GENOMICS **45,** 340–347 (1997)
ARTICLE NO. GE974941

Sequence-Based Exon Prediction around the Synaptophysin Locus Reveals a Gene-Rich Area Containing Novel Genes in Human Proximal Xp

Simon E. Fisher,*[,1,2] Alfredo Ciccodicola,† Karo Tanaka,* Anna Curci,† Sonia Desicato,† Michele D'Urso,† and Ian W. Craig*

*Genetics Laboratory, Department of Biochemistry, Oxford University, South Parks Road, Oxford, OX1 3QU, United Kingdom; and †International Institute of Genetics and Biophysics, CNR, Via Marconi 10, 80125 Naples, Italy

Received April 1, 1997; accepted July 23, 1997

The human Xp11.23–p11.22 interval has been implicated in several inherited diseases including Wiskott–Aldrich syndrome; three forms of X-linked hypercalciuric nephrolithiaisis; and the eye disorders retinitis pigmentosa 2, congenital stationary night blindness, and Åland Island eye disease. In constructing YAC contigs spanning Xp11.23–p11.22, we have previously shown that the region around the synaptophysin (SYP) gene is refractory to cloning in YACs, but highly stable in cosmids. Preliminary analysis of the latter suggested that this might reflect a high density of coding sequences and we therefore undertook the complete sequencing of a SYP-containing cosmid. Sequence data were extensively analyzed using computer programs such as CENSOR (to mask repeats), BLAST (for homology searches), and GRAIL and GENE-ID (to predict exons). This revealed the presence of 29 putative exons, organized into three genes, in addition to the 7 exons of the complete SYP coding region, all mapping within a 44-kb interval. Two genes are novel, one (CACNA1F) showing high homology to α1 subunits of calcium channels, the other (LMO6) encoding a product with significant similarity to LIM-domain proteins. RT-PCR and Northern blot studies confirmed that these loci are indeed transcribed. The third locus is the previously described, but not previously localized, A4 differentiation-dependent gene. Given that the intron–exon boundaries predicted by the analysis are consistent with previous information where available, we have been able to suggest the genomic organization of the novel genes with some confidence. The region has an elevated GC content (>53%), and we identified CpG islands associated with the 5′ ends of SYP, A4, and LMO6. The order of loci was Xpter–A4–LMO6–SYP–CACNA1F–Xcen, with intergenic distances ranging from ~300 bp to ~5 kb. The density of transcribed sequences in this area (>80%) is comparable to that found in the highly gene-rich chromosomal band Xq28. Further studies may aid our understanding of the long-range organization surrounding such gene-enriched regions. © 1997 Academic Press

INTRODUCTION

Positional cloning from human interval Xp11.23–p11.22 has led to the successful isolation of genes for disorders such as Wiskott–Aldrich syndrome (Derry *et al.,* 1994) and X-linked hypercalciuric nephrolithiasis (Fisher *et al.,* 1994; Lloyd *et al.,* 1996). However, the region has also been implicated in several diseases for which the etiology is still unknown; these include the eye disorders retinitis pigmentosa 2 (Meitinger *et al.,* 1989), congenital stationary night blindness (Aldred *et al.,* 1992; Musarella *et al.,* 1992), and Åland Island eye disease (Alitalo *et al.,* 1991), as well as forms of X-linked mental retardation (Lubs *et al.,* 1996). We have previously constructed and characterized YAC contigs spanning several megabases of Xp11.23–p11.22, but discovered that the region around the synaptophysin gene (SYP) is highly refractory to YAC cloning (Fisher *et al.,* 1995). This is most likely due to sequence-specific instability. As a consequence, despite extensive analysis by several groups (Derry *et al.,* 1994; Hagemann *et al.,* 1994; Nelson *et al.,* 1995; Boycott *et al.,* 1996), this region remains poorly characterized. It has been noted that such inherently unstable regions tend to coincide with areas of high GC content that are enriched for transcribed sequences (Palmieri *et al.,* 1994). We therefore postulated that the instability in Xp11.23 may reflect a high density of coding sequences around the SYP locus.

Sequence data from this article have been deposited with the GenBank Data Library under Accession No. U93305.

[1] Present address: Wellcome Trust Centre for Human Genetics, Oxford University, Windmill Road, Headington, Oxford, OX3 7BN, United Kingdom.

[2] To whom correspondence should be addressed. Telephone: +44 1865 740 021. Fax: +44 1865 742 186.

0888-7543/97 $25.00
Copyright © 1997 by Academic Press
All rights of reproduction in any form reserved.

This view was supported by preliminary analysis of novel conserved single-copy sequences from SYP-containing cosmids, which, in contrast to SYP YACs, show no evidence of instability (Fisher, 1995). Furthermore, we have found a clustering of rare-cutter sites in these cosmids, indicating the presence of several CpG islands. In light of these studies, we decided to use a sequence-based strategy to investigate the organization of genes in the vicinity of the SYP locus. We have thereby revealed a gene-rich area containing a large number of novel putative exons, which we have characterized in detail.

MATERIALS AND METHODS

Sequencing strategy. Cosmid DNA was sonicated and ends were repaired with T4 DNA polymerase. Fragments of 1–2 kb were recovered after agarose gel electrophoresis and subcloned into *Sma*I-predigested M13mp18 vector.

