Identification of the complete coding sequence and genomic organization of the Treacher Collins syndrome gene.

J Dixon, S J Edwards, I Anderson, A Brass, P J Scambler and M J Dixon

Genome Res. 1997 7: 223-234
Access the most recent version at doi:10.1101/gr.7.3.223

RESEARCH

Identification of the Complete Coding Sequence and Genomic Organization of the Treacher Collins Syndrome Gene

Jill Dixon,[1] Sara J. Edwards,[1] Isobel Anderson,[2] Andrew Brass,[2] Peter J. Scambler,[3] and Michael J. Dixon[1,4]

[1]School of Biological Sciences and Departments of Dental Medicine and Surgery, and [2]School of Biological Sciences, University of Manchester, Manchester M13 9PT, UK; [3]Molecular Medicine Unit, Institute of Child Health, London WC1N 1EH, UK

Treacher Collins syndrome (TCS) is an autosomal dominant disorder of craniofacial development, the features of which include conductive hearing loss and cleft palate. Recently, the demonstration of a series of 10 mutations within a partial-length cDNA clone have indicated that the TCS gene (*TCOF1*) has been positionally cloned. Although it has been shown that the gene is expressed in a wide variety of fetal and adult tissues, database sequence comparisons have failed to provide significant information on the function of the gene. In the current investigation, a combination of cDNA library screening and rapid amplification of cDNA ends has permitted the isolation of the complete coding sequence of *TCOF1*, which is encoded by 26 exons and predicts a low complexity, serine/alanine-rich protein of ~144 kD. The use of a variety of bioinformatics tools has resulted in the identification of repeated units within the gene, each of which maps onto an individual exon. The predicted protein *Treacle* contains numerous potential phosphorylation sites, a number of which map to similar positions within the repeated units, and shows weak but significant homology to the nucleolar phosphoproteins. Although the precise function of *Treacle* remains unknown, these observations suggest that phosphorylation may be important for its role in early embryonic development and that it may play a role in nucleolar–cytoplasmic shuttling. The information presented in this study will allow continued mutation analysis in families with a history of TCS and should facilitate continued experimentation to shed further light on the function of the gene/protein during development of the craniofacial complex.

[The sequence data described in this paper have been submitted to the GenBank data library under accession nos. U40847 and U79645–U79660.]

Treacher Collins syndrome (TCS) is an autosomal dominant disorder of facial development, which was probably first reported by Thompson (1846), but is named after E. Treacher Collins, who described the essential features of the syndrome in 1900 (Treacher Collins 1900). The incidence of TCS is thought to be ~1/50,000 live births (Fazen et al. 1967; Gorlin et al. 1990), with 60% of cases appearing to arise as the result of a de novo mutation (Jones et al. 1975). TCS displays high penetrance, with only one reported case of nonpenetrance (Dixon et al. 1994a) and a high degree of both inter- and intrafamilial phenotypic variability. The clinical manifestations include the following. (1) Abnormalities of the external ears, which are frequently associated with narrowing of the ear canals and abnormalities of the middle-ear ossicles. Bilateral conductive hearing loss is therefore a common feature of this disorder (Phelps et al. 1981). (2) Hypoplasia of the facial bones, particularly the mandible and zygomatic complex. (3) Downward slanting of the palpebral fissures with colobomas (notching) of the lower eyelids and a lack of eyelashes medial to the defect. (4) Cleft palate. These features are usually bilaterally symmetrical (Kay and Kay 1989); however, because of the variability in expression, it can be extremely difficult to reach a clinical diagnosis and to provide accurate genetic counselling.

The Treacher Collins syndrome locus (*TCOF1*) was initially linked to polymorphic markers from chromosome 5 at 5q31–34 (Dixon et al. 1991). This localization was subsequently confirmed by Jabs et al. (1991), and more recent studies have resulted in

[4]Corresponding author.
E-MAIL mdixon@fs2.scg.man.ac.uk; FAX + 44-161 275 5620.

the creation of a combined genetic and radiation hybrid map around the *TCOF1* locus (Dixon et al. 1993; Loftus et al. 1993). This map allowed a yeast artificial chromosome (YAC) contig of the region to be constructed (Jabs et al. 1993; Dixon et al. 1994b). Transcription mapping of the TCOF1 candidate region eventually led to the identification of a cDNA clone with an open reading frame (ORF) of 4142 bp that did not contain a translation initiation signal, a polyadenylation signal, or a poly(A) tail (Loftus et al. 1996; Treacher Collins Syndrome Collaborative Group 1996). Investigation of this cDNA led to the identification of different mutations in 10 unrelated families, all of which resulted in the introduction of a premature termination codon into the predicted protein product, *Treacle* (Gladwin et al. 1996; Treacher Collins Syndrome Collaborative Group 1996).

