MOLECULAR AND CELLULAR BIOLOGY, Jan. 1996, p. 258–269 Vol. 16, No. 1
0270-7306/96/$04.00+0

Mechanistic Constraints on Diversity in Human V(D)J Recombination

GEORGE H. GAUSS AND MICHAEL R. LIEBER*

Division of Molecular Oncology, Department of Pathology, Washington University School of Medicine, St. Louis, Missouri 63110

Received 2 February 1995/Returned for modification 3 April 1995/Accepted 9 October 1995

We have analyzed a large collection of coding junctions generated in human cells. From this analysis, we infer the following about nucleotide processing at coding joints in human cells. First, the pattern of nucleotide loss from coding ends is influenced by the base composition of the coding end sequences. AT-rich sequences suffer greater loss than do GC-rich sequences. Second, inverted repeats can occur at ends that have undergone nucleolytic processing. Previously, inverted repeats (P nucleotides) have been noted only at coding ends that have not undergone nucleolytic processing, this observation being the basis for a model in which a hairpin intermediate is formed at the coding ends early in the reaction. Here, inverted repeats at processed coding ends were present at approximately twice the number of junctions as P nucleotide additions. Terminal deoxynucleotidyl transferase (TdT) is required for the appearance of the inverted repeats at processed ends (but not full-length coding ends), yet statistical analysis shows that it is virtually impossible for the inverted repeats to be polymerized by TdT. Third, TdT additions are not random. It has long been noted that TdT has a G utilization preference. In addition to the G preference, we find that TdT adds strings of purines or strings of pyrimidines at a highly significant frequency. This tendency suggests that nucleotide-stacking interactions affect TdT polymerization. All three of these features place constraints on the extent of junctional diversity in human V(D)J recombination.

The exons encoding the immunoglobulin and T-cell receptor variable domains are assembled during lymphocyte development by V(D)J recombination (34). In this reaction, two site-specific cuts are made to create four DNA ends, two of which (the coding ends) are joined to generate the exon that encodes the variable domain of the antigen receptor (for reviews, see references 21 and 23). The two signal ends are joined to form a signal joint. Nucleotide loss and addition at the coding ends contribute greatly to the generation of diversity in the immune system repertoire. Nucleotide addition is the result of terminal deoxynucleotidyl transferase (TdT), a DNA polymerase expressed in pre-B and pre-T cells (11, 15, 17). Apart from the preference of TdT to add G nucleotides (1), it has been assumed that TdT polymerizes nucleotides to 3′ termini of DNA in an essentially random fashion (18).

Junctional nucleotides are added sometimes even when TdT is not present, and the source of these nucleotides was unclear until evidence for a second source of junctional nucleotides emerged. This evidence came from chicken (25) and murine (20) systems. Among the coding ends that do not undergo nucleotide loss, there is a pattern of palindromic nucleotide addition. These added nucleotides are termed P (for palindromic) nucleotides. For example, consider a case of a V segment in which the coding region directly adjacent to the recombination signal had the sequence (V)CT-heptamer. The coding end with P insertion would have the sequence (V)CTag, where the added P nucleotides are in lowercase letters. If P nucleotides were added to both V and J coding ends, the complete VJ coding junction might be (V)CTagcaTG (J). The "ag" and the "ca" are the P nucleotide additions to the V and J coding ends, respectively. The initial model for generation of P insertions proposed that the last two nucleotides from one strand were transferred to the other (antiparallel) strand of the same coding end in a two-step mechanism that involved a single-strand endonuclease and a ligation activity (20).

An alternative model for the generation of the inverted repeats at the coding ends involves the formation of a hairpin structure early in the reaction (23). Hairpin formation at the V, D, or J segment coding ends would occur after the recombinase has bound at the heptamer/nonamer signal sequence and made an endonucleolytic cut at the heptamer/coding-end junction. If the endonuclease cuts in the center of the hairpin loop, a blunt end results. If it cuts anywhere else along the hairpin, it generates a 3′ or 5′ overhang with a short inverted repeat (P nucleotides), which could be incorporated into the final coding junction. Variation in the endonucleolytic cut around the hairpin terminus would contribute to the diversity of recombination products.

Roth et al. (30) initially provided evidence for such a hairpin mechanism by demonstrating that some form of covalent linkage seals the coding ends in murine scid cells. More recently, coding-end hairpins have been formed in vitro with pre-B cell nuclear extracts supplemented with recombinant RAG-1 protein (35). It remains unresolved why coding junctions form inefficiently in scid cells. Insertions similar to P nucleotides can occur at recircularization junctions when linear hairpin substrates are transfected into scid cell lines (22), suggesting that scid cells are able to resolve hairpin structures. In addition, scid and wild-type cell lines appear to integrate hairpinned and nonhairpinned DNA at similar ratios (32).

It is critical, for the purposes of this study, to note that inverted repeats (P nucleotides) have been documented at statistically significant frequencies only for coding ends which have not incurred nucleotide loss (20, 25, 26). Because of this, the definition of P nucleotides includes only inverted repeats at

* Corresponding author. Phone: (314) 362-4227. Fax: (314) 362-8756.

full-length coding ends. Studies of hairpins isolated from scid thymocytes indicates that the hairpin is formed at the full-length coding end (39). The documentation of P nucleotides only at full-length coding ends has suggested that the hairpinning process can occur only in the initial steps of the recombination reaction.

Using extrachromosomal V(D)J recombination substrates transfected into human cell lines, we were able to generate large numbers of immunologically unselected coding junctions. Because these junctions are unselected, their analysis yields a more accurate interpretation of the biochemical events which occur at coding ends during V(D)J recombination. Our detailed analysis of coding junctions in human cells yields deeper insights into coding-end processing.

MATERIALS AND METHODS

V(D)J recombination substrates and expression vectors. The coding junctions used in this study were generated by the extrachromosomal V(D)J recombination assay as described previously (14, 24). The five V(D)J recombination substrates used in this study are identical or similar to those described previously. pGG50, pGG51, pGG52, and pML520 are identical to pCJ1, pCJ, pINV, and pINV1, respectively (9). pGG80 is similar to pGG50, except that the 12- and 23-coding-end sequences were changed. The terms 12-end and 23-end refer to the coding ends that are adjacent to the 12- and 12-bp spacer recombination signal sequence on the imput substrate. First, the 66-bp *Bam*HI fragment from pRG17 (containing a different 23-coding-end sequence) was inserted into pGG50, resulting in the cloning intermediate pGG76. Then, the 52-bp *Sal*I fragment from pRG44 (containing a different 12-coding end) was inserted into pGG76, generating the substrate pGG80. Construction of pRG17 and pRG44 has been described previously (10).

All RAG-1, RAG-2, and TdT expression vectors were made by cloning the coding region of these genes into the multipurpose cloning site on pCDM8 (Invitrogen). The wild-type human TdT expression vector, pGG82, and the human mutant TdT expression vector, pGG90, were made starting with baculovirus TdT expression constructs obtained from David Sorscher and Mary Sue Coleman. pGG82 was cloned by removing the wild-type TdT coding region from pAcC4TdT by digestion with *Nco*I and *Bam*HI, blunting with Klenow fragment, and inserting the resulting 1.6-kbp fragment into the Klenow fragment-blunted *Xba*I site of pCDM8ΔXhoI. pGG90 was cloned identically, except that the mutant TdT coding region from pAcC4TdT-D343E was cloned into pCDM8ΔXhoI. The stuffer region of pCDM8 (*Xho*I to *Xho*I) was removed to generate pCDM8Δ XhoI. For biochemical analysis of the D343E mutant, see reference 38.

Three human RAG-1 expression vectors, pCDM8RAG1, pbC-D, and pGG7, were used in this study. pCDM8RAG1 expresses wild-type human RAG-1 protein, and pbC-D expresses the polymorphic variant of human RAG-1 protein with a histidine instead of the nonconserved arginine at residue 249. Both of these vectors have wild-type activity and were provided by Klaus Schwarz (31a). To generate pGG7, the wild-type human RAG-1 coding region was removed from pH36-BSK− (obtained from David G. Schatz) by digestion with *Sma*I and *Sal*I, blunted with Klenow fragment, and inserted into the *Hin*dIII site of pCDM8ΔXhoI after it had been blunted with Klenow enzyme.

Three human RAG-2 expression vectors, pCDM8RAG2, pbC-A, and pGG64, were used in this study. pCDM8RAG2 expresses wild-type human RAG-2, and pbC-A expresses a polymorphic variant of human RAG-2 with an isoleucine instead of the nonconserved valine at residue 8. Both of these vectors have wild-type activity and were provided by Klaus Schwarz (31a). pGG64 was made by first isolating a clone containing the RAG-2 gene by probing a human genomic library (Clonetech HL 1067J) with a 0.6-kbp fragment of human RAG-2 (provided by Laurence A. Turka) (14a). Next, hybridization-positive lambda clones were subcloned by digestion with *Pst*I and *Dra*I and the resulting 1.6-kbp fragment was inserted into pBluescript (Stratagene) digested with *Pst*I and *Eco*RV. This subclone was then digested with *Bam*HI and *Hin*dIII, blunted with Klenow fragment, and inserted into pCDM8-K at the blunted *Xho*I site, to generate pGG64.

pCDM8-K contains a consensus Kozak translational start sequence and alters the predicted amino acid sequence of the wild-type RAG-2 protein by deleting the first two N-terminal residues (MS) and replacing them with seven other residues (MDSRSPG). In many experiments with pGG64, we observed no differences between it and wild-type expression vectors when using extrachromosomal substrates. Other studies confirm that alteration of the RAG-2 translational start sequence does not alter V(D)J recombination in murine cell lines (30a, 31b). Indeed, we initially observed homopolymer tracts, inverted repeats, and sequence effects on nucleotide loss in the coding junctions from Reh and Nalm-6 generated with the endogenously expressed RAG and TdT proteins. In Fig. 3, coding junctions 1 to 7, 38 to 44, and 48 to 61 were generated with pGG7/pGG64, 8 to 16 was generated with pCDM8RAG1/pbC-A, 17 to 27 was generated with pbCD/pCDM8RAG2, 28 to 37 was generated with pCDM8RAG1/pCDM8RAG2, and 45 to 47 was generated with pM2CD/pR2RCD2.

The murine RAG-1 and RAG-2 expression vectors (pM2CD and pR2RCD-2, respectively) were obtained from David G. Schatz. The murine TdT expression vector, pTdt, was obtained from François Rougeon.

