Proc. Natl. Acad. Sci. USA
Vol. 90, pp. 10977–10981, December 1993
Genetics

Transcriptional organization of a 450-kb region of the human X chromosome in Xq28

(CpG islands/cDNA/human genome/human genes)

S. Bione*, F. Tamanini*, E. Maestrini*, C. Tribioli*, A. Poustka†, G. Torri*, S. Rivella*, and D. Toniolo*

*Istituto di Genetica Biochimica ed Evoluzionistica, Consiglio Nazionale delle Ricerche, 27100 Pavia, Italy; and †German Cancer Research Center, 6900 Heidelberg, Germany

Communicated by Renato Dulbecco, August 23, 1993 (received for review June 14, 1993)

ABSTRACT In this paper, we report the transcriptional organization of a 450-kb gene cluster in Xq28, flanked by the glucose-6-phosphate dehydrogenase and the color vision genes. CpG islands previously identified and mapped to distal Xq28 have helped in construction of a continuous contig of cosmids and in identification of cDNAs corresponding to eight transcripts. Thirteen to 16 small genes with CpG islands are clustered in a region of 250–300 kb. Many are highly expressed in muscle or brain and may be the genes responsible for muscle or neurological disorders mapped to distal Xq28. Our analysis indicates that, in this region of the genome, genes not related in sequence are organized in transcriptional domains of 100 kb and that this organization may be important for establishing and regulating gene expression in relation to tissue distribution and X chromosome inactivation.

The construction of physical and transcriptional maps of mammalian genomes, providing clues to a better understanding of genome organization, may also shed light on how chromosomal position can influence gene expression. In viruses and prokaryotes the position of genes is important and often essential for regulating gene expression. In higher organisms, however, the increasing complexity of genomes makes the significance of gene order less obvious. Genes are not dispersed along the genome; rather, they appear to be clustered in specific regions, mainly in the G-positive bands or near the telomeres (1). In some instances (e.g., the homeobox or the globin genes), genes related in function are arranged in groups along the chromosome and in the same topological order in which they are expressed (2, 3). Disruption of such order in the globin gene complex has profound effects on its regulated expression (4). Gene order of apparently unrelated transcripts is also often maintained (5). Evolution via chromosomal rearrangements and genome duplication is the main factor responsible for keeping genes closely linked in different species, but functional relationships may also be important in maintaining gene order. The mammalian X chromosome, whose high conservation in evolution may be related to the mechanism of gene dosage compensation via chromosome inactivation, is the best example of conserved gene association (5), and it may be a special case of the wider phenomenon of genomic imprinting (6, 7).

To study in detail the organization of X chromosome genes, we isolated and localized on the physical map of the human X chromosome a large number of CpG islands (8–10). Since many were clustered in distal Xq28, we could construct a detailed physical map of 450 kb of the CpG island-rich region between the glucose-6-phosphate dehydrogenase and color vision genes and identify cDNAs corresponding to eight additional genes. A transcriptional map of the region could therefore be determined together with the preliminary functional characterization of the newly discovered genes.

MATERIALS AND METHODS

cDNA Isolation. The human fetal brain cDNA library in λZAP was bought from Stratagene. The human NTERA2/D1 cDNA library in λgt11 was a gift of M. G. Persico (Naples). Plaques of each library (2×10^6) were screened by hybridization: cDNAs 2-19, 6-3, 9F, and 1A were obtained from the human fetal brain library; cDNAs G4.5, G4.8, 16A, and STA were obtained from the NTERA2/DI cDNA library. Positive plaques were purified and subcloned in Bluescript.

Cell Lines, Cosmids, and Probes. Cell lines and the human–hamster hybrids used in this work have been described (8). Cosmids and CpG island probes were also described (8–10). Cosmids in the *CV* gene region were obtained by R. Feil and J. L. Mandel (Strasbourg, France). Restriction digestions and PCRs were performed as suggested by the suppliers (Promega/New England Biolabs).