As the structures affected in TCS arise from the first and second pharyngeal arches, which in turn have a significant contribution from the neural crest, it has been proposed that the disorder may be the result of a defect in neural crest cell migration or improper cellular differentiation during development (Poswillo 1975; Wiley et al. 1983). The identification of the *TCOF1* gene has, however, failed to elucidate the biochemical nature of the disorder as initial database comparisons have indicated that it has no strong homologies with previously identified genes, gene families, or protein motifs of classic importance. In this study we report the complete sequence of *TCOF1* and its genomic organization, which has allowed us to identify repeated units within the gene that map onto individual exons. Analysis of *TCOF1/Treacle* using a number of bioinformatics tools has suggested further that protein phosphorylation may be important for its function.

RESULTS

Isolation of the Entire Coding Sequence of *TCOF1*

Screening of a number of cDNA libraries with portions of the original cDNA clone (Treacher Collins Syndrome Collaborative Group 1996) failed to identify any clones that extended the sequence previously presented in a 3′ direction. Extension of the sequence of the *TCOF1* gene in this direction was therefore achieved using rapid amplification of cDNA ends (RACE). Sequence analysis of a 3′ RACE product of ~600 bp isolated using this methodology revealed that it extended the cDNA sequence presented previously by 565 bp, including an additional 40 bp of coding sequence prior to the first in-frame termination codon. Initial screening of cDNA libraries also failed to identify a clone that extended the sequence in a 5′ direction. 5′ RACE produced two PCR products, the smaller of which did not extend the previously reported cDNA sequence. Sequencing of the larger product resulted in the identification of an additional 69 bp of sequence, including a strong Kozak consensus initiation sequence (Kozak 1987a,b). Screening of a human fetal brain cDNA library with the larger of the 5′ RACE products identified three cDNA clones, all of which provided additional sequence information. The longest of these clones contained the start codon and an in-frame termination codon (TAA) 75 bp upstream. RT–PCR analysis of the 5′ end of the gene flanking the initiation codon using RNA extracted from skeletal muscle and a lymphoblastoid cell line yielded a single PCR product of the predicted size, the sequence of which was in perfect agreement with that of the cDNA clone. The combined cDNA cloning and RACE strategies have therefore resulted in the identification of the complete coding sequence of *TCOF1,* which consists of an ORF of 4233 bp, followed by a termination codon and a 3′ untranslated region (UTR) of 507 bp, which contains a single polyadenylation signal. A 5′ UTR of 93 bp has also been identified (Fig. 1). This sequence predicts a 144-kD protein of 1411 amino acids (Fig. 1). The predicted protein is of low complexity with 5 amino acids, alanine (14.86%), serine (13.59%), lysine (11.18%), glutamic acid (9.13%), and proline (9.06%), accounting for the majority of residues.

Genomic Organization of the Gene

Experiments aimed at determining the genomic organization of *TCOF1* indicated that the gene is encoded by 26 exons, ranging in size from 49 to 561 bp (Table 1). Exon 1 contains the translation initiation signal, preceded by the 5′ UTR, which contains a large number of rare-cutter restriction sites, including two *Bss*HII, two *Fse*I, and two *Eag*I sites, within 93 bp. Exon 25 contains the last 24 bp of the coding sequence, the termination codon, and the first 22 bp of the 3′ UTR. The remainder of the 3′ UTR is encoded by exon 26. The intron/exon boundary sequences conform to the published consensus sequences (Table 1) (Breathnach and Chambon 1981), with the exception of exon 14, the splice donor site of which displays the sequence GC rather than the more usual GT. The intron/exon boundaries are of type 0 (splicing occurring between codons) for introns 1, 4, 6–16, 18, 19, 21, and 24;