Cell lines and transfections. Reh and Nalm-6 are human acute lymphoblastic leukemia cell lines which spontaneously recombine their T-cell receptor δ/α locus (13) and transfected V(D)J substrates (9). Reh and Nalm-6 were obtained from the Fuji Saki Cell Center. Daudi, a B-lymphoblast cell line, was obtained from the American Type Culture Collection. PW, an Epstein-Barr virus-immortalized normal human peripheral blood lymphocyte line, was obtained from Alan M. Krensky. 607B, a subclone of GM00607B, an Epstein-Barr virus-transformed human B-lymphoblast cell line, was obtained from Michael L. Cleary. 293-C18, a subclone of the cell line 293, was obtained from Michelle P. Calos and maintains an integrated copy of the EBNA-1 gene.

Reh, Nalm-6, Daudi, PW, and 607B were cultured in RPMI 1640 supplemented with 10% fetal bovine serum, 100 U of penicillin G per ml, 100 U of streptomycin sulfate per ml, and 50 μM 2-mercaptoethanol. 293-C18 cells were cultured in Dulbecco's minimal essential medium with 10% fetal bovine serum, 100 U of penicillin G per ml, and 100 U of streptomycin sulfate per ml. All cell lines were grown at 37°C in a humidified 5% CO_2 incubator.

Reh, Nalm-6, Daudi, PW, and 607B were transfected by the DEAE-dextran electroporation method (8). 293-C18 cells were transfected by calcium phosphate-DNA coprecipitation (31). For 293-C18 transfections, a total of 5 μg of DNA was used for each 100-mm-diameter dish of cells. When RAG expression vectors were used, a 1:3:3 molar ratio of V(D)J substrate, RAG-1, and RAG-2 expression vectors was transfected. When TdT expression vectors were used in cotransfections, a 1:3:3:3 molar ratio of substrate, RAG-1, RAG-2, and TdT vectors was used.

Analysis of junctions. Coding junctions were sequenced with Sequenase version 2.0 as described by the manufacturer (U.S. Biochemical). If two coding junctions with identical sequence were isolated from the same transfection, only one of these junctions was reported. A systematic set of rules was used in scoring the nontemplated junctional additions as either N, P, P_r, or non-TdT N nucleotides. This prevents arbitrary and double assignment of junctional nucleotides into the four classes of junctional addition. All junctional sequences were arranged in the same orientation, with the 23- and the 12-coding ends on the left and right, respectively. The junctional sequences were then compared with the starting substrate for assignment of each nucleotide.

First, the coding-end termini were assigned in each junction. The longest contiguous sequence with perfect homology to the 23- or 12-end of starting substrate was taken as the terminus of the 23- or 12-end for that junction. The sequences of the two termini were then compared. In 50 junctions, the assigned sequences of the 23- and 12-ends overlapped, and these overlapping nucleotides were assigned as junctional microhomology nucleotides. These nucleotides cannot be unequivocally assigned to either coding end and are indicated by boldface italic type (see, e.g., Fig. 1, junction 8).

Second, the junctional additions were scored as N, P, or P_r nucleotides in a two-step process. Initially, all junctional nucleotides were scored as N additions. Then, palindromic N additions were reassigned as either P or P_r nucleotides depending on whether they occurred at a full-length or nucleolytically processed coding end. In cases when the one nucleotide could be scored as part of either a P or P_r, it was scored as a P nucleotide. This occurred in one junction (see Fig. 1, junction 30). In junctions where two mutually exclusive overlapping palindromes were noted, the longer palindrome was used. This occurred in one junction (see Fig. 4A, junction 5).

Statistics. Estimations and tests of statistical significance were done by two different methods. Monte Carlo simulations were performed with a computer program. Monte Carlo simulations are used to model presumed random processes, such as TdT addition (18). Essentially, a Monte Carlo simulation estimates the number of events (inverted repeats or homopolymers) that could be expected from TdT addition by performing the following three steps. (i) The sequences of the coding junctions are saved in the computer. (ii) The sequence of the TdT additions are randomized, and the resulting number of events is counted. (iii) The process of randomization and counting is reiterated 1,000 times. The estimate is computed by averaging the total number of events from all iterations. The probability (P) of the observed number of events in the original data was taken as the quotient N/I, where N is the number of iterations wherein the number of events was greater than or equal to the corresponding number seen in the original data, and I is the total number of iterations.

In our simulations, the randomizing process was made to faithfully reproduce two independent characteristics of the original sequence data. First, the high G+C content of the actual TdT additions was reproduced in the simulations by maintaining the overall A/C/G/T ratios that were present in the original sequence data (G utilization model). Second, base stacking was reproduced in the simulations by maintaining the dinucleotide frequencies of the original data (see Table 1). When both the high G+C content and the stacking were modeled, we abbreviate this as the G+stacking model. We calculated the A/C/G/T ratios excluding P nucleotides. Computer program output is available upon request.

The minimum P value calculated by the Monte Carlo method is limited to the inverse of the number of iterations run by the program. Thus, for 1,000 iterations,

the minimum P value is 0.001. More than 1,000 iterations were not typically performed because of the long computer time required. Instead, a second statistical method, which does not have this limitation, was used.

The second statistical method uses the binomial distribution as described by Meier and Lewis (26) in their analysis of P nucleotides. By summing the terms of the binomial distribution that correspond to a given number of events (either homopolymers or inverted repeats), the P value of observing that given number of events can be calculated, using the equation

$$\sum_{i=n}^{N} \binom{n}{i} (p)^i (1 - p)^{(n-i)}$$

where n is the number of events observed in the data, N is the number of times the event could have occurred (the total number of dinucleotides for RR or YY events, or the number of recessed coding ends with TdT additions for inverted repeats), and p is the Bernoulli probability of an individual event. Thus, we obtained P values representing the statistical significance of P_r additions or homopolymer tracts of a specific number of nucleotides in length.

For P_r additions, we calculated the Bernoulli trial probability, p, by using the independence model for finding common words and patterns between two sets of sequences (16). In this application of the independence model, the sequence of the coding ends represents one set of sequences and the junctional addition (excluding P nucleotides) represents the other set. For P_r additions of length 1 nucleotide (nt), we compiled a list of the frequencies of all recessed terminal nucleotides observed in all junctions which contained an addition at least 1 nt long (189 coding ends). The probability of the corresponding P_r addition for each distinct recessed coding end was then calculated from the observed frequencies of A, C, G, and T seen in all our junctions (excluding P nucleotides, A = 0.240, C = 0.245, G = 0.348, and T = 0.167). The product of each coding-end frequency and its corresponding P_r addition probability was calculated. The sum of these products is, p, the probability for each Bernoulli trial.

Similar probabilities were calculated for P_r inserts of 2, 3, and 4 nt, except that only recessed coding ends which showed addition of at least 2, 3 and 4 nt were included in the calculations (166, 132, and 91 coding ends, respectively). A breakdown of coding-end frequencies and our calculations are available on request.

For homopolymer analysis, the Bernoulli probabilities for the G utilization model assume that each nucleotide addition is independent of the previous nucleotide. Thus, given that G occurs with an overall frequency of 0.348, GGG should occur with a frequency of $0.348 \times 0.348 \times 0.348 = 0.042$. The G+stacking model Bernoulli probabilities for homopolymers of various lengths were calculated by retabulating the observed dinucleotide frequencies as second-nucleotide-addition probabilities. For example, in all dinucleotides beginning with G ($n = 103$), there is a 0.534 probability that a second G will follow the first G (55 GG dinucleotides divided by 103 GN dinucleotides = 0.534). It follows that the Bernoulli probability for the GGG trinucleotide is then $0.348 \times 0.534 \times 0.534 = 0.099$. The initial 0.348 is the overall G nucleotide frequency of the N regions, and the probability that two G nucleotides will follow is $0.534 \times 0.534 = 0.285$. The probabilities for three purines in a row is the sum of the eight individual tripurine probabilities.

Pairwise comparison of the pattern of nucleotide loss between the different coding-end sequences was performed by two-sample Kolmogorov-Smirnov tests with the Systat 5.2 statistical software package. No significant differences between substrates were found when the coding ends with identical sequence were compared, indicating that the pattern of loss depends on the sequence of the coding end and not the recombination substrate. For example, pGG50 and pML520 have identical 12-ends, pGG51 and pGG52 have identical 12-ends, pGG50 and pGG51 have identical 23-ends, and pGG52 and pML520 have identical 23-ends. On comparison of these four pairs of coding ends, their respective Kolmogorov-Smirnov P values were 0.30, 0.76, 0.22, and 0.36.

RESULTS

To gain a better understanding of the biochemical mechanisms of nucleotide processing at human coding junctions, we wanted to examine coding junctions which were free from any nucleotide sequence biases which might be incurred through immunologic selection or might result from PCR amplification preferences. Coding junctions generated with extrachromosomal V(D)J recombination substrates are free from such biases. The human V(D)J recombination substrates used in this study are similar to those described previously (9) and to those used in murine cell lines (14, 24). The substrates bear V(D)J recombination signal sequences and undergo V(D)J recombination upon transfection into pre-B-lymphoid cell lines. The substrates are subsequently recovered from the cells and transformed into *Escherichia coli*, permitting direct molecular cloning of the coding junctions, thus enabling analysis of all V(D)J recombinant products free from selective biases.

Nucleotide addition at coding junctions. Several V(D)J recombination substrates were used to generate coding junctions in six different human cell lines. Two pre-B-cell lines, Reh and Nalm-6 (19, 27), express RAG-1, RAG-2, and TdT and will rearrange transfected V(D)J substrates (9). V(D)J recombination was also examined in three lymphoblastoid cell lines, Daudi, PW, and 607B, and one fibroblastoid cell line, 293-C18. RAG-1 and RAG-2 expression vectors were used to activate recombination in the last four lines. All V(D)J substrates that were used are nearly identical except for their respective coding-end sequences, shown at the top of each column of coding junctions (Fig. 1). Nontemplated nucleotide additions to the junctions (shown between the two columns of coding ends) contain the two previously described types of junctional addition (N nucleotide addition and P nucleotide addition).