RNAs and Northern Blot. RNAs from cell lines were prepared by the guanidine isothiocyanate/CsCl method (11); RNAs from human tissues and poly(A)$^+$ Northern blot were purchased from Clontech. Total RNAs were fractionated in formaldehyde/agarose gels, transferred to nylon filters, and hybridized by standard techniques (11).

Reverse Transcriptase-PCR. Total RNA (1 mg) was reverse transcribed with Moloney murine leukemia virus reverse transcriptase at 37°C in the buffer and conditions suggested by the supplier (BRL), using random hexamers (Promega) as primers or the following oligonucleotides: 16A, A = GGCTCGAGTGCGGATGG; B = TCCCAAACTCAGGGAGC; 2-19, A = GGAGAAGTGGACATGAG; B = CTCATGAAGCAGCCACC. About 1/10th of the reaction mixture was amplified with *Taq* polymerase (Promega) in the buffer supplied with 200 mM dNTP and 0.2–0.5 mM primers. PCRs were 1 min at 94°C, 2 min at 52°C–57°C, and 4 min at 72°C. Primers were as follows: 16A, C = CACTTCTGACCACTTAC; D = AGGGTCACACAATCTGG; 2-19, C = ACCAGTCCAGAGAGCTC; D = TGCTCCCAGAGATGCAC.

Nucleotide Sequence. cDNAs were sequenced from the vector primers as well as from oligonucleotides (17-mers) designed from the sequence using *Taq* polymerase, with the *fmol* DNA sequencing system (Promega) or CircumVent DNA sequencing kit (New England Biolabs).

RESULTS

Physical Map of the Region Between the *G6PD* and *CV* Genes. By hybridization of probes for CpG islands to an Xq28-specific cosmid library, contigs were obtained (10). One (150 kb) contained the *G6PD* gene; a second (150 kb) was placed by pulsed-field gel electrophoresis (PFGE) between

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.

the *G6PD* and the *CV* genes. The sum of the length of the DNA in the two contigs was approximately the distance between the two genes, as determined by PFGE (10, 12). A closer look at the restriction map suggested that the two contigs may be overlapping with themselves and with a cosmid contig in the *CV* gene region constructed by Feil *et al.* (13). Probe 25 from cosmid 18B4-1 (13) was hybridized to the contig and the overlap was verified (Fig. 1). Hybridization of the ends of cosmids D and C (Fig. 1*A*) to a panel of human, hamster, and human–hamster hybrid genomic DNA demonstrated that one of the ends of cosmid D was of hamster origin (boxed in Fig. 1*A*). The overlap between the two cosmids, 10 kb from the rearranged end of D, was confirmed by hybridization of the end fragment from cosmid C to cosmid D (data not shown).

Thus, a continuous DNA contig linked the *G6PD* to the *CV* genes (Fig. 1*A*). In the contig, in addition to the GdX, P3, QM, and ABP-280 genes previously identified (14–16) 11 CpG islands and possible newly identified genes were mapped.

Isolation of Conserved Sequences. CpG islands were separated by 5–20 kb of DNA, a region large enough to code for a transcript. DNA fragments prepared by cutting cosmids with *Xho* I or *Eco*RV, which digested each cosmid only a few times, were fractionated with *Pvu* II or *Sma* I and subcloned. The subclones were hybridized to DNA of different animal species digested with *Eco*RI and many conserved fragments were observed (data not shown). Conserved fragments are shown as solid boxes in Fig. 1*A*.

cDNA Isolation and Characterization. Conserved subclones were hybridized one by one or in groups to two cDNA libraries; one was from human fetal brain, the second was from the undifferentiated human teratocarcinoma cell line NTERA2/D1 (17). One to three cDNAs were isolated with all probes with the exception of probes 2 and 13. Each cDNA was used to probe the corresponding cosmid and genomic DNA from human, hamster, and human–hamster hybrids carrying only the human X chromosome or portions of the X chromosome (8). All the cDNAs were single copy sequences and mapped exclusively to Xq28 (data not shown). In all instances, they hybridized to the corresponding cosmid and were localized between two CpG islands (Fig. 1*B*). This finding confirmed the presence of eight newly discovered genes in the region and of a minimum of 13 transcripts in 250–300 kb of DNA.