AGTGGGGCGCGCGA
GGTCTAAGGGCGCGAGGGAAGTGGCGGGCGGGGACTAAGGCGGGGCGTGCAGGTAGCCGGCCGGCCGGGGGTCGCGGGT

```
   1  ATG GCC GAG GCC AGG AAG CGG CGG GAG CTA CTT CCC CTG ATC TAC CAC CAT CTG CTG CGG   60
   1  M   A   E   A   R   K   R   R   E   L   L   P   L   I   Y   H   H   L   L   R    20

  61  GCT GGC TAT GTG CGT GCG GCG CGG GAA GTG AAG GAG CAG AGC GGC CAG AAG TGT TTC CTG  120
  21  A   G   Y   V   R   A   A   R   E   V   K   E   Q   S   G   Q   K   C   F   L    40

 121  GCT CAG CCC GTA ACC CTT CTG GAC ATC TAT ACA CAC TGG CAA CAA ACC TCA GAG CTT GGT  180
  41  A   Q   P   V   T   L   L   D   I   Y   T   H   W   Q   Q   T   S   E   L   G    60

 181  CGG AAG CGG AAG GCA GAG GAA GAT GCG GCA CTG CAA GCT AAG AAA ACC CGT GTG TCA GAC  240
  61  R   K   R   K   A   E   E   D   A   A   L   Q   A   K   K   T   R   V   S   D    80

 241  CCC ATC AGC ACC TCG GAG AGC TCG GAA GAG GAG GAA GAA GCA GAA GCC GAA ACC GCC AAA  300
  81  P   I   S   T   S   E   S   S   E   E   E   E   E   A   E   A   E   T   A   K   100

 301  GCC ACC CCA AGA CTA GCA TCT ACC AAC TCC TCA GTC CTG GGG GCG GAC TTG CCA TCA AGC  360
 101  A   T   P   R   L   A   S   T   N   S   S   V   L   G   A   D   L   P   S   S   120

 361  ATG AAA GAA AAA GCC AAG GCA GAG ACA GAG AAA GCT GGC AAG ACT GGG AAT TCC ATG CCA  420
 121  M   K   E   K   A   K   A   E   T   E   K   A   G   K   T   G   N   S   M   P   140

 421  CAC CCT GCC ACT GGG AAG ACG GTG GCC AAC CTT CTT TCT GGG AAG TCT CCC AGG AAG TCA  480
 141  H   P   A   T   G   K   T   V   A   N   L   L   S   G   K   S   P   R   K   S   160

 481  GCA GAG CCC TCA GCA AAT ACT ACG TTG GTC TCA GAA ACT GAG GAG GAG GGC AGC GTC CCG  540
 161  A   E   P   S   A   N   T   T   L   V   S   E   T   E   E   E   G   S   V   P   180

 541  GCC TTT GGA GCT GCT GCC AAG CCT GGG ATG GTG TCA GCG GGC CAG GCC GAC AGC TCC AGC  600
 181  A   F   G   A   A   A   K   P   G   M   V   S   A   G   Q   A   D   S   S   S   200

 601  GAG GAC ACC TCC AGC TCC AGT GAT GAG ACA GAC GTG GAG GTA AAG GCC TCT GAA AAA ATT  660
 201  E   D   T   S   S   S   S   D   E   T   D   V   E   V   K   A   S   E   K   I   220

 661  CTC CAG GTC AGA GCT GCC TCA GCC CCT GCC AAG GGG ACC CCT GGG AAA GGG GCT ACC CCA  720
 221  L   Q   V   R   A   A   S   A   P   A   K   G   T   P   G   K   G   A   T   P   240

 721  GCA CCC CCT GGG AAG GCA GGG GCT GTA GCC TCC CAG ACC AAG GCA GGG AAG CCA GAG GAG  780
 241  A   P   P   G   K   A   G   A   V   A   S   Q   T   K   A   G   K   P   E   E   260

 781  GAC TCA GAG AGC AGC AGC GAG GAG TCA TCT GAC AGT GAG GAG GAG ACG CCA GCT GCC AAG  840