M13 cloned DNAs were prepared for random shotgun sequencing by a streamlined manual method (Zollo and Chen, 1994) and sequenced using dye-primer cycle sequencing on 373A automated sequencers (Chen, 1994). Sequence tracts across most of the cosmid were assembled using Applied Biosystem's FACTURA and INHERIT program. Coverage of between six- and eightfold was achieved. Most (~99%) of the sequence was determined on both strands, but at some locations no opposite strand clone had been recovered. In those cases, consensus sequence was derived by sequencing templates in the same orientation with two different sequencing chemistries (dye primer and dye terminator) (Chen, 1994).

Gap closure. In areas that were problematic or covered only once, the assembly program was used to infer clones likely to extend across a gap. The inserts of those M13 clones were amplified by PCR and sequence was then derived with reverse dye-primers from the opposite end of the clone insert (Chen, 1994). A second approach to gap closure required synthesis of additional primers to extend sequences ("walking") from the edges of the contigs using dye-terminator reactions. Any remaining gaps, representing DNA that was not recoverable in M13, were closed using PCR primers to amplify intervening material, which was then sequenced by dye-terminator reactions (Chen, 1994).

Computer-assisted sequence analysis. Programs for sequence analysis were accessed through the HGMP computing facility at Hinxton Hall, Cambridge. TANDEM, CpGPLOT, and BESTFIT are part of the GCG package. Prior to exon prediction, CENSOR was used to identify and mask any repetitive sequences. A range of protein and nucleotide databases was screened with unique portions of the cosmid sequence, using the BLAST algorithm (Altschul *et al.,* 1990). The exon prediction programs GRAIL (Version 2) (Uberbacher and Mural, 1991) and GENE-ID (Guigo *et al.,* 1992) were used to identify coding regions independently without reference to information regarding homology. Data from BLAST and GRAIL/GENE-ID were then compared and combined. Annotated cosmid sequence, including details of known and predicted exons and positions of CpG islands and repetitive elements, can be found in GenBank (Accession No. U93305).

RT-PCR of LMO6. Lymphocytes were isolated from the blood sample of a normal male using Ficoll–Paque (Pharmacia), and total RNA was extracted following the modified guanidine isothiocyanate/acid–phenol method described previously (Chomczynski and Sacchi, 1987). Five micrograms of total RNA was converted into first-strand cDNA using the Super Script Preamplification System (GIBCO BRL). PCR amplification was performed with primers 5′-GTGTGTACCACGTGCCAGG-3′ and 5′-GCAGGCTCGAGAGCAGAAG-3′ using the following conditions: 94°C for 60 s, 60°C for 60 s, and 72°C for 60 s, 30 cycles, generating a 450-bp product spanning exons 5–7 of the LMO6 transcript.

PCR across CACNA1F exons. Three sets of primers were designed to amplify genomic sequence spanning seven of the CACNA1F exons from the cosmid. PCR with 5′-AGGGTATAGAGGGCATACTTGG-3′ and 5′-GGTGTGCCACTGGTGAGG-3′ gave a 536-bp product of which 285 bp corresponded to exons 6 and 7 (see Fig. 4a). Primers 5′-CCTTCTGTTTTGATCTTCAGGG-3′ and 5′-CCTCCCCAACTCCTCCAG-3′ amplified a 405-bp product containing 198 bp from exons 8–9, and primers 5′-TGTCCCCACTTTGTTAGTTTCC-3′ and 5′-CTCATGCCCCACATCCTC-3′ gave an 819-bp product containing 339 bp from exons 10, 11, and 12. PCR conditions for each set of primers were 94°C for 40 s, 60°C for 60 s, and 72°C for 60 s, 36 cycles. These exon-enriched products were pooled for Northern analysis.

Northern blot analysis. Multiple tissue Northern blots (Clontech) were screened with probes from LMO6 and CACNA1F according to the manufacturer's instructions. Subsequent probing with β-actin was used as a control.

RESULTS

Identification of a Region of Interest Around SYP

The human SYP gene was previously characterized and localized to Xp11.23 by Özcelik *et al.* (1990). Physical mapping has since allowed refinement of the position of SYP to the interval between TFE3 and DXS6666 (Fisher *et al.,* 1995). Restriction digest and fluorescence *in situ* hybridization studies of three SYP-containing cosmids (isolated from the ICRF X-specific library) established that they are nonchimeric and unlikely to have rearranged (Colin Cooper, Institute of Cancer Research, Surrey, UK; manuscript in preparation) and enabled them to be ordered and oriented with respect to flanking markers in Xp11.23 (Fig. 1). Analysis of a single-copy 3.1-kb *Eco*RI fragment that was subcloned from SYP-A1236 revealed the presence of a conserved sequence with high homology to calcium channels (see below). In addition, rare-cutter restriction mapping identified several putative CpG islands in the SYP-A1236 cosmid (data not shown). Together, these observations suggested that this cosmid might contain a high density of transcribed sequences.

