Subsequent use of the cloned fragment as a probe provided further evidence that antigenic variation in B. hermsii is associated with DNA rearrangements and with occurrence of expression-linked copies of all, or part, of an antigen-specifying gene.

Introduction

Arthropod-borne members of the eubacterial genus Borrelia cause relapsing fever in humans and analogous diseases in other mammals (reviewed in Felsenfeld, 1971). Relapsing fever was noted as a nosologic entity in antiquity, and epidemics of the disease have continued to affect human societies into the twentieth century. As the disease's name suggests, a person with relapsing fever characteristically has periods of illness spaced by intervals of well-being. When the fevers occur, numerous spirochetes can be found in smears of blood from patients. The borreliae disappear from the blood as the host responds to them with specific antibodies, but they reappear a few days later.

Antigenic variation accounts for the waxing and waning populations of borreliae in the host (Meleney, 1928). Twenty-six antigenic variants or serotypes have been isolated from the progeny of a single organism of Borrelia hermsii strain HS1 that infected a mouse (Stoenner et al., 1982; Barbour and Stoenner, 1985). The rate of switching between serotypes in vivo was estimated to be 10^{-4} to 10^{-3} per cell per generation (Stoenner et al., 1982). The emergence over time of relapse serotypes in mice, although not completely predictable, does follow a loose order with a bias toward a subset of serotypes during the early stages of the infection (Stoenner et al., 1982; Barbour and Stoenner, 1985). A serotype eliminated by neutralizing antibody from the blood of a first host may reappear in the relapse populations of a nonimmune second host (Meleney, 1928; Coffey and Eveland, 1967). The process is, therefore, reversible.

In many of these biological features, the antigenic variation of the prokaryotic borrelia resembles the antigenic variation of the eukaryotic trypanosome (Cross, 1978; Vickerman, 1978; Borst and Cross, 1982). Another similarity between these two vector-borne pathogens is the interrelation between a change in the prevalent surface protein of the cells and the appearance of new serotypes in the blood.

In B. hermsii these serotype-specific antigens have been designated variable major proteins (VMPs); a subscript identifies the serotype in which the VMP is found. The VMPs of all serotypes of B. hermsii HS1 examined to date differ in their molecular weights, peptide maps, and reactivities with serotype-specific polyclonal and monoclonal antibodies (Barbour et al., 1982, 1983). In situ VMPs were cleaved from the cell by proteases (Barbour, 1985) and were radioiodinated under surface-specific labeling conditions (Barbour et al., 1982). In addition, VMP-reactive monoclonal antibodies bound to, and agglutinated, homologous borreliae (Barbour et al., 1984). These results led us to conclude that VMPs are exposed on the cell's surface.

Our studies have focused on three serotypes derived from a single organism of B. hermsii. Serotypes 7 and 21 were recovered from spirochetemic mice; serotype C was found in a population of serotype 7 organisms that had been in broth culture for several passages (Stoenner et al., 1982). Serotype C lacks demonstrable antigenic relatedness to either of the other two serotypes (Barbour et al., 1982; Barbour, 1985). Serotypes 7 and 21, although differing in their reactivities with most antibodies in a typing battery, have in common an epitope recognized by one monoclonal antibody (Barbour, 1985). The apparent molecular weights of VMP_C, VMP_7, and VMP_{21} are 20,000, 39,000, and 38,000, respectively. Comparison of partial amino acid sequences of CNBr peptides from VMP_7 and VMP_{21} revealed short conserved regions as well as highly variable regions (Barstad et al., 1985).

These results indicated that B. hermsii is polymorphic with respect to VMPs. To define the mechanism that generates this degree of antigenic diversity in a bacterium, we examined DNA and RNA of these organisms for changes associated with VMP switching. We demonstrate

‡ To whom reprint requests should be addressed at the Laboratory of Pathobiology, Rocky Mountain Laboratories.

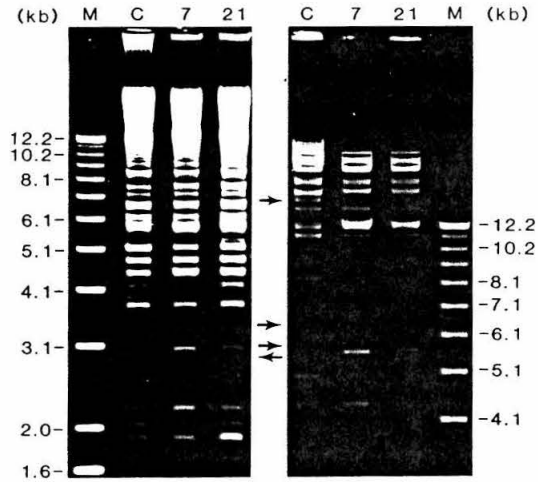


Figure 1. Pst I Digest of Total DNA from Serotypes C, 7, and 21 of Borrelia hermsii HS1

Left, 0.8% agarose gel; right, 0.5% agarose gel. M: "1 kb ladder" markers (Bethesda Research Laboratories). Arrows indicate differences in the restriction patterns in serotype 21 (left) and serotype C (right) DNA.

that DNA sequences of B. hermsii rearrange and that these rearrangements correlate with the differential expression of VMPs.

Results

Nucleic Acid Extractions

Standard methods were inefficient for extraction of DNA from B. hermsii; the crude high molecular weight DNA tended to partition in the phenol phase during extraction of the cell lysate. Two alterations increased the final DNA yield: first, the amphiphilic material at the interface was carefully aspirated when the aqueous phase was removed, and second, the partially purified DNA was treated repeatedly with proteinase K in the presence of SDS. RNA extraction procedures required less modification.

Although many restriction enzymes, including Sau 3A and Dpn I, digested the extracted DNA, the failure of Mbo I to cut the DNA suggested the presence of an adenine methylation system. Another constraint on the use of restriction enzymes was the very low GC (30%) content of this organism's DNA (Schmid et al., 1984); those endonucleases with GC-rich recognition sequences produced patterns of digestion skewed toward the higher molecular weight end of the gel.

An ethidium-bromide-stained gel of Pst I digests of the DNA from the isogenic serotypes C, 7, and 21 is shown in Figure 1. The restriction patterns were essentially the same, but differences were observed. For instance, serotype 21 had a 2.8 kb fragment that was not seen in either serotype C or 7. The restriction pattern obtained using serotype C DNA had unique bands in at least three locations. Thus, even at a gross level, characteristic arrangements of the DNA sequence appear to be associated with each serotype.

Probe α

CB1 amino acid sequence Gly Glu Asn Asp Ala Gln

Oligonucleotide sequence $_{3'}$CCN CTY TTR CTR CGN GT$_{5'}$

Probe β

CB2 amino acid sequence Ala Glu Asn Ala Phe Tyr

Oligonucleotide sequence $_{3'}$CGN CTY TTR CGN AAR AT$_{5'}$

Figure 2. Partial Amino Acid Sequences of the Two Cyanogen Bromide Fragments (CB1 and CB2) of VMP$_7$ of B. hermsii Serotype 7

The CB2 sequence was also found in one of the CNBr fragments of VMP$_{21}$ of serotype 21 (Barstad et al., 1985). The sequences were used to construct mixed oligonucleotide probes α and β. A, adenosine; C, cytosine; G, guanine; T, thymine; N, any nucleotide; R, purine; Y, pyrimidine.

Oligonucleotide Probes

Mixed oligonucleotides that correspond to selected amino acid sequences of VMP$_7$ were synthesized and thereafter served as probes to identify DNA that encodes VMPs. The isolation, CNBr cleavage, and partial sequence analysis of VMP$_7$ and VMP$_{21}$ are reported elsewhere (Barstad et al., 1985). Two regions of the known amino acid sequence of VMP$_7$ were candidates for minimally complex oligonucleotide probes and are shown in Figure 2. One series of residues was located in the first of two CNBr fragments of VMP$_7$; this series did not occur in the known sequence of VMP$_{21}$. Probe α was based on this apparently unique amino acid sequence; consequently, it was assumed to be specific for DNA encoding VMP$_7$. In contrast, probe β was derived from an amino acid sequence that was common to the second CNBr fragment of VMP$_7$ and one of three CNBr fragments of VMP$_{21}$. Probe β was expected to detect DNA sequences coding for VMP$_7$ and VMP$_{21}$. The oligonucleotides were designed 3′ to 5′, i.e., anti-sense, to allow their use for both Southern and Northern hybridization.

To assess its mRNA specificity, probe α was hybridized to a Northern blot of total RNAs from serotypes C, 7, and 21 (Figure 3). The left part of the figure shows an acridine-orange-stained gel of the RNA preparations; the migrations of borrelial 16S and 23S RNAs were similar to those of E. coli 16S and 23S ribosomal RNA (data not shown). The [^{14}C]uridine-labeled borrelial RNA was used for markers in this and subsequent Northern blots. The autoradiograph of the Northern blot after hybridization with probe α is shown in the right part of Figure 3. This probe did not bind to any sequence in the serotype C or 21 RNAs. A single, strong signal was present, however, in the serotype 7 RNA lane. The size of this hybridizing RNA species was estimated to be 1100 bases.

Probe α, specific as it was for a serotype 7 transcript, was then used to probe a Southern blot that comprised total DNAs from the three serotypes (Figure 4). Probe α hybridized to two fragments in the Pst I–cut DNAs. Whereas the fragment of 8.5 kb was common to all three serotypes, a 2.9 kb hybridizing band was detected only in the serotype 7 DNA. Bgl II–digested DNA, when it was

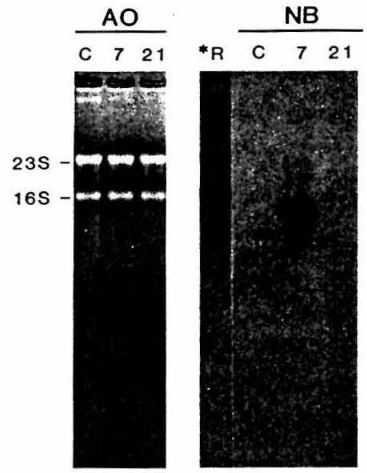


Figure 3. Assessment of Probe α mRNA Specificity

Left (AO): acridine-orange-stained gel (1.8%) of total RNAs from serotypes C, 7, and 21 of B. hermsii HS1. The 23S and 16S ribosomal bands are indicated. Right (NB): autoradiograph of Northern blot hybridized with oligonucleotide probe α that had been labeled with ^{32}P. The hybridization temperature was 24°C. *R: [^{14}C]uridine-labeled RNAs from serotype C; the major bands are 23S and 16S ribosomal RNAs.

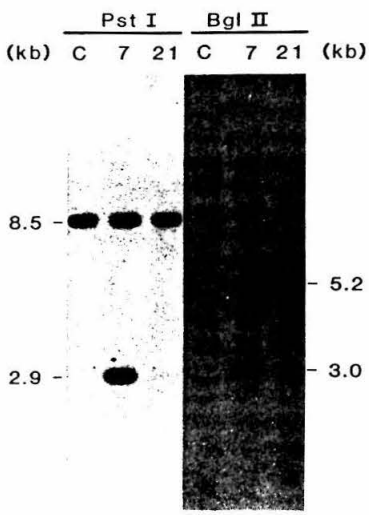


Figure 4. Southern Blot of Pst I and Bgl II Digests of Total DNAs from Serotypes C, 7, and 21

The blot was hybridized with ^{32}P-labeled oligonucleotide probe α. The hybridization temperature was 37°C.

reacted with probe α, produced the same general pattern in a Southern blot: a common fragment of 5.2 kb and a serotype 7 specific band at 3.0 kb. Thus, DNA sequences that corresponded to VMP_7 were present in serotypes C and 21, even though the VMP_7 gene products were apparently not expressed in these latter serotypes. Furthermore, a new form of the VMP_7 sequence occurred only in the DNA of serotype 7.

Probe β was applied in a comparable set of experiments (Figure 5). The left part of the figure shows the hybridization of this oligonucleotide mixture to a Southern blot of Pst I and Bgl II digests. In the Pst I part of the blot, probe β, like probe α, bound to a 8.5 kb fragment common to the three serotypes. While probe β duplicated probe α's hybridization to a 2.9 kb fragment in serotype 7, it alone bound to 2.9 and 2.8 kb fragments in serotype 21. The blot of the Bgl II digest showed two major sets of hybridizing fragments common to the three serotypes: one set at 5.2 kb and a second at 10.3 kb. A unique 3.0 kb hybridizing fragment was again seen in the serotype 7 digest, but there was also a band of 1.9 kb in serotype 21 DNA. In Northern blots with probe β, there was a band of about 1100 bases in serotype 7 RNA and a less strongly hybridizing RNA species of slightly smaller size in serotype 21 (Figure 5). The findings with probe β confirmed the presence of a serotype 7 specific DNA fragment and also indicated that there were DNA sequences unique to serotype 21.

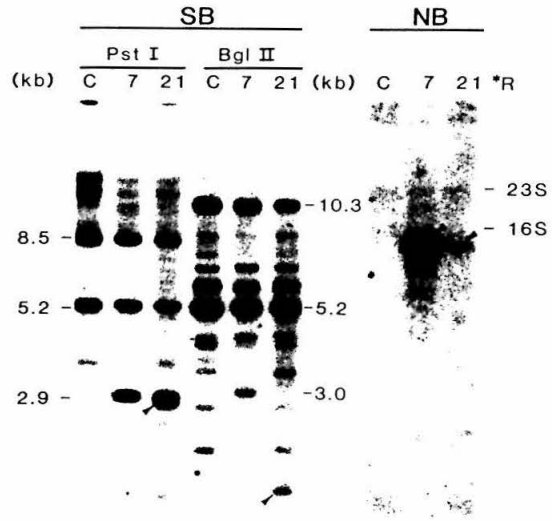


Figure 5. Hybridization of Oligonucleotide Probe β to Southern and Northern Blots

Left (SB): Southern blots of Pst I and Bgl II digests of DNA from serotypes C, 7, and 21. The hybridization temperature was 32°C. Arrow in Pst I digest indicates bands of 2.8 kb and 2.9 kb in serotype 21 DNA. Arrow in Bgl II digest indicates a fragment of approximately 1.9 kb in serotype 21 DNA.

Right (NB): Northern blot performed on total RNAs from the three serotypes. The hybridization temperature was 17°C. Arrows indicate the two hybridizing species. Migration of [^{14}C]uridine-labeled 23S and 16S RNAs are indicated (*R).

Cloning of Borrelial DNA

Our next efforts were toward cloning the 2.9 kb Pst I fragment that hybridized to both oligonucleotide probes and was unique to serotype 7 DNA in this affinity. A pool of restriction fragments approximately 2.7 to 3.2 kb in size were isolated and cloned into pBR322. Two of seven recombinants examined had the same 2.9 kb Pst I fragment. One of the two plasmids was designated pRML7.1; it was recovered in large amounts by standard procedures. The plasmid DNA was cut with Pst I and compared with similarly

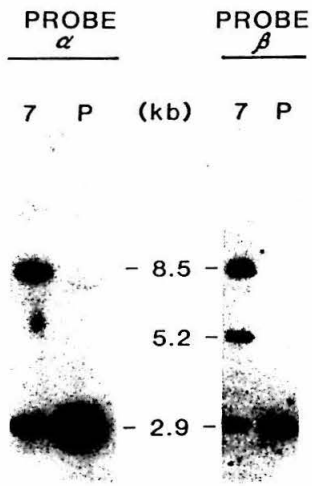


Figure 6. Hybridization of Oligonucleotide Probes α and β to Pst I Digests of Serotype 7 DNA and of pRML7.1

See legends of Figures 4 and 5 for hybridization conditions. Lane 7: serotype 7 DNA. Lane P: pRML7.1.

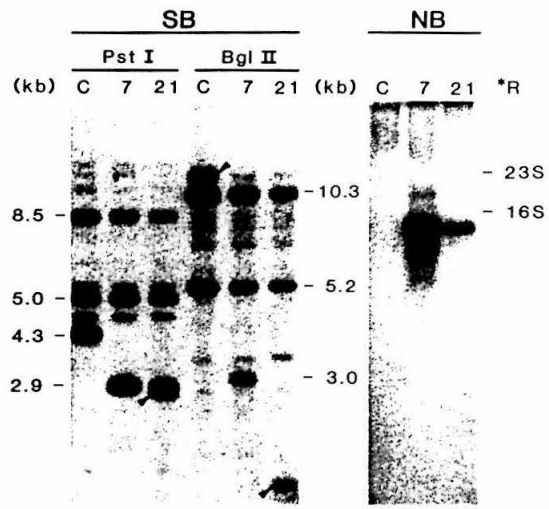


Figure 7. Hybridization Experiments Using the Nick-Translated 2.9 Kb Insert from pRML7.1

Left (SB): Southern blot of Pst I and Bgl II digests from serotypes C, 7, and 21. The hybridization temperature was 68°C. Arrow in the Pst I digest indicates bands of 2.8 and 2.9 kb in serotype 21; arrows within the Bgl II field point to fragments of 12 kb (serotype C) and 1.9 kb (serotype 21). Right (NB): Northern blot using total RNAs. The hybridization temperature was 37°C. Positions of the 23S and 16S RNAs were determined by migration of [^{14}C]uridine-labeled RNA (*R).

digested serotype 7 DNA in a Southern blot (Figure 6). The cloned DNA bound to both oligonucleotide probes and had the same electrophoretic mobility as the original borrelial DNA fragment. Western blot analysis did not reveal any expression product of ED8654 (pRML7.1) that was antigenically related to VMP_7 (data not shown).

Hybridization with Cloned DNA

The borrelial DNA insert in the recombinant plasmid was nick-translated and used in a set of Southern and Northern blots (Figure 7). The Southern hybridizations are seen in the left part of the figure. The patterns of binding of the labeled probe to the enzyme digests were similar to those produced with probe β. In the Pst I digest, a 8.5 kb band was again identified in the three serotypes. Although a 5.2 kb band was common to all serotypes, the hybridization signal was weaker than the signal with probe β, and a new 5.0 kb fragment was the other major common band. In the Bgl II digest Southern blot, common bands of 10.3 and 5.2 kb, which had previously been identified by probes α and β, hybridized with the cloned DNA.

The borrelial DNA probe revealed serotype-specific fragments in each of the three serotypes. In the Pst I digest, serotype 7's 2.9 kb band and serotype 21's 2.8 and 2.9 kb bands were present, but there was also a unique band of 4.3 kb in serotype C DNA. Similar results were obtained with the Bgl II blot. However, in this latter blot the serotype C specific band was approximately 12 kb and comigrated with weakly hybridizing bands in serotypes 7 and 21.

A Northern blot with the nick-translated probe is shown in the right side of Figure 8. The cloned borrelial DNA, like probe β, hybridized to a single band of approximately 1100 bases in serotype 7 RNA and more weakly to a slightly smaller RNA species in serotype 21. The cloned fragment did not detect any homologous sequences in the serotype C RNA.

These data showed that the cloned Pst I fragment included DNA sequences corresponding to VMP_7, and the same or other sequences corresponding to VMP_{21}. Moreover, the cloned fragment contained sequences that rearranged in serotype C.

