Proc. Natl. Acad. Sci. USA
Vol. 81, pp. 4697–4701, August 1984
Biochemistry

Nucleotide sequence 5′ of the chicken c-*myc* coding region: Localization of a noncoding exon that is absent from *myc* transcripts in most avian leukosis virus-induced lymphomas

(DNA sequencing/oncogene/RNA blot analysis/avian leukosis virus integration)

CHENG-KON SHIH*†, MAXINE LINIAL‡, MAUREEN M. GOODENOW*, AND WILLIAM S. HAYWARD*

*Memorial Sloan–Kettering Cancer Center, 1275 York Avenue, New York, NY 10021; †The Rockefeller University, 1230 York Avenue, New York, NY 10021; and ‡Fred Hutchinson Cancer Research Center, 1124 Columbia Street, Seattle, WA 98104

Communicated by Paul A. Marks, April 12, 1984

ABSTRACT **We have determined the nucleotide sequence of the 2.2-kilobase-pair region upstream of the chicken c-*myc* coding exons. Using RNA blot analysis, we have localized a noncoding exon to a region that is separated from the c-*myc* coding sequences by an intron of 700–800 base pairs. In most avian leukosis virus-induced lymphomas proviral integration has occurred within, or downstream of, the first exon, thus presumably displacing the regulatory sequences that normally control c-*myc* expression. More than 70% of the integration sites were clustered in a 250-base-pair region in the first intron, immediately preceding the coding sequences. Sequences from the upstream noncoding exon were absent from the *myc* transcripts in these lymphomas; RNA transcripts from the normal c-*myc* allele were not expressed at detectable levels.**

Structural alterations of the c-*myc* locus have been implicated in the induction of a variety of neoplasms, including avian leukosis virus (ALV)-induced B-cell lymphomas in chicken (1–3), murine plasmacytomas (4–6), and human Burkitt lymphomas (5, 7). In ALV-induced lymphomas, c-*myc* expression is altered by insertion of proviral promoter and regulatory sequences located in the long terminal repeat (LTR) of the integrated provirus. In most cases, integration occurs upstream of the c-*myc* coding sequences and in the same transcriptional orientation, resulting in synthesis of tumor-specific transcripts containing both c-*myc* and viral (LTR) sequences (1). Transcription thus apparently initiates on the viral promoter under control of the viral enhancer. In a minority of tumors, the provirus is in the opposite orientation or downstream of c-*myc* (2). In these tumors, viral enhancer sequences presumably augment transcription from a cellular promoter, because the viral promoter is not appropriately positioned for c-*myc* transcription.

The coding sequences of the chicken c-*myc* locus are organized into at least two exons (8–10). The v-*myc* gene of MC29, which presumably arose by recombination between ALV and cellular *myc* sequences (11), is comprised of the two c-*myc* coding exons, plus an additional 13 nucleotides that are not contiguous with the normal chicken c-*myc* coding sequences (12–14). The coding regions of chicken c-*myc* and MC29 v-*myc* differ by only eight bases (14).

The coding exons of the c-*myc* genes of chicken, mouse, and human are highly conserved (6, 15, 16). An additional exon of approximately 500 base pairs (bp), located 1.6 kilobase pairs (kbp) upstream of the coding exons, has been identified in the c-*myc* locus of both mouse and human (6, 16, 17). This exon is not translated, as the nucleotide sequence reveals the absence of an ATG codon and the presence of multiple termination codons in all three reading frames. The function of this unusually long untranslated exon is unknown, but it has been postulated that it plays a role in the transcriptional and/or post-transcriptional regulation of c-*myc* expression (6, 18, 19). The presence of a noncoding exon in the chicken c-*myc* gene has not been reported previously, although the size of the c-*myc* transcript indicates that sequences, in addition to the coding exons, must be present in the mature mRNA (10).

To characterize the upstream region of the chicken c-*myc* gene and to determine the relationship between proviral integration sites and possible c-*myc* transcriptional control elements, we have determined the nucleotide sequence of the 2.2-kbp region upstream of the c-*myc* coding exons. We have localized an upstream exon in a region separated from the c-*myc* coding sequences by an intron of 700–800 bp. Proviral integrations in most ALV-induced lymphomas were mapped within this intron. Sequences from the upstream exon were not detected in the tumor-specific *myc* transcripts of these lymphomas.

MATERIALS AND METHODS

Nucleotide Sequence Analysis. Molecular clones of the chicken c-*myc* locus were isolated independently in two laboratories (10, 20) from the same chicken genomic library (21) and were subcloned into pBR322 or M13mp10 and -11. The strategies used for sequencing by the Maxam–Gilbert (22) and dideoxy chain-termination methods (23) are illustrated in Fig. 1. Most of the nucleotide sequence was determined by both methods independently. With one possible exception (see legend to Fig. 2), there was complete agreement between the sequences obtained by the two methods.