 261  D   S   E   S   S   S   E   E   S   S   D   S   E   E   E   T   P   A   A   K   280

 841  GCC CTG CTT CAG GCG AAG GCC TCA GGA AAA ACC TCT CAG GTC GGA GCT GCC TCA GCC CCT  900
 281  A   L   L   Q   A   K   A   S   G   K   T   S   Q   V   G   A   A   S   A   P   300

 901  GCC AAG GAG TCC CCC AGG AAA GGA GCT GCC CCA GCG CCC CCT GGG AAG ACA GGG CCT GCA  960
 301  A   K   E   S   P   R   K   G   A   A   P   A   P   P   G   K   T   G   P   A   320

 961  GTT GCC AAG GCC CAG GCG GGG AAG CGG GAG GAG GAC TCG CAG AGC AGC AGC GAG GAA TCG 1020
 321  V   A   K   A   Q   A   G   K   R   E   E   D   S   Q   S   S   S   E   E   S   340

1021  GAC AGT GAG GAG GAG GCG CCT GCT CAG GCG AAG CCT TCA GGG AAG GCC CCC CAG GTC AGA 1080
 341  D   S   E   E   E   A   P   A   Q   A   K   P   S   G   K   A   P   Q   V   R   360

1081  GCC GCC TCG GCC CCT GCC AAG GAG TCC CCC AGG AAA GGG GCT GCC CCA GCA CCT CCT AGG 1140
 361  A   A   S   A   P   A   K   E   S   P   R   K   G   A   A   P   A   P   P   R   380

1141  AAA ACA GGG CCT GCA GCC GCC CAG GTC CAG GTG GGG AAG CAG GAG GAG GAC TCA AGA AGC 1200
 381  K   T   G   P   A   A   A   Q   V   Q   V   G   K   Q   E   E   D   S   R   S   400

1201  AGC AGC GAG GAG TCA GAC AGT GAC AGA GAA GCA CTG GCA GCC ATG AAT GCA GCT CAG GTG 1260
 401  S   S   E   E   S   D   S   D   R   E   A   L   A   A   M   N   A   A   Q   V   420

1261  AAG CCC TTG GGG AAA AGC CCC CAG GTG AAA CCT GCC TCT ACC ATG GGC ATG GGG CCC TTG 1320
 421  K   P   L   G   K   S   P   Q   V   K   P   A   S   T   M   G   M   G   P   L   440

1321  GGG AAA GGC GCC GGC CCA GTG CCA CCT GGG AAG GTG GGG CCT GCA ACC CCC TCA GCC CAG 1380
 441  G   K   G   A   G   P   V   P   P   G   K   V   G   P   A   T   P   S   A   Q   460

1381  GTG GGG AAG TGG GAG GAG GAC TCA GAG AGC AGT AGT GAG GAG TCA TCA GAC AGC AGT GAT 1440
 461  V   G   K   W   E   E   D   S   E   S   S   S   E   E   S   S   D   S   S   D   480

1441  GGA GAG GTG CCC ACA GCT GTG GCC CCG GCT CAG GAA AAG TCC TTG GGG AAC ATC CTC CAG 1500
 481  G   E   V   P   T   A   V   A   P   A   Q   E   K   S   L   G   N   I   L   Q   500

1501  GCC AAA CCC ACC TCC AGT CCT GCC AAG GGG CCC CCT CAG AAG GCA GGG CCT GTA GCC GTC 1560
 501  A   K   P   T   S   S   P   A   K   G   P   P   Q   K   A   G   P   V   A   V   520

1561  CAG GTC AAG GCT GAA AAG CCC ATG GAC AAC TCG GAG AGC AGC GAG GAG TCG TCG GAC AGT 1620
 521  Q   V   K   A   E   K   P   M   D   N   S   E   S   S   E   E   S   S   D   S   540

1621  GCG GAC AGT GAG GAG GCA CCA GCA GCC ATG ACT GCA GCT CAG GCA AAA CCA GCT CTG AAA 1680
 541  A   D   S   E   E   A   P   A   A   M   T   A   A   Q   A   K   P   A   L   K   560
```