Discussion

We have shown that the DNA of B. hermsii rearranges and that application of certain DNA probes in Southern blots enables one to distinguish between antigenic variants. The two mixed oligonucleotide probes, α, and β, were designed from our knowledge of unique (α) and common (β) amino acid sequences in VMP_7 and VMP_{21}. Using probe α, we isolated and cloned a 2.9 kb DNA sequence from serotype 7 cells; this sequence was in turn used as a probe. A schematic summary of hybridization patterns of the three probes to Pst I–generated fragments of serotype C, 7, and 21 DNA is shown in Figure 8.

The results of the Southern blot with probe α suggest that serotype 7 DNA had an additional copy of a VMP_7-specific DNA sequence. The extra copy was in a restriction fragment of different size than the fragment containing a homologous sequence common to serotypes C, 7, and 21. The Northern blot showed that possession of the extra copy was associated with transcription of the sequence. Moreover, because VMP_7 occurs only in serotype 7 cells (Barbour et al., 1982), the extra copy can also be

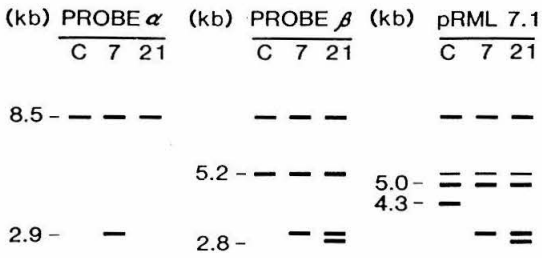


Figure 8. Schematic Summary of the Hybridization Patterns Produced Using Oligonucleotide Probes α and β and the 2.9 Kb Insert in pRML7.1 against Pst I–Digested DNAs from Serotypes C, 7, and 21

linked indirectly to VMP_7 expression. This interpretation brings to mind the mechanisms that trypanosomes (Williams et al., 1979; Hoeijmakers et al., 1980; Pays et al., 1981) and gonococci (Meyer et al., 1982; Stern et al., 1984) use to vary their surface proteins: a copy of all or a part of an antigen-specifying gene is present in a storage form, and expression results from duplicative transposition of it to an expression site.

The hybridization patterns with probe β were also consistent with a mechanism involving an expression-linked copy and a storage copy of a VMP gene. We had synthesized probe β on the basis of primary protein structure known to be common to VMP_7 and VMP_{21}, and, indeed, probe β identified extra copies of a DNA sequence in both serotype 7 and 21. Occurrence of an extra copy of a probe β specific sequence correlated with transcription of the sequence. The stronger hybridization of probe β to the mRNA of serotype 7 than to the mRNA of serotype 21 cannot be accounted for at this time. Because the effective melting temperature of DNA–RNA hybrids incorporating probe β was several degrees lower than the lowest melting temperature predicted for the mixture of oligonucleotides, we cannot rule out, at this time, an error in the protein sequence. Alternatively, a single base difference between the analogous RNA sequences could have produced the inequality of hybridization signals (Wallace et al., 1981). The presence of the two extra copies of the sequence in serotype 21 DNA, instead of the one found in serotype 7, suggests a third explanation: two genes, which were not necessarily identical, may have been transcribed in serotype 21. In any case, there was not detectable transcription in serotype C of a sequence homologous with probe β.

The 2.9 kb cloned insert in pRML7.1 contained sequences that hybridized to both oligonucleotide probes. Use of this pRML7.1 insert as a probe produced Southern hybridization patterns that were similar to those of probe β. In Southern blots with probe β or the plasmid insert there were two sets of strongly hybridizing bands common to the three serotypes. Both probes revealed one unique band in serotype 7 and two unique bands in serotype 21. However, only the borrelial DNA probe identified a serotype-specific sequence in serotype C DNA. By means of hybridization to a flanking region of the VMP gene, the pRML7.1 probe may have revealed a site involved in the expression of VMP_C. The findings with the pRML7.1 and β probes also indicate greater sequence homology between transcripts for VMP_7 and VMP_{21} than between transcripts for VMP_C and either of the other two VMPs.

We have dwelled upon restriction fragments that, by virtue of their hybridization with different DNA probes, allowed discrimination between the three serotypes. But what of the hybridizing bands that were common to the serotypes? The major of these common bands in the Pst I digests are represented in Figure 8. Several other common fragments hybridized weakly with the probes; minor bands such as these were also seen when prob es for variant surface glycoprotein genes were hybridized with trypanosomal DNA digests (Pays et al., 1981). There was no evidence that the common bands denoted transcriptionally active genes, but neither is there evidence the bands represented silent copies of genes, which in different surroundings express functional VMPs. Nevertheless, mindful of the parallels between antigenic variation of relapsing fever borreliae and the salivarian trypanosomes, we find compelling as a mechanism one that invokes an expressed copy from a storage copy. The length to which the analogy between borreliae and trypanosomes extends is unknown; the fundamental differences in organization between prokaryotic and eukaryotic genomes impose limits on that length.

The present series of experiments take us toward an understanding of the molecular basis for antigenic variation in relapsing fever. These findings come more than a century after the identification of blood-borne spirochetes as the agents of the disease (Obermeier, 1873). Despite the initial excitement its discovery produced among pioneering immunologists such as Metchnikoff, Gabritchevsky, and Erhlich (reviewed in Russell, 1936), the borrelia drew little attention for several decades. Now that Stoenner has reawakened interest in the biology of borreliae, the application of available techniques to the molecular biology of relapsing fever is in order. It may be that further definition of the mechanism of borrelial antigenic variation will be germane not only to studies of prokaryotic gene regulation and conversion but also, as some early immunologists thought, to examination of the interface between pathogens and the immune system.

Experimental Procedures

Bacterial Strains

The origin of B. hermsii HS1 (ATCC 35209) and its serotypes C, 7, and 21 has been described previously (Stoenner et al., 1982; Barbour et al., 1982). Cultivation of borreliae, first in irradiated mice and subsequently in broth, was essentially as described previously (Barbour et al., 1983) with the exceptions that the radiation dose per mouse was 650 rads and the medium formula was BSK II (Barbour, 1984). Spirochete cultures in late logarithmic phase (approximately 5×10^7 cells per ml) were harvested by centrifugation (9000 × g for 20 min). Homogeneity of borrelial populations in the harvest was evaluated by indirect immunofluorescence using serotype-specific monoclonal antibodies (Barbour et al., 1982; Barbour, 1985). Only harvests with less than 0.5% cross-contamination were used in these studies. Serotype C, the variant that appears to have a selective advantage in broth medium (Stoenner et al., 1982), was the only detectable contaminant of harvests of serotype 7 and 21 cells. Other serotypes were not detected in serotype C harvests.

E. coli strains ED8654 (Murray et al., 1977) and DH1 (Hanahan,

1983) were used for the isolation and propagation of recombinant plasmids.

Preparation of Nucleic Acids

For DNA preparations, we used either freshly harvested cells or cells that had been kept at −76°C until use (Barbour et al., 1982). Approximately 5 × 10^{10} spirochetes were first washed with 10 mM Tris (pH 8.0), 150 mM NaCl, and 1 mM EDTA and then resuspended in 11 ml of cold 25% (w/v) sucrose in 50 mM Tris (pH 8.0). One milliliter of 0.5 M EDTA was added, and the suspension was placed on ice for 20 min. After addition of lysozyme (25 mg in 2 ml 0.25 M Tris, pH 8.0), the spirochete suspension was incubated at 37°C for 20 min. Cell lysis was achieved by adding first proteinase K (200 μg from a 2 mg/ml stock) and then sodium dodecyl sulfate (SDS; 20 ml from a 10% [w/v] stock). This mixture was placed at 37°C for 1 hr. The crude lysate was extracted twice with phenol:chloroform (50:50) and twice with chloroform alone. Cold ethanol was added to the aqueous phase; the resultant precipitate was collected by centrifugation and dissolved in 10 ml of TE (10 mM Tris, pH 8.0, 1 mM EDTA). This solution was treated with 200 μg of DNAase-free RNAase for 1 hr at 37°C. Proteinase K and then SDS were added for final concentrations of 200 μg/ml and 1%, respectively. Incubation at 37°C was carried out for 1 hr. Phenol and chloroform extractions were then repeated. Addition of 3 vol of cold ethanol to the aqueous phase immediately produced a thread-like precipitate, which was collected with a glass hook. The DNA was dissolved in TE, precipitated with ethanol, and redissolved in TE.

Total RNA was extracted from lysates of freshly harvested borreliae with hot phenol; the method was essentially that of Feramisco et al. (1982). The cell-lysing solution was 4 M guanidine thiocyanate, 25 mM sodium citrate (pH 7.0), 2% 2-mercaptoethanol, and 2% sodium lauryl sarcosinate. The final RNA precipitate was stored in 70% ethanol at −20°C. Prior to use, the RNA was recovered by centrifugation, dried, and suspended in RNAase-free water for a final nucleic acid concentration of 1 mg/ml. To radiolabel the RNA, 10^{10} freshly harvested cells were resuspended in 25 ml of BSK II medium containing 50 μCi of [^{14}C]uridine (ICN Biomedicals); the suspension was incubated at 35°C for 2 hr before harvesting a second time.

Oligonucleotide Synthesis and Labeling

Mixed oligonucleotides were synthesized on an Applied Biosystems model 380A apparatus (Hunkapiller et al., 1984). The crude product was dissolved in TE, mixed with loading buffer, and electrophoresed on a 20% polyacrylamide gel containing 7 M urea. The full-length oligonucleotide mixture was located by shadowing the gel with long wave UV radiation while the gel was on a thin layer chromatography plate (Merck #5507). The topmost band in the gel was cut out, and the oligonucleotide was eluted in 0.1 M NaCl/TE/0.1% SDS. The eluate was extracted first with phenol and chloroform and then with ether. The aqueous phase was adjusted to 20 mM in $MgCl_2$, and the oligonucleotide was precipitated with ethanol. The oligonucleotide was collected by centrifugation, dissolved in TE, extracted with phenol–chloroform, extracted with ether, precipitated with ethanol, and finally dissolved again in TE.

The purified oligonucleotides (0.1 μg) were combined with T4 polynucleotide kinase (2 units; New England Biolabs) and γ-^{32}P-ATP (300 μCi; ICN Biomedicals #35020) in 10 μl of 10 mM Tris (pH 7.4), 20 mM $MgCl_2$, 1 mM 2-mercaptoethanol. After 30 min of incubation of the reaction mixture at 37°C, second aliquots of ATP (100 μCi) and T4 kinase (1 unit) were added, and the reaction was allowed to proceed for an additional 20 min. The labeled oligonucleotide was separated from unincorporated ATP by electrophoresis on a 20% polyacrylamide gel; it was eluted from the gel and passed through a Millipore Millex-GV filter. The specific activity was 1-4 × 10^9 cpm/μg of oligonucleotide.

DNA and RNA Hybridizations

Southern hybridizations were performed essentially as described by Southern (1975) with some modifications. After 2–3 μg of total borrelial DNA was cleaved with a restriction enzyme, the fragments were separated on a 0.7% agarose gel in Tris–borate buffer. Transfer of the DNA to nylon membranes (Biodyne A, 1.2 μm pores; Pall Corp.) occurred in 20× SSC. The prehybridization and hybridization buffers were comprised of 10× Denhardt's, 6× SSC, 5 mM EDTA, and 0.1% SDS. For nick-translated probes, the buffer was supplemented with 100 μg of sonicated calf thymus DNA per ml. Hybridizations with oligonucleotide and nick-translated probes were performed for 10–24 hr; the radioactivity was 5 × 10^6 cpm/ml. After hybridization with oligonucleotide, the blots were washed for 5–10 min in 3× SSC, 5 mM EDTA, and 0.1% SDS at 6°C above the hybridization temperature. Following its incubation with the nick-translated probe, the blot was washed at the hybridization temperature in solutions that always contained 5 mM EDTA and 0.1% SDS but varied in salt concentrations: 1 hr with 4× SSC, 30 min with 2× SSC, 30 min with 1× SSC, and twice for 30 min each with 0.1× SSC (personal communication, M. Koomey). Two washes (15 min each) with 0.1× SSC alone at room temperature completed the procedure. The filters were placed on Kodak X-Omat AR film.

We performed Northern blot transfers using a modification of the method of Seed (1984). The electrophoresis buffer for the 1.8% agarose gels was as follows: 2.2 M formaldehyde; 0.2 M Hepes, sodium salt (pH 7.0); 50 mM sodium acetate; 10 mM EDTA. After completion of electrophoresis, the RNA was transferred by capillary action in 20× SSC to a nylon membrane. The prehybridization, hybridization, and washing buffer was 6× SSC, 50% formamide, 10× Denhardt's, 5 mM EDTA, and 0.1% SDS. The prehybridization temperature was always 37°C. For hybridizations with the nick-translated probe, sonicated calf thymus DNA was added for a final concentration of 100 μg/ml. After hybridization, the Northern blots were washed three times (15 min each) at the hybridization temperature and placed on film.

Cloning Borrelial DNA and Nick-Translating the Cloned Fragment

We modified the method of Dretzen et al. (1981) to isolate from agarose gels the DNA for cloning. After Southern hybridizations identified regions of interest, another restriction digest and electrophoresis were performed under similar conditions. DEAE paper was placed in a slit abutting and parallel to the lane of separated DNA. The gel was then turned 90° to its original orientation, and the DNA was electrophoresed onto the paper. Different pools of restriction fragments within a limited size range were created by cutting the paper into 2 mm widths and then eluting the DNA from these small strips. The pool containing the fragment of interest was identified by a dot blot hybridization with one of the oligonucleotide probes. (The hybridization conditions were those described above.) Ten nanograms of DNA from the chosen pool was mixed with 100 ng of Pst I–cut and dephosphorylated pBR322 (Bethesda Research Laboratories) in a volume of 30 μl and ligated with T4 ligase (BRL) overnight at 15°C. E. coli strain ED8654 was transformed with the ligated DNA by the method of Hanahan (1983). Small preparations of plasmid DNA (Holmes and Quigley, 1981) from tetracycline-resistant, ampicillin-sensitive transformants were digested with Pst I and subjected to Southern blot hybridization with the oligonucleotide probes.

Nick translation of the DNA fragment began with a standard cesium chloride gradient to purify the plasmid. After Pst I digestion of the plasmid, the cloned insert was isolated from an agarose gel by electroelution. Conditions for labeling the fragment with a commercial kit (BRL) and α-^{32}P-dATP (ICN Biomedicals) were those recommended by the manufacturer. A Sephadex G-75 column was used to separate the probe from the reaction mixture. Specific activities were 2–4 × 10^8 cpm/μg of fragment.

Western Blot Analysis

Whole cell lysates of ED8654 harboring either pBR322 or one of the recombinant plasmids were analyzed by Western blotting using a procedure previously described (Barbour et al., 1984). The polyclonal rabbit anti-serotype-7 serum was known to react with VMP_7 in Western blots (Barbour et al., 1982). The two monoclonal antibodies employed, H12436 and H12915, had been shown to bind in Western blots to one or the other of the two CNBr fragments constituting VMP_7 (Barstad et al., 1985).

Acknowledgments

We thank Merry Schrumpf for technical assistance, Suzanna Horvath for synthesis of the oligonucleotides, Bob Evans and Gary Hettrick for photographic work, Betty Kester for preparation of the manuscript, and Paul Barstad, Sven Bergstrom, Willy Burgdorfer, Leonard Mayer,

Ronald Plasterk, Herb Stoenner, and John Swanson for valued discussions and advice. J. T. M. was a predoctoral fellow supported by the National Science Foundation grant PCM 82-09295.

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 February 22, 1985

References

Barbour, A. G. (1984). Isolation and cultivation of Lyme disease spirochetes. Yale J. Biol. Med. *78*, 521–525.

Barbour, A. G. (1985). Clonal polymorphism of surface antigens in a relapsing fever *Borrelia* sp. In Pathogenesis of Bactrial Infection, G. G. Jackson and H. Thomas, eds. (Heidelberg: Springer-Verlag), pp. 235–245.

Barbour, A. G., and Stoenner, H. G. (1985). Antigenic variation of *Borrelia hermsii*. In Genome Rearrangement, I. Herskowitz and M. I. Simon, eds. (New York: Alan R. Liss, Inc.).

Barbour, A. G., Tessier, S. L., and Stoenner, H. G. (1982). Variable major proteins of *Borrelia hermsii*. J. Exp. Med. *156*, 1312–1324.

Barbour, A. G., Barrera, O., and Judd, R. C. (1983). Structural analysis of the variable major proteins of *Borrelia hermsii*. J. Exp. Med. *158*, 2127–2140.

Barbour, A. G., Tessier, S. L., and Hayes, S. F. (1984). Variation in a major surface protein of Lyme disease spirochetes. Inf. Immun. *45*, 94–100.

Barstad, P. A., Coligan, J. E., Raum, M. G., and Barbour, A. G. (1985). Variable major proteins of Borrelia hermsii: epitope mapping and partial sequence analysis of CNBr peptides. J. Exp. Med., in press.

Borst, P., and Cross, S. A. M. (1982). Molecular basis for trypanosome antigenic variation. Cell *29*, 291–303.

Coffey, E. M., and Eveland, W. C. (1967). Experimental relapsing fever initiated by *Borrelia hermsii*. II. Sequential appearance of major serotypes in the rat. J. Infect. Dis. *117*, 29–34.

Cross, G. A. M. (1978). Antigenic variation in trypanosomes. Proc. Roy. Soc. Lond. B. *202*, 55–72.

Dretzen, G., Bellard, M., Sassone-Corsi, P., and Chambon, P. (1981). A reliable method for the recovery of DNA fragments from agarose and acrylamide gels. Anal. Biochem. *112*, 295–298.

Felsenfeld, O. (1971). Borrelia: Strains, Vectors, Human and Animal Borreliosis. (St. Louis: Warren H. Green, Inc.).

Feramisco, J. R., Smart, J. E., Burridge, K., Helfman, D. M., and Thomas, G. P. (1982). Co-existence of vinculin and a vinculin-like protein of higher molecular weight in smooth muscle. J. Biol. Chem. *257*, 11024–11031.

Hanahan, D. (1983). Studies on transformation of *E. coli* with plasmids. J. Mol. Biol. *166*, 557–580.

Hoeijmakers, J. H. J., Frasch, A. C. C., Bernards, A., Borst, P., and Cross, G. A. M. (1980). Novel expression-linked copies of the genes for variant surface antigens in trypanosomes. Nature *284*, 78–80.

Holmes, D. S., and Quigley, M. (1981). A rapid boiling method for the preparation of bacterial plasmids. Anal. Biochem. *114*, 193–197.

Hunkapiller, M., Kent, S., Canthers, M., Dreyer, W., Firca, J., Giffin, C., Horvath, S., Hunkapiller, T., Tempst, P., and Hood, L. (1984). A microchemical facility for the analysis and synthesis of genes and proteins. Nature *310*, 105–111.