RNA Blot Analysis. Poly(A)$^+$ RNAs were prepared from tissues of uninfected 14- to 25-day old chickens (SPAFAS, Norwich, CT) (24) or from lymphomas induced by RAV-1, RAV-2, or td107A (25). RNA was extracted by the guanidinium isothiocyanate method (26). Glyoxal/agarose gel electrophoresis and RNA blot transfer were carried out according to published procedures (27, 28). Probes indicated in Fig. 3 were nick-translated to a specific activity of 10^8 dpm/μg (29). Glyoxalated end-labeled *Hin*dIII fragments of phage λ were used as size markers. The sizes assigned to the RNA species in these studies are somewhat lower than those reported previously (1, 25) based on ribosomal RNA size markers.

RESULTS

Nucleotide Sequence of the Region Upstream of the Chicken c-*myc* Coding Exons. The nucleotide sequence of the 2.2-kbp region upstream of the chicken c-*myc* coding exons is shown

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: ALV, avian leukosis virus; LTR, long terminal repeat; bp and kbp, base pair(s) and kilobase pair(s); kb, kilobase(s).

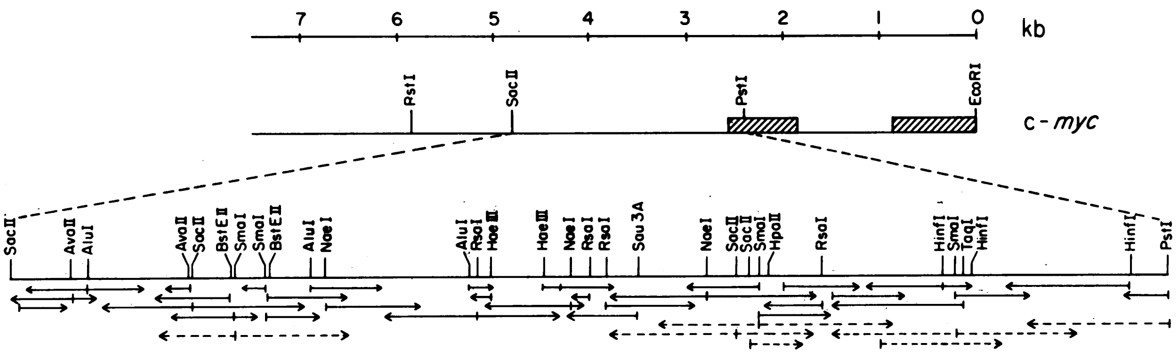

FIG. 1. Strategy for sequencing the 2.2-kbp region upstream of the c-*myc* coding exon. The hatched boxes indicate the two c-*myc* exons that are homologous to the v-*myc* gene. The *Pst* I–*Pst* I fragment was subcloned into pBR322 or M13mp10 and -11; the *Sac* II–*Pst* I region was subjected to sequence analysis. Arrows indicate the direction and extent of sequences determined by the Maxam–Gilbert (⟶) or dideoxy chain-termination (----→) methods. Only the relevent restriction endonuclease sites used in sequencing are shown.

in Fig. 2. The 5′ boundary of the first coding exon (position 2253) and the initiating AUG codon of the open reading frame (position 2268) were determined by aligning our sequence with the published sequence of the chicken c-*myc* coding region (14).