Sequencing and Analysis of SYP-A1236 Cosmid

SYP-A1236, which contains a 44-kb insert, was completely sequenced at high (99.9%) precision using random shotgun methods (Chen, 1991; Wilson *et al.,* 1994; Zollo *et al.,* 1995) supplemented by PCR-based DNA fragment amplification and primer-walking strategies (see Materials and Methods). The assembled sequence was analyzed using a series of computer programs: TANDEM (to search for runs of di-, tri-, tetra-, and pentanucleotide repeats), CENSOR (to identify complex repetitive elements), CpGPLOT (to detect CpG islands), BLAST (for homology searches), and GRAIL2 and GENE-ID (to predict likely exons). This enabled us to elucidate the most likely genomic organization of the region, which is summarized in Fig. 2, and has

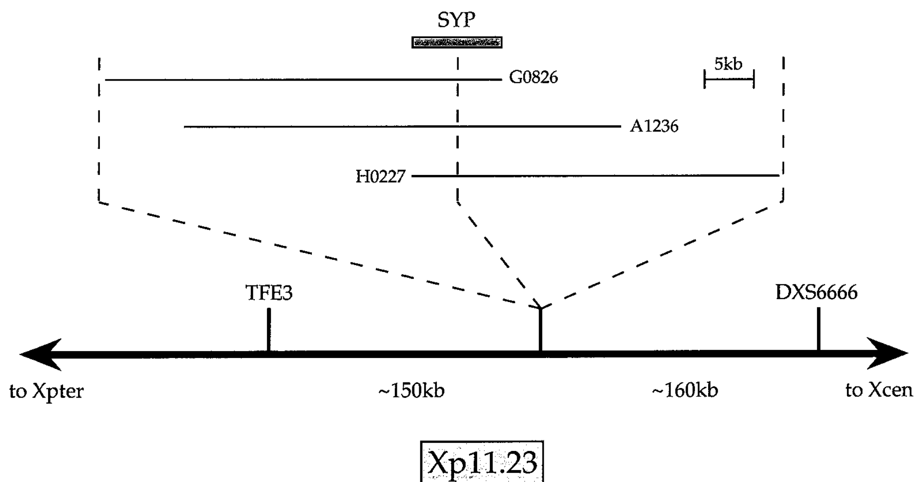

FIG. 1. Organization of three nonchimeric SYP-selected cosmids from the ICRF flow-sorted X-specific library, that have been ordered and oriented with respect to flanking markers in Xp11.23. Identification numbers of cosmids are indicated. Long-range mapping data are taken from a study of the region by Colin Cooper and colleagues at the Institute of Cancer Research (Surrey, UK) (manuscript in preparation) and from the 7th X Chromosome Workshop (October 1996; Sanger Centre, Cambridge, UK).

been submitted, along with the complete sequence, to GenBank (Accession No. U93305).

Repetitive Sequence Content

We searched the sequence for any runs of four or more consecutive di-, tri-, tetra-, or pentanucleotide repeats. A highly TpA-rich stretch was identified at residues 10332–10716 (where residue 1 is at the Xpter end of the cosmid). This stretch contains 1 $(TA)_{10}$ repeat, 3 copies of a $(TA)_6$ repeat, 11 copies of $(TA)_5$, and 4 copies of $(TA)_4$, all within a <400-bp region. Elsewhere, tandem repeats exceeded 5 copies at only three other sites in the cosmid, with $(GT)_8$ at residue 3390, $(TG)_8$ at 12728, and $(CTC)_6$ at 33958. The cosmid contains a high abundance of short interspersed nucleotide repeats (SINEs), but no LINEs. There are 28 partial or complete *Alu* sequences, 8 MIR elements, and 4 MIR2 elements, together constituting almost 20% of the entire sequence (Fig. 2). Details of all these repeats are given in the GenBank entry for the sequence.

Known and Predicted Exons in the Region around SYP

In addition to the 7 exons of the complete SYP coding region (spanning 12.4 kb), we have identified the presence of a further 29 putative exons, organized into three transcriptional units, all mapping within the 44-kb interval cloned in the SYP-A1236 cosmid (see Fig. 2):

(i) A cDNA encoding the A4 differentiation-dependent protein was previously cloned from a human colonic cell line (Oliva *et al.,* 1993), but prior to our study the gene had not been localized. BLAST analysis indicates that the SYP-A cosmid contains the promoter (previously characterized by Oliva *et al.,* 1995) and complete coding sequence of the A4 gene, which is organized into five exons, spanning ~3.5 kb of DNA. The four 5′ exons and their correct splice sites were independently predicted by GRAIL, in the absence of any information regarding homology. The product of this gene is an integral membrane protein, containing four putative transmembrane domains, which may play a role in cell differentiation in the intestinal epithelium (Oliva *et al.,* 1993). We detected a CpG island associated with the 5′ end of the locus.

(ii) A novel locus, consisting of eight exons, lies in the ~12.5-kb interval between the A4 and the SYP loci. This gene (which has been assigned the name LMO6 by the HUGO Nomenclature Committee) is predicted to encode a product of 407 amino acids with significant homology to proteins containing LIM domains, in particular mouse testin (Accession No. S44219; Divecha and Charleston, 1995). The LIM domain is a cysteine-rich sequence motif which binds zinc atoms to form a specific protein-binding interface in protein–protein interactions (Schmeichel and Beckerle, 1994). Like testin, the LMO6 product contains three such domains, each conforming to the revised LIM consensus $CX_2CX_{16-23}HX_{2-4}C/HX_2CX_2CX_{16-21}CX_{2-3}$ (C, H, D). The similarity between the two related proteins is strongest in this triple LIM domain region (48% amino acid identity), with conservation of all the cysteine and histidine residues that are necessary to coordinate zinc atoms (Fig. 3a). LMO6 also showed significant homology to putative coding regions identified in a *Caenorhabditis elegans* sequencing project (Accession No. U52003; Wilson *et al.,* 1994), particularly at its N-terminus, where the amino acid sequences encoded by exons 2 and 3 are ~60% identical to the corresponding region of *C. elegans* product "ZK381.7." CpG islands were found at the 5′ and 3′ ends of this gene.