Meleney, H. E. (1928). Relapse phenomena of *Spironema recurrentis*. J. Exp. Med. *48*, 65–82.

Meyer, T. F., Mlawer, N., and So, M. (1982). Pilus expression in Neisseria gonorrhoeae involves chromosomal rearrangement. Cell *30*, 45–52.

Murray, N. E., Brammar, W. J., and Murray, K. (1977). Lambdoid phages that simplify the recovery of *in vitro* recombinants. Mol. Gen. Genet. *150*, 53–61.

Obermeier, O. (1873). Vorkommen feinster eine Eigenbewegung zeigender Faden im Blute von Rekurrenskranken. Zentralbl. Med. Wissenschaft. *11*, 145–155.

Pays, E., Van MeirVenne, N., LeRay, D., and Steinert, M. (1981). Gene duplication and transposition linked to antigenic variation in *Trypanosoma burcei*. Proc. Natl. Acad. Sci. USA *78*, 2673–2676.

Russell, H. (1936). Observations on immunity in relapsing fever and trypanosomiasis. Trans. Roy. Soc. Trop. Med. Hyg. *30*, 179–190.

Schmid, G. P., Steigerwalt, A. G., Johnson, S. E., Barbour, A. G., Steere, A. C., Robinson, I. M., and Brenner, D. J. (1984). DNA characterization of the spirochete that causes Lyme Disease. J. Clin. Microbiol. *20*, 155–158.

Seed, B. (1984). Attachment of nucleic acids to nitrocellulose and diazonium-substituted supports. In Genetic Engineering, Principles and Methods, Vol. 4, J. K. Settow and A. Hollaender, eds. (New York: Plenum Press), pp. 91–102.

Southern, E. M. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol. *98*, 503–517.

Stern, A., Nickel, P., Meyer, T. F., and So, M. (1984). Opacity determinants of Neisseria gonorrhoeae: gene expression and chromosomal linkage to the gonococcal pilus gene. Cell *37*, 447–456.

Stoenner, H. G., Dodd, T., and Larsen, C. (1982). Antigenic variation in *B. hermsii*. J. Exp. Med. *156*, 1297–1311.

Vickerman, K. (1978). Antigenic variation in trypanosomes. Nature *273*, 613–617.

Wallace, R. B., Johnson, M. J., Hirose, T., Miyake, T., Kawashima, E. H., and Itakura, K. (1981). The use of oligonucleotides as hybridization probes. II. Hybridization of oligonucleotides of mixed sequence to rabbit β-globin DNA. Nucl. Acids Res. *9*, 879–894.

Williams, R. O., Young, J. R., and Majiwa, P. A. O. (1979). Genomic rearrangements correlated with antigenic variation in *Trypanosoma brucei*. Nature *282*, 847–849.

CHAPTER THREE

Borrelia Project Postscript

Future Directions

BORRELIA PROJECT - POSTSCRIPT

Shortly following the publication of this paper and another which built upon its results, the collaboration between Dr. Barbour's laboratory and Dr. Simon's dissolved but I have continued to follow the progress in the field. In particular, there are two results which I find have advanced the understanding of the underlying recombinational processes

Genomic Structure and Vmp gene organization

Reproducible anecdotal evidence from the preparation of total *Borrelia* DNA on CsCl gradients indicated the presence of satellite DNA bands. Subsequent work using pulsed-field gels, denaturing gels and Southern hybridization blots[7,8] revealed that the VMP genes were resident on linear, extra-chromosomal elements ("linear plasmids"). These plasmids, the first of their type reported in eubacteria, were later found to have covalently-closed hairpins at their ends.[9,10] Subsequent mapping of this bacterial genome showed that the entire genetic complement of this organism is linear rather than circular (in contrast to other eubacteria, especially the *Leptospira* and *Treponema*). Most recently, *B. hermsii* and *B. burgdorferi* have been found to contain 20 genome equivalents per cell, arrayed along the length of the cell,[11] a factor which must be accounted for when modelling the switching system.

Structure of the expression site

The initial report describing the difference between the silent copy and an expressed copy of the same gene correctly noted that expression of a particular *vmp* gene is accompanied by a duplication event which places the gene within a particular expression locus.[12] However, they incorrectly identified the location of this site, believing it to be several kilobases distant from the plasmid ends. A more recent study indicates that the location of the vast majority of expression-linked vmp genes were situated less than 0.4 kb from the "telomere."[13] Obviously, having such a highly expressed gene (a Vmp can comprise 20% of the total cellular protein)

without another gene 3' would eliminate problems of transcriptional read-through. Given the gene-conversion nature of the process, this position may also play an important role in that process.

While homology between the sequences of various VMPs is limited (a necessary requirement for antigenic variation), comparison of the regions surrounding active and silent copies of the same gene revealed the presence of both 5' and 3' blocks of homology.[13,14] These blocks, of sufficient length to effectively promote heteroduplexing, would allow the activation of a silent *vmp* gene by correct alignment of the copy next to an active promoter in the expression site. A gene conversion-like event could also explain the loss of expression of the formerly active *vmp* gene, because the fusion of the formerly active gene to a silent locus effectively produces a non-reciprocal event. (Note, however, that this process could occur within any of the 20 genome equivalents; this leaves some interesting questions regarding the propagation of switching information). This "bracketing" of the silent copy with blocks of homology both 5' and 3' (to direct the duplicative transposition) is also found in the trypanosomes.[15]

Having broadly mapped the expression plasmid, Dr. Barbour's group has started to examine this locus more closely One structure immediately identifiable in the 5' sequences on the expression plasmid (and not on the silent copy) was a tract of 16 Ts linked to the promoter.[16] (Only 2 other prokaryotic examples of a T or A tract this long were found in GenBank) Proposals for putative function include transcription enhancement and/or stimulation of recombination due to disruption of B DNA helicity by the homopyrimidine tract. A further examination of the upstream loci also discovered the presence of 3 imperfect repeats of a 2 kb sequence containing inverted repeats of 0.2 kb each at their termini. Heteroduplex analysis by EM visualization demonstrated the ability of these sequences to form stem-and-loop structures. Found to exist only on the expression plasmid, current models suggest these structures may play a role in the directionality which characterizes the *vmp* gene switching.

FUTURE DIRECTIONS

There has been an explosive growth in the amount of information concerning the molecular mechanism of antigenic variation in this spirochete since the publication of Meier *et al.* Despite this, recent developments indicate that a number of fascinating questions remain to be solved. For example, we still do not know what event initiates the switching process, nor do we know the structure or fate of the "non-expression plasmid" product molecule(s) (i.e., the balance of input material). Given a helical array of 20 genomic equivalents within a single organism, does the observed phenotypic switching behavior[17] represent a co-ordinated effort? What is the purpose of such an excess of genetic material? The similarities between the antigenic variation mechanisms in this bacterium and the eukaryotic trypanosomes are striking (especially because both now appear to have multiple telomeric expression sites requiring some method of co-ordination). On a more speculative note, it might be interesting to consider whether any information regarding the process of convergent evolution (operating upon a molecular scale) could be gleaned from a comparison between these two parasites and their non-varying relatives. Put another way, assuming the reasonably stringent requirements for a successful antigenic variation strategy, what does the similarity in mechanisms imply about the immune system's selective pressure on these two disparate pathogens and their ability to exploit its weaknesses?

REFERENCES

1. **Bloom, B. R.** (1979). Games parasites play: how parasites evade immune surveillance. *Nature* **279**, 21 - 26.

2. **Trager, W.** (1986) *Living Together: The Biology of Animal Parasitism.* Plenum Press.

3. **Silverman, M. and M. Simon.** (1984) in *Mobile Genetic Elements*, R. Shapiro, ed., Academic Press.

4. **Girons, I. S. and A.G. Barbour.** (1991). Antigenic variation in *Borrelia*. *Res. Microbiol.* **142**, 711 - 717.

5. The ability to grow even non-specific populations of *Borrelia hermsii in vitro* was itself a strange ocurrance. This spirochete resisted all attempts to culture it until Dr. R. T. Kelley, a gentleman-scientist in Knoxville, Tennessee reasoned that N-acetyl-glucosamine might be a necessary co-factor, due to the importance that molecule plays in the lifecycle of its host, the soft-shelled tick *Ornithodorus*. See (1971) *Science* **173**, 443.

6. **Barbour, A. G., S.L. Tessier, and H.G. Stoenner.** (1982). Variable major proteins of *Borrelia hermsii*. *J. Exp. Med.* **156**, 1312 - 1324.

7. **Meier, J. T., M.I. Simon, and A.G. Barbour.** (1985). Antigenic variation is associated with DNA rearrangements in a relapsing fever *Borrelia*. *Cell* **41**, 403 - 409.

8. **Plasterk, R.H.A., M.I. Simon, and A.G. Barbour.** (1985). Transposition of structural genes to an expression sequence on a linear plasmid causes antigenic variation in the bacterium *Borrrelia hermsii*. *Nature* **318**, 257 - 263.

9. **Barbour, A. G. and C.F. Garon.** (1987). Linear plasmids of the bacterium *Borrelia burgdorferi* have covalently closed ends. *Science* **237**, 409 - 411.

10. **Barbour, A. G. and C.F. Garon.** (1988). The genes encoding major surface proteins of Borrelia burgdorferi are located on a plasmid. *Ann. N. Y. Acad. Sci.* **539**, 144 - 153.

11. **Kitten, T. and A.G. Barbour.** (1992). The relapsing fever agent *Borrelia hermsii* has multiple copies of its chromosome and linear plasmids. *Genetics* **132**, 123 - 128.

12. **Plasterk, R. H. A., M.I. Simon, and A.G. Barbour.** (1985). Transposition of structural genes to an expression sequence on a linear plasmid causes antigenic variation in the bacterium *Borrrelia hermsii. Nature* **318**, 257 - 263.

13. **Kitten, T. and A.G. Barbour.** (1990). Juxtaposition of expressed variable antigen genes with a conserved telomere in the bacterium *Borrelia hermsii. Proc. Natl. Acad. Sci.* **87**, 6077 - 6081.

14. **Barbour, A. G., N. Burman, C.J. Carter, T. Kitten, and S. Bergstrom.** (1991). Variable antigen genes of the relapsing fever agent Borrelia hermsii are activated by promoter addition. *Mol. Microbiol.* 5(2), 489 - 493.

15. **Timmers, H. Th. M., T. de Lange, J.M. Kooter, and P. Borst.** (1987). Coincident multiple activations of the same surface antigen gene in *Trypanosoma brucei. J. Mol. Bio.* **194**, 81 - 90.

16. **Barbour, A.G., C.J. Carter, N. Burman, C. Freitag, C. Garon, and S. Bergstrom.** (1991). *Inf. Imm.* **59**(1), 390 - 397.

17. **Stoenner, H. G., T. Dodd, and C. Larsen.** (1982). Antigenic variation of *Borrelia hermsii. J. Exp. Med.* **156**, 1297 - 1311.

CHAPTER FOUR

V(D)J Recombination

Retroviral Activation of V(D)J Recombination in Fibroblasts

At the time this project commenced, the objective of producing V(D)J recombination in a non-lymphoid cell line had been actively pursued for quite some time. The only measurable success was the (then) recently reported transfection of genomic DNA into fibroblastoid lines.[1] Although that approach led to the isolation of a fragment which was intregrally involved in the acquisition of a functional V(D)J recombination system, the nature of the underlying process remains elusive. In pursuit of a more general method for cloning genes expressed in a stage- or tissue-specific manner and, more specifically, in order to isolate other genes that may be involved in V(D)J recombination, we attempted to activate site-specific recombination activity within a fibroblast cell line via infection with a replication-competent retovirus. The basis for the experiment was the concept that enhancer and promoter elements within the viral long terminal repeats (LTRs) might over-ride endogenous stage- or tissue-specific regulatory elements. Assuming the ability to screen or select for the resulting phenotype, the process of gene isolation would be simplified due to the "viral tag." The retrovirus used here had been designed by Dr. Steven Goff (Columbia University) as a gene-disrupting element, and was molecularly marked to facilitate recognition strategies.[2] [It should be noted, however, that in contrast to those workers, our intention was to promote constitutive expression of the desired gene(s).]

The selection process we devised was based on the acquisition of drug resistance in the event of a V(D)J mediated inversion. A replication-defective retroviral substrate was stably integrated into the genome of an NIH 3T3 fibroblast line, configured so that a site-specific joining event is required in order to activate a drug resistance marker by inversion of the structural gene relative to the orientation of a promoter element.

In order to define the size of the cell population necessary to observe one recombination-positive mutant, we had to make estimates on: (a) the "activation space" of the LTR (i.e., the maximum distance (from the end of the LTR) at which a host gene could be

activated);[2] (b) the number of integrated viral genomes expected per cell; and (c) the frequency of recombination within a cell (and thus the probability of acquiring drug resistance) once the recombination machinery was functional. The results of these calculations led us to expand the NIH 3T3 line to approximately 1×10^9 cells, which were infected at an MOI of 5. After placing in drug selection, 288 resistant colonies were isolated. A set of PCR reactions was performed on each to determine which, if any, had become drug resistant due to the site-specific inversion of the integrated substrate sequences. One clone, designated "L-6" appeared to have inverted the proper region, but without using both of the intended V and J target segments. Further analysis of this line remains to be done.

Molecular Phrenology: P Nucleotide Addition on V(D)J substrates

Current understanding of the ability of the immune system to generate proteins with diverse primary sequences recognizes five major mechanisms:[3,4]

1. Germline Diversity. The genome consists of a large number of independently functional V, D, and J (as well as α, β, γ, and δ) gene segments that can be used to make up a functional protein.
2. Combinatorial Joining. Each functional member of one family can prductively rearrange with any member of the appropriate partner family.[5,6] For example, any of the 250 V_κ gene segments can combine with any of the 4 J_κ gene segments, giving 1000 possible V_κ genes. For heavy chain genes, with the added complexity of D segments, the totals rise even faster: assuming 1000 V_H segments, 10 D_H segments and 4 J_H segments, there are $1000 \times 10 \times 4$ or 40,000 V_H genes possible.
3. Junctional Diversity. Note that the number derived above is solely based upon the number of combinations possible between *complete* gene segments, and does not reflect the large number of different sequences which can arise from even one of the aforementioned combinations when various DNA modifying activities act upon the input termini of the coding sequences involved. These effects *fall into two catagories:*

removal of bases, and the addition of information not coded in the genome (*non-templated base addition*). This latter diversification mechanism can also be divided into two different processes, the first in which base addition appears to be totally random and is correllated with TdT activity (*N region diversity*),[9] and bases added containing sequence data palindromic to the input terminus (*P nucleotide addition*).[8,15] The work in the subsequent chapter is a study of this last phenomenon.

Calculation of the exact numerical effect of this type of diversity has been impossible, but using a conservative estimate of three different amino acids for each junction will bring estimates of combinatorial joining to 3000 genes for V_K genes and 360,000 for V_H genes (containing V-D and D-J junctions). However, two factors serve to decrease the actual numbers of functional proteins. Most importantly, because the immunofunctional (immunoglobulin, T cell ab- and T cell γδ- receptors) proteins are only translated in one reading frame, the flexibility of the joining process necessarily means that the major subset of all possible gene combinations will contain those genes whose gene segments are joined out-of-frame (and thus will be discarded). If joining were totally random in this process, only one third of all rearrangements would be productive, yet estimates of non-productive rearrangements are only 15% for k-chain chromosomes and 45% for heavy chain chromosomes.[12] On a more subtle level, until now the spectrum of possible resultant sequences (in the absence of selection of any kind) has been presumed to arise from activities (nucleolytic and/or base addition) that are insensitive to the sequence acted upon. Our work indicates that the DNA sequence at the terminus of the input gene segments plays a profound role in the fate of that end, and thus may be the ultimate determinant of what sequences are possible at a particular hypervariable region.

4. Combinatorial Association. Unique light and heavy chains can associate with each other; using the numbers derived above, there are 360,000 x 3000 or 1.08×10^9 possible complete antibodies.

5. Somatic Hypermutation. This method of diversification uses fully assembled variable region genes as the substrate. Point mutations are introduced throughout both the coding and the untranslated regions late in B cell development, generating subpopulations of cell with receptors having higher affinities for that antigen.[10] These cells are believed to be selectively expanded under conditions of limiting antigen. The somatic hypermutation event may be induced during the class switch rearrangement.[11] The assembly of functional antibody and TCR genes from their component segments via site-specific recombination has become a paradigm in terms of DNA rearrangements. Recent reviews summerized what was known of the mechanism.[13,14] Two publications at the end of 1989 stimulated renewed interest in the area of non-templated base addition during lymphoid recombination.[8,15] In particular, the work by Lafaille *et al.* proposed the exisence of a previously overlooked process, fundamental to V(D)J recombination, based upon largely upon anecdotal evidence. Given the explicit nature of their hypotheses for both the origin and characteristics of these inserts, Dr. Suzanna M. Lewis and I devised a set of experiments to test those theories. The result of that work is discussed in a paper originally submitted to *Genes and Development* (a revised version was later accepted by *Molecular and Cellular Biology*).

REFERENCES

1. **Schatz, D. and D. Baltimore.** (1988). Stable expression of immunoglobulin gene V(D)J recombinase activity by gene transfer in 3T3 fibroblasts. *Cell*, **53(1)**, 107 - 15.

2. **Lobel, L., M. Patel, W. King, M. Nguyen-Hu, and S. Goff.** (1985). Construction and recoveryof viable retroviralgenomes carrying a bacterial suppressor transfer RNA gene. *Science* **228**, 329 - 331.

3. see for example **Hood, L., I. Weissman, W. Wood, and J. Wilson.** (1984). *Immunology*, 2nd edition, Benjamin Cummings, Menlo Park CA.

4. **Tonegawa, S.** (1983). Somatic generation of antibody diversity. *Nature* **302**, 575 - 581.

5. **Schilling, J., B. Clevinger, J.M. Davie, and L. Hood.** (1980). Amino acid sequence of homogenous antibodies to dextran and DNA rearrangements in heavy chain V-region gene segments. *Nature* **283**, 35 - 40.

6. **Weigert, M., R. Perry, D. Kelley, T. Hunkapiller, J. Schilling, and L. Hood.** (1980). *Nature* **283**, 497 - 499.

7. **Landau, N., D. Schatz, M. Rosa, and D. Baltimore.** (1987). Increased frequency of N-region insertion in a murine pre-B-cell line infected with a terminal deoxynucleotidyl transferase retroviral expression vector. *Mol. Cell Biol.* **7**, 3237 - 3243.

8. **McCormack, W.T., L.W. Troelker, L.M. Carlson, B. Petryniak, C.F. Barth, E.H. Humphries and C.B. Thompson.** (1989). Chicken IgL gene rearrangement involves deletion of a circular episome and addition of single nonrandom nucleotides to both coding segments. *Cell*, **56(5)**, 785 - 91.

9. **Hood** L., *et al., op. cit.* and references therein.