Twelve of the 13 nucleotides at the 5′ end of the v-*myc*

```
SacII      10         20         30         40         50         60         70         80         90        100        110        120
CCGCGGCGCT ATCGGGCGCC GGGAGAGGGC GCGATGGCCC CACGGTAGCG TTGGCCGTGG GAAAGCCGGG CCGCCCCCAG CGCCGGGGAA CCGCAACGGG GGGATGGATG GGGAGGGGGG

       130        140        150        160        170        180        190        200        210        220        230        240
TGCGGGGGTC CTCCCGCCCG GCGATCCGTG TTCCTCCCCG AGCTTCTACG CTTAGAAATG ATACAAATAC TTATAAGTCC GTTTGGTGTG CGTGTGTGTG TGGGGAGGAG GGGGGGGGAG

       250        260        270        280        290        300        310        320        330        340        350        360
GGGGGAGGGG GTGAAGAAAT AAATGCGAAA TAAATAAGAA ATGCATGAGA AATAGGAATG ATATATATGT ATATATCTTA GCGGGGTGCC CGCAGGTGCG GGTATTGCAG CGGGAGGGGC

       370        380        390        400        410        420        430        440        450        460        470        480
GCAGATAGCG GTCCCGCGGG GCCGGGAAGA CGCGATGCGG AGCGCCCCGT CGGGTCTCGG CTCCGCGCAC CTCCGGGGAT GGGTAACGGG GAAGGGGTGA CCCCGGGGTG GGGAAGGAGC

       490        500        510        520        530        540        550        560        570        580        590        600
CGCTGGGTGA GGGGCTGCGG AGCGAGCGGG GGGAGGTGCA GAGCCCCCGG GGGTCACCTT GCAGCCGCTC CCCCCGCAGC CTCCTCCTCC CGTTTAATCC TCCGGGATAA CGAAGCAGCG

       610        620        630        640        650        660        670        680        690        700        710        720
ACACGGGCGG GGGTGCGCGA GCTACGGACG CTCCTTTGTG CCGGTAGGGT AGCCGGCAAC CGCCCCGCCC GCAGCCGCGT TACGGGTGGA CACGGAGCGT GAACCTCCCC TGCCGCCGTC

       730        740        750        760        770        780        790        800        810        820        830        840
GGGGGGCAGC GGGAGGAGGG GGAGCGGAGA GCGAGGAGGG AAGGAGGGAG GGGGGAAGGC AGCGAGGAGG GAAGCAGCGA GGAGCGCCCT TTTCATCCCC GCTCGTATTT TTTTTTTTTT

       850        860        870        880        890        900        910        920        930        940        950 AluI   960
TTTACTATGT TTACTTCCGA CCTCTCCTTG TAGTAGGAAA AAAAACCAAC CGCTGCTCCG CATCGCCTCT CCCCGGCCCC TCTCCCTCCC TCCCTCCCTC CCGCCCGCCC AGCTCCGGCT

       970        980        990       1000       1010       1020       1030       1040       1050       1060       1070       1080
CGCAGTACTC GGGGGGGGGC ACGGAGCCCC TCGGCCGCCC CCTCGCGGCG CGCCCTCCCC GCTCACGGAG CCCGCGCGGA GCCGGGGGCG AGCGGGAGGG AGATGAAGCG GCGACGCGCA

      1090       1100       1110       1120       1130       1140       1150       1160       1170       1180       1190       1200
CCGCGAGAGC GCGCACTCGC GGGGCCCCGC CGTGCCGCTC GTGCTCCCGC CCCCGCTGCA TCTCCCGCCC GCCCCTCGCC GGCGCTTTAA AGACAGCAAA GCAACTTAAC TTCTATTGTA

      1210       1220       1230       1240       1250       1260       1270       1280       1290       1300       1310       1320
CCGGACGGAG CGCGCCGCGC CGCCTTGGAC CGTACAATCT GCCGCACGCC GGGAAGGCGA GCCCTCTCCG CTGTATTTTT TTTCTCATCG TGGTGGGAGG AAGCGATCTA CGTTCTCCCG

      1330       1340       1350       1360       1370       1380       1390       1400       1410       1420       1430       1440
GCGTTTTGGC TCCCTTTTTT CCCCCGTCTT CTCCGGCGGT TCTTTTTATA TTTTTATCTC CAATTTCCTG ATTTTGTTGT TCCCCGCACC GCCCGCAATT ATTGCCTCGC TCCGTCTCGC
                      10
      1450       1460  ↓    1470       1480       1490       1500       1510       1520       1530       1540       1550 SmaI   1560
AGCCGGCGCG TTGGAGGAGC CGGTGAGTGG CGGGGCTCCG CTTTACCCAT CACTCGATCC GCCGCGGTCG GGCTGACAGC GCGGGGCCGC GGGACCGCCG GTGCCTCCCG GGGGACGCGG

      1570       1580       1590       1600       1610       1620       1630       1640       1650       1660       1670       1680
CGCTGCTCCG GGCACGGCGC GCCCCATCCT TCTCCGTGCT CTCGGCTTGG ATATATAATT CTATTTTTTG GAGGGGGGGG GGGGGGTGGG GGAGGGCAAG AAGCATTTGC TTCTCCAGCG

      1690       1700       1710       1720       1730       1740       1750       1760       1770       1780       1790       1800
TACGCACGGC AGGTTATGCT TATTGCACAT ATATACGTAT ATATATGTGC GTGTGTGGAT ATATATGTAT ATATGATAAA TTTGGCAAAG TTTGCCCAGC TCCGTGCACG GCCAGGTGGG

      1810       1820       1830       1840       1850       1860       1870       1880       1890       1900       1910       1920
TGGCTGGGGA GCAGCCCGGC TCTGCGCTGG GAGCTACCGG CCCTTCTCGC CCGGTTCCCG GTGCCGGAGC TGGGCACCGG CTGAGCGCGG CGGCTGCGGG AGCTGTGCCC GAGCGGAGCC
                                                                                        7
      1930       1940       1950       1960       1970       1980       1990       2000  ↓    2010       2020       2030       2040
CCTCCGGAGA GTCGCGGGGA GAGCGCTCCG GGCGTCCCCG GGCGCTGACC CCTCGATGGA CGGGGTCGCG ACTCCCGGTC GCCCGCTGAG CTGGGGAGGG GGTGAGGCGG GGGGCTCACG

      2050       2060       2070       2080       2090       2100       2110       2120       2130       2140       2150       2160
AGGGGTCGTG CTTTTTATTA TTATTATTAT TTTATTATTA TTATTAGTTT ATATATATAT ATATATATAT AAATCAATCT GACGGCGCGG GGTGCCGGAG GAGCGCTGCG TGCCGAGGGT

      2170       2180       2190     * 2200       2210       2220       2230       2240       2250       2260       2270       2280
CGATCTCCCC CGCTATAGGG GCCGGGGGGA GCGGGCCTCG CGGCCCCGAG CGCGGCTCAC CCGGCCCCCC CGTGTCCCCC TCCCGCCCGC AGGCAGCAGC CGCCGCCATG CCGCTCAGCG

      2290       2300       2310       2320       2330       2340       2350       2360       2370       2380       2390  PstI
CCAGCCTCCC CAGCAAGAAC TACGATTACG ACTACGACTC GGTGCAGCCC TACTTCTACT TCGAGGAGGA GGAGGAGAAC TTCTACCTGG CGGCGCAGCA GCGGGGCAGC GAGCTGCAG
```

