4960–4964 *Nucleic Acids Research, 1998, Vol. 26, No. 21*

Comparison of murine *Supt4h* and a nearly identical expressed, processed gene: evidence of sequence conservation through gene conversion extending into the untranslated regions

Pei-Wen Chiang[1,*], Ruobo Zhang[3], Lisa Stubbs[4], Liang Zhang[1], Li Zhu[3] and David M. Kurnit[1,2]

University of Michigan Medical Center, [1]Department of Pediatrics and [2]Department of Human Genetics, 1150 West Medical Center Drive, 3520 MSRB I, Ann Arbor, MI 48109-0650, USA, [3]CLONTECH Laboratories, Palo Alto, CA, USA and [4]Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA

Received June 18, 1998; Revised and Accepted September 2, 1998 DDBJ/EMBL/GenBank accession nos U96809 and U96810

ABSTRACT

We show herein the transcription of a processed gene that originated from a spliced transcript. Recently, we isolated the human and murine homologues of the yeast chromatin protein, SPT4. The *Supt4h* gene is spliced normally from five exons encoded by chromosome 11. Here we show that a related sequence on chromosome 10 encodes *Supt4h2*, a processed intronless gene (with a polyA tail and a tandemly-duplicated 13 bp insertion site in the genome) with a different 5′ control region. Both the spliced gene, *Supt4h*, and the processed gene, *Supt4h2*, are expressed in each of four tissues we examined. *Supt4h2* encodes a 117 amino acid protein nearly identical to the *Supt4h* gene product with only one amino acid difference, indicating extreme conservation of this expressed processed gene with the spliced gene over evolutionary time. This illustrates another potential complexity of the mammalian genome, i.e. the use of a processed gene under the control of a different promoter region than the spliced gene.

INTRODUCTION

The SPT4, SPT5 and SPT6 yeast proteins are a group of architectural factors that function by interacting with histones to assemble a repressive chromatin structure (1). Both histone mutants and SPT4, SPT5 and SPT6 mutants were shown to suppress *swi*/*snf* mutations in yeast (2). The yeast SPT4 gene encodes a 102 amino acid protein involved in the formation and/or maintenance of chromatin structure (1). Both yeast SPT4p and human SUPT4Hp mediate transcription elongation with SUPT4Hp interacting with RNA polymerase II (3,4). Further, Spt4p is required for faithful chromosome transmission and may play a role in kinetochore function (5).

We combined yeast genetics and the expressed sequence tag (EST) project to isolate factors (chromatin structural proteins) involved in global transcription regulation. Since many essential genes are highly conserved during evolution, the mammalian homologues of genes originally defined and isolated in yeast can be isolated by searching the EST collections. Then, full-length cDNAs can be isolated by screening cDNA libraries with the EST clones as probes. Comparative sequencing of these genes from different species then furnishes insights into evolution. Using this strategy, we have isolated the human and murine homologues of yeast SPT4, SPT5 and SPT6 (6–10). SUPT4H and *Supt4h* encode an identical 117 amino acid protein. Both the yeast SPT4 and the derived human and murine SPT4 homologues have a conserved zinc finger domain. Forty-three of the 117 amino acids in the human and murine SPT4 homologues are identical to the yeast SPT4 gene (36.8% identity; 6). In our studies of *Supt4h*, we isolated a highly homologous locus, *Supt4h2*. Here we present the isolation, characterization and sequencing of the highly-conserved and surprisingly expressed processed gene, *Supt4h2*.

MATERIALS AND METHODS

Isolation and sequencing of genomic clones

PCR primers derived from the 3′ non-translated region of the *Supt4h* cDNA sequence were used to screen a murine P1 genomic library. The forward primer was 5′-GCTGAAAAGTCGAGGA-GTGG-3′, and the reverse primer was 5′-GCAAAACCCTGAA-CACAGGT-3′. PCR amplification was conducted under the following conditions: denaturing, 92°C (30 s); annealing and extension, 65°C (90 s) for 29 cycles followed by a 5 min extension at 75°C. The isolated P1 DNAs were digested by *Eco*RI or *Xba*I and subcloned into pBluescriptII (Stratagene). Colony hybridization was performed using a probe derived from the PCR product mentioned above and the blot was hybridized at 65°C in Amersham Quick-Hyb buffer and washed in 0.1× SSC, 0.1% SDS at 65°C. The genomic clones were sequenced automatically (ABI 370A automatic sequencer and PRISM Ready Reaction kit).