10. **Crews, S., J. Griffin, H. Huang, K. Calame, and L. Hood.** (1981). A single V_H gene segment encodes the immune response to phosphorylcholine: somatic mutation is correlated with the class of antibody. *Cell* **25**, 59 - 66.

11. **Clarke, C., J. Berenson, J. Goverman, P. Boyer, S. Crews, G. Siu, and K. Calame.** (1982). An immunoglobulin promoter region is unaltered by NA rearrangement and somatic mutation during B-cell development. *Nuc. Acids Res.* **10**, 7731 - 7749.

12. reviewed in **Rajewsky, K., I. Forster, and A. Cumano.** (1987). Evolutionary and somatic selection of the antibody repetoire in the mouse. *Science* **238**, 1088 - 1093.

13. **Lewis, S. M., and M. Gellert.** (1989). The mechanism of antigen receptor gene assembly. *Cell* **59**, 585 - 588.

14. **Lieber, M.** (1991). Site - specific recombination in the immune system. *FASEB J. (United States)*, **5(14)**, 2934 - 44.

15. **Lafaille, J. J., A. DeCloux, M. Bonneville, Y. Takagaki, and S. Tonegawa.** (1989). Junctional sequences of T cell receptor γδ genes: implications for γδ T cell lineages and for a novel intermediate of V(D)J joining. *Cell* **59**, 859 - 870.

CHAPTER FIVE

P Nucleotides in V(D)J Recombination: A Fine Structure Analysis. (1992). *Molecular and Cellular Biology,* in press.

P Nucleotides in V(D)J Recombination: A Fine-Structure Analysis

Joseph T. Meier and Susanna M. Lewis*

Division of Biology, 156-29

California Institute of Technology

Pasadena CA, 91125

USA

*corresponding author

(818) 356-3904

Fax: (818) 564-8709

Abstract

Antigen receptor genes acquire junctional inserts upon assembly from their component, germline-encoded V, D and J segments. Inserts are generally of random sequence, but a small number of V-D, D-J or V-J junctions are exceptional. In such junctions, one or two added bases inversely repeat the sequence of the abutting germ-line DNA. (For example, a gene segment ending "...AG" might acquire an insert beginning with the residues "CT.." upon joining.) It has been proposed that the non-random residues, termed "P nucleotides", are a consequence of an obligatory end-modification step in V(D)J recombination. P insertion in normal, unselected, V(D)J joining products, however, has not been rigorously established. Here we use an experimentally manipulable system, isolated from immune selection of any kind, to examine the fine-structure of V(D)J junctions formed in wild-type lymphoid cells. Our results, according to statistical tests, show the following: (i) the frequency of P insertion is influenced by the DNA sequence of the joined ends; (ii) P inserts may be longer than two residues in length; and (iii) P inserts are associated with coding ends only. Additionally, a systematic survey of published P nucleotide data shows no evidence for variation in P insertion as a function of genetic locus and/or ontogeny. Together these analyses establish the generality of the P nucleotide pattern within inserts, but do not fully support previous conjecture as to their origin and centrality in the joining reaction.

Introduction

When Ig and TCR genes are assembled from their germ-line components, variable truncation of the coding segments and, often, the introduction of a small number of non-germ-line residues occurs at the recombinant junctions. These sequence alterations fall within the binding domain of the encoded antigen receptor, and constitute an important source of diversity in the immune system. Until fairly recently it was thought that junctional insertion was essentially random, being both variable in length and unpredictable in sequence. Several groups have documented a correlation between the presence of random "N regions" and terminal deoxynucleotidyl transferase activity (18 and cited therein). Nevertheless some junctions exhibited a recurrent pattern of base addition which, as such, were inconsistent with the N region paradigm. A recent proposal integrated these exceptional cases into a consistent formulation (21). The term "P nucleotide" ("P" for "palindrome") was coined to describe insertions occurring exclusively within a subset of junctions that contain at least one non-truncated coding end. ("Coding end" and related terms are defined in Materials and Methods) In such junctions, one or two residues contiguous to the full-length coding segment might conform to the inverse-complementary sequence of that segment (an example is shown in Figure 1A, bottom). P nucleotide residues often occur together with N regions within a single junction.

The distinctive character of P inserts prompted a fresh round of speculation about how V(D)J recombination might work. Several explanations for their origin have been offered. An implicit assumption of these models is that P nucleotide addition is the manifestation of an intrinsic, essential aspect of the V(D)J joining reaction; either a direct consequence of a site-specific cleavage step (27, 31) or a necessary step for subsequent end joining (21, 40). Nevertheless, despite an impressive accumulation of junctional data from rearranging loci in both human and mouse over the last several years, the documentation regarding the existence of P nucleotides is fragmentary and essentially anecdotal. Further, because endogenously-generated V(D)J junctions are subjected to several levels of selection within the immune system, the actual

frequency of P inserts among junctions when they first arise is difficult to assess. According to the sequences of endogenous junctions, the possibility exists that P nucleotide addition only occurs at certain coding end sequences, at particular loci, or even might be attributable, in some cases to immune selection of randomly-generated N region inserts. In the present work, we had two aims: to verify the existence of P nucleotides as a general feature of V(D)J junctions, and through their characterization, to make deductions about the joining mechanism itself.

Materials and Methods

Terminology:

V(D)J joining is used in the generic sense to indicate recombination between two elements as mediated by site-specific recognition of 12-and 23-joining signals. *Coding end* refers to the sequences that abut the 12- and 23-joining signals prior to recombination, whether these sequences constitute V, D, or J segments or instead represent substituted, non-coding sequences (as is often the case with introduced constructs). *Coding joint* refers to the site specific junction between two coding ends. Signal *joint* refers to the site-specific junction between a 12- and a 23-signal.

Cells and transfections

The cell line 204-1-8 (referred to here as 1-8) is an Abelson murine leukemia virus transformed line derived from adult Balb/c bone marrow. Transfections and tissue-culture conditions were as previously described (26, 29).

Plasmids

pSal-Bam is pJH288 (30). The other three substrates in the series were derived by the substitution of synthetic cassettes containing either the 12- or 23-signals for those present in pJH288. In all cases, cassettes were designed to introduce changes only in the sequences adjacent to the signals; not in the signals themselves. For pSpe-Bam, the sequence TCGACTAGTCACAGTGCTACAGACTGGAACAAAAACCG (Sal compatible overhangs, Spe1 site underlined) replaced the 12-signal of pJH288 at its Sal site. Likewise, GATC-CTCGAGCACAGTGGTAGTACTCCACTGTCTGGCTGTACAAAAACCCTCGG (BamH1-compatible overhangs; Xho site underlined) was inserted into the Bam H1 site of pJH288, to create pSal-Xho. pSpe-Xho was constructed by replacing the 23-signal of pSpe-Bam with the "Xho" 23-signal cassette as above.

pDSJ was constructed by replacing the small Sal fragment of pJH200 (16), with that derived from a pJH200 recombinant containing a "precise" signal joint (kindly provided by J. Hesse, N.I.H). The inserted fragment was oriented with the included chloramphenicol gene in reverse transcriptional orientation relative to the pJH200 lac promoter. A "precise" signal joint in an Alu1-Dde 1 fragment originating from recombinant "N" (23) was then introduced at a Sal site of the intermediate. (Here, as in all following steps, recessed ends were filled in as necessary with Klenow fragment of DNA polymerase 1). The signal joint within the Alu1-Dde 1 fragment was composed of the joining signals of V_k21-C and J_k2, and was positioned as indicated in Figure 2B. The other signal joint was composed of the joining signals of V_kL-8 and J_k1 (Figure 2B). To prevent background Cat gene expression, the Cla 1 fragment from pJH288, containing the "oop" terminator (28), was introduced at the remaining Sal 1 site 5' to the inverted Cat gene. The particular combination of signal joints was chosen with care so as to minimize the possibility that inversional *homologous* recombination might occur. Such events would produce inversional rearrangement with two apparently precise junctions, confounding the analysis. Thus all four signals have different spacers; additionally, the J_k2 heptamer has a non-consensus residue in the fifth position (CACACTG).

pMut-2 was derived from p23 (26) by insertion of a Sal cassette containing a change in the second position of the heptamer (CTCAGTG). The sequence of pMut-2 is otherwise identical to p12X23 (26). pMut-27 is identical to pMut-2 except for the absence of the cryptic site 6130 [26 and references therein].

Identification of various classes of recombination product

pSal-Bam: Untrimmed coding joints were identified by the presence of a diagnostic fragment of approximately 245 base pairs following digestion of DNA samples prepared from Cam^r colonies with Sal1 and BamH1 (the exact size of the fragment varied depending upon the structure of the individual junction): all other classes of recombinants however were linearized by this treatment. To determine which of the two coding ends were full-length in the untrimmed junctions thus

identified, they were each tested with Bam H1 alone (for the presence of an approximately 330 base fragment) or Sal 1 only (for the presence of the approximately 245 base fragment). Hybrid recombinants were quantified by digesting 76 randomly selected pSal-Bam isolates with HgiA1. The identified hybrids were then digested with Sal 1 alone to look for linearization; this would indicate an untrimmed coding end. The HgiA1 digestions also revealed any standard recombinants in the sample that contained base additions or deletions at the signal joint.

pSpe-Bam: pSpe-Bam recombinants were tested with Spe1 alone to test for linearization, and with Bam H1 alone (as above), in order to test for the presence of the approximately 330-base pair fragment. Those that were linearized with Spe 1 were subjected to DNA sequence analysis in order to distinguish between standard recombinants and hybrid recombinants containing full length ends. Those with the Bam H1 fragment were scored as untrimmed Bam coding joints. Signal joints were analyzed as for pSal-Bam (above), by screening a randomly selected sample of 75 pSpe-Bam isolates with the enzyme HgiA1.

pSal-Xho: pSal-Xho recombinants were tested by doubly digesting with Sal1 and Xho1. Those with untrimmed coding joints produced an approximately 245 base pair fragment. Candidate recombinants were tested with Sal 1 alone (to test for the presence of the 245 base pair fragment) or Xho alone (to test for linearization) to determine which coding ends were full-length in each case. Hybrid joints and signal joints were analyzed as for pSal-Bam (above), by screening 63 randomly selected pSal-Xho recombinants.

PSpe-Xho: pSpe-Xho recombinants were tested by doubly-digesting with Spe or Xho. Linearization indicated the presence of a full-length coding end. Candidates were then digested with Spe alone. DNA sequence analysis of those that were linearized distinguished between standard and hybrid recombinants. Signal joints among 77 randomly selected pSpe-Xho isolates were analyzed as described for pSal-Bam (above).

For all of the above, the number of standard inversions, (Table 1) as opposed to hybrid deletions and cryptic site 6130 deletions was determined in each case according to the HgiA1

digestions. Fully 1/3 of the total number of recombinants isolated in these experiments were analyzed in this fashion.

pDSJ products: The presence of "precise" junctions in pDSJ was indicated by digestion with ApaL1 (or the isoschizomer, Sno1). However, because imprecise junctions containing two or more P insert residues are likewise ApaL1 sensitive, unambiguous identification of each of the following recombinant classes required DNA sequence analysis. A digestion pattern of six bands indicated the possibility of two precise junctions. The DNA sequences of all such candidates were determined. A pattern with five bands indicated that one of the two junctions was imprecise (with bases added and/or subtracted). 16 of 23 such isolates were analyzed at the DNA sequence level. A pattern with four bands indicated that neither junction consisted of two fused, full-length signals without inserts. The DNA sequence of all 6 examples of this class was determined.

pMut-2, pMut-27 Recombinants: The isolation of the pMut-2 recombinants was carried out as described for p12X23 (26). All chloramphenicol-resistant colonies obtained after transfection/transformation of pMut-2 were picked onto grid arrays in triplicate. Filter lifts were probed with oligos 1,2, and 23-SIG as described (26) . DNA was then prepared from candidate recombinants (positive for the first two and negative with 23-SIG) and digested with HgiA1. All isolates with the proper restriction pattern were then sequenced.

Statistics

The probability of obtaining the observed number of P nucleotide inserts (n) or a larger number, on the basis of randomness was calculated by summing the terms of the binomial distribution from x=n to N, where N is the total number of insert-containing samples examined:

$$\sum_{x=n}^{N} \binom{N}{x} p^x q^{N-x}.$$

The expected frequency for a particular nucleotide, "p", (and thus for the other three, q=1-p) was first determined by totaling the number of A,C,G, and T residues found as inserts using the

extrachromosomal assay system here and previously (26, 30). All such junctions (coding, hybrid or open-and-shut) were oriented so that to the left and right were the coding ends originally associated with the 12- and 23-signals, respectively. The inserts within signal joints were tallied with the 23-signal on the right. Accordingly (out of a total of 484 inserted residues scored) the frequencies of A, C, G, and T were 0.16, 0.31, 0.38 and 0.15 respectively. The expected P-nucleotide probabilities were then calculated using these values for p, the value "o" (observed) shown in Table 2 for n, and likewise "N" for N. For example, in pSal-Bam, a two base P insert adjacent to the Sal end would have the sequence "GT" (see Table 2A, column 2). On the basis of randomness and independence, the expected frequency for GT is 0.38 x 0.15, or .057. As shown in Table 2, there were 10 junctions that had an insert of two or more residues (so that N=10); among these, 6 began with GT (i.e., n=6). Summing the binomial distribution from n=6 to N=10 gives the *P* value shown in the table.

To specifically evaluate the probability of observing P inserts of greater than two, as observed in the Sal end-containing coding joints, the calculation was as follows. The number of P inserts of a length of three or greater (six) was substituted for "n" and the number of inserts of three bases or more that *already* contained at least two P nucleotides was substituted for "N" (nine). The value 0.38 (the probability of observing a G at any position within an insert , as described above) was substituted for "p" and the expansion was solved to give $P \leq 0.03$.

Analysis of endogenous junctions

Assembled collections (1, 2, 4, 9-12, 14)(1, 2, 4, 8-11, 13, and J. Teale, unpublished) were first checked for data reported in more than one publication. Also, any junctions derived more than once in a PCR reaction or from an individual animal were discounted. Each junction was examined for full-length ends, for the presence of inserts and finally for the presence of P nucleotides. No end appeared in the final calculations unless the germline sequence up to, and including, the joining signal is known. (We note that in some studies, the term "germline" does not refer to sequences derived from germline DNA, but instead has been used incorrectly to refer

to consensus sequences deduced from cDNA.) The only exception was that for the IgH analysis, we accepted two consensus D_h sequences described by A. Feeney (10) as being highly likely to represent germline elements.

Each VDJ junction was scored by its parts: 3'V, 5'D, 3'D and 5'J. No putative D segments shorter than three residues were scored as such. No D-D junction involving a D segment shorter than four residues was scored as such. Junctions were compared to germline elements and the maximum number of bases was assigned to each end. (This means that in some of the cases where residues might have been contributed by either of two segments, - the so-called ambiguous assignments -, bases infrequently were assigned to both in order to count the maximum number of full-length ends.) Where three or more D segment residues were present but could represent either an internal or terminal D segment sequence, they were scored as a terminal fragment. We accepted single base mismatches no closer than two bases from a putative segment terminus. If the mismatch occurred at the penultimate base, both the mismatched base and the one following were scored as junctional inserts.

DNA sequence analysis

DNA sequence analysis was carried out using a "Sequenase", v. 2 kit (U.S. Biochemical). The "Lac-1" oligo (25) was used as a primer for sequencing the coding joints in pSal-Bam, in its derivatives and in pMut-2 and pMut-27. Hybrid joints were sequenced using Lac-1 or "JH33" (26). The oligos JH33, or Ter-1(26) were used to sequence signal joints. For pDSJ, "right" junctions were analyzed with the Ter-1 primer; the "left" junctions were sequenced with the Lac-1 oligo.

Results

P nucleotides are added to coding ends

To investigate P insertion at coding ends, we used the extrachromosomal plasmid assay developed by Hesse *et al.* (16, 29). Briefly, the assay entails transfection of recombination substrates into a pre B-like cell line, 1-8, that is active for V(D)J joining. During residence in the 1-8 cells, some of the transfected molecules become rearranged in a site-specific manner (Figure 2A). This is detected by re-isolating the plasmid DNA from transfectants about 48 hours later, and using it to transform *E. coli* cells. Recombined molecules confer chloramphenicol resistance to *E. coli* due to activation of chloramphenicol acetyl transferase (cat) expression (Figure 2A).

This approach has several important features. By making use of plasmid substrates that contain minimal recombination recognition sites, and no locus-specific sequences(16), we can look at junction products in the absence of locus-specific influences. Furthermore, all recombinants are generated in a single clonal cell line, providing a standardized cellular context with which to compare results between experiments. Finally, because recombined molecules are not isolated until after they are introduced into *E. coli*, and because the recombinant junctions (coding and signal joints) are extraneous to the coding sequences of the selectable marker (see Figure 2A), we can expect to analyze the full diversity of junctions formed by the V(D)J joining machinery without the selective bias that occurs within an intact immune system.

By this approach we tested whether the alteration of the sequence of a coding end would alter the identity of the associated junctional inserts in a pattern predicted by the P nucleotide theory. To a first approximation, if P addition is an integral part of the joining process, then P inserts would be expected to arise at statistically significant frequencies regardless of the sequence of a given coding end.

We generated a series of four closely-related recombination substrates that differed only at their "coding ends" (this and related terms are defined in Materials and Methods). In the parental construct, pJH288 (30), and in our three variants (pSpe-Bam, pSal-Xho and pSpe-Xho),

restriction enzyme recognition sequences were located next to the 12- and 23-joining signals, in a position equivalent to native V, D or J coding elements. This design facilitated the later isolation of non-truncated junctions. As their names indicate, pSal-Bam and the three derivatives had either Sal 1 or Spe 1 recognition sites abutting their 12-signals and either Bam H1 or Xho 1 sites adjoining their 23-signals.

After transfection, DNA samples were prepared from over 800 chloramphenicol resistant recombinants and screened for the presence of full-length coding ends by digestion with the appropriate restriction enzymes. Representative numbers of "untrimmed" junctions (those containing at least one non-truncated coding end) were then subjected to DNA sequence analysis. A general summary of the experiment and our results are given in Table 1; details of the method of analysis are provided in Materials and Methods.

Upon recombination, each of the four coding ends was found to have acquired P nucleotide additions in at least some instances. DNA sequences are shown in Figure 3. For each coding joint, residues that can be attributed to either of the input ends were so assigned (to the left or right) and those residues that do not correspond to either end, as shown in the middle, were scored as inserts (for details see legend to Figure 3). P nucleotides were noted as indicated (bold type-face, and underlined).

We wished to rule out the possibility that some or all putative P inserts might in fact simply be N regions with a fortuitous inverse-complementary match to the coding end. To do so, we applied a statistical test. The probabilities of obtaining the observed frequencies of P nucleotide inserts through random N insertion were calculated according to the binomial distribution (Table 2 A-E: details provided in Materials and Methods). Each end was evaluated, and in each case, P inserts of various lengths were considered.

As summarized in Table 2G, all four coding ends gave rise to junctions containing P inserts. In each instance the match between observed inserts and P nucleotide sequences, as specified by the coding end, was highly significant: $P<0.01$. Thus, the observed P inserts were clearly distinct from N addition.