In the junctions from Reh and Nalm-6 (Fig. 1), N nucleotide addition is observed in 78% (62 of 79) of the junctions and accounts for 74% of the added nucleotides. The majority of these N additions are due to TdT activity (11, 17). N additions were present at 76% (44 of 58) of the junctions recovered from fibroblasts when TdT, RAG-1, and RAG-2 expression vectors were used (Fig. 2; also see Fig. 4A). They were largely but not entirely absent when the TdT expression vector was not used (Fig. 3), and they were absent in experiments in which a polymerase-defective mutant TdT expression vector was used (Fig. 4B). Similarly, rare N additions are observed in $TdT^{-/-}$ mice (11). The origin of the rare N additions occurring in the absence of TdT is unknown. We refer to them as non-TdT N nucleotides.

The second type of junctional addition, P nucleotides, are observed at 15% (30 of 206) of the junctions and account for 10% (45 of 456) of the total junctional nucleotides in this study. P nucleotides are defined as inverted repeats that occur at coding ends that show no nucleotide loss and appear to be the result of the resolution of hairpin structures at full-length coding ends (22, 23, 30).

N additions include frequent homopolymer tracts. Inspection of the junctions reveals that many N regions contain tracts of nucleotide repeats. For example, junction 15 from Reh (Fig. 1) contains a homopolymer tract consisting of five contiguous C nucleotides, and junction 17 contains a tract of four G nucleotides. Similar homopolymer tracts were observed in experiments in which TdT expression vectors supplied the TdT enzyme (Fig. 2 and 4A). Numerous homopolymer tracts consisting of the other bases are also present in the junctions.

TdT has a well-documented bias for addition of dGTP in preference to the other nucleotides (1). We wondered if this G utilization bias could account for the abundance of these homopolymer tracts. To test this hypothesis, we examined the dinucleotide frequencies of all TdT additions (Table 1). On the basis of the overall frequency of G mononucleotides among N additions (excluding P nucleotides, G accounts for 34.8% of N additions), one would expect the dinucleotide GG to occur with a frequency of $0.348 \times 0.348 = 0.121$. Therefore, GG should occur approximately 36 times among the 295 dinucleotides present in all N regions ($295 \times 0.121 = 35.7$). However, GG occurs 55 times in the junctions, which is significantly more often than expected ($P \approx 0.0008$). Analysis of the other dinucleotides (Table 1) also suggested that the TdT G utilization bias would not, by itself, account for the number of homopolymer tracts observed. It is noteworthy that without exception, all purine-purine (RR) and pyrimidine-pyrimidine (YY) dinucleotides occurred more often than expected (Table

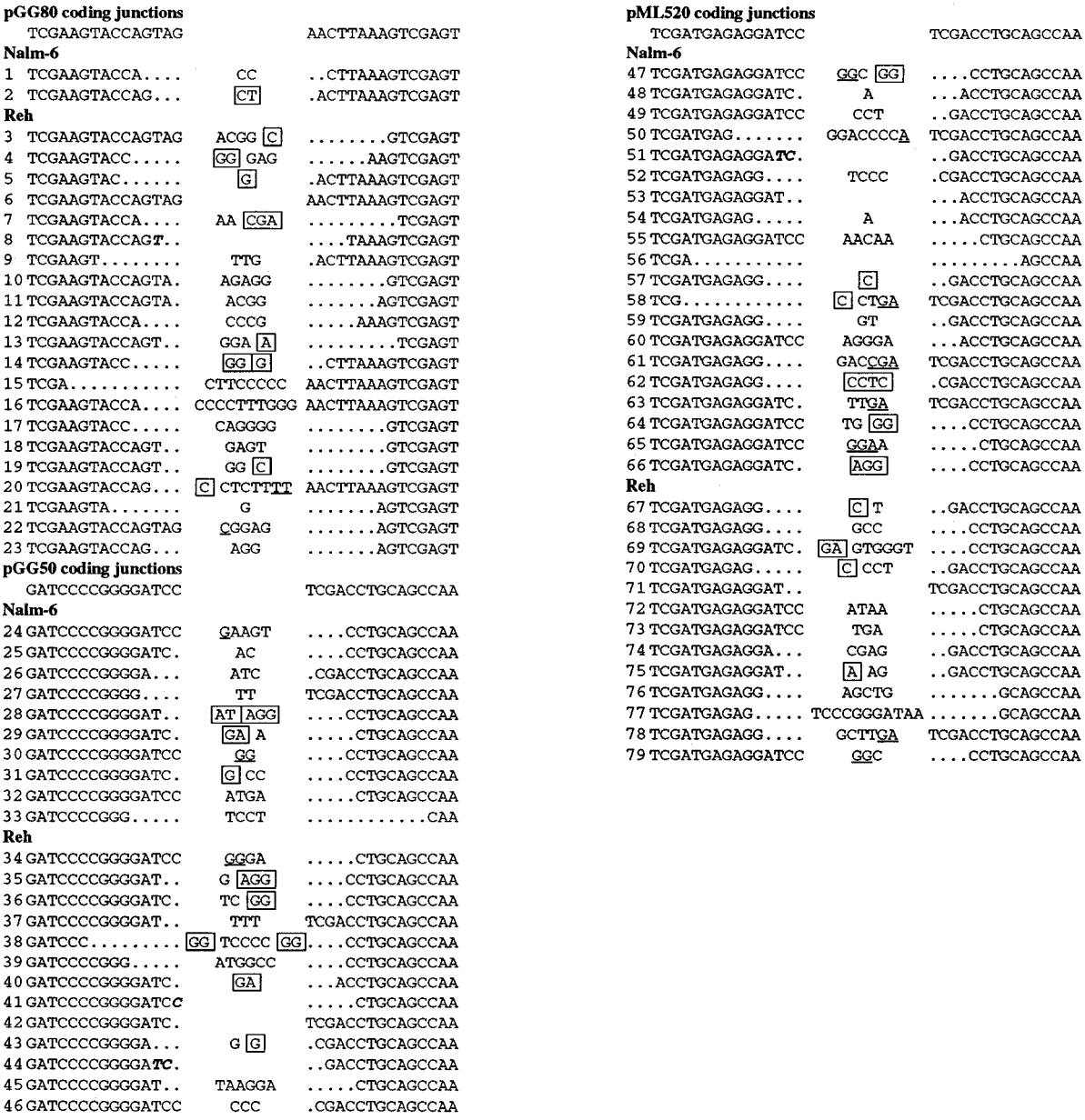


FIG. 1. Coding junctions from human pre-B-lymphoid cell lines. The substrates were transfected into the cell lines indicated (in boldface type on the left), and the recombinant products were subsequently recovered and sequenced. The nucleotide sequence of the coding ends in the starting substrates is shown at the top. The combined sequence of these full-length coding ends would be identical to a hypothetical coding junction that had not undergone any nucleotide loss or any nucleotide addition. Nucleotides lost from the coding ends are indicated by dots, and junctional addition is placed between the two coding ends. P nucleotide additions are underlined, P_r additions are boxed, and nucleotides that cannot be unequivocally assigned to either coding end are noted in boldface italics (microhomology use).

1). Conversely, all purine-pyrimidine (RY) and pyrimidine-purine (YR) dinucleotides occurred less often than expected. This suggested to us that there might be a second TdT bias which results in the many homopurine and homopyrimidine tracts observed. We shall tentatively refer to this second bias as the TdT homopolymer bias.

Biophysical studies indicate that bases in solution or linked as polynucleotides will form stacked secondary structures which are stabilized by hydrophobic and van der Waals interactions (3, 29). These interactions might influence the sequence of TdT additions. For example, TdT may preferentially add a dATP to a DNA substrate which has an A at the 3′ end, leading to frequent AA dinucleotides. This phenomenon could be extended to explain the observed homopolymer tracts. It also predicts more frequent purine tracts and pyrimidine tracts than the TdT G utilization bias would have predicted. This prediction is supported by the overabundance of RR and YY dinucleotides relative to RY and YR in Table 1. To quantitatively examine this prediction, we estimated the number of purine and pyrimidine tracts that could be explained by (i) the G utilization bias alone and (ii) the combination of the G utilization bias and homopolymer bias (designated G+stacking). Using the observed base composition and dinucleotide frequencies of the sequence data, we made estimates from the binomial distribution. The model incorporating both G utilization and homopolymer biases (G+stacking model) is more

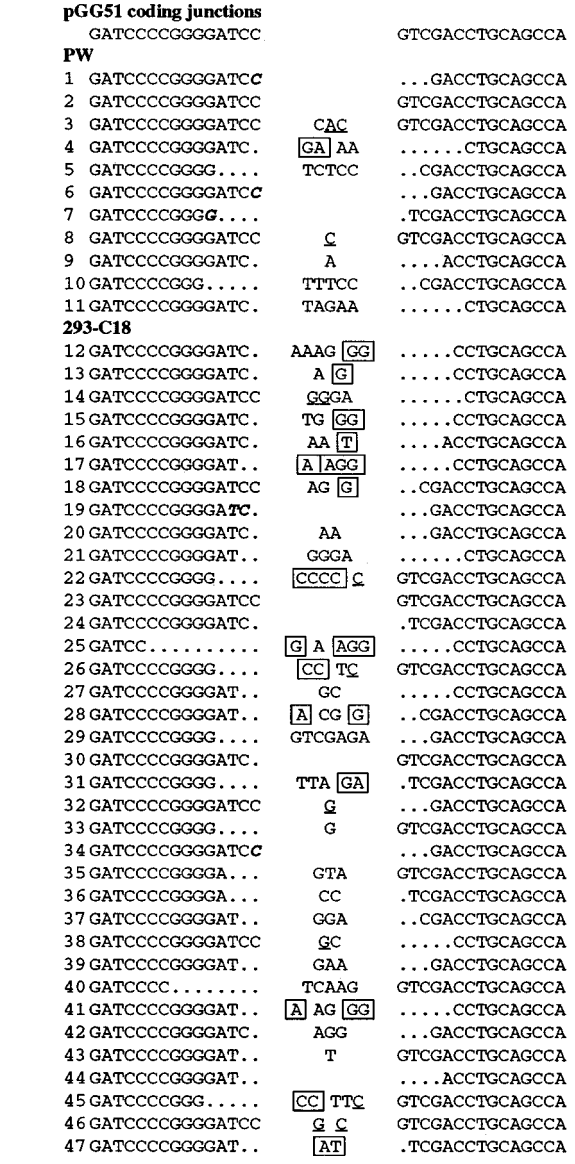


FIG. 2. Junctions using a human TdT expression vector. Coding junctions are presented as described in Fig. 1. P nucleotide additions are underlined, P_r additions are boxed, and microhomology nucleotides are noted in boldface italics. Nucleotides lost from coding ends are indicated by dots. Cell lines were transfected with pGG51 along with wild-type human TdT, RAG-1, and RAG-2 expression vectors to activate V(D)J recombination.

consistent with the numbers of homopolymers in the observed data than the model based on G utilization bias alone.