*To whom correspondence should be addressed. Tel: +1 313 647 4747; Fax: +1 313 936 9353; Email: pwchiang@umich.edu

Interspecific backcross mapping

The B6xSpret (*C57BL/6J* × *Mus spretus*) backcross mapping panels were obtained from Jackson Laboratories. Oligonucleotide primers for the polymorphic *Supt4h* locus, from a region which has no homology with *Supt4h2*, are F: 5′-ACAGCTGGGTCT-CCAAGTGG-3′ and R: 5′-ACGGACACAGCATATACACC-3′. To map the *Supt4h2* gene, we used a PCR product derived from the 5′ region of the *Supt4h2* gene, which has no homology to the *Supt4h* gene. The forward primer was 5′-CTCTGCTCCTAGGT-CACTTG-3′ and the reverse primer was 5′-GTCAATCTTAAA-GATGCATG-3′. PCR amplification was conducted under the following conditions: denaturing, 92°C (30 s); annealing and extension, 65°C (90 s) for 29 cycles followed by a 5 min extension at 75°C. The probe was hybridized to *Eco*RV-digested DNA of 160 IB progeny (11) [(*C3Hf/R1-Mgf*$^{Sl-ENURg}$/+ × *M.spretus*) × *C3Hf/Rl*]. The segregation of a variant 6.2 kb *M.spretus* restriction fragment (compared to a corresponding *Mus musculus* fragment 4.2 kb in length) was used to map *Supt4h2*. *Myb* was traced in this cross through analysis of a 4.4 kb variant fragment detected in *M.spretus Taq*I digests; a PCR-generated fragment, corresponding to exon 5 of the gene was used as probe. *Hk1* was followed using a variant *Bam*HI fragment (8 kb) detected by a mouse embryo cDNA probe (IMAGE clone ID 437155; obtained from Research Genetics, Inc.). This cDNA was identified by its close match between an EST derived from the clone and the published sequence of full length mouse hexokinase type I cDNA (DDBJ/EMBL/GenBank accession no. J05277). DNA preparations, Southern blots and hybridizations were carried out as previously described (12). Mapping data were stored and map positions, with standard errors, were calculated using standard statistical methods (13) with the aid of the Map Manager data analysis program (14).

Analysis of tissue expression pattern of *Supt4h* and *Supt4h2*

CLONTECH's Quick-Clone cDNAs (derived from poly A+ RNAs) from mouse brain, heart, liver and 17-day embryo were used. Long PCR was performed using CLONTECH's Advantage *KlenTaq* DNA Polymerase mix that contains antibody against *Taq* DNA polymerase for mimicking a 'hot-start' PCR reaction. In the 50 µl PCR reaction mixtures, we used 1.5 µl of Quick-Clone cDNA, 1 µl of DNA polymerase mix, 1 µl of dNTPs at 10 mM each, 5 µl of 10× PCR reaction buffer containing 100 mM Tris–HCl (pH 8.3), 500 mM KCl, and 15 mM $MgCl_2$. PCR primers were designed from the common region of *Supt4h* and *Supt4h2*. The forward primer was 5′-AACCTGTGTTGTCGCATCGG-3′ and the reverse primer was 5′-CCAAAGTCCTGCCCATAGAC-3′. This primer pair amplified a 649 bp fragment from both *Supt4h* and *Supt4h2*. The PCR amplification was conducted under the following conditions: 94°C for 30 s, followed by 65°C for 3 min for a total of 32 cycles. After PCR, the amplified product was directly digested with either *Age*I or *Bst*BI (New England BioLabs). We used 5 µl of PCR product for digestion in a 20 µl volume for 2 h at 37°C. The digestion was analyzed on a 2% agarose gel.