Figure 1 *(See p. 227 for legend.)*

```
1681  ATT CCT CAG ACC AAG GCC TGC CCA AAG AAA ACC AAT ACC ACT GCA TCT GCC AAG GTC GCC  1740
 561  I   P   Q   T   K   A   C   P   K   K   T   N   T   T   A   S   A   K   V   A    580

1741  CCT GTG CGA GTG GGC ACC CAA CCC CCC CGG AAA GCA GGA ACT GCG ACT TCT CCA GCA GGC  1800
 581  P   V   R   V   G   T   Q   P   P   R   K   A   G   T   A   T   S   P   A   G    600

1801  TCA TCC CCA GCT GTG GCT GGG GGC ACC CAG AGA CCA GCA GAG GAT TCT TCA AGC AGT GAG  1860
 601  S   S   P   A   V   A   G   G   T   Q   R   P   A   E   D   S   S   S   S   E    620

1861  GAA TCA GAT AGT GAG GAA GAG AAG ACA GGT CTT GCA GTA ACC GTG GGA CAG GCA AAG TCT  1920
 621  E   S   D   S   E   E   E   K   T   G   L   A   V   T   V   G   Q   A   K   S    640

1921  GTG GGG AAA GGC CTC CAG GTG AAA GCA GCC TCA GTG CCT GTC AAG GGG TCC TTG GGG CAA  1980
 641  V   G   K   G   L   Q   V   K   A   A   S   V   P   V   K   G   S   L   G   Q    660

1981  GGG ACT GCT CCA GTA CTC CCT GGG AAG ACG GGG CCT ACA GTC ACC CAG GTG AAA GCT GAA  2040
 661  G   T   A   P   V   L   P   G   K   T   G   P   T   V   T   Q   V   K   A   E    680

2041  AAG CAG GAA GAC TCT GAG AGC AGT GAG GAG GAA TCA GAC AGT GAG GAA GCA GCT GCA TCT  2100
 681  K   Q   E   D   S   E   S   S   E   E   E   S   D   S   E   E   A   A   A   S    700

2101  CCA GCA CAG GTG AAA ACC TCA GTA AAG AAA ACC CAG GCC AAA GCC AAC CCA GCT GCC GCC  2160
 701  P   A   Q   V   K   T   S   V   K   K   T   Q   A   K   A   N   P   A   A   A    720

2161  AGA GCA CCT TCA GCA AAA GGG ACA ATT TCA GCC CCT GGA AAA GTT GTC ACT GCA GCT GCT  2220
 721  R   A   P   S   A   K   G   T   I   S   A   P   G   K   V   V   T   A   A   A    740

2221  CAA GCC AAG CAG AGG TCT CCA TCC AAG GTG AAG CCA CCA GTG AGA AAC CCC CAG AAC AGT  2280
 741  Q   A   K   Q   R   S   P   S   K   V   K   P   P   V   R   N   P   Q   N   S    760

2281  ACC GTC TTG GCG AGG GGC CCA GCA TCT GTG CCA TCT GTG GGG AAG GCC GTG GCT ACA GCA  2340
 761  T   V   L   A   R   G   P   A   S   V   P   S   V   G   K   A   V   A   T   A    780

2341  GCT CAG GCC CAG ACA GGG CCA GAG GAG GAC TCA GGG AGC AGT GAG GAG GAG TCA GAC AGT  2400
 781  A   Q   A   Q   T   G   P   E   E   D   S   G   S   S   E   E   E   S   D   S    800

2401  GAG GAG GAG GCG GAG ACG CTG GCT CAG GCG AAG CCT TCA GGG AAG ACC CAC CAG ATC AGA  2460
 801  E   E   E   A   E   T   L   A   Q   A   K   P   S   G   K   T   H   Q   I   R    820

2461  GCT GCC TTG GCT CCT GCC AAG GAG TCC CCC AGG AAA GGG GCT GCC CCA ACA CCT CCT GGG  2520
 821  A   A   L   A   P   A   K   E   S   P   R   K   G   A   A   P   T   P   P   G    840

2521  AAG ACA GGG CCT TCG GCT GCC CAG GCA GGG AAG CAG GAT GAC TCA GGG AGC AGC AGC GAG  2580