Organization of the Genes. To study the organization of this large number of genes their direction of transcription was established.

The partial nucleotide sequence of the cDNAs (data not shown) demonstrated in some of them (STA, G4.5, 1A, and 9F) a poly(A) tail that identified their 3′ end (Fig. 1*B*).

Two cDNAs, 2-19 and 16A, were oriented by PCR. From a partial nucleotide sequence, oligonucleotide primers were designed (primers A and B in Fig. 2*B*) and used separately to prime reverse transcription of total RNA of HeLa (for 2-19) or intestine (for 16A) cell. The products of the two reactions were amplified by using internal primers (C and D in Fig. 2*B*), the sense strand being the one from which a cDNA could be amplified (Fig. 2*A*).

The two remaining cDNAs, G4.8 and 6.3, were oriented by sequencing the cDNA and the corresponding cosmid. From the nucleotide sequence of the cDNAs, oligonucleotides

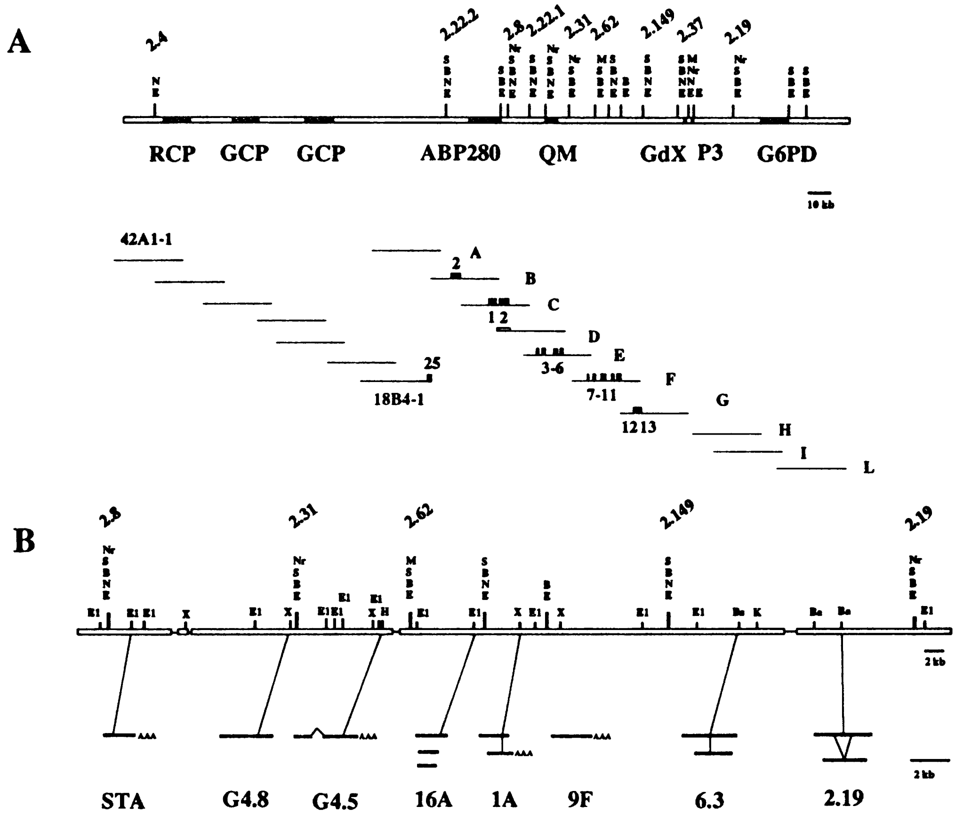

FIG. 1. (*A*) Physical map of the genomic region between *CV* and *G6PD* genes. Cosmids A–L were isolated in our laboratory from an Xq28-specific cosmid library (10). Cosmids in the RCP/GCP region are from Feil *et al.* (13). Probe 25, used to verify the overlap, is indicated above the cosmid 18B4-1. Solid boxes numbered 1–13 above the other cosmids are conserved regions. The position of known genes is indicated by slashed boxes. Above are the CpG islands identified either by a probe name (8) or by the presence of a cluster of rare cutter sites only. Only clustered rare cutter sites that correspond to CpG islands are indicated, but additional sites are present in the region. E, *Eag* I; N, *Not* I; B, *Bss*HII; Nr, *Nru* I; S, *Sac* II; M, *Mlu* I. (*B*) Schematic representation and localization of the cDNAs below an enlargement of the physical map of the region. In addition to the rare cutter sites some of the *Eco*RI (E1), *Xho* I (X), *Bst*EII (Bs), and *Bam*HI (Ba) sites used to localize and orient the cDNAs are indicated. Below the physical map is the position of the cDNAs. The poly(A) tail is indicated by (AAA) at one end of some of the cDNAs.