FIG. 2. Nucleotide sequence of the 2.2-kbp region upstream of the chicken c-*myc* coding exon. The asterisk indicates a possible sequence polymorphism (guanosine or adenosine, position 2195), based on a sequence difference in clones derived independently in two laboratories. Dashed lines indicate sequences homologous to the v-*myc* gene of MC29 virus (12–14). The coding exon of the c-*myc* gene starts at position 2253. The boxed ATG triplet in the coding exon specifies the beginning of an open reading frame. A consensus splice acceptor site (___┌‾‾) is located at the 5′ boundary of the first coding exon. Potential splice donor sites (‾‾┐__), showing at least 70% homology with published consensus sequences (30, 31), are indicated throughout the 2.2-kbp sequence. Arrows show the sites of proviral integrations in tumor 7 (position 2001) and tumor 10 (position 1461), localized by aligning the previously published nucleotide sequences of cloned virus–cell junction fragments (10). Restriction endonuclease sites, *Alu* I and *Sma* I, denote the boundaries of fragment E, which hybridizes to normal c-*myc* RNA, but not to tumor *myc*-related transcripts (see text and Fig. 3).

gene of MC29 virus, which are not contiguous with the c-*myc* coding sequences (13, 14), were located 458 nucleotides upstream of the 5′ boundary of the c-*myc* coding exons (position 1784–1795). The 12 nucleotides are adjacent to a consensus splice-donor sequence, which may have been used to join these nucleotides with the c-*myc* coding exons during the generation of MC29 virus. The origin of the 13th nucleotide (located at the extreme 5′ boundary of the v-*myc* gene) is not clear, since it does not match either the cellular c-*myc* sequence or the published viral *gag* gene sequence (32). It is likely that the mismatch reflects a polymorphism at this position in the *gag* gene, since the published sequence of this gene is derived from the Prague strain of Rous sarcoma virus and not from the (unknown) parent of MC29 virus.

Localization of an Upstream Exon of the Chicken c-*myc* Gene. The 5′ noncoding c-*myc* exon is highly conserved between human and mouse (33). However, we could find no significant sequence homology between the human or mouse first exons and the 2.2-kbp upstream region of chicken c-*myc* (data not shown). We therefore adopted the following approach to identify the upstream exon.

Six probes from the region upstream of the c-*myc* coding exons (probes A–F) and one probe from the 3′ c-*myc* coding exon (probe G) were used for RNA blot analysis of poly(A)$^+$ RNA from normal chicken thymus (Fig. 3). Probe G hybridized to an abundant 2.4-kilobase (kb) RNA, as well as to minor transcripts of 4.0, 3.4, and 3.2 kb. The same transcripts were detected in RNA from bursa and spleen (data not shown).

Probe A was derived from the 1-kbp region immediately upstream of the region that was sequenced, and probes B, C, and D were generated from the first 1 kbp of the sequenced region. None of these probes (A–D) hybridized to the chicken c-*myc* RNAs (Fig. 3).

Probe E (*Alu* I to *Sma* I) hybridized to the transcripts detected with probe G (Fig. 3). These results indicate the presence of an additional exon within a 600-nucleotide region (position 951–1547) that is located 700 bp upstream of the c-*myc* coding exons. Probe F, which includes the 12 nucleotides from the 5′ end of the MC29 v-*myc* gene, did not hybridize to the 2.4-kb c-*myc* RNA, although it did hybridize to the less abundant 4.0-kb c-*myc* RNA (Fig. 3). This result suggests that the region included in probe F is part of an intron that is spliced out to generate the 2.4-kb RNA and raises the possibility that the 4.0-kb RNA is the primary transcript that serves as precursor to the c-*myc* mRNA (2.4 kb).