RESULTS

Isolation and sequencing *Supt4h* and *Supt4h2* genomic clones

To isolate genomic sequences encompassing *Supt4h*, three independent P1 clones were isolated using PCR primers derived from the 3′ non-translated region of *Supt4h*. One of the P1 clones showed a different restriction enzyme digestion pattern than the other two P1 clones (data not shown). Portions of the P1 clones containing murine genomic DNA related to *Supt4h* were sequenced and analyzed using the BLAST (homology searching; 15,16) and TSSW programs (17; recognition of human PolII promoter region and transcription start site). Sequences of 11 327 bp (*Eco*RI fragment) and 3968 bp (*Xba*I fragment) were obtained from the two different P1 clones (Materials and Methods). The 11 327 bp murine genomic DNA sequence contained five exons of the *Supt4h* gene (Fig. 1). Exon 1 is 105 bp, exon 2 is 107 bp, exon 3 is 56 bp, exon 4 is 54 bp and exon 5 is 371 bp. The TSSW program successfully predicted a promoter region from this 11 327 bp sequence at nt 1794 (Fig. 1). The 3968 bp fragment contained a region which is almost identical to the *Supt4h* cDNA sequence without introns and with only six nucleotide differences in the coding and non-coding regions together (underlined in Fig. 1).

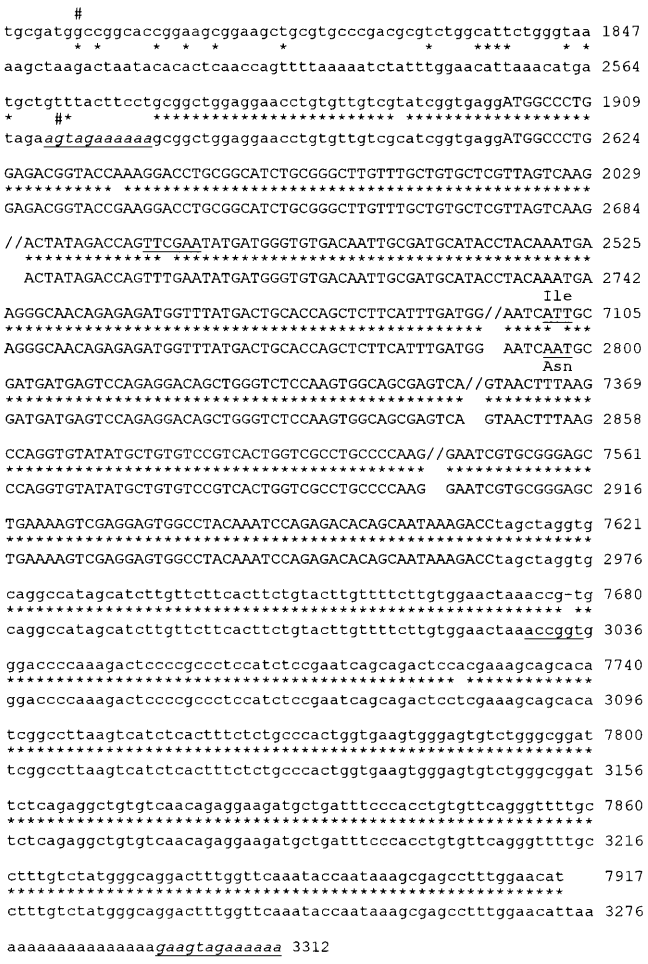

Figure 1. Sequence comparison and promoter predictions for the *Supt4h* (upper) and *Supt4h2* (lower) genes. The identical nucleotides between these two sequences are marked by *. The exon–intron boundaries for *Supt4h* are marked by (//). The 13 nucleotide tandem repeats at the 5′ and 3′ ends of *Supt4h2* are underlined and italicized. The start of the promoter for *Supt4h* was predicted by TSSW at nt 1794 (marked by # on top of the sequence) and the start of the promoter for *Supt4h2* was predicted by TSSW at nt 2570 (marked by # on top of the sequence). The unique *Age*I and *Bst*BI restriction sites are underlined. The protein coding region is capitalized and the only amino acid difference between *Supt4h* and *Supt4h2* is underlined.