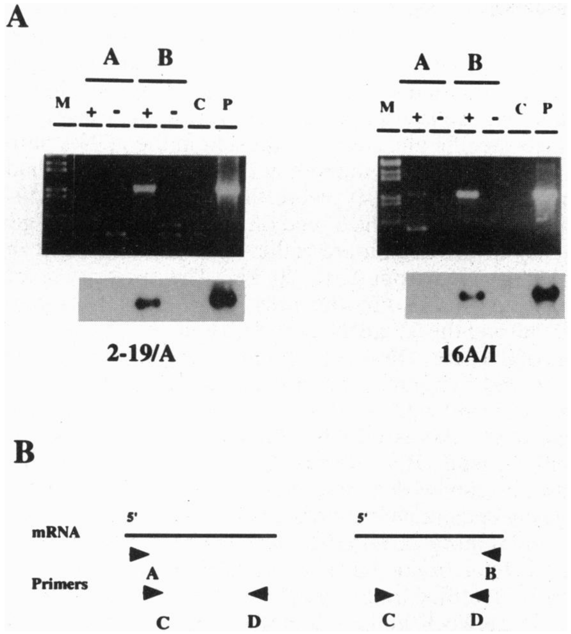

FIG. 2. Determination of the 5′ end of the cDNAs 2-19 and 16A by reverse transcriptase (RT)-PCR. (*A*) PCR amplification with primers C and D, prepared from the nucleotide sequence of each cDNA, of the RT reaction mixture primed with primer A or B. Lanes: +, reaction with RT; −, control without RT; C, PCR control without DNA; P, PCR product from the cDNA; M, phage λ digested with *Eco*RI and *Hin*dIII. Below the ethidium bromide-stained gel is hybridization of the same gel, transferred to filter, to the corresponding cDNA. (*B*) Scheme of the experiments and of the primers used.

were synthesized to prime sequence reactions from the cDNAs and from cosmid DNA. The procedure was repeated until the two DNA sequences diverged. From the sequence of the cDNA, past the splice junction, an oligonucleotide was derived to prime nucleotide sequencing in the opposite direction and thus sequence both 5′ and 3′ intron splice sequences. The sequence at the two splice junctions (shown in Fig. 3) established the 5′ end of the cDNAs.

The 5′ end of each cDNA was mapped with respect to flanking CpG islands by restriction analysis and hybridization. A unique restriction site was used to prepare asymmetric fragments that were hybridized to the appropriate digestion of the cosmid. The 5′ end of each cDNA corresponded to a different CpG island, with the exception of the 2-31 CpG island, which was at the 5′ end of two cDNAs, G4.5 and G4.8.

In Fig. 4 the direction of transcription of the newly discovered genes and of the genes previously identified is schematically shown. Genes with the same direction of transcription are not randomly distributed in the region but they are grouped in DNA traits of ≈100 kb.