 841  K   T   G   P   S   A   A   Q   A   G   K   Q   D   D   S   G   S   S   S   E    860

2581  GAA TCA GAC AGT GAT GGG GAG GCA CCG GCA GCT GTG ACC TCT GCC CAG GTG ATT AAA CCC  2640
 861  E   S   D   S   D   G   E   A   P   A   A   V   T   S   A   Q   V   I   K   P    880

2641  CCC CTG ATT TTT GTC GAC CCT AAT CGT AGT CCA GCT GGC CCA GCT GCT ACA CCC GCA CAA  2700
 881  P   L   I   F   V   D   P   N   R   S   P   A   G   P   A   A   T   P   A   Q    900

2701  GCC CAG GCT GCA AGC ACC CCG AGG AAG GCC CGA GCC TCG GAG AGC ACA GCC AGG AGC TCC  2760
 901  A   Q   A   A   S   T   P   R   K   A   R   A   S   E   S   T   A   R   S   S    920

2761  TCC TCC GAG AGC GAG GAT GAG GAC GTG ATC CCC GCT ACA CAA TGC TTG ACT CCT GGC ATC  2820
 921  S   S   E   S   E   D   E   D   V   I   P   A   T   Q   C   L   T   P   G   I    940

2821  AGA ACC AAT GTG GTG ACC ATG CCC ACT GCC CAC CCA AGA ATA GCC CCC AAA GCC AGC ATG  2880
 941  R   T   N   V   V   T   M   P   T   A   H   P   R   I   A   P   K   A   S   M    960

2881  GCT GGG GCC AGC AGC AGC AAG GAG TCC AGT CGG ATA TCA GAT GGC AAG AAA CAG GAG GGA  2940
 961  A   G   A   S   S   S   K   E   S   S   R   I   S   D   G   K   K   Q   E   G    980

2941  CCA GCC ACT CAG GTG TCA AAG AAG AAC CCA GCT TCC CTC CCA CTG ACC CAG GCT GCC CTG  3000
 981  P   A   T   Q   V   S   K   K   N   P   A   S   L   P   L   T   Q   A   A   L   1000

3001  AAG GTC CTC GCC CAG AAA GCC AGT GAG GCT CAG CCT CCT GTT GCC AGG ACC CAG CCT TCA  3060
1001  K   V   L   A   Q   K   A   S   E   A   Q   P   P   V   A   R   T   Q   P   S   1020

3061  AGT GGG GTT GAC AGT GCT GTG GGA ACA CTC CCT GCA ACA AGT CCC CAG AGC ACC TCC GTC  3120
1021  S   G   V   D   S   A   V   G   T   L   P   A   T   S   P   Q   S   T   S   V   1040

3121  CAG GCC AAA GGG ACC AAC AAG CTC AGA AAA CCT AAG CTT CCT GAG GTC CAG CAG GCC ACC  3180
1041  Q   A   K   G   T   N   K   L   R   K   P   K   L   P   E   V   Q   Q   A   T   1060

3181  AAA GCC CCT GAG AGC TCA GAT GAC AGT GAG GAC AGC AGC GAC AGT TCT TCA GGG AGT GAG  3240
1061  K   A   P   E   S   S   D   D   S   E   D   S   S   D   S   S   S   G   S   E   1080

3241  GAA GAT GGT GAA GGG CCC CAG GGG GCC AAG TCA GCC CAC ACG CTG GGT CCC ACC CCC TCC  3300
1081  E   D   G   E   G   P   Q   G   A   K   S   A   H   T   L   G   P   T   P   S   1100

3301  AGG ACA GAG ACC CTG GTG GAG GAG ACC GCA GCA GAG TCC AGC GAG GAT GAT GTG GTG GCG  3360
1101  R   T   E   T   L   V   E   E   T   A   A   E   S   S   E   D   D   V   V   A   1120

3361  CCA TCC CAG TCT CTC CTC TCA GGT TAT ATG ACC CCT GGA CTA ACC CCA GCC AAT TCC CAG  3420
1121  P   S   Q   S   L   L   S   G   Y   M   T   P   G   L   T   P   A   N   S   Q   1140
```