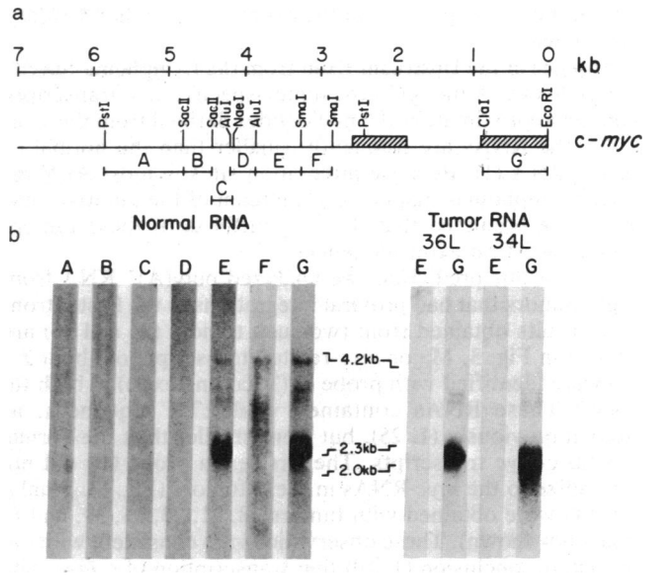

FIG. 3. Presence or absence of upstream sequences in c-*myc* mRNA from normal chicken thymus and from ALV-induced B-cell lymphomas. (*a*) Fragments used as probes and relevant restriction endonuclease sites used in generating these fragments. (*b*) Approximately 5–7 μg of poly(A)$^+$ RNA of normal or tumor tissues was used in each lane for gel electrophoresis. Early thymus was used as a source of normal c-*myc* mRNA, because c-*myc* mRNA is approximately 10-fold more abundant in this tissue than in most normal adult tissues (24). Band intensities do not necessarily reflect mRNA concentrations, as exposure times (up to 3 weeks) were adjusted to give similar levels of sensitivity for each of the RNA samples.

Proviral Integration Sites in the ALV-Induced Lymphomas. In nearly all of the ALV-induced lymphomas that we have examined, proviral sequences were located within the 1- to 1.5-kbp region upstream of the c-*myc* coding exons (see Fig. 4). A majority of the integrations (≈70%) were clustered within the 250-bp region immediately upstream of the coding exons (Fig. 4). With one exception (tumor 11), there were no proviral integrations in the remaining 450-bp intron region in this group of tumors. In five tumors, the proviruses were located either in or close to the upstream exon. Proviruses were integrated in the same transcriptional orientation as

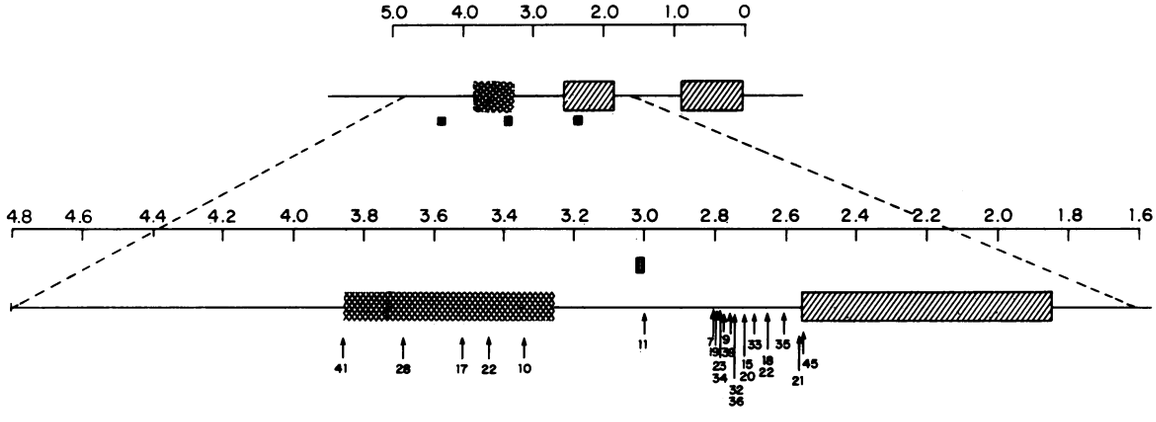

FIG. 4. Sites of proviral integrations in avian B-cell lymphomas. The two c-*myc* coding exons are designated by hatched boxes; the 12 additional nucleotides present at the 5′ end of MC29 v-*myc* are indicated above the enlarged map at 3.0 kb. The precise boundaries of the first exon have not been firmly established. The cross-hatched box shows the location of the restriction fragment used as probe to detect the first exon. A sequence within the second exon and two upstream regions that demonstrate complementarity to this sequence (positions 437–512, 1435–1544, Fig. 2) are designated by black boxes beneath the upper c-*myc* map (see *Discussion*). Computer-assisted calculations (34) indicated that the complementarity was significant (free energy of −150 to −250 kcal/mol; 1 cal = 4.18 J). Positioning of proviral integration sites in different tumors (indicated by arrows) was based on sizes of the *Eco*RI restriction fragments, as described (25). Mapping of proviruses in tumors 7 and 10 by this method was in agreement with the localizations based on sequence analysis (Fig. 2). Scale is indicated in kbp.

that of the c-*myc* gene, with the exception of tumor 41 (data not shown).

Absence of the Upstream Exon from the Lymphoma-Specific *myc* RNAs. Although most tumor-specific *myc* transcripts contain approximately 100 nucleotides derived from the viral LTR, the RNAs are frequently smaller than the normal c-*myc* mRNA (1). Because integration sites in most ALV-induced lymphomas mapped downstream of the putative first exon, we predicted that the *myc* mRNAs of these tumors would lack first exon sequences.