A

```
cDNA G4.8   AGAGGAGGAG----------------------------GCCCTCAAC
Cos E       AGAGGAGGAGgtgaaggg--------attcccagGCCCTCAAC
```

B

```
cDNA 6.3    GTTCAGTCTC----------------------------GTGTACCAG
Cos F       GTTCAGTCTCgtgcgtga---------ctccccagGTGTACCAG
```

FIG. 3. Nucleotide sequence around a splice junction of the cDNAs G4.8 and 6.3 and of the corresponding cosmids E and F. Intron sequences are in lowercase letters.

Northern Blot Analysis. To gain information on the function of the new genes, cDNAs were hybridized to total RNA from 10 different human cell lines and tumors, and from 10 different human tissues. The results of some of the hybridizations are shown in Fig. 5.

9F and 2-19 cDNAs were expressed in similar amounts in all cell lines and tissues (data not shown). RNA hybridizing to STA had a similar distribution, but a higher amount was present in muscle (Fig. 5*A*). The remaining cDNAs were expressed in most cell lines but to a different extent: different sized RNA bands hybridized to G4.8 and G4.5 and the distribution of the bands was different in different cell lines (data not shown). When total RNAs from human tissues were probed with the same cDNAs, greater differences were observed (Fig. 5). (*i*) Both RNA forms hybridizing to G4.8 were highly expressed in skeletal muscle and somewhat less in heart. G4.8 RNA was barely detectable in liver and brain while in other tissues (testis and placenta) one of the two forms was prevalent. (*ii*) mRNA corresponding to cDNA 1A was highly expressed in fetal and adult brain. (*iii*) cDNA 16A, G4.5, and 6.3 did not hybridize to total RNA from tissues. They were hybridized to a panel of poly(A)$^+$ RNAs from eight human tissues (Fig. 5*B*). 16A hybridized to one band of 2.3 kb in adult brain and pancreas. G4.5 demonstrated a pattern of hybridization similar to the one found in cell lines, but a smaller (1.2 kb) and more abundant RNA was present in muscle and heart RNAs only. cDNA 6.3 hybridized to an RNA of ≈8 kb, ubiquitously present but found in higher amounts in brain. Smaller bands may also be present in some tissues.

DISCUSSION

In this paper we report the structural organization of a gene cluster of the human X chromosome in Xq28. Starting from the X chromosome-specific CpG islands identified and mapped in our laboratory (9, 10), a contig of cosmids of 450

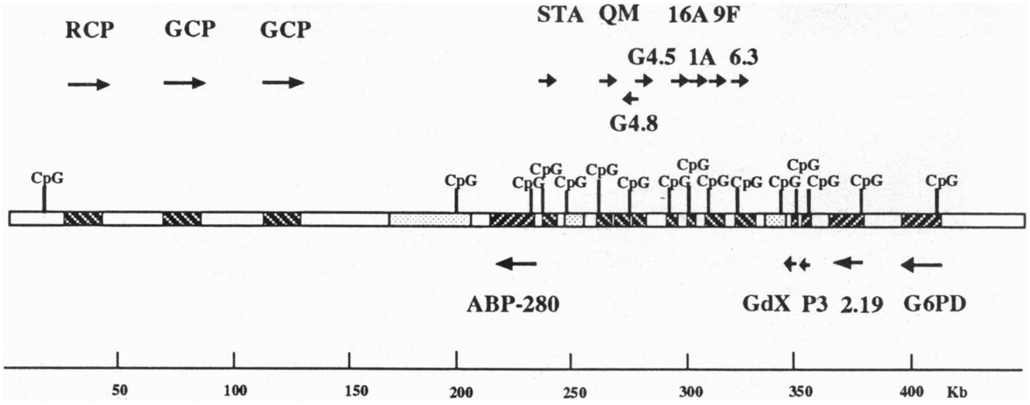

FIG. 4. Schematic representation of the transcriptional organization of the genes in the region. The genomic region occupied by each of the cDNAs is indicated by hatched boxes. The regions where we expect to find additional transcripts are indicated by stippled boxes. Arrows indicate direction of transcription of each gene.

