Cutting Edge: DGYW/WRCH Is a Better Predictor of Mutability at G:C Bases in Ig

Hypermutation Than the Widely Accepted

RGYW/WRCY Motif and Probably Reflects a Two-Step Activation-Induced Cytidine

Deaminase-Triggered Process

Igor B. Rogozin and Marilyn Diaz

J Immunol 2004; 172:3382-3384; doi: 10.4049/jimmunol.172.6.3382

http://www.jimmunol.org/content/172/6/3382
Cutting Edge: DGYW/WRCH Is a Better Predictor of Mutability at G:C Bases in Ig Hypermutation Than the Widely Accepted RGYW/WRCY Motif and Probably Reflects a Two-Step Activation-Induced Cytidine Deaminase-Triggered Process

Igor B. Rogozin* and Marilyn Diaz†‡

A feature of Ig hypermutation is the presence of hypermutable DNA sequences that are preferentially found in the V regions of Ig genes. Among these, RGYW/WRCY is the most pronounced motif (G:C is a mutable position; R = A/G, Y = C/T, and W = A/T). However, a molecular basis for the high mutability of RGYW was not known until recently. The discovery that activation-induced cytidine deaminase targets the DNA encoding V regions, has enabled the analysis of its targeting properties when expressed outside of the context of hypermutation. We analyzed these data and found evidence that activation-induced cytidine deaminase is the major source of the RGYW mutable motif, but with a new twist: DGYW/WRCH (G:C is the mutable position; D = A/G/T, H = T/C/A) is a better descriptor of the Ig mutation hotspot than RGYW/WRCY. We also found evidence that a DNA repair enzyme may play a role in modifying the sequence of hypermutation hotspots. The Journal of Immunology, 2004, 172: 3382–3384.

Although diversity of the B cell repertoire in adaptive immunity is generated through V(D)J recombination, specificity to a particular foreign Ag is attained via the combined forces of Ig somatic hypermutation (SHM) and cellular selection for the high-affinity variants generated by the SHM machinery (1, 2). In this process, mutations are introduced into the DNA encoding the V domains of the Ig receptor at a rate close to six orders of magnitude greater than the spontaneous mutation rate (3, 4). The mutations generated by the SHM mechanism are not randomly distributed. Mutation hotspots in V regions occur primarily within two DNA sequence motifs: RGYW/WRCY (R = A/G, Y = C/T, and W = A/T; the hotspot in G:C is underlined), which is found in both strands, and WA/TW (the hotspot in A:T is underlined), preferentially found in one strand (5–7). DNA polymerase η probably contributes to the WA hotspot: in vitro analysis of mutation spectra of various DNA polymerases revealed a correlation between the WA/TW motif and the error specificity of human DNA polymerase η (7), and xeroderma pigmentosum patients with a defect in this polymerase display fewer A-T mutations in Ig genes (8). A candidate for the predominant RGYW mutator is a targeting complex containing activation-induced cytidine deaminase (AID). Recent evidence suggests that AID deaminates C bases in DNA (9–15). Although cytosine in other contexts such as dCMP in the nucleotide pool or cytosine in RNA were originally suspected, the nucleotide pool in Escherichia coli cells expressing AID is unaltered (16), and direct evidence for an RNA target remains lacking. The best evidence that AID targets the DNA of IgV regions in vivo comes from the finding that mice deficient in uracil DNA glycosylase (ung) have a dramatic increase in transitions from G:C base pairs in IgV genes undergoing hypermutation (17). Thus, it appears that AID deamidates cytosines in the DNA of IgV regions, and this lesion triggers the full SHM process including other factors and possibly multiple DNA polymerases (7, 8, 17–20).

AID can deaminate ssDNA in vitro. Goodman and colleagues (11) examined the spectra of mutations generated in vitro and found a potential new variant of the RGYW hotspot motif: WRC/GYW. This observation was supported by Lieber and colleagues (12). This contradicts earlier analysis where a highly significant correlation was found between the identity of the bases at the first/last positions of the RGYW/WRCY motif and mutability, particularly in Ig loci (5–7, 21–24). Strikingly, both groups observed a high frequency of C:G>T:A mutations in WRC/GGYW motifs (11, 12). Goodman and colleagues (11) reasonably attributed these differences from the RGYW motif to their assay target being a prokaryotic gene (lacZ). Because of these discrepancies, we re-examined the mutability of the RGYW motif and its potential variants using both new and...
Materials and Methods

Seven spectra of somatic mutations in different pro- and eukaryotic genes described by Milstein et al. (6) were analyzed. The data are available upon request from I. Rogozin (rogozin@ncbi.nlm.nih.gov). A sequence where mutations were revealed is referred to as a target sequence.