Figure 5 illustrates that the G+stacking model is more accurate at predicting the number of homopurine and homopyrimidine tracts than the G utilization model alone. Each curve in Fig. 5 represents a probability histogram for the number of tracts that one would expect from either of the two models. The best estimate for each model is given by the peak of the respective curve, and the area under the curve to the right of the observed number of tracts is equal to the *P* value for the observation (number of tracts). The actual number of tracts observed in the junctions lies within the bell-shaped region of the G+stacking curves, indicating that the G+stacking model makes reasonable predictions for number of homopolymers that TdT would be expected to generate. In contrast, the observed number of tracts lies well to the right of the G-alone curves, indicating that the G-alone model always underestimates the actual frequency of tracts in the junctions. Thus, consideration of stacking significantly improves our ability to estimate the frequency of homopolymer tracts and is much more consistent with the frequency of purine and pyrimidine tracts that occur in the data. We infer that stacking may be the explanation for frequent homopolymer tracts in N regions.

We examined published endogenous sequences for purine and pyrimidine tracts and found that they were also more consistent with the G+stacking model (data not shown). This supports our suggestion that N-region additions deviate from random as a result of two factors: (i) G utilization preference and (ii) a propensity to form homopolymer tracts.

Inverted repeats at nucleolytically processed coding termini. In our analysis of the sequence of N regions, we noted a second nonrandom feature of the junctional additions. We noticed that many inverted repeats were present in the N regions at nucleolytically processed coding termini. These inverted repeats are analogous to P nucleotide additions, except that the inverted repeats occur at nucleolytically processed termini whereas P additions are defined as inverted repeats at full-length termini. For the purposes of this study, this new type of inverted-repeat addition is designated P_r inserts (subscript r for recessed or nucleolytically processed coding termini). Thus, P_r inserts are defined as junctional additions which are palindromic to coding ends that have undergone nucleotide loss. In the 79 junctions from pre-B cells (Fig. 1), there are 30 such P_r inserts, ranging from 1 to 4 nt in length.

It could be argued that P_r inserts are in fact TdT additions which are, by chance, palindromic to the coding ends. We wondered if P_r inserts would appear in coding junctions generated in the absence of TdT. To test this hypothesis, we transfected our recombination substrates into four human cell lines which do not express TdT: one fibroblast cell line, 293-C18, and three other lines from hematopoietic lineages, Daudi, 607B, and PW. Human RAG-1 and RAG-2 expression vectors, similar to those used in murine cells (28), were cotransfected with the substrates to activate V(D)J recombination in these human cell lines. Consistent with data from TdT knockout mice (11, 17), some P additions were observed in these junctions (Fig. 3). In contrast, P_r additions were not seen at high frequency in junctions generated without TdT (Fig. 3). When a human TdT expression vector was transfected into these cells in parallel series of experiments, P_r additions were seen in many coding junctions. With the human TdT expression vector (Fig. 2), P_r inserts occur at 26% (19 of 72) of coding ends which have been nucleolytically processed, compared with P inserts, which occur at 45% (10 of 22) of full-length coding ends.

As noted above, two junctions generated in the absence of TdT contained additions that were not P nucleotides (Fig. 3, junctions 17 and 19). This is consistent with other studies, as these rare non-TdT nucleotide additions have been found in both endogenous and extrachromosomal substrate junctions (11, 17, 33). Two of the non-TdT nucleotides in Fig. 3 are inverted repeats of a nucleolyticially processed coding end and therefore could be classified as P_r inserts. However, the number of these inserts is not large enough to be statistically meaningful.

P_r formation is not specific to human TdT, RAG-1, and RAG-2. Given that P_r additions appear to be dependent on TdT expression, it was perplexing why P_r additions are less abundant in murine junctions. We hypothesized that perhaps the human TdT but not the murine TdT is somehow necessary for P_r addition. To test this, we transfected murine TdT, murine RAG-1, and murine RAG-2 expression vectors along with

pGG52 coding junctions

TCGATGAGAGGATCC		GTCGACCTGCAGCCA
293-C18		
1 TCGATGAGAGGATC.		GTCGACCTGCAGCCA
2 TCGATGAGAGGATCC		..CGACCTGCAGCCA
3 TCGATGAGAGGA***TC***.		...GACCTGCAGCCA
pGG51 coding junctions		
GATCCCCGGGGATCC		GTCGACCTGCAGCCA
293-C18		
4 GATCCCCGGGGA***TC***.		...GACCTGCAGCCA
5 GATCCCCGGGGA...	C	GTCGACCTGCAGCCA
6 GATCCCCGGGGAT***CC***		.......TGCAGCCA
7 GATCCCCGGGGA***TC***.		...GACCTGCAGCCA
8 GATCCCC***CG***.....		....ACCTGCAGCCA
9 GATCCCCGGGGAT***C***.		.......TGCAGCCA
10 GATCCCCGGGGATC***C***		...GACCTGCAGCCA
11 GATCCCCGGGGAT***CC***		.......TGCAGCCA
12 GATCCCCGGG***G***....		.TCGACCTGCAGCCA
13 GATCCCCGGGG....		..CGACCTGCAGCCA
14 GATCCCCGGGGA***TC***.		...GACCTGCAGCCA
15 GATCCCCGGGGATC.		....ACCTGCAGCCA
16 GATCCCCGGG***GA***...		.....CCTGCAGCCA
17 GATCCCCGGGGAT..	AA	......CTGCAGCCA
18 GATCCCCGGGGATC***C***		...GACCTGCAGCCA
19 GATCCCCGGGGATC.	G	.....CCTGCAGCCA
20 GATCCCCGGGGATC***C***		......CTGCAGCCA
21 GATCCCCG***G***.....		....ACCTGCAGCCA
22 GATCCCCGGGGAT..		.......TGCAGCCA
23 GATCCCCGGGGAT..		....ACCTGCAGCCA
24 GATCCCCGGGGATC.		....ACCTGCAGCCA
25 GATCCCCGGG***GA***...		.....CCTGCAGCCA
26 GATCCCCGGGGAT..		...GACCTGCAGCCA
27 GATCCCCGGGGATC.		GTCGACCTGCAGCCA
28 GATCCCCGGGGA...	AC	GTCGACCTGCAGCCA
29 GATCCCCGGGGATCC		......CTGCAGCCA
30 GATCCCCGGGGATCC		....ACCTGCAGCCA
31 GATCCCCGGG***GA***...		.....CCTGCAGCCA
32 GATCCCCGGGG***G***....		.TCGACCTGCAGCCA
33 GATCCCCGGGGA***TC***.		...GACCTGCAGCCA
34 GATCCCCGGGGATC***C***		...GACCTGCAGCCA
35 GATCCCCGGGGATCC		.....CCTGCAGCCA
36 GATCCCCGGGGA...	C	GTCGACCTGCAGCCA
37 GATCCCCGGGGA***TC***.		...GACCTGCAGCCA
38 GATCCCCGGGGAT..		.......TGCAGCCA
39 GATCCCCGGGGATC.		GTCGACCTGCAGCCA
40 GATCCCCGGGGATC***C***		...GACCTGCAGCCA
41 GATCCCCGGG***GA***...		.....CCTGCAGCCA
42 GATCCCCGGGGAT..		....ACCTGCAGCCA
43 GATCCCCGGGGATC.		....ACCTGCAGCCA
44 GATCCCCGGGGATC.		.....AGCCA
45 GATCCCCGGGGATC***C***		...GACCTGCAGCCA
46 GATCCCCGGGGAT***CC***		.......TGCAGCCA
47 GATCCCCGGG***GA***...		.....CCTGCAGCCA
pGG51 coding junctions		
GATCCCCGGGGATCC		GTCGACCTGCAGCCA
607B		
48 GATCCCCGGGGAT***CC***		.......TGCAGCCA
49 GATCCCCGGGGATC.		....ACCTGCAGCCA
50 GATCCCCGGGGA***TC***.		...GACCTGCAGCCA
51 GATCCCCGGG***GA***...		.....CCTGCAGCCA
52 GATCCCCGGGGATC.		....ACCTGCAGCCA
53 GATCCCCGGGGAT***CC***		.......TGCAGCCA
54 GATCCCCGGGGA***TC***.		...GACCTGCAGCCA
55 GATCCCCGGG***GA***...		.....CCTGCAGCCA
PW		
56 GATCCCCGGGGA...	C	GTCGACCTGCAGCCA
57 GATCCCCGGGGATCC		GTCGACCTGCAGCCA
58 GATCCCCGGGGA...		...GACCTGCAGCCA
Daudi		
59 GATCCCCGGG***GA***...		.....CCTGCAGCCA
60 GATCCCCGGGGAT***CC***		.......TGCAGCCA
61 GATCCCCGGG***GA***...		.....CCTGCAGCCA

FIG. 3. Coding junctions from cell lines with no TdT expression. Coding junctions are presented as described in the legend to Fig. 1. Microhomology nucleotides are noted in boldface italics. These nucleotides indicate short regions of homology between the two coding ends that are at the junction of the two coding ends. Cell lines were transfected with V(D)J substrates pGG51 or pGG52 along with human RAG-1 and RAG-2 expression vectors (junctions 1 to 44 and 48 to 61) or murine RAG-1 and RAG-2 (junctions 45 to 47) to activate V(D)J recombination.

our V(D)J substrate into human cell lines. We found that murine TdT from an expression vector is sufficient for P_r addition in human cell lines (Fig. 4A). Junctions generated with murine RAG-1, RAG-2, and TdT expression vectors were similar to those obtained with human RAG-1, RAG-2, and TdT expression vectors (compare Fig. 2 and 4A). Thus, generation of P_r inserts does not appear to be a property limited to human TdT.