To test this prediction, we analyzed poly(A)$^+$ RNA from eight tumors that had proviral integrations in the first intron. The results obtained from two such tumors (34 and 36) are shown in Fig. 3. Major *myc*-related transcripts of about 2.1 kb were identified with probe G (3′ coding exon) in both tumors. These RNAs contained viral LTR sequences, as shown previously (1, 25), but were smaller than the normal 2.4-kb c-*myc* transcripts. The first exon probe (E) did not hybridize to the *myc* RNAs in these tumors (Fig. 3). Similar results were obtained with tumors 18, 21, 23, 35, 39, and 45 (data not shown). These observations are consistent with our previous conclusion (1, 10) that transcription of c-*myc* initiates on the viral promoter.

Probe E would be expected to hybridize to transcripts from the normal c-*myc* allele in ALV-induced lymphomas. The failure to detect such transcripts (Fig. 3) suggests that the normal allele is expressed at low levels (if at all) compared with the allele altered by ALV integration.

DISCUSSION

Structure of the Chicken c-*myc* Gene. The avian c-*myc* gene is composed of two coding exons interrupted by an intron of approximately 1000 bases (8–10). We have identified an additional exon included within a 600-bp region (position 951–1547) located 700 bp upstream of the coding exons. Probes derived from sequences further upstream, or immediately downstream, of this region did not hybridize to the 2.4-kb c-*myc* mRNA, suggesting that there are no other exons. Experiments in which S1 nuclease protection and nuclear runoff transcription assays were used also support this conclusion (35). If transcription initiates near the 5′ end of the region including the upstream exon, the primary c-*myc* transcript would be approximately 4 kb [assuming several hundred nucleotides of poly(A)]. This is in agreement with our observation that the largest detectable c-*myc* poly(A)$^+$ RNA is 4 kb (Fig. 3). Although we cannot rigorously exclude the possibility that additional exons, too small to be detected by any of these methods, may be present in c-*myc* mRNA, it seems probable that the chicken c-*myc* gene is composed of three exons and two introns, analogous to the structure of the human and mouse c-*myc* gene (6, 17, 18).

Two potential initiation sites, located approximately 1 and 2 kb upstream of the coding exons, were identified by *in vitro* transcription of cloned c-*myc* DNA (36). Both of these sites correspond to major DNAse I-hypersensitive sites (20). Our localization of the upstream exon suggests that the more proximal initiation site functions *in vivo*, at least in the tissue examined (normal thymus). This would place the initiation site near potential "CAT" and "TATA" sequences (37, 38), located at positions 1138–1143 (G-C-A-T-C-T) and 1166–1171 (T-T-T-A-A-A) (Fig. 2). There is no ATG codon for nearly 500 bp downstream of this putative promoter region. Furthermore, there are termination codons in all three reading frames, suggesting that the upstream exon of the chicken c-*myc* gene is noncoding and that protein synthesis initiates in the second exon, at position 2268. A possible splice donor site (30, 31) is located at position 1462–1467. The sequence at this site (G-G-T-G-A-G) is identical to that at the splice donor site that functions to join the two c-*myc* coding exons (14). If this site is used, the intron separating the noncoding and coding exons would be 790 bp. There are several other candidate splice donor sites in this region (e.g., positions 1540 and 1645), but these show lower homology with the published consensus sequences.

The nucleotide sequences of the first exons of mouse and human c-*myc* are approximately 70% homologous—a surprisingly high level of conservation for a noncoding region (18). This has led several investigators to propose that the first exon plays some critical role for which there is a strong selective pressure. However, this conservation does not extend to the chicken gene. We could find no significant homology between the noncoding exon of chicken and those of either mouse or human. In fact, no significant homology was found between the human or mouse first exons and the entire 2.2-kbp upstream region of chicken c-*myc*.

Proviral Integration Sites in ALV-Induced Lymphomas. In nearly all chicken ALV-induced lymphomas, proviral integrations are located within 1–1.5 kbp upstream of the coding sequences (Fig. 4; refs. 2, 3, and 39). [Exceptions include one tumor in which integration was downstream from c-*myc* (2), two cell lines in which integration was 2–2.5 kbp upstream of the coding exons (40), and several lymphomas with integrations even further upstream (H. L. Robinson, personal communication).]

Proviral–cellular junctions from five different lymphomas have been cloned and sequenced (refs. 10 and 41; R. A. Swift and H.-J. Kung, personal communication). By comparing these sequences with the sequence of the upstream region (Fig. 2), these junctions can be precisely localized to positions 1461, 1740,[§] 1769, 2001, and 2102 (Fig. 2).