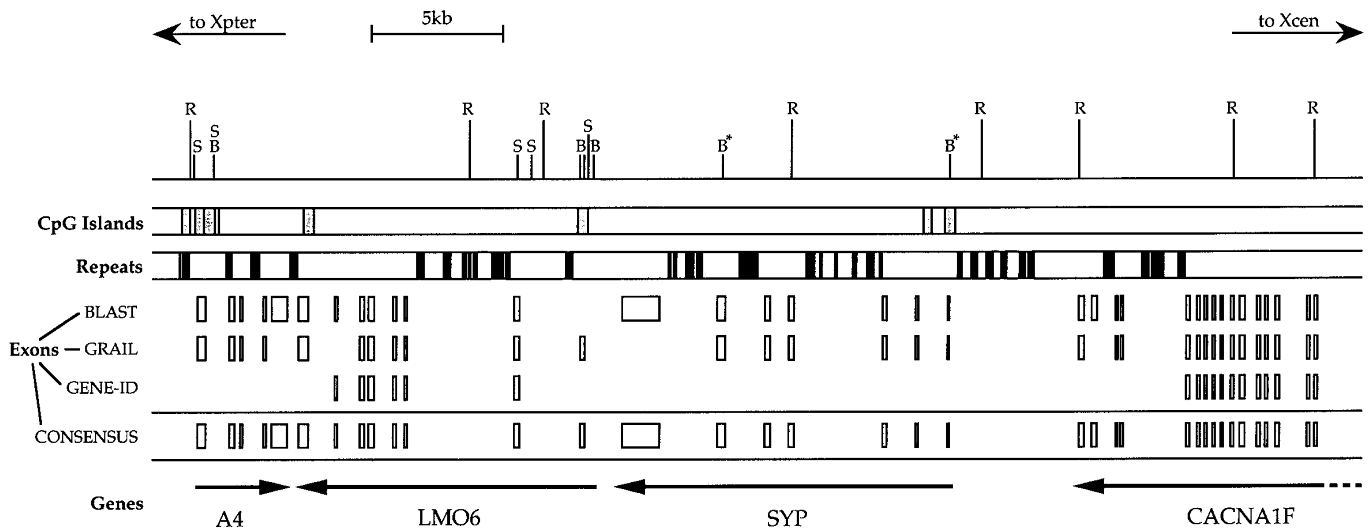

FIG. 2. Genomic organization of a 44-kb region around the SYP locus, as deduced from sequence analysis. Restriction sites are given for *Eco*RI (R), and the rare cutters *Bss*HII (B) and *Sst*II (S). Sites marked with an asterisk are double sites separated by 2 bp. The positions of putative CpG islands (shaded boxes), repetitive elements (black boxes), and predicted exons (open boxes) are shown aligned below the restriction map. The four main transcription units predicted by this analysis are indicated by arrows pointing from 5′ to 3′. A substantial proportion of the 5′ end of CACNA1F lies outside the region of study. Note that GENE-ID was used only for the analysis of the novel loci. Orientation of this region with respect to the X chromosome is given.

Using an RT-PCR approach, we were able to amplify a 450-bp product spanning exons 5–7 of the LMO6 transcript from lymphoblastoid RNA (see Fig. 3a). This confirmed that the LMO6 locus is indeed transcribed and unlikely to be a pseudogene. In addition the sequence of this RT-PCR product agreed with that suggested by the computer analysis of the SYP-A cosmid, thus verifying the proposed exon–intron organization and predicted amino acid sequence for the triple LIM domain region of the gene. LMO6 expression was further investigated by using the RT-PCR product to probe a Northern blot containing poly(A)-selected mRNA from several human tissues (Fig. 3b). A predominant band of ~3.4 kb was observed in all tissues, with strongest expression in skeletal muscle, heart, placenta, and pancreas. In addition, a weaker 2.3-kb transcript was detected in most tissues. Since the predicted coding region of LMO6 only spans ~1.2 kb, the transcripts must contain a significant amount of 5′ and/or 3′ untranslated sequence. Such transcribed, but untranslated, regions will often go undetected by programs such as GRAIL and GENE-ID.

The multiple transcripts detected on the Northern may correspond to alternatively spliced variants from the LMO6 locus. A comparable situation was previously shown for mouse testin, where a single gene gives rise to two transcripts, of 1.4 and 2.8 kb, which appear to encode proteins that are identical in the LIM domain region, but differ at their N-termini (Divecha and Charleston, 1995).

(iii) On the centromeric side of SYP, we have identified 16 exons from the 3′ end of a locus (approved symbol CACNA1F) that encodes a novel L-type calcium channel $\alpha 1$ subunit. Several tissue-specific isoforms of $\alpha 1$ subunits have previously been identified, and they range in size from 1873 to 2424 amino acid residues (Seino *et al.,* 1992). All share a common structure, involving four internal units of homology (I–IV), each composed of six alpha-helical membrane-spanning segments (S1–S6) (Tanabe *et al.,* 1987). CACNA1F shows strongest homology to the pancreatic isoform (Accession No. A38198; Seino *et al.,* 1992) with 5 of the putative exons showing >90% identity at the amino acid level (Fig. 4a). The 3′ part of the CACNA1F locus contained in the SYP-A cosmid codes for 646 amino acid residues at the carboxy-terminus of the $\alpha 1$ subunit, including transmembrane domains IVS2–IVS6 (Fig. 4a). A substantial proportion of the 5′ end of this gene is therefore expected to lie centromeric to the characterized region.