Can TdT account for the high incidence of P_r additions? The fact that P_r additions were seen at high frequency only when TdT was present would seem to suggest that they are indeed fortuitous TdT additions. We estimated how often TdT would generate these fortuitous additions based on the G+ stacking model for TdT addition. The G+stacking model takes into account the tendency of TdT to form homopolymer tracts in addition to its tendency to preferentially add G nucleotides. Two independent estimates were made from the G+stacking model; the first estimate used the binomial distribution, and the second used Monte Carlo simulations of the TdT additions. The two methods produced very similar estimates (Fig. 6). Both predicted that TdT would generate approximately 34 inverted repeats 1 nt in length. The methods were in close agreement at other lengths as well, predicting 9, 1, and 0 inverted repeats of 2, 3, and 4 nt in length, respectively. The actual number of inverted repeats in the junctions is 55, 33, 11, and 2 for 1-, 2-, 3-, and 4-nt repeats, respectively. The probability that TdT would generate this number of inverted repeats on the basis of the G+stacking model is extremely low ($P < 0.001$, $P < 0.001$, $P < 0.001$, and $P = 0.026$ for inverted repeats 1, 2, 3, and 4 nt in length, respectively). The low P values imply that the G utilization bias in combination with the homopolymer bias for TdT addition cannot account for the high frequency of inverted repeats.

This is not entirely unexpected, because inverted repeats must include at least one RY or YR dinucleotide. The central dinucleotide of all inverted repeats is either a RY or YR, and these mixed dinucleotides are underrepresented in the junctional additions (Table 1). One possibility for the inverted repeats is that the stacking interactions are perturbed by specific sequences at certain DNA 3′ termini. Such perturbations might allow more frequent formation of RY or YR dinucleotides at these termini, resulting in more frequent inverted repeats at these termini (see Discussion). However, barring complex stacking perturbations, neither the G-alone nor the G+stacking model can explain the high incidence of inverted repeats in the junctions.

A mutant human TdT is not sufficient for P_r additions. Given that both the murine and human TdT proteins are sufficient for P_r additions, one hypothesis we considered is that perhaps only the presence of TdT, but not necessarily the process of DNA polymerization, is necessary for P_r formation. We constructed an expression vector that produced a mutant TdT that has no detectable polymerization activity in vitro but still has affinity to bind DNA (38). The mutant TdT protein has a single point mutation in the active site (D343E); except for this change, this mutant and the wild-type TdT expression vector are identical. No P_r addition was seen in experiments with the mutant TdT (Fig. 4B), and the junctions from this experiment were similar to others in which TdT was absent (compare Fig. 2 and 4B). Therefore, DNA synthesis by TdT appears to be required for P_r formation.

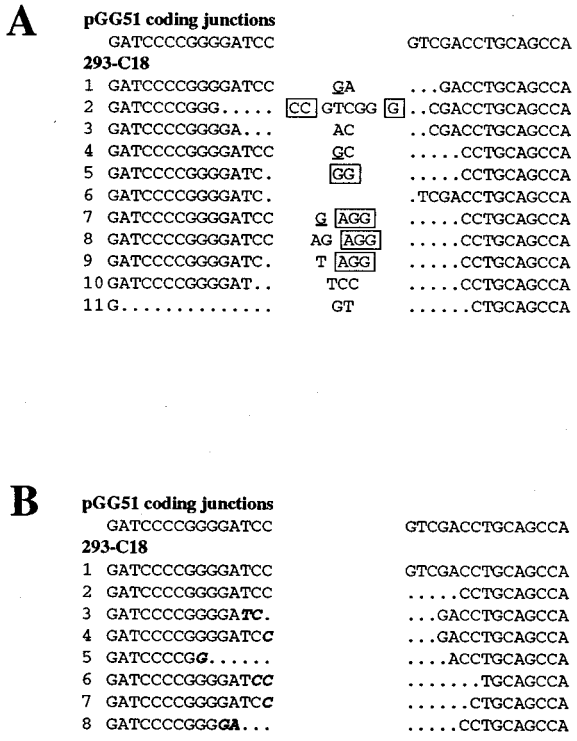


FIG. 4. Junctions using murine TdT and a mutant human TdT. (A) pGG51 coding junctions from 293-C18 cells transfected with murine TdT, murine RAG-1, and murine RAG-2 expression vectors. (B) Junctions from 293-C18 cells transfected with a polymerase-defective point mutant of human TdT (D343E), human RAG-1, and human RAG-2 expression vectors. Junctions are presented as described in the legends to Fig. 1 to 3.

Microhomology use in human coding junctions. Microhomology use is observed at endogenous coding junctions from neonatal mice when TdT is not present at high levels (5, 12). Notably, 57% (35 of 61) of the junctions in Fig. 3 show 1 or 2 nt of homology at the junction between the two coding ends. The utilization of these short blocks of homology extends this feature of the V(D)J recombination reaction to human cells.

The extent of nucleotide loss is affected by coding-end sequence. In previous studies, it has been proposed that coding ends with different nucleotide sequences consistently showed different amounts of nucleotide loss (2, 6). We have tested this hypothesis by comparing the different coding-end sequences among our recombination substrates. AT-rich sequences appear to be subject to more nucleolytic processing than do GC-rich sequences. In Fig. 1, for example, the AT-rich 12-end (right coding end) of pGG80 generally appears to have more nucleotide loss than the other coding ends. Of all the junctions with this coding end, 57% (13 of 23) show nucleotide loss of almost all of the AT-rich portion of the sequence. In contrast, nucleotide loss appears to be limited with GC-rich coding ends. For example, nucleotide loss from the 23-end (left coding end) of pGG50 and pGG51 appears to be limited predominantly to the first 4 nt from the terminus and seldom extends far into the GC-rich portion of the coding end.

Another example of what appears to be the effect of sequence on nucleotide loss is most clearly illustrated in Fig. 4A. Of the 11 junctions in Fig. 4A, 7 show loss of 5 to 6 nt from the 12-end, leaving a coding end with the terminal sequence 5′-CCTG or 5′-CTG. Except for pGG80, all of the other substrates used in this study use the same 12-end or one almost identical to it: pGG51 and pGG52 terminate with 5′-GTCGA, whereas pGG50 and pML520 have the same sequence without the 5′-terminal G nucleotide. Of the 184 junctions sequenced with these 12-ends, 58 (32%) show similar nucleotide loss, leaving the same terminal sequence: 5′-CCTG or 5′-CTG. The large fraction of junctions with similar nucleotide loss suggests that the end point of nucleotide loss is influenced by the sequence of the coding ends.

Statistical analysis of our junctions adds quantitative support to the hypothesis that nucleotide sequence affects the pattern of nucleotide loss from coding ends. Other studies have shown that there is no consistent difference between nucleolytic processing of 12- and 23-coding ends with the same sequence (1). To control for the possibility that the different cell lines we used express different levels of nucleolytic activity, we did a pairwise comparison of the different coding ends by using only junctions generated in Reh. When the different coding-end sequences were compared by the two-sample Kolmogorov-Smirnov test, they showed statistically significant differences in the pattern of nucleotide loss. In Reh, nucleotide loss from the AT-rich 12-end of pGG80 was significantly different from that from the 12-end of pGG50 and pML520 ($P \approx 0.005$), the 23-end of pGG52 and pML520 ($P \approx 0.005$), and the 23-end of pGG52 and pML520 ($P \approx 0.02$). No significant difference was found between the 12-end of pGG50 and the 12-end of pML520 (the sequence of the 12-ends on these substrates is identical), suggesting that the pattern of loss depends on the sequence of the coding end and not the recombination substrate.

Next, we did a comparison after pooling the sequence data from all the cell lines (Fig. 7A). Again, significant differences

TABLE 1. N-region dinucleotide frequencies

Dinucleotide	No. observed (obs)	No. expected (exp)[a]	obs/exp ratio	P[b]
AA	22	17.0	1.3	0.14
AC	9	17.3	0.5	0.99§
AG	32	24.6	1.3	0.08
AT	8	11.8	0.7	0.91
CA	3	17.3	0.2	0.999997§
CC	42	17.7	2.4	2×10^{-7}§§
CG	13	25.2	0.5	0.998§
CT	15	12.1	1.2	0.23
GA	30	24.6	1.2	0.16
GC	8	25.2	0.3	0.999998§
GG	55	35.7	1.5	0.0008§§
GT	10	17.1	0.6	0.98§
TA	8	11.8	0.7	0.91
TC	17	12.1	1.4	0.10
TG	9	17.1	0.5	0.99§
TT	14	8.2	1.7	0.04§§

[a] Expected frequencies were calculated from the overall base composition of all N regions (A = 0.240, C = 0.245, G = 0.348, T = 0.167) and the total number of dinucleotides in the junctions (n = 295). P nucleotides and non-TdT N additions were not included in the calculations of the N region base composition or in the tabulation of the observed dinucleotide frequencies.

[b] P values (probabilities) were calculated from the binomial distribution, using the expected frequencies and the total number of dinucleotides. §, Significantly underrepresented with respect to the expected number; §§, significantly overrepresented. Note that all eight RR and YY frequencies are above the expected number, even though only three are significantly so. All eight RY and YR frequencies are below the expected number, and six of these reductions are significant.