Although in two of these cases integration occurred within a sequence of alternating A-T residues, no cellular sequences were common to all five integration sites. This suggests that proviral integration does not involve recognition of a specific nucleotide sequence. However, the distribution of integration sites in the upstream region is not random (Fig. 4). More than two-thirds of the integrations are clustered within a 250-bp region immediately upstream of the coding exons. A remarkable feature of this 250-bp sequence is a stretch of 62 A-T residues (position 2052–2114), interrupted only by a single guanosine residue. This region also corresponds to a DNase I-hypersensitive site (20). Either the low melting A-T sequence or the more open chromatin structure (or both) may make this region more accessible to proviral integrations. There are also a number of palindromic sequences in this region that might play some role in integration.

Alternatively, the nonrandom distribution of integration sites in lymphomas may reflect a selection for integration events that lead to neoplastic growth. For example, integrations distal to the coding sequences might generate functional c-*myc* mRNA only if they occur upstream of a potential splice donor sequence, which would permit removal of unnecessary (and perhaps inhibitory) intron sequences (10). All of the distal integrations that have been precisely localized (at positions 1461, 1740, 1769, and 2001) are located just upstream of consensus splice donor sequences (Fig. 2).

Altered c-*myc* Expression in Oncogenesis. In most ALV-induced lymphomas, integration has occurred in or downstream of the first exon. The normal transcriptional control sequences of the c-*myc* gene, presumably located upstream of the first exon, would thus be displaced by the viral regulatory signals in the LTR. In tumors carrying truncated c-*myc* mRNA, we could demonstrate that the c-*myc* allele altered by proviral integration was expressed at levels at least 100-

[§]The inducing virus was chicken syncytial virus, rather than ALV (H.-J. Kung, personal communication).

fold higher than the normal c-*myc* allele in the same cells (Fig. 3).

Several groups have postulated that the 5′ noncoding exon plays a regulatory role, controlling c-*myc* expression at either the transcriptional or post-transcriptional level (6, 18, 19). One model, based on a proposed secondary structure resulting from complementarity between sequences in the first and second exons, suggests a translational control of human c-*myc* gene expression (19). The corresponding second exon region of the chicken c-*myc* gene shows significant complementarity to two regions within the 2.2-kbp upstream region (Fig. 4). The significance of the complementarity, however, is unclear. One of the complementary sequences (position 1435–1544) may be partially included in the first exon, but the other (437–512) is located further upstream. Because the tumor-specific *myc* mRNAs of most ALV-induced lymphomas lack the first exon, they would not be subject to any regulation that might be conferred by the 5′ noncoding exon. Loss of the first exon is not an absolute requirement for tumor induction by ALV, however, since tumors with proviruses integrated further upstream retain first exon sequences. Similarly, in the murine plasmacytoma and human Burkitt lymphoma systems, loss of the first exon is often, but not invariably, associated with translocations involving c-*myc* (refs. 18, 35, 42).

The ways in which cellular genes may be altered in oncogenesis vary from gene to gene. In some cases, somatic mutations in the coding sequences of cellular oncogenes may result in the production of variant gene products, for example the c-*ras* gene (43). In contrast, altered regulation of gene expression most likely accounts for the transformed phenotype in other tumors. Disruption of the complex regulation of normal c-*myc* expression—by proviral integration (1–3), chromosomal translocation (4–7), or gene amplification (44–46)—appears to be the common feature of tumors in which c-*myc* has been implicated.

We would like to thank L. O'Connor for help in preparation of this manuscript, E. Prediger for assistance in analysis of nucleic acid secondary structure, N. Goldberg and K. Lewison for technical assistance, and S. Henikoff for invaluable assistance with dideoxy chain-termination sequencing. This work was supported by grants from the National Institutes of Health (CA34502), Bristol Meyers, and the Kleberg Foundation (W.S.H.) and by National Institutes of Health Grant CA18282 (M.L.). M.M.G. is the recipient of a postdoctoral fellowship from the Damon Runyon–Walter Winchell Fund.