Mutation hotspots are defined using a threshold for the number of mutations at a site. The threshold is established by analyzing the frequency distribution derived from a mutation spectrum using CLUSTERM program (www.itb.cnr.it/webmutation/) (25, 26). CLUSTERM identifies several homogeneous classes of sites from a mutation spectrum. Each class of sites is approximated by a binomial (or Poisson) distribution. The probability of mutation is assumed to be the same for all sites in a class, so variation in mutation frequency for sites of the same class is assumed random and not statistically significant. In contrast, statistically significant differences in mutation frequency for sites from different classes reflect mutation hotspots. A class with the highest mutation frequency is called a hotspot class. A hotspot site is defined as a permanent in vivo SHM data, and noting that context differences between eukaryotic and prokaryotic genes may influence the pattern of SHM.

Table I. Average number of mutations in variants of GYW/WRC

<table>
<thead>
<tr>
<th>Gene</th>
<th>gpt</th>
<th>neo</th>
<th>Globin</th>
<th>VJα3 intron</th>
<th>JH4</th>
<th>VαOX1Jε5</th>
<th>VH26</th>
<th>Σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGYW/WRC</td>
<td>11.0+</td>
<td>1.6+</td>
<td>5.3+</td>
<td>2.7+</td>
<td>1.9+</td>
<td>10.1+</td>
<td>10.3+</td>
<td>42.9</td>
</tr>
<tr>
<td>GGYW/WRC</td>
<td>3.3</td>
<td>1.1+</td>
<td>2.1+</td>
<td>2.1+</td>
<td>0.8</td>
<td>6.8+</td>
<td>4.8+</td>
<td>20.8</td>
</tr>
<tr>
<td>TGYW/WRGA</td>
<td>3.8</td>
<td>0.4</td>
<td>1.9</td>
<td>1.5+</td>
<td>1.5+</td>
<td>6.0+</td>
<td>3.1</td>
<td>18.4</td>
</tr>
<tr>
<td>CGYW/WRGC</td>
<td>1.8</td>
<td>0.7</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>3.0</td>
<td>3.8</td>
<td>9.3</td>
</tr>
</tbody>
</table>

* Numbers of mutations were calculated only at the underlined G-C bases. Cytosines in both DNA strands are probable targets of AID. +, Indicates a significant excess of mutations in a mutable motif (in other words, significant correlation ($P_{W < W_{\text{random}} < 0.05}$) between motifs and somatic mutations). The significance of correlations was measured with a Monte Carlo procedure (the CONSEN program). CGYW/WRGC motifs were found in all studied spectra.

Results and Discussion

To determine the mutability of the CGYW motif, we analyzed seven mutation spectra in various Ig and non-Ig genes, in hypermutating B cells. The numbers of mutations in GGYW and TGYW are similar, but the average number of mutations was much higher in AGYW (Table I). Strikingly, and in contradiction to the in vitro data, the average number of mutations in the CGYW motif was significantly lower (Table I). These combined data indicate that a key motif of Ig hypermutation is DGYW ($D = A/G/T$). Because the GGYW motif displays a higher frequency of mutations than the TGYW motif in V regions (e.g., in the Vh26 gene; Table I), the RGYW may be adequate for IgV gene targets. However, DGYW is a far better descriptor of the intrinsic targeting of the SHM machinery (Table I).

Context analysis of somatic mutation hotspots predicted by the CLUSTERM program (25) also suggests that DGYW is the best mutable motif instead of the widely accepted RGYW motif. We found 35 AGYW, 12 GGYW, and 17 TGYW hotspots in this analysis. However, no mutation hotspots were found in the CGYW motif. Although there were fewer CGYW target motifs (28 sites) than DGYW motifs (295 sites), the lack of mutations at CGYW motifs could not be attributed to its lesser prevalence, because, given the number of sites, attaining such mutation frequencies at random was unlikely (0 of 64 vs 28 of 295; $p = 0.008$ by the Fisher exact test). This result was confirmed using the algorithm of the CONSEN program (5) wherein no significant targeting of somatic mutations was found within CGYW (Table I), whereas AGYW, GGYW, and TGYW displayed a significant correlation ($P_{W < W_{\text{random}} < 0.05}$) in the studied spectra (Table I).