Figure 1 (*Continued*)

type 1 (splicing occurring after the first base of the codon) for introns 3, 5, 17, 20, 22, and 23; and type 2 (splicing occurring after the second base of the codon) for intron 2. In many cases, the introns were sequenced in their entirety, which permitted their sizes to be determined accurately, whereas the sizes of other introns were estimated by PCR (Table 1). Interestingly, sequence analysis of the 3′ UTR indicated that it was not single exonic, an intron of ~730 bp, as determined by the PCR, being present between exons 25 and 26. Introns 1, 3, 6, 13, and 16 consistently could not be amplified using PCR, and thus these introns were considered to be out of the range of PCR amplification under the conditions used in this study. The entire sequence data generated in this study have been submitted to GenBank under the following accession numbers: U79645 (exon 1); U79646 (exon 2); U79647 (exon 3); U79648 (exon 4); U79649 (exon 5); U79650 (exon 6); U79651 (exon 7 to exon 13); U79652 (exon 14 to exon 16); U79653 (exon 17); U79654 (exon 18); U79655 (exons 19 and 20); U79656 (exon 21); U79657 (exon 22); U79658 (exon 23); U79659 (exons 24 and 25); and U79660 (exon 26). All available intronic sequence has been submitted under the appropriate accession number.

Bioinformatics Analysis

As initial database sequence comparisons failed to show any strong homologies between *TCOF1* and previously identified genes, gene families, or motifs of classic importance, a number of bioinformatics programs were employed in the current study. The use of dot plots allowed the identification of repeated units within the gene, each of which was

```
3421  GCC TCA AAA GCC ACT CCC AAG CTA GAT TCC AGC CCC TCA GTT TCC TCT ACT CTG GCC GCC  3480
1141  A   S   K   A   T   P   K   L   D   S   S   P   S   V   S   S   T   L   A   A    1160

3481  AAA GAT GAC CCA GAT GGC AAG CAG GAG GCA AAG CCC CAA CAG GCA GCA GGC ATG TTG TCC  3540
1161  K   D   D   P   D   G   K   Q   E   A   K   P   Q   Q   A   A   G   M   L   S    1180

3541  CCT AAA ACA GGT GGA AAA GAG GCT GCT TCA GGC ACC ACA CCT CAG AAG TCC CGG AAG CCC  3600
1181  P   K   T   G   G   K   E   A   A   S   G   T   T   P   Q   K   S   R   K   P    1200

3601  AAG AAA GGG GCT GGG AAC CCC CAA GCC TCA ACC CTG GCG CTG CAA AGC AAC ATC ACC CAG  3660
1201  K   K   G   A   G   N   P   Q   A   S   T   L   A   L   Q   S   N   I   T   Q    1220

3661  TGC CTC CTG GGC CAA CCC TGG CCC CTG AAT GAG GCC CAG GTG CAG GCC TCA GTG GTG AAG  3720
1221  C   L   L   G   Q   P   W   P   L   N   E   A   Q   V   Q   A   S   V   V   K    1240

3721  GTC CTG ACT GAG CTG CTG GAA CAG GAA AGA AAG AAG GTG GTG GAC ACC ACC AAG GAG AGC  3780
1241  V   L   T   E   L   L   E   Q   E   R   K   K   V   V   D   T   T   K   E   S    1260

3781  AGC AGG AAG GGC TGG GAG AGC CGC AAG CGG AAG CTA TCG GGA GAC CAG CCA GCT GCC AGG  3840
1261  S   R   K   G   W   E   S   R   K   R   K   L   S   G   D   Q   P   A   A   R    1280

3841  ACC CCC AGG AGC AAG AAG AAG AAG AAG CTG GGG GCC GGG GAA GGT GGG GAG GCC TCT GTT  3900
1281  T   P   R   S   K   K   K   K   K   L   G   A   G   E   G   G   E   A   S   V    1300

3901  TCC CCA GAA AAG ACC TCC ACG ACT TCC AAG GGG AAA GCA AAG AGA GAC AAA GCA AGT GGT  3960
1301  S   P   E   K   T   S   T   T   S   K   G   K   A   K   R   D   K   A   S   G    1320

3961  GAT GTC AAG GAG AAG AAA GGG AAG GGG TCT CTT GGC TCC CAA GGG GCC AAG GAC GAG CCA  4020
1321  D   V   K   E   K   K   G   K   G   S   L   G   S   Q   G   A   K   D   E   P    1340

4021  GAA GAG GAG CTT CAG AAG GGG ATG GGG ACG GTT GAA GGT GGA GAT CAA AGC AAC CCA AAG  4080
1341  E   E   E   L   Q   K   G   M   G   T   V   E   G   G   D   Q   S   N   P   K    1360

4081  AGC AAG AAG GAG AAG AAG AAA TCC GAC AAG AGA AAA AAA GAC AAA GAA AAA AAA GAA AAG  4140
1361  S   K   K   E   K   K   K   S   D   K   R   K   K   D   K   E   K   K   E   K    1380

4141  AAG AAG AAA GCA AAA AAG GCC TCA ACC AAA GAT TCT GAG TCA CCG TCC CAG AAG AAA AAG  4200
1381  K   K   K   A   K   K   A   S   T   K   D   S   E   S   P   S   Q   K   K   K    1400

4201  AAG AAA AAG AAG AAG ACA GCA GAG CAG ACT GTA TGA                                  4236
1401  K   K   K   K   K   T   A   E   Q   T   V   *                                    1412

4237  CGAGCACCAGCACCAGGCACAGGGATTTCCTAGCCGAGCAGTGGCCATCCCCATGCCTCTGACCTCCACCGACCTCTGC  4316
4317  CCACCATGGGTTGGAACTAAACTGTTACCTTCCCTCGCTCCACAGAAGAAGACAGCCAGCTTCAGGGGTCCCTGTGCTG  4395
4396  GCCAAGCCAGTGAGCCTGCGGGGAGGCTGGTCCAAGGAGAAAGTGGACCAGCTCCCATGACCTCACCCCACTCCCCCAA  4474
4475  CACAGGACGCTTCATATAGATGTGTACAGTATATGTATTTTTTTAAGTGACCTCCTCTCCTTCCACAGACCCCACATGC  4553
4556  CCAAAGGCCTCGGGACTTCCCACCACCTTGCTCCACAGATCCAGCTAGGCCTGACCTGTGCCTCATCCCGTGCCGCTCG  4632
4633  GTCTCTGGCTGATCCCGAGGCTTTGTCTTCCTCTCGTCAGTTCTTTTGGTTGTGTTTTTTGTTTTTTTTTAATAACTCA  4711
4712  AAAAAAAATAAAAGACTTGGAGGAAGGGTGAAAAAAAAAAAAAAAAAAA                               4760
```

Figure 1 Complete nucleotide sequence of *TCOF1* (U40847) and the derived amino acid sequence of the protein Treacle. Position 1 is taken as the first base of the initiation codon. The single polyadenylation signal is underlined. The new sequence data (nucleotide and amino acid) are in boldface type.