1. Hayward, W. S., Neel, B. G. & Astrin, S. M. (1981) *Nature (London)* **290,** 475–480.
2. Payne, G. S., Bishop, J. M. & Varmus, H. E. (1982) *Nature (London)* **295,** 209–214.
3. Fung, Y.-K. T., Crittenden, L. B. & Kung, H.-J. (1982) *J. Virol.* **44,** 742–746.
4. Shen-Ong, G. L. C., Keath, E. J., Piccoli, S. P. & Cole, M. D. (1982) *Cell* **31,** 443–452.
5. Taub, R., Kirsch, I., Morton, C., Lenoir, G., Swan, D., Tronick, S., Aaronson, S. & Leder, P. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 7837–7841.
6. Stanton, L. W., Watt, R. & Marcu, K. B. (1983) *Nature (London)* **303,** 401–406.
7. Dalla-Favera, R., Bregni, M., Erikson, J., Patterson, D., Gallo, R. C. & Croce, C. M. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 7824–7827.
8. Vennstrom, B., Sheiness, D., Zabielski, J. & Bishop, J. M. (1982) *J. Virol.* **42,** 773–779.
9. Robins, T., Bister, K., Garon, C., Papas, T. & Duesberg, P. (1982) *J. Virol.* **41,** 635–642.
10. Neel, B. G., Gasic, G. P., Rogler, C. E., Skalka, A. M., Ju, G., Hishinuma, F., Papas, T., Astrin, S. M. & Hayward, W. S. (1982) *J. Virol.* **44,** 158–166.
11. Sheiness, D. & Bishop, J. M. (1979) *J. Virol.* **31,** 514–521.
12. Alitalo, K., Bishop, J. M., Smith, D. H., Chen, E. Y., Colby, W. W. & Levinson, A. D. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 100–104.
13. Reddy, E. P., Reynolds, R. K., Watson, D. K., Schultz, R. A., Lautenberger, J. & Papas, T. S. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 2500–2504.
14. Watson, D. K., Reddy, E. P., Duesberg, P. H. & Papas, T. S. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 2146–2150.
15. Watson, D. K., Psallidopoulos, E. C., Samuel, K. P., Dalla-Favera, R. & Papas, T. S. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 3642–3645.
16. Colby, W. W., Chen, E. Y., Smith, D. H. & Levinson, A. D. (1983) *Nature (London)* **301,** 722–725.
17. Watt, R., Nishikura, K., Sorrentino, J., Ar-Rushdi, A., Croce, C. M. & Rovera, G. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 6307–6311.
18. Battey, J., Moulding, C., Taub, R., Murphy, W., Stewart, T., Potter, H., Lenoir, G. & Leder, P. (1983) *Cell* **34,** 779–787.
19. Saito, H., Hayday, A. C., Wiman, K., Hayward, W. S. & Tonegawa, S. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 7476–7480.
20. Schubach, W. & Groudine, M. (1984) *Nature (London),* **307,** 702–708.
21. Dodgson, J. B., Strommer, J. & Engel, J. D. (1979) *Cell* **17,** 879–887.
22. Maxam, A. & Gilbert, W. (1980) *Methods Enzymol.* **65,** 499–560.
23. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 5463–5467.
24. Hayward, W. S., Shih, C.-K. & Moscovici, C. (1983) in *Cetus–UCLA Symposium on Tumor Viruses and Differentiation,* eds. Scolnick, E. M. & Levine, A. J. (Liss, New York), pp. 279–287.
25. Neel, B. G., Hayward, W. S., Robinson, H. L., Fang, J. & Astrin, S. M. (1981) *Cell* **23,** 323–334.
26. Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY).
27. McMaster, G. K. & Carmichael, G. G. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 4835–4838.
28. Thomas, P. S. (1980) *Proc. Natl. Acad. Sci. USA* **77,** 5201–5205.
29. Rigby, P. W. J., Dieckmann, M., Rhodes, C. & Berg, P. (1977) *J. Mol. Biol.* **113,** 237–251.
30. Lerner, M. R., Boyle, J. A., Mount, S. M., Wolin, S. L. & Steitz, J. (1980) *Nature (London)* **283,** 220–224.
31. Sharp, P. A. (1981) *Cell* **23,** 643–646.
32. Schwartz, D. E., Tizard, R. & Gilbert, W. (1983) *Cell* **32,** 853–869.
33. Neuberger, M. S. & Calabi, F. (1983) *Nature (London)* **305,** 240–243.
34. Zuker, M. & Stiegler, P. (1981) *Nucleic Acids Res.* **9,** 133–148.
35. Wiman, K. G., Clarkson, B., Hayday, A. C., Saito, H., Tonegawa, S. & Hayward, W. S. (1984) *Proc. Natl. Acad. Sci. USA,* in press.
36. Neel, B. G. (1982) Dissertation (The Rockefeller University, New York).
37. Benoist, C., O'Hare, K., Breathnach, R. & Chambon, P. (1980) *Nucleic Acids Res.* **8,** 127–142.
38. Grosveld, G. C., de Boer, E., Shewmaker, C. K. & Flavell, R. A. (1982) *Nature (London)* **295,** 120–126.
39. Rovigatti, V., Royler, C., Neel, B., Hayward, W. & Astrin, S. (1982) in *Tumor Cell Heterogeneity,* eds. Owens, A. H., Coffey, D. S. & Baylin, S. B. (Academic, New York), pp. 319–330.
40. Pachl, C., Schuback, W., Eisenman, R. & Linial, M. (1983) *Cell* **33,** 335–344.
41. Westaway, D., Payne, G. & Varmus, H. E. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 843–847.
42. Erikson, J., Finan, J., Nowell, P. & Croce, C. M. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 5611–5615.
43. Tabin, C. J., Bradly, S. M., Bargmann, C. I., Weinberg, R. A., Papageorge, A. G., Scolnick, E. M., Dhar, R., Lowy, D. L. & Chang, E. H. (1982) *Nature (London)* **300,** 143–149.
44. Dalla-Favera, R., Wong-Staal, F. & Gallo, R. C. (1982) *Nature (London)* **299,** 61–63.
45. Collins, S. & Groudine, M. (1982) *Nature (London)* **298,** 679–681.
46. Alitalo, K., Schwab, M., Lin, C. C., Varmus, H. E. & Bishop, J. M. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 1707–1711.