Several other properties of P insertion were established by the statistical analysis. The first was that P nucleotides were associated with some coding ends more often than with others (Figure 3, Table 2). For example, overall, the *P* value for P inserts at Sal ends was 10^{-7}, while that for the Bam end was 0.005 (Table 2G). Differences were evident when various endwise combinations were tested, as well as for P inserts of various lengths (Table 2A,B). The second observation was that P nucleotide inserts were short or long depending upon the particular coding end involved. For example, at the Sal end, 6 of 27 P inserts were three or more residues in length, whereas this was true of none of the P inserts (31 total) characterized at the Xho end (Table 2G; compare the numbers for P inserts $\geq$3 to P inserts $\geq$1 in each case). A third, related observation is that for the Sal end in particular, a *significant* number of the inserts were longer than two basepairs. We considered the possibility that three-base P inserts were actually two-base P inserts that happened by chance to appear longer due to the addition of N region sequence. The probability of obtaining a third P residue at the observed frequency through random addition ($P \leq 0.03$, calculated as described in Materials and Methods) indicated that P inserts longer than two residues were not created by random N addition.

P nucleotides are not detected at signal ends

Lafaille *et al.* (21), noted that P inserts appeared adjacent to coding ends but not signal ends, and concluded that the P addition mechanism operated only upon the former. In the present study, a large number of junctions that incorporate signal ends were assayed for P inserts. These were of two kinds: signal joints and hybrid joints. Signal joints are reciprocally related to coding joints (24), and can be isolated (as in the present study) with substrates in which recombination sites are oriented so as to promote inversional rearrangement (e.g., Figure 2). In a signal joint, a 12-spacer signal and a 23-spacer signal are fused; but the ends are almost always connected without truncation and/or base addition. Signal joint formation thus creates an HgiA1 (Sno1 or ApaL1) recognition site (30). This feature facilitates identification of the rare signal joint with either an insert and/or loss of residues. The other signal-containing junction mentioned

above, a hybrid joint, is an alternative V(D)J joining product in which coding-to-signal end fusions occur (25)(22, and cited therein). With the present constructs, one of two possible hybrid joint conformations is recoverable; that in which a coding end (represented by the Sal or Spe end, as the case may be) has become connected to a 23-signal (Figure 2A). The restriction pattern of these recombinants, obtained by HgiA1 digestion is also distinctive.

A collection of "imprecise" signal joints as well as a large number of hybrid junctions were identified by digesting approximately 300 of the DNA samples with HgiA1. All independent isolates that contained at least one untrimmed end according to DNA sequence analysis, are shown in Figure 4 (A,B). The results of our statistical analysis were unambiguous. P inserts were absent from signal ends, whether such ends occurred within a signal joint or hybrid joint. The *P* values associated with the few apparent P nucleotides that were observed in each case were 0.8 and 0.5, respectively (Table 2E, F). P inserts were demonstrated, however, at the coding ends of hybrid joints (Figure 4A, Table 2A, C; $P \leq 0.02$ for all cases). These data prove the coding end specificity of P inserts noted by Lafaille *et al.* (21).

Is evidence of P addition hidden at signal joints?

On the basis of the above analysis, one might conclude that the P insertion mechanism only modifies coding ends. However when considered, proposed models for P addition (21, 27), lead to the prediction that P nucleotide addition to signal ends, were it to occur, might likely be hidden at signal joints.

Two models have been put forward to account for the presence of P nucleotides. In one model (21), a di-nucleotide is removed from one strand of the cut coding end and transferred to the other strand (Figure 1A, left). In a second, (27) the two strands of the coding terminus are first joined to one another to form a sealed hairpin, following which the hairpin is opened by nicking one strand at a position a few bases in from the end (shown to the right). The proposed order of single strand nicks and ligations differ between these two models, but both posit an intermediate structure with staggered ends. In this intermediate, one strand has been extended at the expense

of the other (Figure 1A, middle). Thereafter, according to either model, the staggered termini meet with varying fates: when ligated directly, P nucleotides appear as junctional inserts appended to full-length ends (Figure 1A, bottom), but more often other operations intervene which remove the added bases before end-joining can occur .

12- and 23-signal ends, which are palindromic, bear an inverse-complementary relationship to one another as oriented for joining. By either of the above proposals, P nucleotide addition at signal ends would generate a pair of termini with complementary single-stranded overhangs (Figure 1B). Annealing of complementary overhangs (arguably a favored event in V(D)J joining;(14)) followed by sealing of two single-strand nicks, would reproducibly create a junction without insertion or truncation at the crossover site. The "precise" stereotyped structure typical of signal joints therefore might in fact arise *as a consequence* of P nucleotide addition (Figure 1A, B), but P *inserts* would be consistently absent from these junctions.

We tested for hidden P addition by manipulating the degree of complementarity between joined ends. In one experiment, we decreased the complementarity between signal ends; in a second, we increased it at coding ends. Reduction of the base-pairing potential between proposed P-overhang intermediates (Figure 1C) should radically alter the characteristic precision of signal joints, if they are usually formed by the pathway shown in Figure 1B. Likewise, by creating the opportunity for annealing between the overhang intermediates of coding ends (as in Figure 1B), we might detect a corresponding increase in precision in the coding joint connections formed between such ends.

Substrates (pMut-2, pMut-27-Figure 5A) in which base-pairing of putative intermediate structures is disrupted were constructed in the course of a separate study (S. M. Lewis, in preparation). The pertinent difference between these plasmids and pSal-Bam was the alteration of the second base of the 12-signal heptamer from CACTGTG to CACTGAG (Figure 1C). The one-base change eliminates two out of four possible base pairing interactions upon attempted annealing of the postulated overhang intermediates (Figure 1C, middle).

The DNA sequences of all verifiably-independent isolates recovered upon transfection of pMut-2 and pMut-27 are shown in Figure 5A. To test our hypothesis, it was necessary to introduce the base change within two bases of the heptamer border, placing the alteration within a region of the heptamer that is critical for joining signal function (17). As a result, only one or two (sometimes no) recombinants were recovered per transfection. Despite the severe depression in recombination frequency, and the loss of potential base-pairing between postulated P nucleotide-modified intermediates, there were no examples in which the signal ends were truncated. The mutant joining signal eliminated efficient recombination without causing the signal joints to acquire "coding joint character."

In a reciprocal experiment a plasmid, pDSJ, was constructed in which the two coding ends were replaced by signal ends. In this four-signal plasmid (Figure 2B), *both* junctions arising from V(D)J joining (not just the signal joint) are created from ends that possess inverse complementarity. We wished to determine whether the "coding joints" resulting from pDSJ rearrangement would assume a signal joint-like structure, lacking N regions, overt signs of P addition in the form of P inserts, and evidence of end-truncation. pDSJ was transfected into 1-8 cells, and analyzed by a diagnostic ApaL1 digestion as detailed in Materials and Methods. The DNA sequences of 26 of a total of 32 isolates were determined (Figure 5B).

Because of the symmetry of the pDSJ substrate, we could not anticipate whether the "right junction" or the "left junction" would become the coding joint, or whether, in fact two types of junction would be distinguishable. What we found was that in almost all isolates one junction had been processed as a coding joint while the other had the characteristics of a signal joint. To be specific, the fine structure of all of the "left junctions" shown above the dashed line in Figure 5B was consistent with that of signal joints: all contained two untrimmed signal ends (with occasional base inserts). The corresponding "right junctions" all exhibited base loss and/or addition, as is typical of coding joints. Below the line in Figure 5B are isolates in which these identities were reversed, with the left junctions corresponding to coding joints and the right

junctions appearing to be the signal joints. In one recombinant (D57, below the solid line, bottom), neither junction contained two untrimmed ends.

Thus, in this experiment the joining reaction displayed a fundamental asymmetry despite the symmetry of the input substrate. There were no examples of a pDSJ recombinant in which both junctions were precisely joined (i.e., with neither base loss nor addition). The two isolates, D26 and D70, in which truncation had not occurred had, in each case, acquired a P insert within one of their two product junctions. Thus, the D26 and D70 junctions could not have formed via the pathway in Figure 1B, because had this been the case, P residues, as inserts, should not have been observed.

Taken together, these experiments are consistent with a P addition process that targets coding, not signal, ends during V(D)J joining.

P nucleotides at endogenous loci

To complement our studies with artificial substrates, we wished to determine whether any consistent variation in the frequency of P inserts, e.g., between loci or during development, would come to light upon examination of endogenously generated V(D)J joining products. In the case of N region insertion, an absence early in ontogeny gives way to a steady increase during development, suggestive of developmental regulation of Tdt activity (1, 2, 10, 11, 14); McVay, 1991 #81 (1, 2, 10, 11, 14, 32). Possibly, the incidence of P inserts might also fluctuate in some informative pattern. Further, it was of interest to establish the frequency with which P inserts longer than two residues are observed in the physiological context.

A preliminary analysis focused on murine light chain genes because rearranged kappa and lambda genes usually lack junctional inserts of any kind, N or P. In the collection of Kabat *et al.* (18), we counted 77 murine light chain junctions (disregarding redundant listings of somatic variants) that are derived from known germline V and J segments, and which contain an untrimmed end. Not one contained a P insert. However, this absence of P insertion was not consistent with the fine structures of kappa gene recombinants present upon the circular DNA

molecules excised by repeated rounds of V_kJ_k joining. While the Kabat collection is largely limited to expressed, functional, light chain genes, excision products contain coding joints that have presumably been passed over by positive selection. Among 16 excision products analyzed by Harada and Yamagishi (15), one of two untrimmed V segments was appended by a three-base P insert, and one of 6 untrimmed J segments had a two-base P insert. One and two-base P inserts have also been found in human kappa light chain gene junctions(33). These data strongly suggest that V_kJ_k junctions containing P inserts are generated with some regularity at the light chain locus (as reflected by excision products), but that inserts are subsequently counter-selected in some unknown fashion. Selection against N regions may be a general problem in the analysis of endogenous data, as it has been suggested to occur at other loci as well (2, 5).

Without , as was hoped, finding a situation where P inserts are clearly absent, we turned to an in-depth analysis of two loci, the TCRb locus and the IgH locus. In both cases, a large number of precursor germline sequences had been fully defined, and the junction data were extensive enough to permit comparisons between populations grouped according to several different parameters. We surveyed eight large-scale collections of IgH and TCRb junctions. In two cases (1, 4) unpublished and/or formatted sequences were generously provided by the authors.

P nucleotide addition was of peripheral concern in some of the studies, and the format in which junction sequences were displayed was very different from one publication to the next. In order to provide a consistent basis for comparison we found it necessary, (though tedious), to examine the data junction by junction. Full length ends, inserts, and P residues were scored according to a set of rules based upon consistent assignments of residues to each of these categories (complete details are given in Materials and Methods). There was variable agreement between our determinations and those of different authors. This affirmed the importance of re-evaluating all sequences according to standardized criteria prior to any comparison between studies. The data in Table 3 summarize the results of our analysis.

Between 2% and 15% of all ends incorporated into endogenous V(D)J junctions were associated with P inserts. The proportion of untrimmed ends associated with P inserts ranged from 11 to 78 percent. Although 95% of these P nucleotide inserts were one to two residues in length, 5% were 3 bases long, and there was one example of a four base P insert (Table 4). For both the TCRb and IgH loci, fetal/neonatal samples exhibited a lower level of P insertion than the equivalent adult sample. The largest ontogenetic differences were associated with 3' D ends in each case.

Before concluding that P insertion varies during ontogeny, other explanations for the observed differences must be considered. First, the number of full-length D ends within fetal/neonatal samples might have been inflated, (leading to lowered P insert ratios) because fetal junctions tend to lack N regions and crossover sites are often located within regions of short homology between component segments (1, 10, 14). Therefore, in a large number of the fetal (but not adult) junctions, therefore, there is uncertainty in the assignment of residues within the junctional region. As with the introduced substrate data, in order to conform to a consistent set of rules, residues were assigned so that, where possible, a full length end would be scored. As a consequence, the apparent P insert-to-untrimmed end ratio for the fetal/neonatal sample may have been depressed relative to its actual value. We suspect this was in fact the case, as illustrated by the alternative set of calculations shown in Table 3 (first and last column, parentheses). When all ambiguous 3'D-to-5'J junctions were excluded from the analysis, the discrepancy between fetal/neonatal and adult was much reduced for the IgH locus and disappeared completely at the TCRb locus (Table 3).

Another pronounced difference between the fetal/neonatal sample and that of the adult was the P insert frequency associated with the 3' Vb coding end (78% in the case of the adult, as compared to 45% for the fetal/neonatal sample). The 3'Vb coding end data were dominated by a single Vb segment Vb17a. This segment accounted for the majority of the adult 3'Vb ends, and about half of those in the fetal/neonatal collection (2, 4). There were no homologies at the Vb17a-Db junctions in either the adult or fetal collections that could account for these differences.

However, because the Vb17a data were derived predominantly from TCR surface receptor-positive cells, either positive or negative immune selection could have biased the samples. In order to compare samples that are as similar as possible except for the age of the mouse, a comparison was limited to thymocyte-derived Vb17a-containing junctions isolated from only one strain of mice (2). In this case the sample sizes are small but suggest that significant differences may not exist (8 of 14 untrimmed neonatal junctions had P inserts, vs. 11 of 15 untrimmed adult junctions).

We conclude that any actual ontogenic fluctuation in P insert frequency is too subtle to be revealed even with the extensive junctional data included in this survey. By way of contrast, a very striking variation in N insertion occurs through ontogeny, (1, 2, 10, 11, 14, 32). Thus P insertion is not as tightly regulated as N addition, if at all.

Discussion

In other site specific recombination systems, cutting and joining are energetically coupled operations carried out by a single site-specific enzyme (8). By contrast, V(D)J recombination appears to accomplish these operations through an ill-defined collaboration between several activities (reviewed in reference 22). Such complexity may well account for the lack of success in recreating the V(D)J joining reaction in a test-tube, despite nearly a decade of effort in a number of laboratories. The only known enzyme to be implicated (through molecular genetic analyses) in the reaction is terminal deoxynucleotidyl transferase [reference 19 and cited therein]. Though likely critical for N insertion, terminal deoxynucleotidyl transferase activity is not essential for V(D)J joining. By all appearances, V(D)J rearrangement is not the culmination of an orderly stringing-together of nucleolytic, polymerization, and ligation operations, nor (judging from the phenotype of SCID mice) should we expect that all functions playing a role in V(D)J joining are specifically dedicated to the process. Instead, some activities that participate in V(D)J joining may also feature in generalized recombination/repair mechanisms in the eucaryotic cell. There being no *a priori* guidelines in the interpretation of the fine structure of V(D)J junctions, some aspects may be indicative of an intrinsic, necessary, property of the joining mechanism and some may not. We interpret our present findings in this context.

The specificity of P nucleotide addition

We have demonstrated a statistically highly significant correlation between the sequences at the tip of the coding ends and the predicted P nucleotide pattern within inserts (Table 2G). Thus our data prove the hypothesis that P inserts exist (21). An unanticipated result was that the P nucleotide frequency within coding joints was variable. This variation was end-specific: for example, overall, P inserts were more frequently seen at the Sal end than at the Bam end (Figure 3, Table 1, Table 2G). End-associated differences in P insert frequency were apparent even between the two coding ends of the same construct (e.g., pSal-Bam; see Figure 3A, Table

1A, Table 2A, C). Interestingly, a study reported by Kallenbach *et al* . (19) employed a plasmid substrate in which the 12 and 23- signals associated with Bam and Sal coding ends were opposite to the arrangement in pSal-Bam. The *P* values we calculate from their data were surprisingly close to what was found here (0.83 for the Bam end, and 0.0003 for the Sal end, in comparison to 0.4 and 0.0002, respectively -Table 2A,C). This rules out the possibility that observed differences in P insert frequency were a function of the signal arrangement. Further, in our experiments, care was taken so that all recognized variables of the system (cell line, substrate structure, etc.) were held constant: we conclude that the DNA sequence of the coding end itself is the primary determinant of the observed P insert variation.

We have avoided the assertion that P nucleotide *addition* varies among the four substrates tested, because it is possible that the different frequencies with which we observe P nucleotides in the final products instead reflect differential eradication of pre-existing P residues. This distinction is important: if P nucleotide addition is irregular, then the presence of P residues must not be central to the joining reaction. If, instead, P nucleotides are always added to ends, (following which they are subject to sequence-influenced removal), P residues may indeed play an important role in joining.

Of the two possibilities, we first consider that P addition is consistent, but the added nucleotides are only variably preserved in the final junction products. Our data in fact suggest that different coding ends are subject to conspicuous and reproducible differences in the degree of truncation they exhibit upon incorporation into coding joints. As a consequence, one might imagine that the frequency of P insertion would also vary according to end identity. The two extreme cases were the Spe and Xho ends. Only 5% of coding joints exhibited non-truncated Spe ends, whereas 46% of coding joints had non-truncated Xho ends (Table 1B). Not surprisingly, the percentage of *all* coding joints that contained P inserts was correspondingly lower for the Spe end (2%) than for the Xho end (15%).

However, the observed sequence-influenced truncation was an unlikely explanation for the sequence-dependent frequency of P inserts at *untrimmed* ends (Table 2). As shown in Table

1B, neither a positive nor negative correlation existed between the frequency with which an end remained full-length and the likelihood of discovering P inserts among those examples that escaped truncation. Untrimmed Bam ends for example have fewer P inserts; if this is caused by P removal, then there must be a second distinct truncation activity with unusual discriminatory properties that allow it to subtract only P nucleotides from a recombination intermediate. The single stranded region of proposed recombination intermediates (Figure 1) includes non-P bases: but the agent that eradicates P nucleotides must somehow be able to distinguish P from non-P in order to account for a fluctuation in the ratio of P inserts to untrimmed ends. This type of explanation may be correct, but at present requires overly-intricate and unprecedented activities.

Another way in which P additions might be underrepresented in certain junctions is if the unpaired P nucleotides at an end were to anneal to the opposite strand of the partner end and create favored alignments for joining (27). The resulting junctions would lack evidence of any insert. We can evaluate this possibility by looking for a deficit of P inserts within junctions whose endpoints are consistent with potential "P overlap alignment." According to the data shown in Figure 3, while 47% of all junctions have P inserts, 31% of the junctions with P overlap alignments likewise have P inserts. Although this deficit suggests that alignments using P extension may occur, the effect is not particularly striking. Moreover, the results with pDSJ (Figure 5), a substrate that maximizes the opportunity for overlap alignment (Figure 1C), show that this does not take place in a predictable, recurrent fashion. As can be appreciated from both the frequency and *length* of P inserts where they are favored (in the case of Sal ends, Figure 3A, for example), it is unlikely that a propensity for P overlap alignments in some cases and not others is a significant factor in the variable observance of P inserts.

By far, the simplest explanation for our results is that P nucleotides are *not* introduced at all ends prior to joining. Rather than supposing that in certain cases, P additions are either lost of hidden at higher frequencies, our data indicate that the initial acquisition of P nucleotides is an irregular, sequence-specific, event.