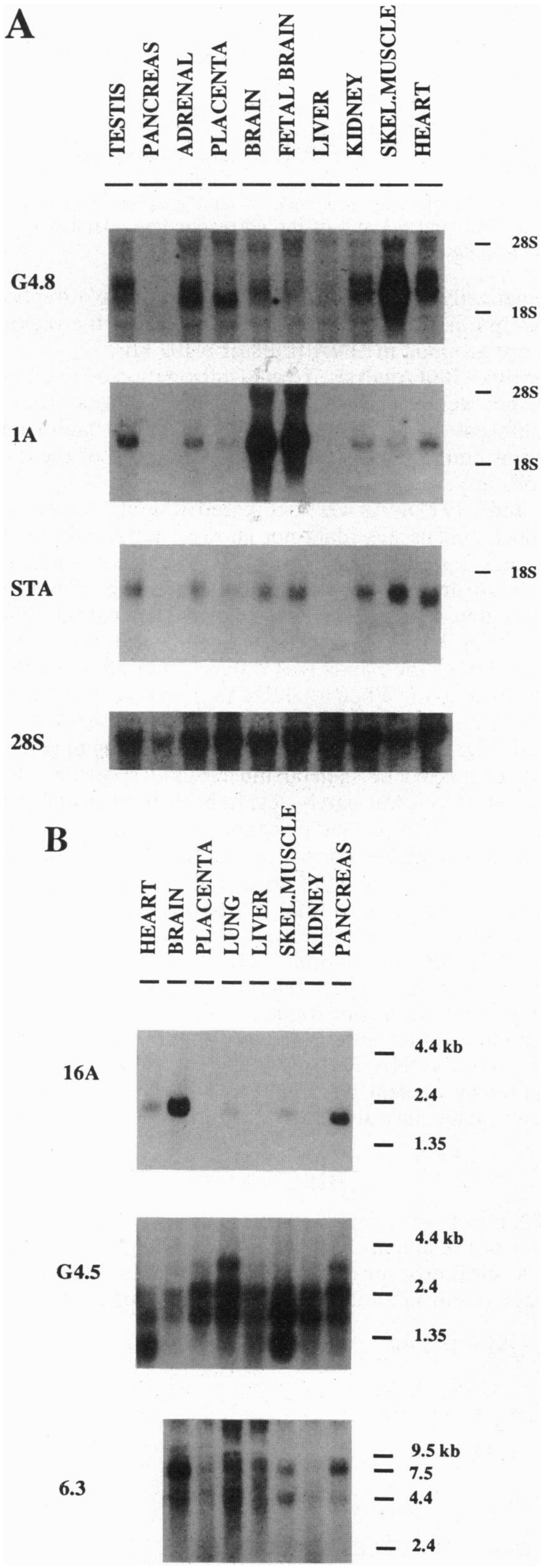

FIG. 5. Northern blots hybridized to the cDNAs identified in the region. (*A*) Total RNAs (15 μg) from normal human tissues indicated. (*B*) Poly(A)$^+$ RNA (2 μg) from the normal human tissues indicated. Hybridizations were at 65°C and hybridization washings were in 0.2× standard saline citrate at 65°C. Hybridization to 28S RNA was used to demonstrate the RNA content of each lane in *A*.

kb, linking the *G6PD* and the *CV* genes, was constructed. Using conserved DNA fragments flanking the newly identified CpG islands to screen cDNA libraries, eight cDNAs were isolated. Only cDNAs corresponding to three CpG islands have not yet been identified. However, they are flanked by conserved DNAs (probes 2, which is repeated twice, and 13) and it is likely that they represent the 5′ end of genes whose transcripts are very rare and/or highly tissue or stage specific and were not present in the cDNA libraries analyzed. A minimum number of 13 genes, 8 of which had not been mapped previously, were thus mapped in the 250–300 kb of DNA between the 3′ end of the *G6PD* and *GCP* genes.