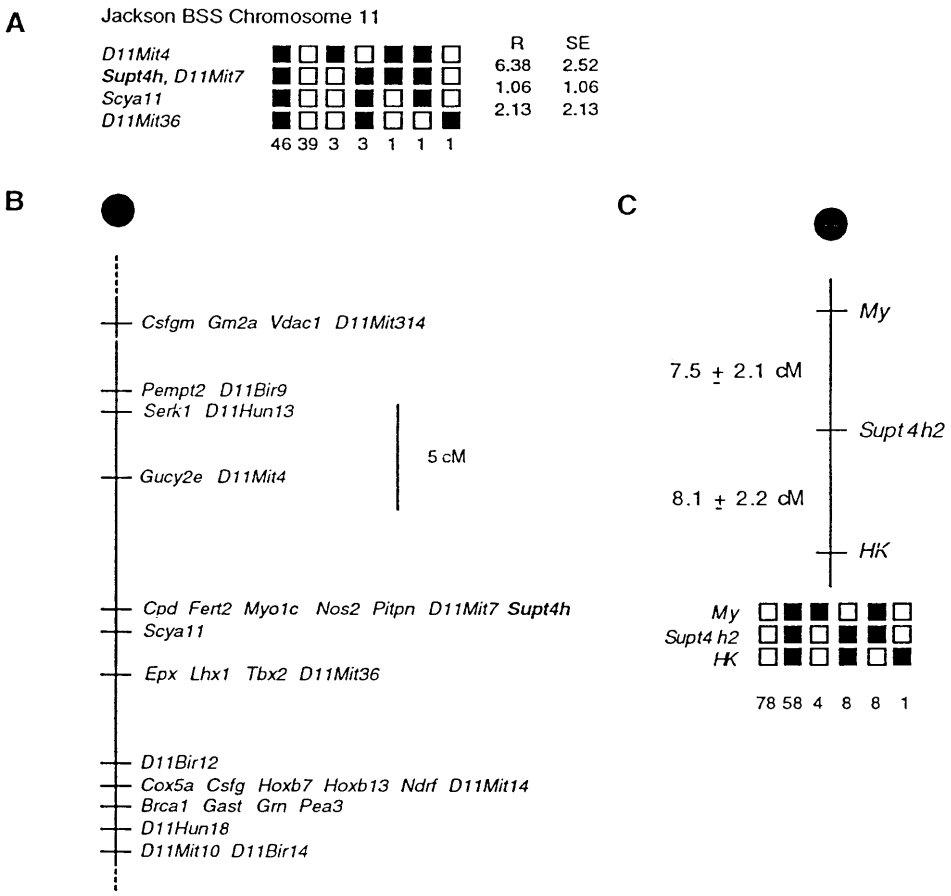

Figure 2. (**A**) Haplotype figure from The Jackson BSS backcross showing part of chromosome 11 with loci linked to *Supt4h*. Loci are listed in order with the most proximal at the top. The black boxes represent the *C57BL6/JEi* allele and the white boxes the *SPRET/Ei* allele. The number of animals with each haplotype is given at the bottom of each column of boxes. The percent recombination (R) between adjacent loci is given to the right of the figure, with the standard error (SE) for each R. Missing typings were inferred from surrounding data where assignment was unambiguous. (**B**) Map figure of central chromosome 11, using data from the Jackson BSS cross. The map is depicted with the centromere toward the top. A 5 cM scale bar is shown to the right of the figure. Loci mapping to the same position are listed in arbitrary order. Missing typings were inferred from surrounding data where assignment was unambiguous. Raw data from The Jackson Laboratory were obtained from the World Wide Web address http://www.jax.org/resources/documents/cmdata . (**C**) Assignment of the processed *Supt4h2* gene copy to proximal mouse chromosome 10. The haplotype figure represents the types of chromosomes observed in the interspecific backcross progeny, with numbers below each column representing numbers of animals observed to carry each type, as described above. Black boxes represent *M.spretus* alleles for a particular gene, while white boxes represent alleles from the C3Hf parent of this cross. The position deduced for *Supt4h2* from these data are depicted above, with calculated distances (given with standard errors in parentheses) between each pair of probes shown at left of the map.

The second, third and fourth nucleotide differences are located in the coding region, and only the fourth of these yields an amino acid difference (asparagine to isoleucine). The sequence encoded an uninterrupted open reading frame, suggesting that *Supt4h2* was expressed rather than being a pseudogene. *Supt4h2* potentially encodes the same size protein (117 amino acids) as *Supt4h*, with only one amino acid difference between the two gene copies. Fortuitously, the third and fifth DNA sequence variations create differences in restriction enzyme digestion patterns between *Supt4h* and *Supt4h2*. The nucleotide at position 2699 of *Supt4h2* is a T instead of a C (the third nucleotide in the six different nucleotides underlined in Fig. 1). This change deletes a *Bst*BI recognition site (5′-TTCGAA-3′) that is present in the *Supt4h* sequence. *Supt4h2* also contains an extra G at nt 3033, in the untranslated region (the fifth nucleotide in the six different nucleotides underlined in Fig. 1), creating an *Age*I site (5′-ACCGGT-3′). Therefore, the *Bst*BI site is unique to the *Supt4h* cDNA and the *Age*I site is unique to the *Supt4h2* cDNA.