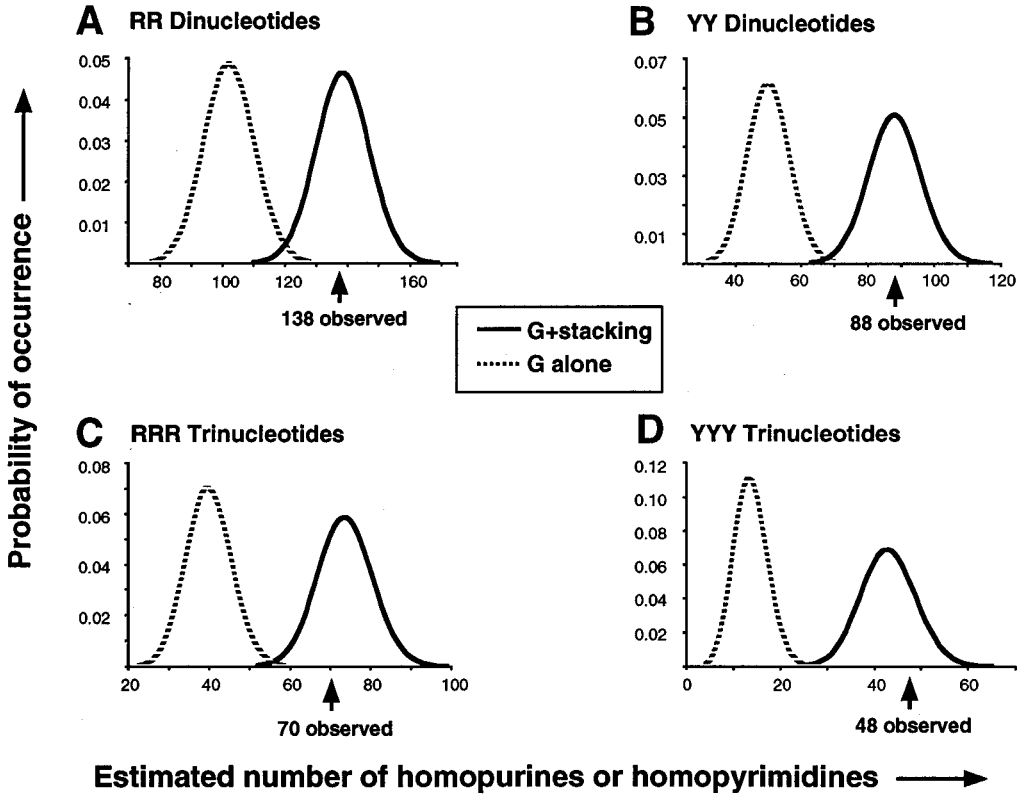


FIG. 5. Number of homopurine and homopyrimidine tracts expected from TdT addition. (A) Purine-purine (RR) dinucleotides. Curves represent theoretical probability histograms for the number of RR dinucleotides that TdT would be expected to generate within the combined junctions in Fig. 1, 2, and 4A. The peak of each curve is centered at the best estimate for the number of RR dinucleotides that TdT would be expected to generate by using either the G bias model (broken curve) or the G+stacking model (solid curve). The actual number of RR dinucleotides in the data is 138 and is noted by the arrow on the *x* axis. The G bias model is based on the preference for TdT to add G nucleotides and assumes that the 3′ terminus of the DNA substrate does not influence the choice of the next nucleotide to be added to the DNA substrate. The G+stacking model is likewise based on the G utilization preference but, in addition, assumes that the choice of the next base added by TdT is influenced by base-stacking interactions between the 3′-terminal nucleotide on the DNA substrate and the nucleotide that will be added by TdT. The curves were generated from the binomial distribution, using the total number of dinucleotides in the N regions (excluding P nucleotides) in Fig. 1, 2, and 4A (n = 295). The probability that any individual dinucleotide will be an RR dinucleotide is given by either (i) the observed N-region base composition in Fig. 1, 2, and 4A (G bias model) or (ii) the combination of the observed N-region base composition and N-region dinucleotide frequencies (G+stacking model). (B) Pyrimidine-pyrimidine (YY) dinucleotides. As in the legend to panel A, the curves represent theoretical probability histograms for the number of YY dinucleotides that TdT would be expected to generate within the junctions. The actual number of YY dinucleotides in the data is 88. (C) Purine (RRR) trinucleotides. The actual number of RRR trinucleotides in the junctions is 70. For a description of the curves, see the legend to panel A. (D) Pyrimidine (YYY) trinucleotides. The actual number of YYY trinucleotides in the junctions is 48. For a description of the curves, see the legend to panel A.

were found in the pattern of loss between the different coding ends. The AT-rich 12-end of pGG80 was significantly different from all other coding ends except the 23-end of pGG80. The GC-rich 23-end of pGG50 and pGG51 was significantly different from all other coding ends except the 23-end of pGG52 and pML520 (Fig. 7B).

DISCUSSION

In this study, we observed three features which appear to constrain the diversity of coding junction formation. First, the sequence of nucleotide additions by TdT appeared to be influenced by base-stacking interactions, resulting in frequent homopolymer tracts. Second, inverted repeats were observed at high frequencies not only at full-length coding-end termini but also at coding ends that have undergone nucleolytic processing. These inverted repeats are dependent on the presence of wild-type TdT, but statistical models indicate that it is unlikely that these inverted repeats are actually synthesized by TdT. Third, the degree of nucleolytic processing was significantly different between coding termini which vary in sequence.

These observations suggest that constraints on the potential diversity of coding junctions are inherent in the mechanism of V(D)J recombination (Fig. 8). TdT additions increase diversity, but G utilization and homopolymer biases place some constraints on the diversity of these additions. Variable nucleotide loss from coding ends increases the diversity of junctions, but loss is not entirely random and is constrained by the sequence of the coding ends. Inverted repeats (P and P_r addition) add junctional nucleotides, thus increasing diversity, but the sequence of these additional nucleotides is constrained precisely because they are inverted repeats of the coding termini. Non-TdT nucleotide additions also increase diversity, but they occur at only very few junctions (≈5%).

Base stacking and TdT addition. It generally has been assumed that the sequence of TdT additions, apart from their high G+C content, is largely random. Our observation that frequent homopolymer tracts in junctions suggests that this notion of the random nature of TdT additions is not accurate. Purine-purine and pyrimidine-pyrimidine dinucleotides occur more often than random addition would predict. Conversely, purine-pyrimidine and pyrimidine-purine dinucleotides occur

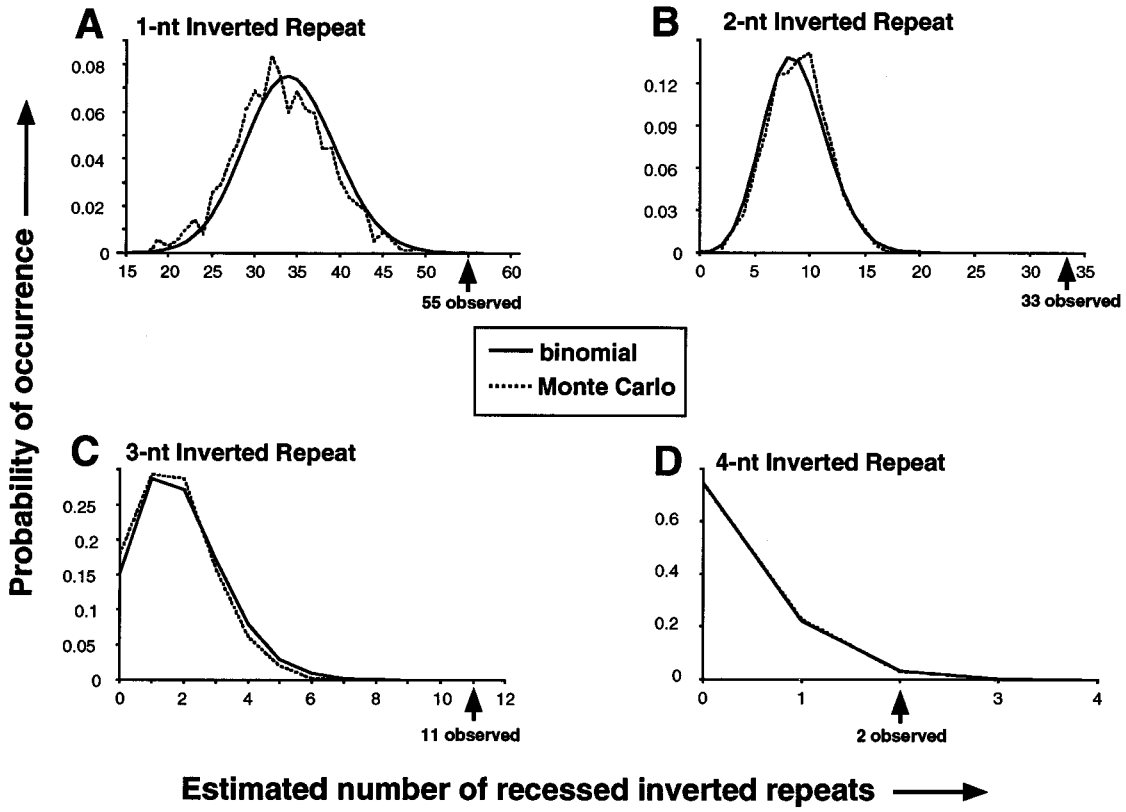


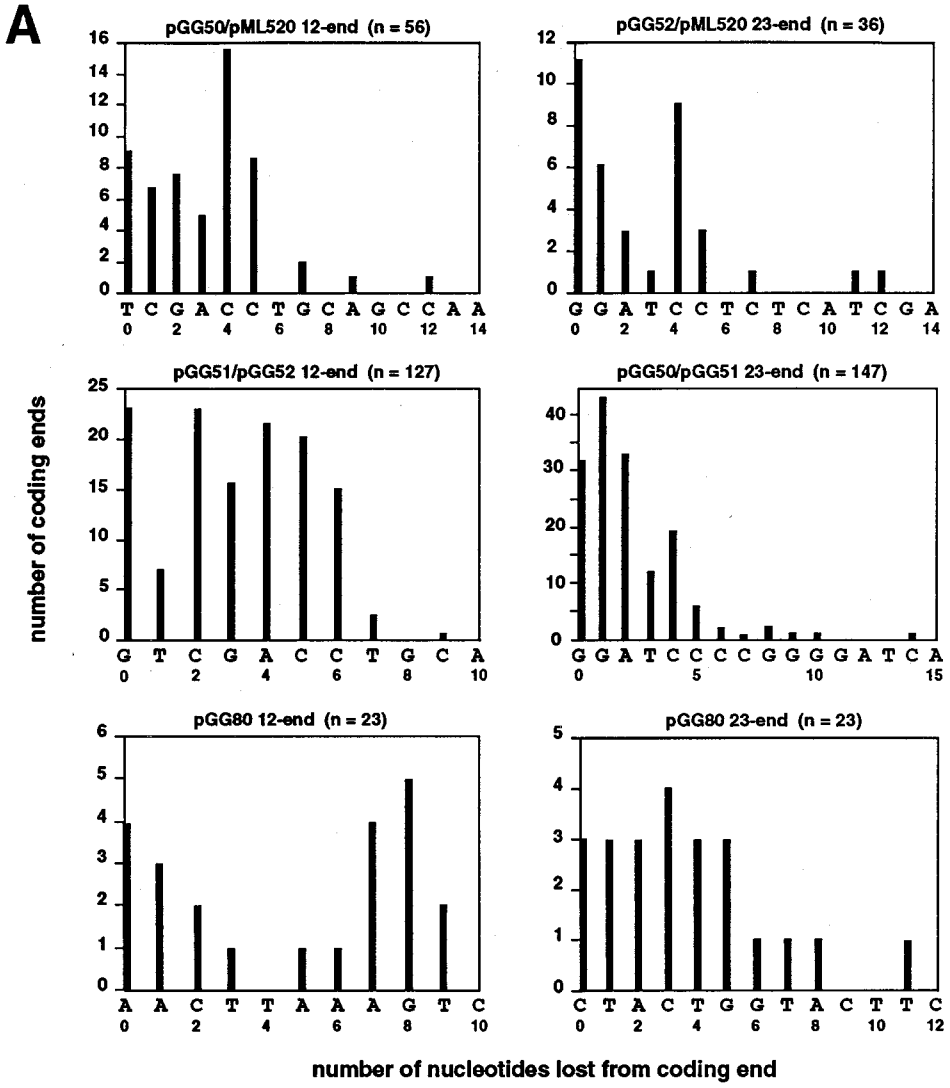
FIG. 6. Number of recessed inverted repeats expected from TdT addition. (A) Recessed inverted repeats 1 nt in length. The curves represent theoretical probability histograms of the number of inverted repeats that TdT would be expected to generate at the nucleolytically processed coding termini in Fig. 1, 2, and 4A. All curves were generated from the G+stacking model by using either the binomial distribution (solid curves) or Monte Carlo simulations of TdT addition at the coding junctions (broken curves). The actual number of recessed inverted repeats 1 nt in length in the junctions is 55, as noted by the arrow on the *x* axis. (B) Recessed inverted repeats 2 nt in length. The actual number of recessed inverted repeats 2 nt in length in the junctions is 33. (C) Recessed inverted repeats 3 nt in length. The actual number of recessed inverted repeats 3 nt in length in the junctions is 11. (D) Recessed inverted repeats 4 nt in length. The actual number of recessed inverted repeats 4 nt in length in the junctions is 2. For a description of the curves in panels B to D, see the legend to panel A.