Genes in this region are rather crowded and quite small, ranging in size from 5 to 20 kb. The majority encodes transcripts of 2–3 kb: the only large transcripts are the ABP-280 and the 6.3 mRNAs, >8 kb long (ref. 18; this work). Some of the new cDNAs were ubiquitously expressed, others were found in higher amounts mainly in two tissues, brain (16A, 1A) and muscle (G4.8). cDNA G4.5 hybridized to ubiquitously expressed RNA bands as well as to a muscle-specific smaller RNA form. All forms are rare in human tissues, as can be detected only in poly(A)$^+$ RNA, but they are easily detectable in Northern blots of total RNA from cell lines and tumors. RNA G4.5 may also be expressed at a high level in fetal brain, since many cDNA clones have been recently identified from a fetal brain cDNA library made from a 17- to 18-week human embryo (data not shown): RNAs encoded by G4.5 may thus be required in dividing and fetal cells. In summary, our studies confirm the idea that CpG islands may be frequently found at the 5′ end of genes with a tissue-limited distribution, as has been shown in a recent survey of sequenced human genes from the EMBL data bank (19).

The availability of cDNAs for many clustered genes has prompted us to establish a transcriptional map of the region. Our data show that transcripts in this region of the genome are not randomly oriented but that 100-kb transcriptional domains may be defined where genes have the same direction of transcription. Moreover, the pattern of expression of the genes in each domain suggests that the transcriptional order we have defined may have a functional role. Genes *G6PD*, 2-19, GdX, and P3 which belong to the same transcriptional domain are expressed in most tissues. A second domain is defined by the tissue-specific *CV* genes. In a third domain, cDNA STA, 6-3, 1A, 16A, G4.5, and G4.8 were found in a higher amount or with a specific mRNA pattern in brain and/or muscle. In addition, most transcripts in this domain are not found in liver (Fig. 5*A*). Thus, a common tissue distribution is shared by many of the genes in this domain and cannot be explained by simply postulating tissue-specific regional control element(s). Rather, in the absence of sequence similarity, they may share common functions in those tissues or at some developmental stage. Some kind of temporal and/or regional control of their expression may open them to transcription at the same time.

Until now only the structural organization of special gene families has been studied in detail. In mammals, the only genomic region with a similar concentration of genes, not all related in sequence, is the major histocompatibility complex (MHC) locus. Forty megabases of the MHC locus have been cloned, encompassing some 80 genes, and in the class III region, close to one gene every 20 kb has been found (20, 21). In the mouse, 12 transcription units were identified in the H-2K region and all were expressed in testis and/or embryo (22). For some of the genes in the MHC complex the direction of transcription has been determined and transcriptional domains like the one reported in this paper can be recognized: a complete transcriptional map is available, however, only for regions not much larger than 100 kb. The study of additional chromosomal regions with similar density of apparently unrelated genes will establish whether the gene organization we have described in Xq28 is the rule in the human genome and will be the basis for functional studies on how chromosomal position may affect gene expression. This would be particularly important for X chromosome-linked or imprinted genes on autosomes that are subject to chromosomal or locus control of their expression.

Many genes responsible for inherited disorders are mapped to distal Xq28 (23). The results of our studies on the expression of the new genes in cell lines and tissues suggest that some of them may be candidates for those diseases. The muscle-specific expression of genes G4.8 and G4.5 points to the two genes as candidates for the muscle disorders mapped to this region—Emery–Dreifuss muscular dystrophy and Barth syndrome (24). Accordingly, the brain-specific expression of 16A and 1A makes them candidates for X chromosome-linked mental retardation or neurological disorders.

We thank Dr. G. Camerino, S. Riva, and G. Della Valle for critical reading of the manuscript and R. Feil and J. L. Mandel for CV region cosmids and the probe 25 DNA. This work was supported by grants from Progetti Finalizzati Consiglio Nazionale delle Ricerche "Ingegneria Genetica" e "Biotecnologie e Biostrumentazione" and by "Telethon Italy."