The sequence comparison also demonstrates the presence of a duplicated region (13 bp) at the 5′ and 3′ ends of the *Supt4h2* sequence and a 25 nt poly (A) stretch in front of the 3′ duplicated region. This suggests that *Supt4h2* arose as reintegration of a processed gene encoded by *Supt4h* (Fig. 1).

Mapping of *Supt4h* and *Supt4h2*

The *Supt4h* gene was mapped using a microsatellite marker located in intron 3 (Fig. 2A and B). Oligonucleotide primers flanking the microsatellite identified a DNA-length polymorphism between *C57BL/6J* and *M.spretus*. The chromosomal location of the *Supt4h* gene was then obtained using an interspecific backcross [(*C57BL/6* × *Mus spretus*)F1 × *Mus spretus*; the BSS panel from the Jackson Laboratories] consisting of 94 first-generation backcross progeny. The result of this experiment indicates that *Supt4h* is located in the central region of mouse chromosome 11, since no recombination events were identified

between *Supt4h* and *D11Mit36*. This location placed the *Supt4h* gene ~6.5 cM centromeric to the *Brca1* gene (BSS chromosome 11 map).

Since no microsatellites were identified near the processed *Supt4h2* gene copy, we mapped this sequence by hybridization of a probe derived from the 5′ region of *Supt4h2*, which has no homology to the *Supt4h* gene (Fig. 2C). The *Supt4h2* probe was hybridized to *Eco*RV-digested DNA samples representing a third interspecific backcross (IB; 11). The results of this analysis localized *Supt4h2* to mouse chromosome 10, between the previously-mapped genes, *Myb* and *Hk1*, ~10 and 7 cM from those two genes, respectively. *Myb* is located in the center of a large region of syntenic homology to human chromosome 6q21–q24, while *Hk1* maps to the proximal region of a neighboring interval that is related to human chromosome 10q21–q22 (11,18,19).

Expression pattern analysis of *Supt4h* and *Supt4h2*

Based solely on the prediction of the TSSW program, there is a putative promoter region for *Supt4h2* starting at nt 2570 (Fig. 1). There is no similarity between the promoter regions of *Supt4h* and *Supt4h2* (Fig. 1), so that the two sequences could have different expression profiles. To determine if *Supt4h2* is indeed expressed, we analyzed Marathon cDNAs from day 17 mouse embryos and from adult mouse brain, heart and liver. The Marathon cDNAs were generated from poly A+ RNA isolated from the respective mouse tissues. PCR primers were derived from the common regions of *Supt4h* and *Supt4h2*, yielding a 649 bp product from both genes. PCR was performed using CLONTECH's Advantage *KlenTaq* DNA polymerase (which has proofreading activity) and followed by diagnostic restriction enzyme digestion to determine which sequences were expressed. As shown in Figure 3, both *Supt4h* and *Supt4h2* are expressed in each tissue. Each of the cDNAs is digested by both *Age*I and *Bst*BI. As expected, if both *Supt4h* and *Supt4h2* were expressed, only partial digestion was achieved with each enzyme. Thus, a detectable fraction of the PCR product showed the predicted *Supt4h2* digestion pattern in all four tissues.

To ensure that the restriction digests indeed detected transcripts from both *Supt4h* and *Supt4h2*, we cloned the PCR products derived from mouse heart into the TA cloning vector, pCRII (Invitrogen). Multiple cDNA clones were isolated from the heart tissue and PCR was performed to amplify the inserts from each of these clones. Some inserts were digested solely by *Age*I and some solely by *Bst*BI. Sequence analysis of the clones giving the *Bst*BI-only or the *Age*I-only restriction digest patterns confirmed that the nucleotide differences observed indeed represented transcription from *Supt4h* and *Supt4h2*.

DISCUSSION

The presence of rare functional, but diverged, processed genes has been documented. The human testis-specific PGK gene (PGK-2) is a functional processed autosomal gene which shows 85 and 87% homology, respectively, to the X-linked constitutively-expressed PGK gene (PGK-1) at the nucleotide and amino acid levels, respectively (20). A processed N-myc gene (N-myc2) was isolated from woodchuck and ground squirrel with ~80% identity to the N-myc1 gene at the protein level (21). In both cases, the processed genes diverged from the original genes.