An RT-PCR approach was unable to amplify predicted CACNA1F exons from lymphoblastoid RNA, suggesting that the gene is not expressed in this tissue. In the absence of additional information regarding tissue specificity of CACNA1F expression, genomic PCR was used to generate probes spanning seven exons of CACNA1F (exons 6–12 encoding amino acids 161–445 in Fig. 4a). Screening of a multiple tissue Northern blot with these probes revealed a major ~6.8-kb transcript in skeletal muscle (Fig. 4b). Although a human skeletal muscle isoform has previously been identified (CACNL1A3; Accession No. L33798; Hogan *et al.,* 1994) it shows only 78% nucleotide identity to CACNA1F in the coding regions used to probe the Northern; this result is therefore unlikely to be a consequence of cross-hybridization to CACNL1A3. A 1.4-kb transcript was also detected at high levels in heart and at weaker levels in several other tissues. Given that a transcript of this

a

```
                          exon 2                    ↓                   exon 3
  1 MCRLISDFQRHSISDDDSGCASEEYAWVPPGLKPEQVYQFFSCLPEDKVPYVNSPGEKYRIKQLLHQLPP  70
     | :  : ..  ...: :... .|.|.||. ... . |::  ||.:| | ..|.|..|| ||| .|||:
 92 YKRNVMILTNPVAAKKNVSINTVTYEWAPPVQNQALARQYMQMLPKEKQPVAGSEGAQYRKKQLAKQLPA 161

       ↓              exon 4
 71 HDSEAQYCTALEEEEKKELRAFSQQRKRENLGRGIVRIFP.............................V 111
    || ::  | .|...| ||: .| .. |.|.|| | |:: .
162 HDQDPSKCHELSPKEVKEMEQFVKKYKSEALGVGDVKFPSEMNAQGDKVHNCGNRHAPAAVASKDKSAES 231

          ↓                             exon 5
112 TITGASCEECGKQIGGGDIAVFASRAGLGACWHPQCFVCTTCQELLVDLIYFYHVGKVYCGRHHAECLRP 181
    . |. ||  | . ..:|: |::|.|||.:  |||.||:|.||.|||||:|||:. ||:||||||..:: :|
232 KKTQYSCYCCKHTTNEGEPAIYAERAGYDKLWHPACFICSTCGELLVDMIYFWKNGKLYCGRHYCDSEKP 301
          CXXCXXXXXXXXXXXXXXXXXXXXXXHXXCXXCXXCXXXXXXXXXXXXXXXXXXCXXH
                                   LIM-I

        ↓                              exon 6                             ↓
182 RCQACDEIIFSPECTEAEGRHWHMDHFCCFECEASLGGQRYVMRQSRPHCCACYEARHAEYCDGCGEHIG 251
    ||.:|||:||| |:|:||..:||:.||||||:|:  |:|. |||  .:| | :||  .|| |:|| : |:
302 RCAGCDELIFSNEYTQAENQNWHLKHFCCFDCDHILAGKIYVMVTDKPVCKPCYVKNHAVVCQGCHNAID 371
     CXXCXXXXXXXXXXXXXXXXXXHXXHXXCXXCXXXXXXXXXXXXXXXXXXXCXXC      CXXCXXXXX
                         LIM-II

                          exon 7
252 LDQGQMAYEGQHWHASDRCFCCSRCGRALLGRPFLPRRGLIFCSRAC 298
     : ..:.|:.  ||||. || || |::.|:|..|:|  |::||| .|
372 PEVQRVTYNNFSWHASTECFLCSCCSKCLIGQKFMPVEGMVFCSVEC 418
    XXXXXXXXXXXXXXHXXXXCXXCXXCXXXXXXXXXXXXXXXXXXCXXXC
                       LIM-III
```

b

heart
brain
placenta
lung
liver
skeletal muscle
kidney
pancreas

9.5 —
7.5 —
4.5 —
2.4 —
1.35 —

FIG. 3. Analysis of the LMO6 locus. (**a**) Alignment between predicted LMO6 product (top line) and the mouse testin product S44219 (bottom line) in a region of significant homology. Adapted from the output of BESTFIT (GCG package). This program uses a scoring matrix based on the evolutionary distance between amino acids (Gribskov and Burgess, 1986), with "|" indicating identity, ":" indicating a comparison value of ⩾0.5, and "." indicating a value of ⩾0.1. Positions of LMO6 splice sites relative to its predicted product are shown by arrows above the amino acid sequence. (Note that exon 1 is predicted to contain only 5′ untranslated sequence.) The three conserved LIM domains are indicated. (**b**) Multiple tissue Northern blot probed with an RT-PCR product spanning exons 5–7 of LMO6. Transcripts of 3.4 and 2.3 kb are detected in most tissues. Larger bands observed in skeletal muscle and heart may correspond to partially spliced transcripts. Positions and sizes (in kb) of RNA marker bands are indicated.