Table II. CGYW is a coldspot in vivo but a hotspot in vitro

<table>
<thead>
<tr>
<th>Motif</th>
<th>Vh26</th>
<th>VαOX1Jε5</th>
<th>GFP in Fibroblasts</th>
<th>lacZ In Vitro</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site</td>
<td>No. of mutations</td>
<td>Site</td>
<td>No. of mutations</td>
<td>Site</td>
</tr>
<tr>
<td>AGYW</td>
<td>AGCT</td>
<td>28</td>
<td>AGTT</td>
<td>28</td>
</tr>
<tr>
<td>GGYW</td>
<td>GCTA</td>
<td>17</td>
<td>GTTA</td>
<td>13</td>
</tr>
<tr>
<td>TGYW</td>
<td>TGCT</td>
<td>11</td>
<td>TGCA</td>
<td>17</td>
</tr>
<tr>
<td>CGYW</td>
<td>CGCT</td>
<td>9</td>
<td>CGCT</td>
<td>3</td>
</tr>
</tbody>
</table>

* The polarity (strand) of all sequences was chosen such that the potentially mutable base (underlined) is a G. A motif with the highest frequency of mutations was chosen for each variant of NGYW ($N = A/T/G/C$). The highest ratio of mutation frequencies at AGYW vs CGYW was observed in the Vh26 spectrum among all studies of in vivo spectra (28 mutations vs 9 mutations). GFP, Green fluorescent protein.
In the seven analyzed in vivo spectra, the most mutable individual sequences were the following: AGCT (the gpt, neo, and Vh26 sequences), AGCA (globin), AGTA (a intron), TGTT (IgH4), and AGTT (VκOx1) (Table II). Thus, six AGYW/WGCG and one TGYW/WGCC motifs were found in vivo. In the in vitro study, wherein AID intrinsic targeting was tested, the two most mutable sequences were TAGGC/CGTA and TGCT/AGCA (11) (Table II), indicating that the CGYW/WGCC motif (TAGGC/CGTA) is highly mutable in vitro. Again, such exceptionally high mutability of CGYW/WGCC motifs has never been observed in SHM spectra examined so far (5–7, 21–24). In fact, even in the case wherein unique components of the Ig hypermutation machinery are not expected to be present (i.e., fibroblasts), DGYW remained the hotspot motif, clearly suggesting that AID alone targets these motifs (28) (Table II).

Why is CGYW/WRCG mutable in vitro but not in vivo? We propose that AID targets WRCG/WGY in vivo. Given the fact that these simple motifs are probably found throughout the genome, it is likely that AID is targeted to IgY regions by a cofactor, and it locally either scans for, or has highest affinity to the WRCG/YW motif. We further propose that when AID targets C in the WRCG motif, an unknown DNA repair enzyme correctly repairs the lesion. Because this motif contains the GC dinucleotide characteristic of CpG islands, and because CG motifs are known to be highly mutable in the vertebrate genome (29), it is likely that many repair enzymes have evolved to repair lesions at CG dinucleotides, including those originating from cytosine deamination. Consistent with this idea, AID expression in fibroblasts is also characterized by the absence of mutations at the frequently found CGYW (WGCY) motif of the green fluorescent protein locus target (28) (Table II). Because Ig hypermutation does not occur in fibroblasts, the paucity of mutations at CGYW/WRCG in these cells further indicates that the putative repair enzyme is ubiquitous. Armed with this information, we examined the possibility that the enzyme is ung by analyzing the ung−/− spectra (17). Indeed we found 2.6 more mutations at CGs in the ung−/− spectrum, suggesting a role for this glycosylase in modifying AID’s intrinsic targeting characteristics. However, the idea that ung is the enzyme processing the deaminated cytosine in the WRGC/GGY motif is speculative without direct experimental evidence, and thus, it remains possible that other DNA repair enzymes are involved. Thus, we propose that AID targets WRGN (N = A/G/C/T), but when N = G, a DNA repair enzyme, perhaps ung, efficiently processes the lesion and thus eliminates mutations from WRGC/GGY motifs in vivo. The combined action of AID and removal of mutations from the WRGC/GGY motif leaves us with the newly refined AID-driven SHM motif of DGYW/WRCG.

Acknowledgments

We thank Jan Drake, Eugene Koonin, and Tom Kunkel for comments on the manuscript. Also, we thank Youri Pavlov and Claude-Agnes Reynaud for stimulating conversations about SHM and potential variants of the RGYW motifs.

References