less often than expected. These observations suggest that each nucleotide addition is not random and independent of the previous addition. In N regions, there is a tendency for G to follow G or A; similarly, C tends to follow C or T (of course, YY tracts on one strand may actually reflect RR tracts on the complementary strand or vice versa). Base-stacking interactions may affect TdT polymerization to favor the formation of homopolymers. Supporting this hypothesis, binomial simulations of TdT additions that assumed both base stacking and G preference were far more consistent with the observed data than simulations assuming G preference alone.

Inverted repeats. Inverted repeats of significant length and statistically significant incidence have previously been analyzed only at full-length coding ends (26). In contrast, this study shows that in human cell lines expressing TdT, unusually high frequencies of inverted repeats can occur at coding ends that have incurred nucleotide loss as well as at full-length coding ends. The incidence of inverted repeats is not consistent with statistical estimates for TdT addition. We have considered several alternative mechanisms by which inverted repeats might be formed. A possible explanation is that the P_r inverted repeats are the result of a hairpin intermediate similar to that suggested for P nucleotides (23, 30). We are currently investigating this hypothesis by using the recently described hairpinning activity (35).

Significance of recessed inverted repeats. Our statistical analysis shows that there is a highly skewed distribution for inverted repeats at coding termini. Overall, P_r inserts occurred at a comparable frequency to that of P inserts: P inserts occur at 42% (30 of 72) of all full-length coding ends, and P_r inserts occur at 25% (55 of 219) of nucleolytically processed coding ends (when wild-type TdT is expressed). Moreover, in these junctions, long ($\geq$3-nt) P_r additions are more frequent than long P additions. Together, P and P_r inserts were present in approximately half of all coding junctions, and together, they account for almost one-third of the junctional nucleotides (Table 2).

Given that murine RAG-1, RAG-2, and TdT are able to generate P_r additions in human fibroblasts, it is puzzling why P_r inserts are less prevalent in murine junctions, both in endogenous junctions and in those from extrachromosomal substrates. The frequency of recessed inverted repeats in one transgenic study (4) was somewhat higher than that expected by TdT addition but could still be attributed to fortuitous TdT addition ($P \approx 0.1$ [our calculation]). We also analyzed a set of 197 junctions generated with extrachromosomal substrates transfected into murine and hamster cell lines (10a, 25). In this analysis, both P and P_r inserts were found. P inserts were present at statistically significant frequencies, but P_r inserts were not (data not shown). We do not take this as evidence that there is a fundamental difference between the murine and human V(D)J recombination reactions, only that there may be subtleties in the reaction mechanism which, for unknown reasons, allow more P_r formation in human cells. The propen-



B

	pGG50/pML520 12-end	pGG51/pGG52 12-end	pGG80 12-end	pGG52/pML520 23-end	pGG50/pGG51 23-end
pGG51/pGG52 12-end	0.3067				
pGG80 12-end	**0.0018**	**0.0004**			
pGG52/pML520 23-end	0.2422	**0.0149**	**0.0058**		
pGG50/pGG51 23-end	**0.0013**	**< .0001**	**0.0001**	0.1350	
pGG80 23-end	0.9401	0.8723	0.1080	0.2735	**0.0128**

FIG. 7. Comparison of nucleotide loss from different coding-end sequences. (A) Distribution of nucleotide loss from the different coding ends. A histogram is shown for each different coding-end sequence, showing the number of coding ends along the *y* axis and the number of nucleotides lost on the *x* axis. *n* is the total number of coding ends scored for each histogram. The sequence of the full-length coding end (5′- to 3′, left to right) is indicated along the bottom of each histogram. The coding ends are displayed as if the heptamer of the V(D)J signal were on the left. Numbers below the sequence indicate the number of nucleotides lost from each coding end. When the precise number of nucleotides lost from a coding end was ambiguous because of microhomology use at the junction, the nucleotide loss from each end was calculated as half of the total nucleotides lost from the junction, and loss was then evenly distributed among all potential positions within the individual coding ends. To generate a histogram for each coding end, sequence data were pooled from all cell lines transfected with V(D)J substrates containing that coding end. (B) Statistical comparison of the distribution of nucleotide loss from the different coding ends shown in panel A. Pairwise comparison of the pattern of nucleotide loss between the different coding ends was performed by two-sample Kolmogorov-Smirnov tests, and the *P* values for each pairwise comparison are indicated within the matrix. *P* values of <0.05 are indicated in boldface type.

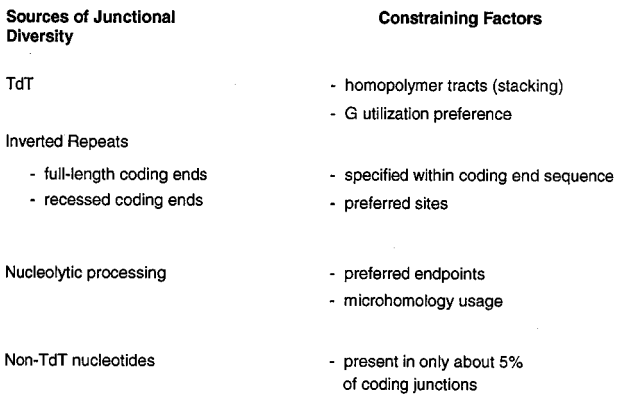


FIG. 8. Sources of junctional diversity and constraining factors. Sources of junctional diversity in V(D)J recombination are listed on the left, and aspects of each that constrain diversity are listed on the right. Junctional diversity does not include all potential sources of antigen receptor diversity but includes only diversity generated after the a given pair of recombining segments have entered into a productive recombination reaction. Prior to this, there is the combinatorial diversity due to multiple V, multiple D, and multiple J segments.

sity for P_r formation may be attributable to components of the recombination complex that are expressed to higher levels in human cells. One illustration of this possibility is DNA-dependent protein kinase (DNA-PK), the component defective in murine scid cells. DNA-PK is reported to be present at 50- to 100-fold-higher levels in human cells than in murine cells (7). There are many other possibilities.

Immunologic selection may play a role in the scarcity of P_r additions at endogenous junctions. P_r inserts that are generated by the recombination complex may not survive immunologic selection in the organism. Thus, the endogenous junctions that are eventually sequenced may have less than a statistically significant of incidence of P_r inserts. Examination of published human and murine endogenous junctions reveals that some P_r additions are present at these junctions (4, 11, 13, 17, 36, 37), but they generally occur at frequencies statistically indistinguishable from those of TdT additions. In one of these studies (13), we found three P_r inserts within a collection of 12 unique endogenous junctions that had TdT additions. Although this is a small dataset, statistical analysis indicates that the two longest P_r inserts are, in fact, statistically significant ($P < 0.05$, our calculation).

Nucleolytic processing of coding ends. The junctions presented in this study statistically confirm that the profile of nucleotide loss can be very different for coding ends which differ in sequence, as has been suggested from analyses of murine endogenous junctions (6) and extrachromosomal substrates in murine cell lines (2). Sequences with high A+T content appear to suffer more loss than sequences containing a greater number of G · C base pairs. A high G+C content may confer nucleolytic resistance, perhaps through alterations in the structure of the DNA helix or by limiting the extent of helix unwinding.

Preferences in the end points of nucleotide loss and sites of inverted-repeat formation may result from a common mechanism. It is notable that approximately 33% of the P_r additions in our junctions occur at the same sequence in the 12-end of the substrates pGG50, pGG51, and pML520 (see examples in junctions 35 and 36 [Fig. 1] and junctions 7 to 9 [Fig. 4A]). The coding-end sequence at this position is 5′-CCTG. Among the processed coding ends that terminate unambiguously with this sequence, 60% (18 of 30) have P_r additions. As noted in Results, this particular sequence may cause perturbations in the stacking interactions, favoring inverted-repeat additions by TdT.

Alternatively, if P_r additions are the result of a hairpin intermediate, the high incidence of P_r inserts at this sequence could result from asymmetric hairpin opening caused by the specific DNA sequence. This is consistent with the observation that P inserts occur more frequently at certain coding ends (2, 6, 26). Boubnov et al. (2) suggest that P nucleotides are more common at full-length coding ends terminating with G or C homopolymers compared with A or T homopolymers. Feeney and colleagues (6) observe that P inserts occur at approximately 25% of murine endogenous junctions containing J_H3 whereas they occur at only 5% of junctions containing J_H4. It is intriguing that J_H3 terminates with 5′-CCTG, the same sequence seen for 33% of our P_r additions. Hairpins with this sequence may be opened asymmetrically more often than other sequences. This sequence within the coding ends of our V(D)J substrates may have fortuitously favored inverted repeat formation either through a similar hairpin mechanism or by perturbing the stacking interactions which influence TdT nucleotide addition preference.