a

```
      ↓                    exon 1               ↓                    exon 2              ↓
   1 HYEQTAPFNYAMDILNMVFTGLFTIVMVLKIIAFKPKHYFTDAWNTFDALIVVGSIVDIAVTEVNS.............. 66
     ||||.  || ||||||||||:||: ||||:||||| ||.|||||||.|||:|||:|:|:.|.:.
1246 HYEQSKMFNDAMDILNMVFTGVFTVEMVLKVIAFKPKGYFSDAWNTFDSLIVIGSIIDVALSEADPTESENVPVPTATPG 1325
                    <=======IVS2=======>          <=======IVS3=======>

                        exon 3                          ↓          exon 4          ↓
  67 .SEDSSRISITFFRLFRVMRLVKLLSKGEGIRTLLWTFIKSFQALPYVALLIAMIFFIYAVIGMQMFGKVALQDGTQINR 145
     ||:|.||||||||||||||||||||:|||||||||||||||||||||||||||:|||||||||||||||||:.|..||||
1326 NSEESNRISITFFRLFRVMRLVKLLSRGEGIRTLLWTFIKSFQALPYVALLIAMLFFIYAVIGMQMFGKVAMRDNNQINR 1405
               <======IVS4=======>                       <=======IVS5=======>

      exon 5               ↓                       exon 6                            ↓
 146 NNNFQTFPQAVLLLFRCATGEAWQEIMLASLPGNRCDPESDFGPGEEFTCGSNFAIAYFISFFMLCAFLIINLFVAVIMD 225
     |||||||||||||||||||||||||||||||:|||. ||||||:.||||:||||||||.|||||:||||||||||||||||
1406 NNNFQTFPQAVLLLFRCATGEAWQEIMLACLPGKLCDPESDYNPGEEYTCGSNFAIVYFISFYMLCAFLIINLFVAVIMD 1485
                                                                  <==========IVS6=========>

               exon 7                 ↓                   exon 8                  ↓
 226 NFDYLTRDWSILGPHHLDEFKRIWSEYDPGAKGRIKHLDVVALLRRIQPPLGFGKLCPHRVACKRLVAMNMPLNSDGTVT 305
     |||||||||||||||||||||||||||||:|||||||||||.||||||||||||||||||||||||||||||||||||||
1486 NFDYLTRDWSILGPHHLDEFKRIWSEYDPEAKGRIKHLDVVTLLRRIQPPLGFGKLCPHRVACKRLVAMNMPLNSDGTVM 1565

          exon 9          ↓                exon 10                ↓                 exon 11
 306 FNATLFALVRTSLKIKTEGNLEQANQELRIVIKKIWKRMKQKLLDEVIPPPDEEEVTVGKFYATFLIQDYFRKFRRRKEK 385
     |||||||||||.|||||||||||||:||| |||||||: . ||||:|:||::::||||||||||||||||||||::|||.
1566 FNATLFALVRTALKIKTEGNLEQANEELRAVIKKIWKKTSMKLLDQVVPPAGDDEVTVGKFYATFLIQDYFRKFKKRKEQ 1645

                           ↓                    exon 12                          ↓
 386 GLLGNDAAPSTSSALQAGLRSLQDLGPEMRQALTCDTEEEEEEGQEGVEEEDEKDLETNK 445
     ||:|. :|..|. |||||||.|:|:|||:|.|:.|| :::|.|  |.  ||::. :. |
1646 GLVGKYPAKNTTIALQAGLRTLHDIGPEIRRAISCDLQDDEPE..ETKREEEDDVFKRNG 1703
```

b

heart brain placenta lung liver skeletal muscle kidney pancreas

9.5 —
7.5 —
4.5 —
2.4 —
1.35 —

FIG. 4. Analysis of the CACNA1F locus. (**a**) Alignment between predicted CACNA1F product (top line) and the human pancreatic calcium channel $\alpha 1$ subunit A38198 (bottom line) in region of highest homology. Adapted from output of BESTFIT program. Symbols used are the same as those of Fig. 3a. Arrows indicate positions of CACNA1F splice sites relative to its predicted product. Note that, since a substantial proportion of the 5′ part of CACNA1F has not yet been identified, the numbering of amino acids and exons used here is provisional, and will be revised once the complete locus has been characterized. Transmembrane domains IVS2–IVS6 are indicated. The shorter IVS3–IVS4 linker found in CACNA1F has also been seen in other $\alpha 1$ subunits (Seino *et al.*, 1992). (**b**) Multiple tissue Northern blot screened with a probe enriched for exons 6–12 of CACNA1F, revealing a 6.8-kb transcript in skeletal muscle and 1.4-kb bands in several tissues. Positions and sizes (in kb) of RNA marker bands are indicated.

size is not large enough to encode a calcium channel $\alpha1$ subunit, it is unclear at this stage how this mRNA species relates to the CACNA1F exons.

CpG island positions, exon–intron boundaries, and complete amino acid sequence information for all of the above are fully detailed in GenBank Accession No. U93305.

DISCUSSION

We have identified a gene-rich area around the SYP locus in Xp11.23. Detailed computer analysis of 44 kb has indicated that >80% of this region is transcribed. With an elevated GC content (53%), placing it in the 3–4% of the genome considered GC rich, and an abundance of *Alu* sequences, but no LINE elements, the characteristics of the region are comparable to the gene-dense subtelomeric band, Xq28 (Chen *et al.,* 1996). In such regions, sequence-based strategies are highly effective for the positional cloning of novel loci and, as confirmed here, will often identify coding regions that have previously gone undetected by alternative methods of transcript mapping.

Using gene prediction methods and homology searches, we have discovered two new genes and confirmed that they are expressed by probing Northern blots with putative exonic sequences. LMO6 encodes a product containing three LIM domains and is expressed in a wide range of tissues. It is thus likely to have an important cellular function, involving interaction with other proteins, but its precise role remains to be established. Although the putative LMO6 product shows significant homology to mouse testin for much of its length, there is sufficient sequence divergence to indicate that the former is not the human homologue of the latter, but is instead a novel related protein. CACNA1F codes for a new isoform of L-type calcium channel $\alpha1$ subunit. Northern analysis indicates that it is expressed in skeletal muscle, and it shows only ~78% homology to the nucleotide sequence of a previously cloned human skeletal muscle specific isoform (Accession No. L33798). Following completion of our analysis of the SYP cosmid genomic organization, we recently rescreened the latest editions of the genome databases using BLAST and identified newly entered ESTs showing almost 100% nucleotide homology to portions of LMO6 (Accession No. AA161256) and CACNA1F (Accession Nos. AA019974 and AA019975). This provides further evidence that these loci are indeed transcribed.