P nucleotide addition at endogenous loci

To more fully characterize P insertion, we took advantage of the fact that a number of large-scale studies of the junctional diversity among assembled TCR and Ig genes have been published within the last two years. These include collections from both surface receptor positive and negative cells, from precursor as well as mature lymphocytes, and from cells at various stages of ontogeny in both the mouse and human. A comparative analysis of P insertion upon rearrangement of physiological substrates, could yield unique information. While the sequence-specific differences detected with the plasmid assay would be averaged out, other significant variation, or lack thereof, might emerge.

We computed the numbers of P nucleotides within junctions following a set of rules outlined in Materials and Methods. While the percentages of P inserts varied, no systematic differences emerged (Table 3). As detailed in Results, the perceived age-specific differences could well have arisen post-recombination, through selection, or as a secondary consequence of age-related differences in junction structure. Further, there was no consistent variation in P insertion with regard to gene segment identity (e.g., whether an end is a 3'V or 5'J). Finally, in accord with introduced substrate data, (this study and reference 19) coding ends originally associated with one type of signal (e.g., a 12-signal as opposed to 23-signal) did not preferentially acquire inserts.

These results support the view that P insertion is a general feature of V(D)J joining: although low and variable in appearance, P inserts were detected at all ends examined in detail. The percent of endogenous P inserts at untrimmed ends was comparable to the ratio of P inserts to untrimmed junctions we obtained using introduced substrates (compare Table 4, last column, to Table 1B).

P nucleotides and N regions

Both P and N inserts must initially be created by addition of nucleotides onto a free end during joining. However it was not known whether N residues could be added onto ends that already have a P nucleotide extension. Junctions with "composite" inserts containing both P and N residues might arise in either of two ways: either a P-modified end acquired an N region prior to joining, or P and N nucleotides were each contributed by different ends. We therefore made note of all junctions in which both ends were untrimmed, and that also contained inserts. Seven of 27 joints from the endogenous survey of the IgH locus that fell into this category had composite inserts in which an N region was sandwiched between two P inserts. One of these junctions (sequence # 109 from Decker *et al.* (9)), had *two* P residues at *both* borders of the insert, with a three base N region between. With introduced substrates, we here obtained one example, out of 10 doubly-untrimmed, insert-containing junctions, in which an N region was bounded by P nucleotides on *both* sides of the insert (72.1.3, Figure 3C). Among the pSal-Bam recombinants reported by Lieber *et al.* (30) one of two candidate junctions has a *two*-base P insert on *both* sides of a three-base N region. Thus such "sandwich" junctions are encountered with regularity among doubly untrimmed, insert-containing recombinants and, in particular, the existence of junctions in which the P inserts on *both* sides are longer than one residue, strongly suggests that N regions can be added to ends that possess pre-existing P residues. Thus P residue-modified termini must persist long enough to be exposed to the action of terminal transferase, or a similar type of activity. This excludes one class of models in which, for example, hairpin ends are not opened for resolution until immediately prior to joining.

The length of P inserts

Very long P nucleotide stretches (as long as 15 bp) have been documented in TCR g and d junctions derived from mice with severe combined immune deficiency (SCID) (6, 13, 20, 40). Although the defect in SCID mice has not yet been defined at the molecular level, it all but

eliminates productive V(D)J joining. Significantly, SCID mice exhibit a general DNA-repair defect, in addition to an inability to form V(D)J junctions (reviewed in reference 3).

The long P inserts within SCID junctions are anomalous, and are not necessarily indicative of length variation in non-mutant cells. In normal, non-SCID cells, P inserts of greater than two bases have been reported (20, 36, 40), and here, in a large-scale analysis, we find that 5% of all P nucleotide inserts in normal TCRb and IgH junctions were 3 or more bases long (Table 4). This frequency is not so high that extended P inserts in non-SCID junctions could be considered a typical variation in the P addition step. An alternative possibility is that the longer inserts are actually composites formed from a two-base P plus fortuitous N region sequence.

Results from the plasmid assay were enlightening in this regard. Our data show (see Results) that "excess" P nucleotides at Sal coding ends were not likely to have been introduced through random N addition and should indeed be regarded as P inserts of greater than two residues. Further, the appearance of long P inserts was end-specific. The Xho end, for example was never found associated with a P insert of greater than two residues. We draw two conclusions from these results: 1) P inserts of greater than two bases arise in non-SCID cells at a statistically significant frequency, and 2) the length of P inserts may be directly influenced by the sequence to which they are appended.

Asymmetry of V(D)J joining

The asymmetry of V(D)J joining has been clearly demonstrated by the pDSJ results reported here. A "symmetrical" substrate, in which each of the four ends have signal identity, forms two distinctive junctions when rearranged. One junction had the properties of a signal joint, while the other had the fine-structural features of a coding joint. The "coding joint" exhibited features including base loss and addition as well as P insertion.

The differences observed between the grossly reciprocal structures of coding and signal joints, is usually interpreted to mean that mechanistically distinct cutting and joining operations create the signal and coding joints in turn (3, 21, 27, 31, 39). This view cannot be fully correct,

because the existence of coding-to-signal connections, as in a hybrid and open and shut junctions [reference 25 and cited therein] is difficult to thereby rationalize. If P addition is the step at which asymmetry is introduced into V(D)J joining, the joining process is probably thereafter "symmetrical." That is, if ligation of signal ends is carried out by a site-specific ligase, while coding end formation is accomplished by non-specific cellular end-joining machinery, it is hard to conceive how the two operations can mix so seamlessly in hybrid joint formation. It is far more probable that both coding and signal joint formation are created by one and the same joining mechanism, whether a specific or non-specific enzymatic machinery is responsible.

An alternative view is that the apparent asymmetry in V(D)J joining may be the result of differential occlusion of coding and signal ends by site-specific components of the V(D)J joining machinery. Although in pDSJ, all four ends have signal identity, only two of the four ends are site-specifically engaged by the recombination apparatus, and only those two ends may be protected from base loss and addition, and P nucleotide modification.

Mechanistic Implications

Our results, based on the evaluation of both introduced and endogenous joining substrates, affirm the generality of P nucleotide insertion, and are consistent with the view that P nucleotide addition occurs along with other end-processing operations during coding joint formation. However, at the same time, we have found that junctions constructed from certain coding ends are more apt to contain P nucleotide inserts than others, certain ends are prone to truncation, and that the length of P inserts can vary. These observations are inconsistent with a mechanism in which a dinucleotide transfer is an obligatory part of the joining operation (Figure 1A, left; ref. 2). Instead, the initial P nucleotide addition step may be stochastic, and the number of nucleotides involved is not fixed.

A hairpin intermediate, as suggested by Lieber (27), can accommodate the data. If, as proposed, hairpins form only at coding ends in the course of site-specific cleavage, the length of a P insert (from zero to more than two) might then be dictated by the position of the single-strand nick that opens the hairpin. This nicking step might be influenced by the DNA sequence of the

hairpin end, or perhaps by some aspect of tertiary structure. Depending upon whether a terminus with a 3' , 5', or no overhang is created upon opening the hairpin, the probability of incorporating P inserts into junctions upon joining ought to vary. By this model, a sequence-dependent variation in P insert frequency and the mechanism(s) that cause sequence-specific differences in truncation ought to be independent of one another, which is in fact what we observe (Table 1B).

There is precedent for hairpin formation in other site-directed DNA transactions. A hairpin intermediate has been proposed in the excision of plant transposable elements (7, 35) and has been demonstrated as alternative product in l site-specific recombination under conditions where normal strand exchange is blocked (34). We note, however, that there has been no systematic analysis of insertion at non-lymphoid junctions that would serve to rule out the possibility that P inserts (or hairpins) may occur in the context of more general, illegitimate recombination. In one such study (37) a number of inserts appear to fit the P insert pattern; however, the data are not amenable to statistical analysis, so the question remains open.

We favor a more general model in which hairpin formation, and thus P nucleotide addition, plays a role in DNA damage-repair. Without being intrinsic to the V(D)J joining operation, P nucleotides may be observed by virtue of their introduction through an incidental enzymatic activity that intervenes during the recombination process. It could be that certain types of broken ends are sealed by hairpin formation in order to prevent exonucleolytic loss and/or to delay potentially disruptive interactions between broken ends and other chromosomal sites until the breaks can be mended. In fact, the SCID phenotype (reviewed in ref. 29) suggests that a link between P insert activity, general DNA repair, and the V(D)J joining operation does exist. This link could be hairpin formation and/or processing. As examples, the SCID defect could cause excessive hairpin formation or interfere with their eventual resolution. An exciting recent development is the physical detection of hairpin coding ends in SCID thymocytes (38).

Acknowledgments

The authors would like to thank L. Czyzyk for superlative technical assistance, Drs. E.B. Lewis, D. Mathog, and H. Lipshitz for advice on the statistical analysis, and Drs. B. Wold, J. Kobori, L. Hood, P. Fahnestock, H. Lipshitz and P. Bjorkman for comments on the manuscript. We thank Dr. M. Gellert for comments, and for communicating his results prior to publication. We thank Dr. J. Teale for providing unpublished sequence data. J.T.M gratefully acknowledges the support of Dr. Leroy Hood, and that of NIH grant GM40867 to Dr. Hood. This work was funded by research grant IM-599 from the American Cancer Society to S.M.L.

References

1. **Bangs, L. A., I. Sanz and J. M. Teale.** 1991. Comparison of D, J_H and Junctional diversity in the fetal, adult and aged B cell repertoires.

2. **Bogue, M., S. Candéias, C. Benoist and D. Mathis.** 1991. A special repertoire of a:b T cells in neonatal mice. EMBO J. **10**:3647-3654.

3. **Bosma, M. J. and A. M. Carroll.** 1991. The Scid mouse mutant: Definition, characterization, and potential uses. Annu. Rev. Immunol. **9**:323-50.

4. **Candéias, S., C. Waltzinger, C. Benoist and D. Mathis.** 1991. The Vb17+ T cell repertoire: Skewed Jb usage after thymic selection; dissimilar CDR3s in CD4+ Versus CD8+ Cells. J. Exp. Med **174**:989-1000.

5. **Carlsson, L., C. Övermo and D. Holmerg.** 1992. Selection against N-region diversity in immunoglobulin heavy chain variable regions during the development of pre-immune B cell repertoires. Int'l Immunol. **4**:549-553.

6. **Carroll, A. M. and M. J. Bosma.** 1991. T-lymphocyte development in SCID mice is arrested shortly after the initiation of T-cell receptor d gene recombination. Genes Dev. **5**:1357-1366.

7. **Coen, E. S., R. Carpenter and C. Martin.** 1986. Transposable elements generate novel spatial patterns of gene expression in Antirrhinum majus. Cell **47**:285-296.

8. **Craig, N. L.** 1988. The Mechanism of Conservative Site-Specific Recombination. Annu. Rev. Genet. **22**:77-105.

9. **Decker, D. J., N. E. Boyle, J. A. Koziol and N. R. Klinman.** 1991. The expression of the Ig H chain repertoire in developing bone marrow B lineage cells. J. Immunol. **146**:350-361.

10. **Feeney, A. J.** 1990. Lack of N regions in fetal and neonatal mouse immunoglobulin V-D-J junctional sequences. J. Exp. Med **172**:1377-1390.

11. **Feeney, A. J.** 1991. Junctional sequences of fetal T cell receptor b chains have few N regions. J. Exp. Med **174**:115-124.

12. **Feeney, A. J.** 1991. Predominance of the prototypic T15 anti-phosphorylcholine junctional sequence in neonatal pre-B cells. J. Immunol. **147**:4343-4350.

13. **Ferrier, P., L. R. Covey, S. C. Li, H. Suh, B. A. Malynn, T. K. Blackwell, M. A. Morrow and F. W. Alt.** 1990. Normal Recombination substrate VH to DJH rearrangements in pre-B cell lines from scid mice. J. Exp. Med **171**:1909-1918.

14. **Gu, H., I. Förster and K. Rajewsky.** 1990. Sequence homologies, N sequence insertion and J_H gene utililization in $V_HD_HJ_H$ joining: Implications for the joining mechanism and the ontogenetic timing of Ly1 B cell and B-CLL progenitor generation. EMBO J. **9**:2133-2140.

15. **Harada, K. and H. Yamagishi.** 1991. Lack of feedback inhibition of V-kappa gene rearrangement by productively rearranged alleles. J. Exp. Med **173**:409-415.

16. **Hesse, J. E., M. R. Lieber, M. Gellert and K. Mizuuchi.** 1987. Extrachromosomal DNA substrates in pre-B cells undergo inversion of delection at immunoglobulin V-(D)-J joining signals. Cell **49**:775-783.

17. **Hesse, J. E., M. R. Lieber, K. Mizuuchi and M. Gellert.** 1989. V(D)J recombination: a functional definition of the joining signals. Genes Dev. **3**:1053-1067.

18. **Kabat, E. A., T. T. Wu, H. M. Perry, K. S. Gottesman and C. Foeller,** *In* E.A. Kabat (ed.) Sequences of Proteins of Immunological Interest, Fifth ed. NIH publication No. 91-3242, . Vol. 2. 1991, U.S. Department of Health and Human Services. Bethesda, MD.

19. **Kallenbach, S., N. Doyen, M. F. D'Andon and F. Rougeon.** 1992. Three lymphoid-specific factors account for all junctional diversity characteristic of somatic assembly of T-cell receptor and immunoglobulin genes. Proc. Natl. Acad. Sci. USA **89**:2799-2803.

20. **Kienker, L. J., W. A. Kuziel, B. A. Garni-Wagner, V. Kumar and P. W. Tucker.** 1991. T cell receptor g and d gene rearrangements in SCID thymocytes: similarity to those in normal thymocytes. J. Immunol. **147**:4351-4359.

21. **Lafaille, J. J., A. DeCloux, M. Bonneville, Y. Takagaki and S. Tonegawa.** 1989. Junctional sequences of T cell receptor gd genes: implications for gd T cell lineages and for a novel intermediate of V-(D)-J joining. Cell **59**:859-870.

22. **Lewis, S. and M. Gellert.** 1989. The mechanism of antigen receptor gene assembly. Cell **58**:585-588.

23. **Lewis, S., A. Gifford and D. Baltimore.** 1985. DNA elements are asymmetrically joined during the site-specific recombination of kappa immunoglobulin genes. Science **228**:677-685.

24. **Lewis, S., A. Gifford and D. Bltimore.** 1984. Joining of Vk to Jk gene segments in a retroviral vector introduced into lymphoid cells. Nature (London) **308**:425-428.

25. **Lewis, S., J. E. Hesse, K. Mizuuchi and M. Gellert.** 1988. Novel strand exchanges in V(D)J recombination. Cell **55**:1099-1107.

26. **Lewis, S. M. and J. E. Hesse.** 1991. Cutting and closing without recombination in V(D)J joining. EMBO J. **10**:3631-3639.

27. **Lieber, M. R.** 1991. Site-specific recombination in the immune system. FASEB J. **4**:2934-2944.

28. **Lieber, M. R., J. E. Hesse, S. Lewis, G. C. Bosma, N. R. Rosenberg, K. Mizuuchi, M. J. Bosma and M. Gellert.** 1988. The defect in murine severe combined immune deficiency: joining of signal sequences but not coding segments in V(D)J recombination. Cell **55**:7-16.

29. **Lieber, M. R., J. E. Hesse, K. Mizuuchi and M. Gellert.** 1987. Developmental stage specificity of the lymphoid V(D)J recombination activity. Genes Dev. **1**:751-751.

30. **Lieber, M. R., J. E. Hesse, K. Mizuuchi and M. Gellert.** 1988. Lymphoid V(D)J recombination: nucleotide insertion at signal joints as well as coding joints. Proc. Natl. Acad. Sci. USA **85**:8588-8592.

31. **McCormack, W. T., L. W. Tjoelker, L. M. Carlson, B. Petryniak, C. Barth, E. Humphries and C. B. Thompson.** 1989. Chicken IgL gene rearrangement involves

deletion of a circular episome and addition of single nonrandom nucleotides to both coding segments. Cell **56**:785-791.

32. **McVay, L. D., S. R. Carding, K. Bottomly and A. C. Hayday.** 1991. Regulated expression and structure of T cell receptor g/d transcripts in human thyumic ontogeny. EMBO J. **10**:83-91.

33. **Milstein, C., J. Even, J. M. Jarvis, A. Gonzalez-Fernandez and E. Gherardi.** 1992. Non-random features of the repertoire expressed by the memebers of one Vk gene family and of the V-J recombination. Eur. J. Immunol. **22:**

34. **Nash, H. A. and C. A. Robertson.** 1989. Heteroduplex substrates for bacteriophage lambda site-specific recombination: Cleavage and strand transfer products. EMBO **11**:3523-3533.

35. **Peacock, W. J., E. S. Dennis, W. L. Gerlach, M. M. Sachs and D. Schwartz.** 1984. Insertion and excision of *Ds* controlling elements in Maize. Cold Spring Harbor Symp. Quant. Biol. **45**:347-354.

36. **Reynaud, C., V. Anquez and J. Weill.** 1991. The chicken D locus and its contribution to the immunoglobulin heavy chain repertoire. Eur. J. Immunol. **21**:2661-2670.

37. **Roth, D. B., X.-B. Chang and J. Wilson.** 1989. Comparison of filler DNA at immune, nonimmune, and oncogenic rearrangements suggestys multiple mechanisms of formation. Mol. Cell. Biol. **9**:3049-3057.

38. **Roth, D. B., J. P. Menetski, P. Nakajima, M. J. Bosma and M. Gellert.** 1992. V(D)J recombination: Broken DNA molecules with covalently sealed (hairpin) Coding ends in SCID mouse thymocytes. Cell **in press:**

39. **Roth, D. B., P. B. Nakajima, J. P. Menetski, M. J. Bosma and M. Gellert.** 1992. V(D)J recombination in mouse thymocytes: double-strand breaks near T cell receptor d rearrangement signals. Cell **69**:41-53.

40. **Schuler, W., N. R. Reutsch, M. Amsler and M. J. Bosma.** 1991. Coding joint formation of endogenous T cell receptor genes in lymphoid cells from SCID mice: unusual P-nucleotide additions in VJ-coding joints. Eur. J. Immunol. **21**:598-596.

Figure Legends

Figure 1: Hypothetical origin of P nucleotide inserts

Panel A) Two models for coding joint formation. Coding ends only (arbitrarily derived from pSal-Bam) are shown. Top: Blunt-ended termini are indicated; however, the exact structure of the cleavage products is unknown. Upper middle: P nucleotide addition. To the left is the proposal of Lafaille et al.(21), to the right is that of Lieber (27). The Lafaille proposition is that a di-nucleotide is obligatorily removed from one strand of the coding end and then ligated to the other. The Lieber model suggests that the two strands of the coding end become connected to form a hairpin, following which a nick opens the hairpin up. The operations are proposed to affect both coding termini prior to joining. Lower middle: In either case, overhang intermediates are proposed as shown. Thereafter the ends of the overhangs are altered by an unknown number of activities in an unknown order. Modifications include N insertion, truncation, and filling-in of recessed termini. Bottom: A product coding joint. (The sequence shown corresponds to isolate 10.1.11 from pSal-Bam; Figure 3) Junctional inserts at the crossover site are indicated by spaces on either side: P nucleotides (bold type-face, underlined), as well as an N insert (plain type) are shown.