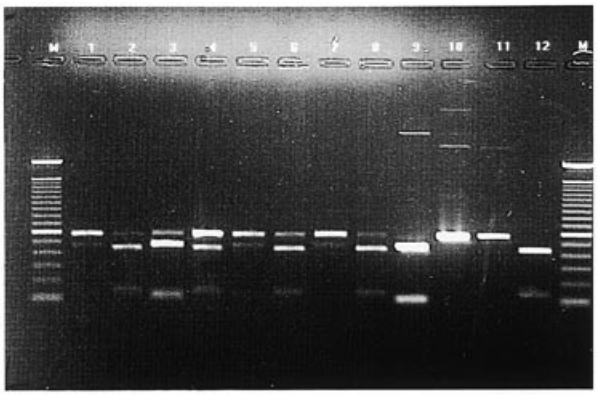

Figure 3. Expression pattern analysis of *Supt4h* and *Supt4h2*. The PCR products were amplified from CLONTECH's Quick-Clone cDNAs from adult mouse brain (lanes 1 and 2), heart (lanes 3 and 4), liver (lanes 5 and 6) and mouse 17-day embryo (lanes 7 and 8). Lanes 9–12 are controls with *Supt4h* PCR product (lanes 9–10) and *Supt4h2* PCR product (lanes 11–12). The PCR products from *Supt4h* and *Supt4h2* were cloned into pCRII vector and the identity of the isolated clones were confirmed by sequencing. Both controls were derived by PCR using the isolated pCRII clones mentioned above. The *Age*I site is unique to the *Supt4h2* cDNA (lanes 2, 4, 6, 8 and 12) and the *Bst*BI site is unique to the *Supt4h* cDNA (lanes 1, 3, 5, 7 and 9). The smaller *Bst*BI digestion product could be seen after darker staining (data not shown). The relative intensity of the two bands cannot be taken to reflect the transcription status *in vivo*.

In contrast, the extreme conservation between *Supt4h* and *Supt4h2*, extending from the 5′-untranslated region through the poly A tail in the 3′ untranslated region, resembles the high degree of conservation that marks the translated portion of the histone gene families (22,23). It is possible that the high demand of histone mRNA during S phase requires multiple copies of histone genes. Although the coding regions of the histone genes are almost identical between different copies, the non-coding regions are significantly more divergent. For example, the three histone H2b genes in mice differ by two amino acids from each other (24). This similarity of the nucleotide sequences in the coding region of the histones likely results from gene conversion events targeted at the coding region. The near identity between *Supt4h* and *Supt4h2* presumably also results from gene conversion. However, unlike the case for *Supt4h* and *Supt4h2*, the homology amongst the histone genes does not extend to the untranslated regions and there is no expression of a processed gene similar to *Supt4h2*. The conservation of the untranslated regions suggests that these regions are important for *Supt4h* function.

We do not know the significance of the relative frequency of messages derived from both the spliced *Supt4h* gene and the processed *Supt4h2* gene in different tissues. Nevertheless, we have observed the remarkable finding that a processed gene is both conserved and expressed in all four tissues we examined. It is possible that the one amino acid difference between *Supt4hp* and *Supt4h2p* is significant.

In summary, a processed gene, *Supt4h2*, shares virtually total homology with the spliced murine *Supt4h* gene. Computer comparison of diverged sequence just 5′ to the genes shows that the original *Supt4h* gene and the processed *Supt4h2* gene have different control sequences. Both genes are expressed, albeit at varying levels, in different tissues. The extreme conservation between the two genes demonstrates that gene conversion and

selection have exerted tight control to prevent significant diversion. The serendipitous placement of a highly conserved processed gene under the control of a different promoter than used by the spliced gene highlights the potential complexity of the mammalian genome.

ACKNOWLEDGEMENTS

L.S. was supported by a grant from the US Department of Energy (awarded under contract DE-AC05-96OR22464 with Lockheed-Martin Energy Systems, Inc.). P.-W.C. and D.M.K. were supported by NIH grants R01 HL50025 and R42 CA77235. We thank Beverly Selmer for expert technical assistance, and Dr Micheal Mucenski and Hsi-Hsien Lin for kindly providing the *Myb* gene probe. We also thank L. Rowe and M. Barter for Figure 2 and the University of Michigan DNA sequencing core.