In addition, our analysis of the cosmid sequence has enabled us to localize the A4 differentiation-dependent gene to the SYP region and determine its exon–intron organization for the first time. The order of loci was found to be Xpter–A4–LMO6–SYP–CACNA1F–Xcen, with intergenic distances ranging from ~300 bp to 5 kb. We identified CpG islands associated with the 5′ ends of A4, SYP, and LMO6.

In general, for all four loci present in the cosmid, the exon–intron boundaries identified by the exon prediction programs were highly consistent with those suggested by the BLAST analysis. We have therefore been able to propose the likely genome organization of the novel loci with some confidence. However, it should be noted that, while we have shown the SYP-A cosmid to be highly stable, any results based on genomic sequencing alone can be sensitive to subtle sequence rearrangements. In addition, alternative splicing patterns may go undetected by computer-based methods. Isolation and analysis of LMO6 and CACNA1F cDNA clones will be important for verification of the data presented here, particularly in regions of lowest homology.

In conclusion, our studies have revealed a high density of transcribed sequences in Xp11.23 and enabled characterization of genomic organization in a region that has previously eluded analysis. Further investigations will establish whether any of the aforementioned genes are involved in the diseases that map to the region. They may also allow insights into the long-range organization surrounding such gene-enriched regions.

ACKNOWLEDGMENTS

This work was supported by an E.C. grant under BIOMED2, coordinated by Professors K. E. Davies and I. W. Craig, Oxford. A.C. was funded by Telethon–Italy (Grant E546).

REFERENCES

Aldred, M. A., Dry, L. L., Sharp, D. M., Van Dorp, D. B., Brown, J., Hardwick, L. J., Lester, D. H., Pryde, F. E., Teague, P. W., Jay, M., Bird, A. C., Jay, B., and Wright, A. F. (1992). Linkage analysis in X-linked congenital stationary night blindness. *Genomics* **14:** 99–104.

Alitalo, T., Kruse, T. A., Forsius, H., Eriksson, A. W., and de la Chapelle, A. (1991). Localization of the Åland Island eye disease locus to the pericentromeric region of the X chromosome by linkage analysis. *Am. J. Hum. Genet.* **48:** 31–38.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Boycott, K. M., Halley, G. R., Schlessinger, D., and Bech-Hansen, N. T. (1996). A 2-megabase physical contig incorporating 43 DNA markers on the human X chromosome at p11.23–p11.22 from ZNF21 to DXS255. *Genomics* **33:** 488–497.

Chen, E. Y. (1991). Overview of manual and automated DNA sequencing by the dideoxy chain termination method. *In* "Methods, a Companion to Methods in Enzymology," Vol. 3, pp. 3–19, Academic Press, San Diego.

Chen, E. Y. (1994). Automated DNA sequencing and analysis. *In* "Automated DNA Sequencing and Analysis Techniques" (M. D. Adams, C. Fields, and J. C. Venter, Eds.), pp. 3–10, Academic Press, San Diego.

Chen, E. Y., Zollo, M., Mazzarella, R., Ciccodicola, A., Chen, C. N., Zuo, L., Heiner, C., Burough, F., Repetto, M., Schlessinger, D., and D'Urso, M. (1996). Long-range sequence analysis in Xq28: Thirteen known and six candidate genes in 219.4kb of high GC DNA between the RCP/GCP and G6PD loci. *Hum. Mol. Genet.* **5:** 659–668.

Chomczynski, P., and Sacchi, N. (1987). Single-step method of RNA

isolation by acid guanidinium thiocyanate–phenol–chloroform extraction. *Anal. Biochem.* **162:** 156–159.

Derry, J. M. J., Ochs, H. D., and Francke, U. (1994). Isolation of a novel gene mutated in Wiskott–Aldrich syndrome. *Cell* **78:** 635–644.

Divecha, N., and Charleston, B. (1995). Cloning and characterization of two new cDNAs encoding murine triple LIM domains. *Gene* **1565:** 283–286.

Fisher, S. E. (1995). Positional cloning of the gene responsible for Dent's disease. D.Phil. thesis, Oxford University, Oxford, UK.

Fisher, S. E., Black, G. C. M., Lloyd, S. E., Hatchwell, E., Wrong, O., Thakker, R. V., and Craig, I. W. (1994). Isolation and partial characterization of a chloride channel gene which is expressed in kidney and is a candidate for Dent's disease (an X-linked hereditary nephrolithiasis). *Hum. Mol. Genet.* **3:** 2053–2059.

Fisher, S. E., Hatchwell, E., Chand, A., Ockendon, N., Monaco, A. P., and Craig, I. W. (1995). Construction of two YAC contigs in human Xp11.23–p11.22, one encompassing the loci OATL1, GATA, TFE3, and SYP, the other linking DXS255 to DXS146. *Genomics* **29:** 496–502.

Gribskov, M., and Burgess, R. R. (1986). Sigma factors from E. coli, B. subtilis, phage SP01, and phage T4 are homologous proteins. *Nucleic Acids Res.* **14:** 6745–6763.

Guigo, R., Knudsen, S., Drake, N., and Smith, T. (1992). Prediction of gene structure. *J. Mol. Biol.* **226:** 141–157.