For simplicity, two aspects of the Lieber model are not explicitly represented: as proposed, the hairpin is created after only one of two strands has been cleaved. Also, following hairpin formation, the location of the nick that opens the hairpin is variable. (For details see Discussion and Lieber 1991.)

Panels B and C) P Nucleotides May Figure in the Formation of "Precise" Signal Joints. The hypothetical fate of signal ends, (or of coding ends with an inverse-complementary relationship), is indicated in panel B. Top: Two signal ends (or, alternatively the complementary coding ends present in pDSJ) are represented after cleavage. Middle: Overhang intermediates produced after P addition. Note the complementary relationship between the single-strand protrusions Bottom: "Precise" junction formed by direct annealing of the complementary overhangs followed by

ligation. The vertical bar represents the apparent crossover site. Panel C: The fate of non-complementary signal ends. Top: Non inverse-complementary signal ends (pMut-2,27). Middle: Overhang intermediates. Note the extensions are not longer complementary , and direct annealing is not possible. Bottom: the "imprecise" outcome. An invented sequence exhibiting truncation and N and P insertion is shown.

Figure 2: Constructs and Products.

Relevant regions of the constructs used in this study are shown, along with possible products. Open boxes represent Sal 1 or Spe 1 sites, open triangles represent 12-spacer joining signals. Shaded boxes and triangles represent Bam or Xho sites and 23-spacer signals, respectively. B, S and A indicate Bam H1, Sal 1 and ApaL1 sites, respectively. (All sites for these enzymes are shown excepting ApaL1, which cleaves at several positions in the plasmid outside the illustrated region.) "Stop" indicates the prokaryotic transcription terminator, and "P", the promoter that drives cat gene expression after rearrangement.

A) Structures of products obtained with the related constructs, pSal-Bam, pSpe-Bam, pSal-Xho, pSpe-Xho, pMut-2 and pMut-27. A standard reaction result in an inversion, with a coding joint at one boundary, and a signal joint at the other (shown in the middle). A hybrid deletion (shown at the bottom) connects a coding end (here, Sal or Spe) to the 23-signal. Variable coding and hybrid junctions are indicated by jagged lines, the "precise" signal joints is shown with a straight edge.

B) pDSJ and possible products. pDSJ contains four signals arranged in two signal joints as follows from left to right.: the 12-signal from $V_{k21\text{-}C}$, connected to the 23-signal from J_{k2}, and the 23-signal from J_{k1} abutting the 12-signal from $V_{kL\text{-}8}$. Four possible outcomes in which one, both, or neither product junction has the typical signal joint structure are possible. Variable and precise junctions are indicated as in panel A. Predicted sensitivity of each class of junction to Apa L1 digestion, is shown.

Figure 3: P Inserts Within Untrimmed Coding Joints

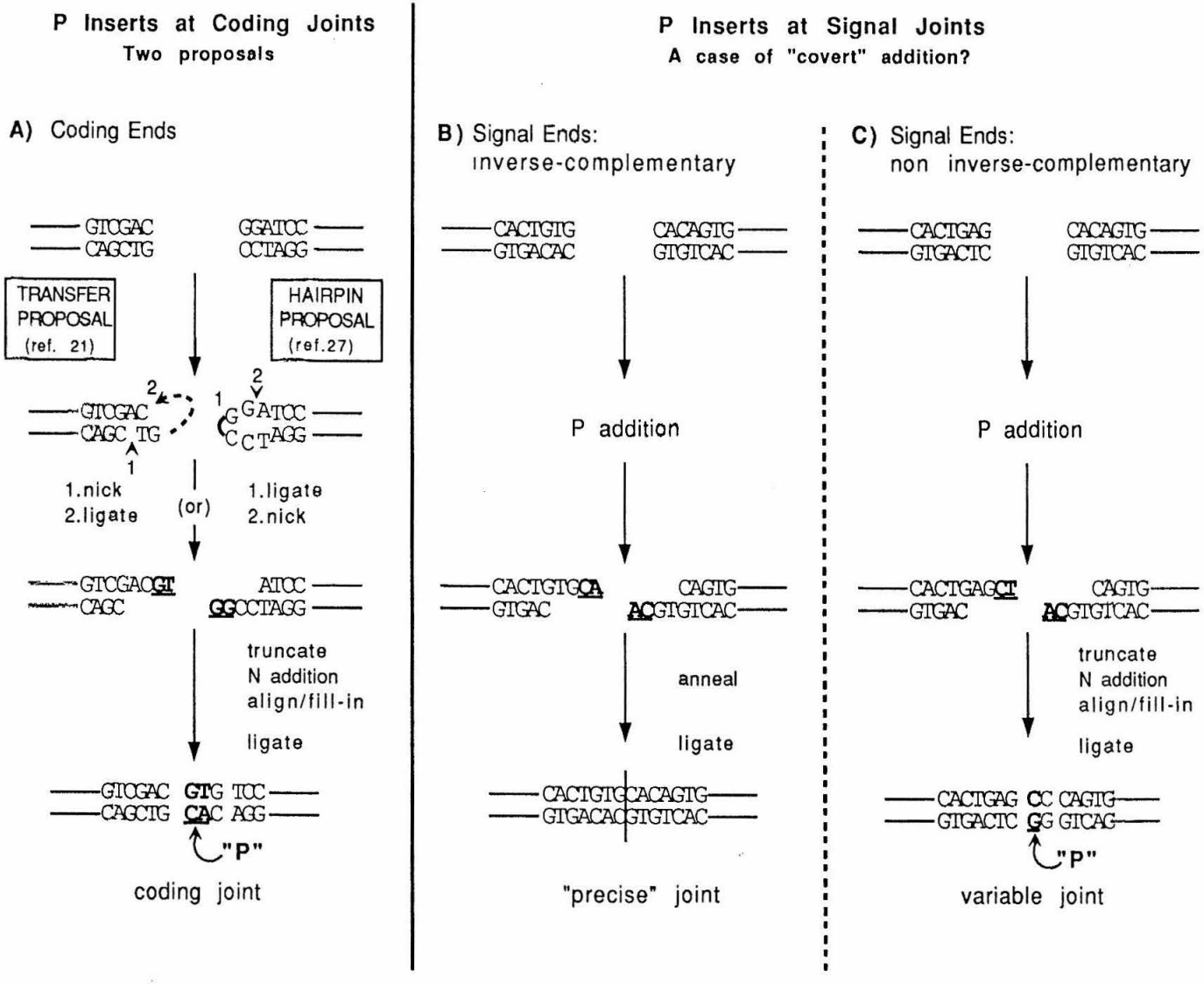
Panels A, B, C, and D: products isolated with pSal-Bam, pSpe-Bam, pSal-Xho and pSpe-Xho as indicated. The first line in each panel shows the sequences of the two relevant, non-truncated coding ends. The recombinants containing untrimmed Sal or Spe coding ends are shown at the top portion of each panel, those in which both coding ends are intact are grouped in the middle, and those with intact Bam or Xho coding ends are shown at the bottom. For consistency, where a residue might be assigned to either of the two coding ends it was assigned to the untrimmed coding end. Inserted residues are shown in the center: for each junction P inserts are underlined and in bold typeface, N residues are in plain type. All recombinants listed are independent isolates.

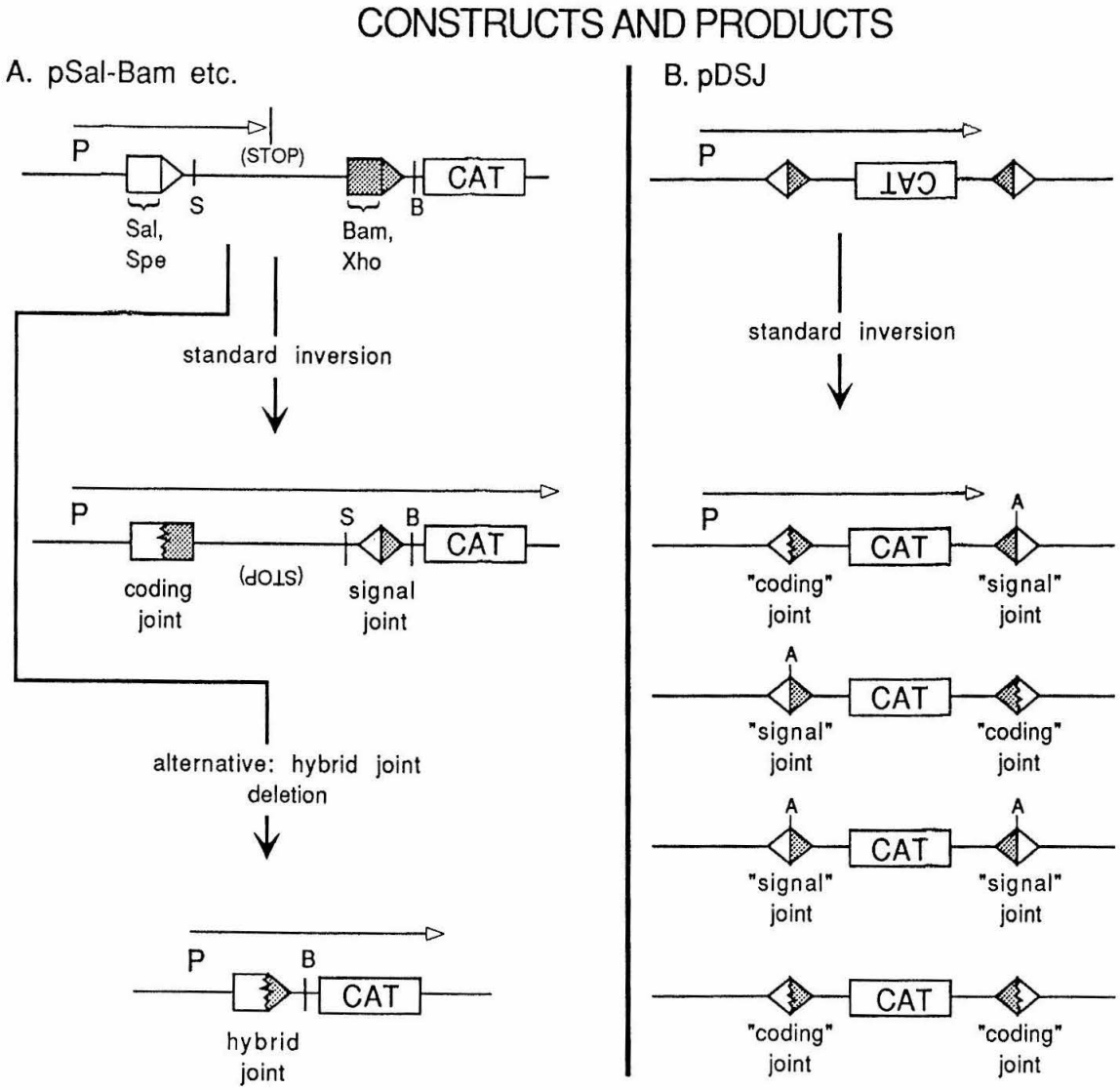
Figure 4: P Inserts Within Hybrid and Signal Joints

Panel A) Top: Hybrids in which the Sal coding end is connected to the 23-signal. Products were isolated from pSalBam and pSalXho. Panel A) Bottom: Hybrids in which the Spe end is joined to the 23-signal. Products were isolated from pSpeBam and pSpeXho. Panel B): Signal joints exhibiting base loss or addition. All details are as described in the legend to Figure 3.

Figure 5: Analysis of P Inserts: Complementary vs.. Non-Complementary Signal Ends

Panel A) Sequences of pMut-2 and 27 recombinants. Panel B) pDSJ recombinants. The first line in each panel gives the sequence of untrimmed end comprising the junctions in each construct . Ambiguous residues were assigned as in Figure 3. Inserts, and P nucleotides, are designated as described in the legend to Figure 3. Recombinants with the same sequence were listed only if derived from independent transfections.


P Inserts at Coding Joints
Two proposals
A) Coding Ends
TRANSFER PROPOSAL (ref. 21)
HAIRPIN PROPOSAL (ref.27)
1.nick
2.ligate
(or)
1.ligate
2.nick
truncate
N addition
align/fill-in
ligate
"P"
coding joint
P Inserts at Signal Joints
A case of "covert" addition?
B) Signal Ends:
inverse-complementary
P addition
anneal
ligate
"precise" joint
C) Signal Ends:
non inverse-complementary
P addition
truncate
N addition
align/fill-in
ligate
"P"
variable joint


CONSTRUCTS AND PRODUCTS
A. pSal-Bam etc.
P
(STOP)
CAT
S
B
Sal,
Spe
Bam,
Xho
standard inversion
P
S
B
CAT
coding
joint
(STOP)
signal
joint
alternative: hybrid joint
deletion
P
B
CAT
hybrid
joint
B. pDSJ
P
CAT
standard inversion
P
A
CAT
"coding"
joint
"signal"
joint
A
CAT
"signal"
joint
"coding"
joint
A
A
CAT
"signal"
joint
"signal"
joint
CAT
"coding"
joint
"coding"
joint

3A)	Sal		Bam
pSal-Bam	TGCAGGTCGAC		GGATCCTCTCA
4.1.20	TGCAGGTCGAC		CCTCTCA
5.1.19	TGCAGGTCGAC		CCTCTCA
5.1.17	TGCAGGTCGAC	**GTCG**G	CCTCTCA
8.7.1	TGCAGGTCGAC	**G**	TCCTCTCA
10.1.11	TGCAGGTCGAC	**GT**G	TCCTCTCA
8.3.5	TGCAGGTCGAC	**GTC**	ATCCTCTCA
8.5.10	TGCAGGTCGAC	**GTC**CC	ATCCTCTCA
11.1.1	TGCAGGTCGAC		GATCCTCTCA
10.1.9	TGCAGGTCGAC		GATCCTCTCA
8.6.4	TGCAGGTCGAC	**GT**	GATCCTCTCA
8.1.1	TGCAGGTCGAC	**G**CT	GATCCTCTCA
8.2.9	TGCAGGTCGAC	**GTC**	GATCCTCTCA
10.1.6	TGCAGGTCGAC	**G**GA	GATCCTCTCA
8.7.12	TGCAGGTCGAC	**CC**	GGATCCTCTCA
8.7.11	TGCAGGTCGAC	**G**	GGATCCTCTCA
5.1.33	TGCAGGTCGA		GGATCCTCTCA
8.1.13	TGCAGGTCG		GGATCCTCTCA
4.1.24	TGCAGGTC		GGATCCTCTCA
5.1.28	TGCAGGTC		GGATCCTCTCA
8.1.5	TGCAGGTC		GGATCCTCTCA
10.1.3	TGCAGGTC		GGATCCTCTCA
5.1.32	TGCAGGT		GGATCCTCTCA
8.6.2	TGCAGGT		GGATCCTCTCA
8.6.8	TGCAGGT	AA	GGATCCTCTCA
8.1.4	TGCAGG		GGATCCTCTCA
4.1.60	TGCAGG		GGATCCTCTCA
8.1.11	TGCAGG	G	GGATCCTCTCA
8.7.9	TGCAGG	**CC**	GGATCCTCTCA
4.1.75	TGCAGG	GGG	GGATCCTCTCA
8.1.10	TGCAG		GGATCCTCTCA
5.1.36	TGCAG	T	GGATCCTCTCA
5.1.18	TGCA		GGATCCTCTCA
8.8.12	TGC	**C**	GGATCCTCTCA
10.1.4	TGC	**C**	GGATCCTCTCA
10.1.8	TGC	**CC**	GGATCCTCTCA

3B)	Spe		Bam
pSpe-Bam	GGTCGACTAGT		GGATCCTCTCA
34.2.6	GGTCGACTAGT		CCTCTCA
35.4.6	GGTCGACTAGT		ATCCTCTCA
32.3.6	GGTCGACTAGT	**A**A	GATCCTCTCA
42.1.2	GGTCGACTAG	G	GGATCCTCTCA
42.2.12	GGTCGACTAG	**C**	GGATCCTCTCA
32.3.7	GGTCGACTA		GGATCCTCTCA
34.2.3	GGTCGACTA		GGATCCTCTCA
35.4.3	GGTCGACTA		GGATCCTCTCA
45.4.3	GGTCGACTA		GGATCCTCTCA
33.3.11	GGTCGACTA	**C**	GGATCCTCTCA
43.1.2	GGTCGACTA	CG	GGATCCTCTCA
34.3.12	GGTCGACT		GGATCCTCTCA
45.2.7	GGTCGACT	CC**CC**	GGATCCTCTCA
32.3.5	GGTCGAC	CA	GGATCCTCTCA
18.1.1	GGTCGAC	GTA	GGATCCTCTCA
33.3.10	GGTCGAC		GGATCCTCTCA
44.5.1	GGTCGAC		GGATCCTCTCA
45.2.1	GGTCGAC		GGATCCTCTCA
33.3.5	GGTCGAC	**C**	GGATCCTCTCA
43.2.6	GGTCGAC	**C**	GGATCCTCTCA
44.2.12	GGTCGAC	CAA	GGATCCTCTCA
42.5.1	GGTCGA	**TCC**	GGATCCTCTCA
33.2.11	GGTCGA	T**TCC**	GGATCCTCTCA
45.2.9	GGTCG		GGATCCTCTCA
32.2.4	GGTCG	T**C**	GGATCCTCTCA
32.2.5	GGTCG	**CC**	GGATCCTCTCA
33.2.1	GGTCG	CC**CC**	GGATCCTCTCA
33.3.6	GGTC		GGATCCTCTCA
44.1.4	GGTC		GGATCCTCTCA
33.2.6	GG		GGATCCTCTCA
44.5.4	-11(A)		GGATCCTCTCA
42.5.5	-11(A)	**C**	GGATCCTCTCA
35.3.4	-16(G)	T	GGATCCTCTCA
27.1.10	-17(G)		GGATCCTCTCA