REFERENCES

1 Winston,F. (1992) In McKnight,S.L. and Yamamoto,K.R. (eds), *Transcriptional Regulation.* Cold Spring Harbor Laboratory Press, NY, pp. 1271–1293.
2 Winston,F. and Carlson,M. (1992) *Trends Genet.*, **8**, 387–391.
3 Hartzog,G.A., Wada,T., Handa,H. and Winston,F. (1998) *Genes Dev.*, **12**, 357–369.
4 Wada,T., Takagi,T., Yamaguchi,Y., Ferdous,A., Imai,T., Hirose,S., Sugimoto,S., Yano,K., Hartzog,G.A., Winston,F., Buratowski,S. and Handa,H. (1998) *Genes Dev.*, **12**, 343–356.
5 Basrai,M.A., Kingsbury,J., Koshland,D., Spencer,F. and Hieter,P. (1996) *Mol. Cell. Biol.*, **16**, 2838–2847.
6 Chiang,P.-W., Wang,S.-Q., Smithivas,P., Song,W.-J., Crombez,E., Akhtar,A., Im,R., Greenfield,J., Ramamoorthy,S., Van Keuren,M. *et al.* (1996) *Genomics*, **34**, 368–375.
7 Chiang,P.-W., Qu,X., Jackson,C.L., Wang,S.-Q. and Kurnit,D.M. (1996) *Genomics*, **38**, 421–424.
8 Chiang,P.-W., Wang,S.-Q., Smithivas,P., Song,W.-J., Ramamoorthy,S., Hillman,J., Puett,S., Van Keuren,M.L., Crombez,E. Kumar,A. *et al.* (1996) *Genomics*, **34**, 368–375.
9 Chiang,P.-W., Baldacci,P.A., Babinet,C., Camper,S.A., Watkins-Chow,D., Baker,D.D., Tsai,C.H., Ramamoorthy,S., King,E., Slack,A.C. *et al.* (1996) *Mammal. Genome*, **7**, 459–460.
10 Chiang,P.-W., Stubbs,L., Zhang,L. and Kurnit,D.M. (1998) *Genomics*, **47**, 426–428.
11 Stubbs,L., Carver,E.A., Shannon,M.E., Kim,J. and Geisler,J. (1996) *Genomics*, **35**, 499–508.
12 Stubbs,L., Poustka,A., Baron,A., Lehrach,H., Lonai,P. and Duboule,D. (1990) *Genomics*, **7**, 422–427.
13 Silver,J. (1985) *J. Heredity*, **76**, 436–440.
14 Manly,K.F.A. (1993) *Mammal. Genome*, **4**, 303–313.
15 Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) *J. Mol. Biol.*, **215**, 403–410.
16 Gish,W. and States,D.J. (1993) *Nature Genet.*, **3**, 266–272.
17 Solovyev,V.V., Salamov,A.A. and Lawrence,C.B. (1998) *ISMB*, **3**, 367–375.
18 DeBry,R.W. and Seldin,M.F. (1996) *Genomics*, **33**, 337–351.
19 Taylor,B.A., Burmeister,M. and Bryda,E.C. (1997) Chromosome Committee Reports: Mouse chromosome 10. Mouse Genome Database (MGD), Mouse Genome Informatics, The Jackson Laboratory, Bar Harbor, ME. (http://www. informatics. jax. org/)
20 McCarrey,J.R. and Thomas,K. (1987) *Nature*, **326**, 501–505.
21 Quignon,F., Renard,C.A., Tiollais,P., Buendia,M.A. and Transy,C. (1996) *Oncogene*, **12**, 2011–2017.
22 Wang,Z.-F., Krasikov,T., Frey,M.R., Wang,J., Matera,A.G. and Marzluff,W.F. (1996) *Genome Res.*, **6**, 688–701.
23 Wang,Z.-F., Tisovec,R., Debry,R.W., Frey,M.R., Matera,A.G. and Marzluff,W.F. (1996) *Genome Res.*, **6**, 702–714.
24 Liu,T.-J., Liu,L. and Marzluff,W.F. (1987) *Nucleic Acids Res.*, **15**, 3023–3039.