1. Bickmore, W. A. & Sumner, A. T. (1989) *Trends Genet.* **5,** 144–148.
2. Boncinelli, E., Simeone, A., Acampora, D. & Mavilio, F. (1991) *Trends Genet.* **7,** 329–334.
3. Collins, F. S. & Weissman, S. M. (1984) *Prog. Nucleic Acid Res. Mol. Biol.* **31,** 315–462.
4. Kim, C. G., Epner, E. M., Forrester, W. C. & Groudine, M. (1992) *Genes Dev.* **6,** 928–938.
5. Eleventh International Workshop on Human Gene Mapping (1991) *Cytogenet. Cell Genet.* **58,** 1.
6. Surani, M. A., Kothari, R., Allen, N. D., Singh, P. B., Fundele, R., Ferguson-Smith, A. C. & Barton, S. C. (1990) *Development Suppl.* **89,** 98.
7. Reik, W. (1989) *Trends Genet.* **5,** 331–336.
8. Maestrini, E., Rivella, S., Tribioli, C., Purtilo, D., Rocchi, M., Archidiacono, N. & Toniolo, D. (1990) *Genomics* **8,** 664–670.
9. Tribioli, C., Tamanini, F., Patrosso, C., Milanesi, L., Villa, A., Pergolizzi, R., Maestrini, E., Rivella, S., Bione, S., Mancini, M., Vezzoni, P. & Toniolo, D. (1992) *Nucleic Acids Res.* **20,** 727–733.
10. Maestrini, E., Tamanini, F., Kioschis, P., Gimbo, E., Marinelli, P., Tribioli, C., D'Urso, M., Palmieri, G., Poustka, A. & Toniolo, D. (1992) *Hum. Mol. Genet.* **1,** 275–280.
11. Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Lab. Press, Plainview, NY), 2nd Ed.
12. Poustka, A., Dietrich, A., Lengenstein, G., Toniolo, D., Warren, S. T. & Lehrach, H. (1991) *Proc. Natl. Acad. Sci. USA* **88,** 8302–8306.
13. Feil, R., Aubourg, P., Heilig, R. & Mandel, J. L. (1990) *Genomics* **6,** 367–373.
14. Alcalay, M. & Toniolo, D. (1988) *Nucleic Acids Res.* **16,** 9527–9543.
15. van den Ouweland, A., Kioschis, P., Verdijk, M., Tamanini, F., Toniolo, D., Poustka, A. & van Oost, B. (1992) *Hum. Mol. Genet.* **1,** 269–273.
16. Maestrini, E., Patrosso, C., Mancini, M., Rivella, S., Rocchi, M., Repetto, M., Frattini, A., Zoppè, M., Vezzoni, P. & Toniolo, D. (1993) *Hum. Mol. Genet.*, **9,** 761–766.
17. Andrews, P., Damjanov, I., Simon, D., Banting, G. S., Carlin, C., Dracopoli, B. & Fogh, J. (1984) *Lab. Invest.* **50,** 147–162.
18. Gorlin, J. B., Yamin, R., Egan, S., Steward, M., Stossel, T. P., Kwiatkowski, D. J. & Hartwig, J. H. (1990) *J. Cell Biol.* **111,** 1089–1105.
19. Larsen, F., Gundersen, G., Lopez, R. & Prydz, H. (1992) *Genomics* **13,** 1095–1107.
20. Trowsdale, J., Ragoussis, J. & Campbell, D. R. (1991) *Immunol. Today* **12,** 443–446.
21. Trowsdale, J. & Powis, S. H. (1992) *Curr. Opin. Genet. Dev.* **2,** 492–497.
22. Yeom, Y. I., Abe, K., Bennett, D. & Artzt, K. (1991) *Proc. Natl. Acad. Sci. USA* **89,** 773–777.
23. Mandel, J. L., Monaco, A. P., Nelson, D. L., Schlessinger, D. & Willard, H. (1992) *Science* **258,** 103–109.
24. McKusick, V. A. (1990) *Mendelian Inheritance in Man* (Johns Hopkins Univ. Press, Baltimore), 9th Ed.