Concluding remarks. The data presented in this report make up the first large unselected set of coding junctions from human cell lines. Fine-structure analysis of the junctions reveals some points not apparent in other studies of junctional diversity and may provide some insight into the biochemical mechanism of V(D)J recombination. First, the base-stacking interactions may influence TdT additions, favoring the formation of homopolymer tracts in N regions. Second, inverted repeats similar to P nucleotides are observed with high frequency at many coding ends that have undergone nucleolytic processing. Third, the use of several different coding-

TABLE 2. Distribution of junctional nucleotides

Category of junctional addition	No. of inserts	No. of nucleotides	% of total nucleotides	No. of junctions	% of junctions[a]	% of coding ends
N	307	307	67	100	73[b]	
P	30	45	10	30	15[c]	42[c]
P_r	55	101	22	47	34[b]	25[d]
Non-TdT N	3	3	1	2	5[e]	

[a] Many junctions contain multiple types of addition; thus, the percentages in column 6 total >100%.
[b] Because N and P_r inserts are dependent on TdT, data for N and P_r are expressed as a percentage of the total number of junctions from experiments wherein wild-type TdT was expressed ($n = 137$).
[c] P inserts are independent of TdT; thus, the percentage of total junctions with P nucleotides is expressed as a percentage of the total number of junctions in the study (30 of 206 = 15%). P inserts occur at 30 of 72 = 42% of full-length coding ends (coding ends using microhomology excluded because of ambiguity of coding termini).
[d] P_r inserts occur at 55 of 219 = 25% of recessed ends from experiments wherein wild-type TdT was expressed.
[e] Non-TdT N additions occur at 2 of 69 = 5% of junctions from experiments wherein wild-type TdT was not expressed.

end sequences indicates that the extent of nucleotide loss at coding junctions is influenced by the sequence of the coding ends.

ACKNOWLEDGMENTS

We thank Samuel Karlin and Tod Klinger for helpful discussions concerning statistical methodology and Ravi Konchigeri and Grant Weber for their programming expertise.

G.H.G. is supported by PHS grant 5T32CA09302 awarded by the National Cancer Institute through the Stanford University Program in Cancer Biology. This work was supported by NIH grants and, in part, by a grant from the Council for Tobacco Research. M.R.L. is a Leukemia Society of America Scholar and a Cancer Research Institute Investigator.

REFERENCES

1. **Basu, M., V. M. Hegde, and M. J. Modak.** 1983. Synthesis of compositionally unique DNA by terminal deoxynucleotidyl transferase. Biochem. Biophys. Res. Commun. **111:**1105–1112.
2. **Boubnov, N. V., Z. P. Wills, and D. T. Weaver.** 1993. V(D)J recombination coding junction formation without DNA homology: processing of coding termini. Mol. Cell. Biol. **13:**6957–6968.
3. **Broom, A. D., M. P. Schweizer, and P. O. Ts'o.** 1967. Interaction and association of bases and nucleosides in aqueous solution. J. Am. Chem. Soc. **89:** 3612–3622.
4. **Engler, P., E. Klotz, and U. Storb.** 1992. N region diversity of a transgenic substrate in fetal and adult lymphoid cells. J. Exp. Med. **176:**1399–1404.
5. **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.
6. **Feeney, A. J., K. D. Victor, K. Vu, B. Nadel, and R. U. Chukwuocha.** 1994. Influence of the V(D)J recombination mechanism on the formation of the primary T and B cell repertoires. Semin. Immunol. **6:**143–163.
7. **Finnie, N. J., T. M. Gottlieb, T. Blunt, P. A. Jeggo, and S. P. Jackson.** 1995. DNA-dependent protein kinase activity is absent in xrs-6 cells: implications for site-specific recombination and DNA double-strand break repair. Proc. Natl. Acad. Sci. USA **92:**320–324.
8. **Gauss, G. H., and M. R. Lieber.** 1992. DEAE-dextran enhances electroporation of mammalian cells. Nucleic Acids Res. **20:**6739–6740.
9. **Gauss, G. H., and M. R. Lieber.** 1993. Unequal signal and coding joint formation in human V(D)J recombination. Mol. Cell. Biol. **13:**3900–3906.
10. **Gerstein, R. M., and M. R. Lieber.** 1993. Coding end sequence can markedly affect the initiation of V(D)J recombination. Genes Dev. **7:**1459–1469.
11. **Gilfillan, S., A. Dierich, M. Lemeur, C. Benoist, and D. Mathis.** 1993. Mice lacking TdT: mature animals with an immature lymphocyte repertoire. Science **261:**1175–1178.
12. **Gu, H., I. Forster, and K. Rajewsky.** 1990. Sequence homologies, N sequence insertion and JH gene utilization in VHDJH joining: implications for the joining mechanism and the ontogenetic timing of Ly1 B cell and B-CLL progenitor generation. EMBO J. **9:**2133–2140.
13. **Hansen-Hagge, T. E., S. Yokota, H. J. Reuter, K. Schwarz, and C. R. Bartram.** 1992. Human common acute lymphoblastic leukemia-derived cell lines are competent to recombine their T-cell receptor delta/alpha regions along a hierarchically ordered pathway. Blood **80:**2353–2362.
14. **Hesse, J. E., M. R. Lieber, M. Gellert, and K. Mizuuchi.** 1987. Extrachromosomal DNA substrates in pre-B cells undergo inversion or deletion at immunoglobulin V-(D)-J joining signals. Cell **49:**775–783.
15. **Kallenbach, S., N. Doyen, D. M. Fanton, 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.
16. **Karlin, S., F. Ost, and B. E. Blaisdell.** 1989. Patterns in DNA and amino acids and their statistical significance, p. 149–157. *In* M. S. Waterman (ed.), Mathematical methods for DNA sequences. CRC Press, Inc., Boca Raton, Fla.
17. **Komori, T., A. Okada, V. Stewart, and F. W. Alt.** 1993. Lack of N regions in antigen receptor variable region genes of TdT-deficient lymphocytes. Science **261:**1171–1175.
18. **Kornberg, A., and T. Baker.** 1992. DNA replication, 2nd ed. W. H. Freeman & Co., New York.
19. **Korsmeyer, S. J., A. Arnold, A. Bakhshi, J. V. Ravetch, U. Siebenlist, P. A. Hieter, S. O. Sharrow, T. W. LeBien, J. H. Kersey, D. G. Poplack, P. Leder, and T. A. Waldmann.** 1983. Immunoglobulin gene rearrangement and cell surface antigen expression in acute lymphocytic leukemias of T cell and B cell precursor origins. J. Clin. Invest. **71:**301–313.
20. **Lafaille, J. J., A. DeCloux, M. Bonneville, Y. Takagaki, and S. Tonegawa.** 1989. Junctional sequences of T cell receptor gamma delta genes: implications for gamma delta T cell lineages and for a novel intermediate of V-(D)-J joining. Cell **59:**859–870.
21. **Lewis, S. M.** 1994. The mechanism of V(D)J joining: lessons from molecular, immunological, and comparative analyses. Adv. Immunol. **56:**27–150.
22. **Lewis, S. M.** 1994. P nucleotide insertions and the resolution of hairpin DNA structures in mammalian cells. Proc. Natl. Acad. Sci. USA **91:**1332–1336.
23. **Lieber, M. R.** 1991. Site-specific recombination in the immune system. FASEB J. **5:**2934–2944.
24. **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:**751–761.
25. **McCormack, W. T., L. W. Tjoelker, 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:**785–791.
26. **Meier, J. T., and S. M. Lewis.** 1993. P nucleotides in V(D)J recombination: a fine-structure analysis. Mol. Cell. Biol. **13:**1078–1092.
27. **Miyoshi, I., S. Hiraki, T. Tsubota, I. Kubonishi, Y. Matsuda, T. Nakayama, H. Kishimoto, I. Kimura, and H. Masuji.** 1977. Human B cell, T cell and null cell leukaemic cell lines derived from acute lymphoblastic leukaemias. Nature (London) **267:**843–844.
28. **Oettinger, M. A., D. G. Schatz, C. Gorka, and D. Baltimore.** 1990. RAG-1 and RAG-2, adjacent genes that synergistically activate V(D)J recombination. Science **248:**1517–1523.
29. **Powell, J. T., E. G. Richards, and W. B. Gratzer.** 1972. The nature of stacking equilibria in polynucleotides. Biopolymers **11:**235–250.
30. **Roth, D. B., J. P. Menetski, P. B. 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 **70:**983–991.

30a. **Sadofsky, M. J., J. E. Hesse, and M. Gellert.** 1994. Definition of a core region of RAG-2 that is functional in V(D)J recombination. Nucleic Acids Res. **22:**1805-1809.

31. **Sambrook, J., E. F. Fritsch, and T. Maniatis.** 1989. Molecular cloning: a laboratory manual, 2nd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

31a. **Schwarz, K.** Unpublished data.

31b. **Silver, D. P., E. Spanopoulou, R. C. Mulligan, and D. Baltimore.** 1993. Dispensable sequence motifs in the RAG-1 and RAG-2 genes for plasmid V(D)J recombination. Proc. Natl. Acad. Sci. USA **90:**6100–6104.

32. **Staunton, J. E., and D. T. Weaver.** 1994. scid cells efficiently integrate hairpin and linear DNA substrates. Mol. Cell. Biol. **14:**3876–3883.
33. **Taccioli, G. E., G. Rathbun, E. Oltz, T. Stamato, P. Jeggo, and F. W. Alt.** 1993. Impairment of V(D)J recombination in double-strand break repair mutants. Science **260:**207–210.
34. **Tonegawa, S.** 1983. Somatic generation of antibody diversity. Nature (London) **302:**575–581.
35. **van Gent, D. C., J. F. McBlane, D. A. Ramsden, M. J. Sadofsky, J. E. Hesse, and M. Gellert.** 1995. Initiation of V(D)J recombination in a cell-free system. Cell **81:**925–934.
36. **Wasserman, R., N. Galili, Y. Ito, B. A. Reichard, S. Shane, and G. Rovera.** 1992. Predominance of fetal type DJH joining in young children with B precursor lymphoblastic leukemia as evidence for an in utero transforming event. J. Exp. Med. **176:**1577–1581.
37. **Yamada, M., R. Wasserman, B. A. Reichard, S. Shane, A. J. Caton, and G. Rovera.** 1991. Preferential utilization of specific immunoglobulin heavy chain diversity and joining segments in adult human peripheral blood B lymphocytes. J. Exp. Med. **173:**395–407.
38. **Yang, B., K. N. Gathy, and M. S. Coleman.** 1994. Mutational analysis of residues in the nucleotide binding domain of human terminal deoxynucleotidyl transferase. J. Biol. Chem. **269:**11859–11868.
39. **Zhu, C., and D. B. Roth.** 1995. Characterization of coding ends in thymocytes of scid mice: implications for the mechanism of V(D)J recombination. Immunity **2:**101–112.