Hagemann, T., Surosky, R., Monaco, A. P., Lehrach, H., Rosen, F. S., and Kwan, S.-P. (1994). Physical mapping in a YAC contig of 11 markers on the human X chromosome in Xp11.23. *Genomics* **21:** 262–265.

Hogan, K., Powers, P. A., and Gregg, R. G. (1994). Cloning of the human skeletal muscle alpha 1 subunit of the dihydropyridine-sensitive L-type calcium channel (CACNL1A3). *Genomics* **24:** 608–609.

Lloyd, S. E., Pearce, S. H., Fisher, S. E., Steinmeyer, K., Schwappach, B., Scheinman, S. J., Harding, B., Bolino, A., Devoto, M., Goodyer, P., Rigden, S. P., Wrong, O., Jentsch, T. J., Craig, I. W., and Thakker, R. V. (1996). A common molecular basis for three inherited kidney stone diseases. *Nature* **379:** 445–449.

Lubs, H. A., Chiurazzi, P., Arena, J. F., Schwartz, C., Tranebjaerg, L., and Neri, G. (1996). XLMR genes: Update 1996. *Am. J. Med. Genet.* **64:** 147–157.

Meitinger, T., Fraser, N. A., Lorenz, B., Zrenner, E., Murken, J., and Craig, I. W. (1989). Linkage of X-linked retinitis pigmentosa to the hypervariable DNA marker M27beta (DXS255). *Hum. Genet.* **81:** 283–286.

Musarella, M. A., Kirschgessner, C., Trofatter, J., Miyake, Y., Mets, M., and De Gennaro, L. (1992). Assignment of the gene for incomplete congenital stationary night blindness (CSNB2) to proximal Xp. *Am. J. Hum Genet.* **51:** 772.

Nelson, D. L., Ballabio, A., Cremers, F., Monaco, A. P., and Schlessinger, D. (1995). Report of the sixth international workshop on X chromosome mapping. *Cytogenet. Cell Genet.* **71:** 307–342.

Oliva, M. M., Wu, T. C., and Yang, V. W. (1993). Isolation and characterization of a differentiation-dependent gene in the human colonic cell line HT29-18. *Arch. Biochem. Biophys.* **302:** 183–192.

Oliva, M. M., Cortese, J. F., and Yang, V. W. (1995). Promoter regulation of a differentially expressed gene in the human colonic epithelial cell lines HT29-18 and HT29-18-C-1. *Gene* **159:** 151–157.

Özcelik, T., Lafreniere, R. G., Archer, B. T., III, Johnston, P. A., Willard, H. F., Francke, U., and Sudhof, T. C. (1990). Synaptophysin: Structure of the human gene and assignment to the X chromosome in man and mouse. *Am. J. Hum. Genet.* **47:** 551–561.

Palmieri, G., Romano, G., Ciccodicola, A., Casamassimi, A., Campanile, C., Esposito, T., Cappa, V., Lania, A., Johnson, S., and Reinbold, R. (1994). YAC contig organization and CpG island analysis in Xq28. *Genomics* **24:** 149–158.

Schmeichel, K. L., and Beckerle, M. C. (1994). The LIM domain is a modular protein-binding interface. *Cell* **79:** 211–219.

Seino, S., Chen, L., Seino, M., Blondel, O., Takeda, J., Johnson, J. H., and Bell, G. I. (1992). Cloning of the alpha-1 subunit of a voltage-dependent calcium channel expressed in pancreatic beta cells. *Proc. Natl. Acad. Sci. USA* **89:** 584–588.

Tanabe, T., Takeshima, H., Mikami, A., Flockerzi, V., Takahashi, H., Kangawa, K., Kojima, M., Matsuo, H., Hirose, T., and Numa, S. (1987). Primary structure of the receptor for calcium channel blockers from skeletal muscle. *Nature* **328:** 313–318.

Uberbacher, E. C., and Mural, R. J. (1991). Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc. Natl. Acad. Sci. USA* **88:** 11261–11265.

Wilson, R., Ainscough, R., Anderson, K., Baynes, C., Berks, M., Bonfield, J., Burton, J., Connell, M., Copsey, T., Cooper, J., Coulson, A., Craxton, M., Dear, S., Du, Z., Durbin, R., Favello, A., Fulton, L., Gardner, A., Green, P., Hawkins, T., Hillier, L., Jier, M., Johnston, L., Jones, M., Kershaw, J., Kirsten, J., Laister, N., Latreille, P., Lightning, J., Lloyd, C., McMurray, A., Mortimore, B., O'Callaghan, M., Parsons, J., Percy, C., Rifken, L., Roopra, A., Saunders, D., Shownkeen, R., Smaldon, N., Smith, A., Sonnhammer, E., Staden, R., Sulston, J., Thierry-Mieg, J., Thomas, K., Vaudin, M., Vaughan, K., Waterston, R., Watson, A., Weinstock, L., Wilkinson-Sproat, J., and Wohldman, P. (1994). 2.2 Mb of contiguous nucleotide sequence from chromosome III of C. elegans. *Nature* **368:** 32–38.

Zollo, M., and Chen, E. Y. (1994). A manual high-throughput M13 DNA preparation. *BioTechniques* **16:** 370–372.

Zollo, M., Mazzarella, R., Bione, S., Toniolo, D., Schlessinger, D., D'Urso, M., and Chen, E. Y. (1995). Sequence and gene content in 52kb including and centromeric to the G6PD gene in Xq28. *DNA Seq.* **6:** 1–11.