3C)	Sal		Xho
pSal-Xho	TGCAGGTCGAC		CTCGAGGATCC
28.3.11	TGCAGGTCGAC		GGATCC
38.4.8	TGCAGGTCGAC		GGATCC
73.1.5	TGCAGGTCGAC	**GT**AA	GAGGATCC
28.2.17	TGCAGGTCGAC		GAGGATCC
29.4.5	TGCAGGTCGAC		GAGGATCC
72.1.7	TGCAGGTCGAC	**G**	CGAGGATCC
73.1.7	TGCAGGTCGAC	**G**GG	CGAGGATCC
64.1.1	TGCAGGTCGAC	**GTC**TC	CGAGGATCC
65.1.6	TGCAGGTCGAC	**G**	TCGAGGATCC
28.2.12	TGCAGGTCGAC	**G**	TCGAGGATCC
73.1.1	TGCAGGTCGAC		CTCGAGGATCC
28.2.8	TGCAGGTCGAC	**G**	CTCGAGGATCC
72.2.6	TGCAGGTCGAC	**G**	CTCGAGGATCC
72.1.1	TGCAGGTCGAC	CC**G**	CTCGAGGATCC
73.2.5	TGCAGGTCGAC	**GTCAG**	CTCGAGGATCC
72.1.3	TGCAGGTCGAC	**GT**GAG**G**	CTCGAGGATCC
73.2.7	TGCAGGTCGAC	TGAGA**AG**	CTCGAGGATCC
22.1.4	TGCAGGTCGA	**G**	CTCGAGGATCC
28.2.15	TGCAGGTCG		CTCGAGGATCC
29.4.11	TGCAGGTCG		CTCGAGGATCC
29.2.10	TGCAGGTCG	**G**	CTCGAGGATCC
64.1.4	TGCAGGTCG	CT	CTCGAGGATCC
23.2.8	TGCAGGTC	CC	CTCGAGGATCC
29.3.8	TGCAGGT	A	CTCGAGGATCC
28.4.9	TGCAGGT	**G**	CTCGAGGATCC
38.2.12	TGCAGGT		CTCGAGGATCC
38.3.2	TGCAGG	**G**	CTCGAGGATCC
38.3.5	TGCAGG		CTCGAGGATCC
28.3.1	TGCAG		CTCGAGGATCC
29.2.17	TGCAG		CTCGAGGATCC
38.3.7	TGCAG		CTCGAGGATCC
64.1.6	TGCAG		CTCGAGGATCC
28.4.6	TGCAG	C	CTCGAGGATCC
23.2.2	TGCA	C	CTCGAGGATCC
23.2.9	TGCA	A	CTCGAGGATCC
28.2.19	TG		CTCGAGGATCC

3D)	Spe		Xho
pSpe-Xho	GGTCGACTAGT		CTCGAGGATCC
62.2.2	GGTCGACTAGT		AGGATCC
47.1.4	GGTCGACTAGT	**AC**	AGGATCC
49.3.1	GGTCGACTAGT	**A**	CGAGGATCC
63.5.1	GGTCGACTAGT	**A**	CGAGGATCC
49.1.5	GGTCGACTAGT	**AC**C	CGAGGATCC
46.1.2	GGTCGACTAGT	CG**G**	CTCGAGGATCC
63.2.9	GGTCGACTAGT	**A**	CTCGAGGATCC
47.1.3	GGTCGACTAG	**AG**	CTCGAGGATCC
63.2.10	GGTCGACTAG	**AG**	CTCGAGGATCC
49.1.2	GGTCGACTAG	AATC	CTCGAGGATCC
62.5.9	GGTCGACTAG		CTCGAGGATCC
63.2.12	GGTCGACTAG		CTCGAGGATCC
53.1.7	GGTCGACT	**G**	CTCGAGGATCC
62.2.1	GGTCGACT	**G**	CTCGAGGATCC
63.2.5	GGTCGACT	**G**	CTCGAGGATCC
63.3.4	GGTCGACT	C**G**	CTCGAGGATCC
49.3.6	GGTCGAC	CC**G**	CTCGAGGATCC
62.2.12	GGTCGA	**G**	CTCGAGGATCC
63.2.4	GGTCGA	**G**	CTCGAGGATCC
47.1.5	GGTCGA		CTCGAGGATCC
48.1.1	GGTCGA		CTCGAGGATCC
62.2.6	GGTCGA		CTCGAGGATCC
63.4.4	GGTCGA		CTCGAGGATCC
48.1.4	GGTCG	CCCAT	CTCGAGGATCC
62.4.11	GGTCG	**G**	CTCGAGGATCC
63.4.2	GG	AGA	CTCGAGGATCC
63.4.8	GG	**AG**	CTCGAGGATCC
62.4.5	G		CTCGAGGATCC

4A)

	SalI		**23-signal**
	TGCAGGTCGAC		CACAGTGGTAG
29.6.5	TGCAGGTCGAC	**G**G	G
26.2.4	TGCAGGTCGAC		AGTGGTAG
29.6.9	TGCAGGTCGAC		CACAGTGGTAG
38.4.5	TGCAGGTCGAC		CACAGTGGTAG
28.2.10	TGCAGGTCGAC	**G**	CACAGTGGTAG
29.5.6	TGCAGGTCGAC	A	CACAGTGGTAG
29.6.2	TGCAGGTCGAC	C	CACAGTGGTAG
28.5.12	TGCAGGTCGAC	**G**A	CACAGTGGTAG
29.4.12	TGCAGGTCGAC	**GT**	CACAGTGGTAG
72.1.5	TGCAGGTCGAC	**GT**GGAA	CACAGTGGTAG
28.5.5	TGCAGGTCGA		CACAGTGGTAG
29.4.3	TGCAGGTCGA		CACAGTGGTAG
29.2.5	TGCAGGTCG		CACAGTGGTAG
64.1.2	TGCAGGTCG		CACAGTGGTAG
28.5.1	TGCAGGTCG	T	CACAGTGGTAG
8.1.6	TGCAGGTCG	GG**G**	CACAGTGGTAG
8.2.1	TGCAGGTCG	CA	CACAGTGGTAG
4.1.18	TGCAGGT	AGGA	CACAGTGGTAG
4.1.41	TGCAGGT		CACAGTGGTAG
8.6.5	TGCAG	TC	CACAGTGGTAG

	Spe		**23-signal**
	GGTCGACTAGT		CACAGTGGTAG
63.4.11	GGTCGACTAGT	C	CACAGTGGTAG
63.3.7	GGTCGACTAGT		CACAGTGGTAG
46.1.4	GGTCGACTAGT	**AC**	CACAGTGGTAG
49.3.2	GGTCGACTAGT	**A**	CACAGTGGTAG
53.1.12	GGTCGACTAG	G**G**	CACAGTGGTAG
44.2.1	GGTCGACTAG		CACAGTGGTAG
44.4.6	GGTCGACT	CCC	CACAGTGGTAG
45.3.12	GGTCGACT		CACAGTGGTAG
62.4.7	GGTCGACT	A	CACAGTGGTAG
62.2.8	GGTCGACT		CACAGTGGTAG
48.3.6	GGTCGAC	GGA	CACAGTGGTAG
62.2.3	GGTCGAC		CACAGTGGTAG
27.2.1	GGTCGAC		CACAGTGGTAG
27.1.4	GGTCGA		CACAGTGGTAG
32.2.6	GGTCGA		CACAGTGGTAG
62.4.1	GGTCGA		CACAGTGGTAG
32.2.10	GGTCG		CACAGTGGTAG
64.1.2	GGTCG		CACAGTGGTAG
32.3.9	-13(G)	TA	CACAGTGGTAG

4B)	12 signal		23 signal
	GTAGCACTGTG		CACAGTGGTAG
43.2.9	GTAGCACTGTG	GGTC	CACAGTGGTAG
5.1.2	GTAGCACTGTG	GTT	CACAGTGGTAG
8.1.3	GTAGCACTGTG	TGGA	CACAGTGGTAG
8.1.8	GTAGCACTGTG	T	CACAGTGGTAG
8.5.8	GTAGCACTGTG	AGG**G**	CACAGTGGTAG
8.5.9	GTAGCACTGTG	GG**G**	CACAGTGGTAG
8.8.7	GTAGCACTGTG	**C**C	CACAGTGGTAG
29.4.9	GTAGCACTGTG	**C**C	CACAGTGGTAG
27.1.5	GTAGCACTGT		CACAGTGGTAG
43.4.3	GTAGCACTG		CACAGTGGTAG
5.1.7	GTAGCACTG	C**G**	CACAGTGGTAG
63.3.5	GTAGC		CACAGTGGTAG
28.5.3	-59(G)		CACAGTGGTAG

5A) pMut2, pMut27 RECOMBINANTS

	CODING JOINT			SIGNAL JOINT		
	CTGCAGGTCGA		GATCCTCTCAT	GTACCACTGAG		CACAGTGATCC
241	CTGCAGGTCGA		TCCTCTCAT	GTACCACTGAG		CACAGTGATCC
027-6	CTGCAGGTCGA		TCCTCTCAT	GTACCACTGAG	**G**	CACAGTGATCC
027-1	CTGCAGGTCGA	AA	CCTCTCAT	GTACCACTGAG	**T**	CACAGTGATCC
LC20-3	CTGCAGGTCGA		TCCTCTCAT	GTACCACTGAG		CACAGTGATCC

5B) pDSJ RECOMBINANTS

	LEFT JUNCTION			RIGHT JUNCTION		
	GGAGCACTGTG		CACAGTGGTAG	ACACCAGTGTG		CACAGTGCTAC
D1	GGAGCACTGTG		CACAGTGGTAG	ACACCAGT	A	AGTGCTAC
D4	GGAGCACTGTG		CACAGTGGTAG	ACACCAGTGT	CGG	TAC
D5	GGAGCACTGTG		CACAGTGGTAG	ACACCAGTGTG		CTAC
D7	GGAGCACTGTG		CACAGTGGTAG	ACAC		CAGTGCTAC
D22	GGAGCACTGTG		CACAGTGGTAG	ACACCAGTGTG	G	GTGCTAC
D26	GGAGCACTGTG		CACAGTGGTAG	ACACCAGTGTG	**CA**	CACAGTGCTAC
D60	GGAGCACTGTG		CACAGTGGTAG	ACACCAGT		GTGCTAC
D61	GGAGCACTGTG		CACAGTGGTAG	ACACCA	GAACA	CACAGTGCTAC
D66	GGAGCACTGTG		CACAGTGGTAG	ACAC	G	CAGTGCTAC
D68	GGAGCACTGTG	**G**	CACAGTGGTAG	ACACCAG	A	GCTAC
D67	GGAGCACTGTG	ACTT	CACAGTGGTAG	ACAC		CAGTGCTAC
D70	GGAGCACTGTG		CACAGTGGTAG	ACACCAGTGTG	**C**CT	CACAGTGCTAC
D72	GGAGCACTGTG		CACAGTGGTAG	ACACCAGTGT		-11(A)
D78	GGAGCACTGTG		CACAGTGGTAG	ACACC	C	GCTAC
D80	GGAGCACTGTG		CACAGTGGTAG	ACACCAGTGTG	**CAC**TG	C
D82	GGAGCACTGTG	G**G**	CACAGTGGTAG	ACACCAG	A	-11(A)
D84	GGAGCACTGTG		CACAGTGGTAG	ACACCAGTGTG		CAGTGCTAC
D86	GGAGCACTGTG	AGGGC	CACAGTGGTAG	ACAC		CAGTGCTAC
D88	GGAGCACTGTG		CACAGTGGTAG	ACAC		CAGTGCTAC
D65	GGAGC	**TGTG**	CACAGTGGTAG	ACACCAGTGTG	**C**C	CACAGTGCTAC
D89	GGAGC	**G**	CACAGTGGTAG	ACACCAGTGTG	**G**	CACAGTGCTAC
D57	-12(C)		AG	ACAC	T	GTGCTAC

TABLE 1: P Nucleotides in coding joints

	# of trfx's[a]	Cam^r screened[b]	# coding joints[c]	# untrimmed junctions[d]		(seq.)[e]	untrimmed *vs.* total (%)[f]	P insert-containing junctions: *vs.* total (%)[g]	*vs.* untrimmed (%)[h]
A) Summary acc. to construct									
pSal-Bam	5	141	69	Sal	16	(16)	23	16	73
				Bam	36	(33)	52	8	15
pSpe-Bam	9	262	128	Spe	3	(3)	2	[0.8]	[33]
				Bam	34	(34)	27	9	32
pSal-Xho	9	226	111	Sal	25	(19)	23	15	65
				Xho	42	(34)	38	13	35
pSpe-Xho	9	190	93	Spe	8	(8)	4	3	75
				Xho	52	(37)	27	14	53
B) Summary acc. to end									
Sal			180		41	(35)	23		69
Bam			197		70	(67)	35		24
Spe			221		11	(11)	5		64
Xho			204		94	(71)	46		44

a Number of transfections performed per construct

b Number of Cam^r colonies picked.

c Number of standard recombinants analyzed (estimated as described in Materials and Methods).

d Total number of untrimmed coding joints for each of the two ends within a given construct (direct determination).

e Number of untrimmed ends subjected to DNA sequence analysis.

f Truncation index: percentage of untrimmed end among all coding joints.

g Percentage of P insert-containing junctions among all coding joints (calculated on the basis of estimate in *c*).

h Percentage of P insert-containing junctions among untrimmed junctions (as determined directly from the DNA sequence analysis).

Square brackets indicate a sole insert-containing junction was observed.

TABLE 2: Statistical Analysis of P Insert Frequencies

Insert length:		**≥1**	**≥2**	**≥3**	**≥4**	**≥5**
A) SAL		**G**	**GT**	**GTC**	**GTCG**	**GTCGA**
to Bam:	o=	11	6	4	1	0
(pSal-Bam)	N=	12	10	9	2	2
	P<	0.0002	10^{-5}	10^{-4}	0.01	
to Xho:	o=	11	4	2	0	0
(pSal-Xho)	N=	13	7	7	5	4
	P<	0.0008	0.0004	0.007		
to 23-signal:	o=	5	2	0		
(pSal-Bam &	N=	7	4	1		
pSal-Xho)	*P*<	0.08	0.02			
B) BAM		**C**	**CC**	**TCC**	**ATCC**	
to Sal:	o=	5	3	0		
(pSal-Bam)	N=	13	5	2		
	P<	0.4	0.008			
to Spe:	o=	12	5	2	0	
(pSpe-Bam)	N=	18	10	6	3	
	P<	0.002	0.002	0.003		
C) SPE		**A**	**AC**	**ACT**		
to Bam:	o=	1	0			
(pSpe-Bam)	N=	1	1			
	P<	0.2				
to Xho:	o=	6	2	0		
(pSpe-Xho)	N=	8	4	3		
	P<	0.0004	0.02			
to 12-signal:	o=	3	1			
(pSpe-Bam&	N=	4	1			
(pSpe-Xho)	*P*<	0.02	0.05			
D) Xho		**G**	**AG**	**GAG**		
to Sal:	o=	12	2	0		
(pSal-Xho)	N=	20	8	4		
	P<	0.04	0.08			
to Spe:	o=	19	3	0		
pSpe-Xho)	N=	23	10	5		
	P<	10^{-4}	0.02			
E) 23-SIGNAL		**G**	**TG**			
to Sal:	o=	2	0			
(pSal-Bam &	N=	11	4			
pSal-Xho)	*P*<	0.96				
to Spe:	o=	0	0			
(pSpe-Bam&	N=	7	3			
pSal-Xho)	*P*<					
to 12-sig:	o=	5	0			
(all	N=	12	11			
constructs)	*P*<	0.50				
F) 12-SIGNAL		**C**	**CA**			
(to 23-sig)	o=	2	0			
all constructs	N=	9	8			
	P<	0.82				

(continued)

Table 2 (cont.)

G) POOLED DATA

SAL		**G**	**GT**	**GTC**	**GTCG**	**GTCGA**
	o=	27	12	6	1	0
	N=	32	21	16	7	2
	P<	10^{-7}	10^{-9}	10^{-6}	.05	
BAM		**C**	**CC**	**TCC**	**ATCC**	
	o=	17	8	2	0	
	N=	31	15	8	3	
	P<	0.005	10^{-4}	.006		
SPE		**A**	**AC**	**ACT**		
	o=	10	3	0		
	N=	13	6	3		
	P<	10^{-8}	.003			
XHO		**G**	**AG**	**GAG**		
	o=	31	5	0		
	N=	43	18	9		
	P<	10^{-5}	.004			
23-SIGNAL		**G**	**TG**			
	o=	7	0			
	N=	30	18			
	P<	0.97				
12-SIGNAL		**C**	**CA**			
	o=	2	0			
	N=	9	8			
	P<	0.82				

A-F provides a statistical analysis of P insert frequencies for each end, displayed by junction. (E.g. as shown in A, the Sal end was assayed in three different junctions: with the Bam end in pSal-Bam, with the Xho end in pSal-Xho and with the 23-signal in hybrid joints.)
G pools the results for each end as given in parts A-F.

"N" : Number of junctions with inserts, P or otherwise, of length $\leq$ 1,2,3 etc. as indicated by the column headings.
"o": Observed number of untrimmed junctions with **P inserts** of length $\leq$ 1,2,3 etc. The sequence of such inserts is given in each sub-heading. The absence of an entry indicates there were no junctions containing inserts - P or otherwise - of the specified length.
"*P* ": The probability associated with each observed value, calcuated as described in Material and Methods.

TABLE 3: P Nucleotides within endogenously-generated junctions

	# ends[a]	# untrimmed ends[b]	# w/ inserts[c]	# P[d]	P=1[e]	P=2[f]	P=3[g]	P=4[h]	untrimmed total ends (%)[i]	P total ends (%)[j]	P untrimmed (%)[k]
A) IgH locus											
Adult											
3'V	103	26	25	10	3	6	1		25	10	38
5'D	474	43	38	22	10	10	1	1	9	5	51
3'D	584	214 (165)	140	83	62	14	7		37	14	39 (50)
5'J	562	119 (85)	79	44	24	19	1		21	8	37 (52)
Fetal/neonatal											
3'V	139	20	5	3	1	2			14	2	12
5'D	214	11	6	5	5				5	2	45
3'D	225	135 (40)	15	15	10	5			60	7	11 (37)
5'J	308	81 (33)	23	16	8	8			26	5	20 (48)
B) TCR locus											
Adult											
3'V	687	91	80	70	49	17	4		13	10	78
5'D	655	186	144	100	41	45	16		28	15	54
3'D	655	74 (73)	51	28	18	10			11	4	38 (38)
5'J	700	174 (173)	132	78	34	36	8		18	11	45 (45)
Fetal/neonatal											
3'V	387	108	54	49	35	13	1		28	13	45
5'D	365	86	48	40	23	15	2		24	11	47
3'D	330	75 (48)	26	17	12	5			23	5	13 (35)
5'J	415	90 (78)	34	25	9	15	1		22	6	28 (67)

(continued)

(Table 3 continued)

A) Analysis of IgH junctions (1,9,10,12,14 and J. Teale, personal communication). B) Analysis of TCRβ data (2, 4, 11). Scoring is described in Materials and Methods.

[a]: Number of ends analyzed.

[b]: Number of untrimmed ends; the values for 3'D and 5'J given in parentheses discount ambiguous junctions (see text).

[c]: Number of untrimmed junctions with inserts.

[d]: Number of full-length ends with P nucleotides.

[e-h]: A breakdown of the data in the previous column ([d]), according to P insert length.

[i]: Percentage of untrimmed junctions, calculated without the correction (in parentheses) shown in [b].

[j]: Percentage of P inserts among all junctions.

[k]: Percentage of P inserts among untrimmed junctions only: calculations for 3'D and 5'J that discount ambiguous junctions ([b]) are shown in parentheses.