Table 1. Genomic Organization of the *TCOF1* Locus

Exon	Size (bp)	Splice Donor	Splice Acceptor	Nucleotide Position	Intron Size
1	-	AGC GGC CAG gtaagcgttc	ctctctgcag AAG TGT TTC	5′ UTR - 108	-
2	56	TGG CAA CA gtaagtggtg	tgtcctgcag A ACC TCA	109-164	2.7kb (P)
3	140	AAA GCC A gtaagagcct	tttcttgcag CC CCA AGA	165-304	-
4	74	AAA GCC AAG gtgagtggga	ttctctgtag GCA GAG ACA	305-378	860bp (P)
5	187	AAG CCT G gtaagaagtc	cgatcctcag GG ATG GTG	379-565	600bp (P)
6	74	GAC GTG GAG gtaattgcca	ttttcaccag GTA AAG GCC	566-639	-
7	213	CTG CTT CAG gtgaggcctg	gtttctccag GCG AAG GCC	640-852	230bp (S)
8	195	CCT GCT CAG gtgaggcaga	ctcactccag GCG AAG CCT	853-1047	142bp (S)
9	210	GCA GCT CAG gtgaggctgg	tgtctcccag GTG AAG CCC	1048-1257	175bp (S)
10	216	CCG GCT CAG gtgaggcccc	ctcactccag GAA AAG TCC	1258-1473	166bp (S)
11	189	GCA GCT CAG gtgaggcctg	gtcccctcag GCA AAA CCA	1474-1662	171bp (S)
12	249	GTG GGA CAG gtgaggcctg	gtcatcccag GCA AAG TCT	1663-1911	93bp (S)
13	198	CCA GCA CAG gtgaggccta	ctccactcag GTG AAA ACC	1912-2109	-
14	138	CCA TCC AAG gcaagtgggg	tgcaattcag GTG AAG CCA	2110-2247	185bp (S)
15	180	CTG GCT CAG gtgaggggga	ctccctccag GCG AAG CCT	2248-2427	118bp (S)
16	201	TCT GCC CAG gtaagacttg	gtttttcaag GTG ATT AAA	2428-2628	-
17	187	ACT CCT G gtgagcgacc	tccatttcag GC ATC AGA	2629-2815	2kb (P)
18	137	GCC ACT CAG gtacctggtg	ccacccacag GTG TCA AAG	2816-2952	1.6kb (P)
19	114	TCA AGT GGG gtgagcttcc	accgaattag GTT GAC AGT	2953-3066	298bp (S)
20	220	ACG CTG G gtgagggtgc	tctccagtag GT CCC ACC	3067-3286	540bp (P)
21	83	CCA TCC CAG gtaactgcaa	gcttcttcag TCT CTC CTC	3287-3369	570bp (P)
22	181	AAA ACA G gtaagttaag	ctctccatag GT GGA AAA	3370-3550	2.5kb (P)
23	561	GAC AAG A gtgagtgacc	cttcccttag GA AAA AAA	3551-4111	1.6kb (P)
24	98	AAG AAA AAG gtagagagtt	ctcctcacag AAG AAG ACA	4112-4209	570bp (S)
25	49	GGCACAG gtacgcttcc	cttcccctag GGATTTCC	4210-4258	730bp (P)
26		3′ untranslated region		4259-4740	

The cDNA position of the intron/exon boundaries is defined relative to the sequence presented in this study with position 1 being the first base of the initiation codon. The intronic sequence is indicated in lowercase; the exonic sequence in uppercase. The codon usage is also indicated. The intron sizes were either determined by sequencing (S) or by PCR analysis (P). The genomic sequence information has been submitted to GenBank under accession nos. U79645–U79660, inclusive.

shown to map onto an individual exon. With the exception of exon 14, the repeated units extended from exon 7 to 16. In this region all of the intron/exon boundaries are of type 0 and the splice donor sites are all of the consensus sequence /gtgagg, with the exception of the atypical intron 14 (Table 1). Moreover, all of the introns in this region, with the exception of intron 13, which precedes the atypical exon 14, have been sequenced and shown to be small, ranging in size from 93 to 230 bp (Table 1). The repeat units were multiply aligned and shown to be highly conserved (Fig. 2). Exons 7, 8, 9, 10, 13,

```
 8 - AKASGKTSQV GAASAPAKES PRKGAAPA.. ...PPGKTGP AVAKAQ.... ..AGKREEDS QSSSEE.SD. SEEEA..... PAQ
 9 - AKPSGKAPQV RAASAPAKES PRKGAAPA.. ...PPRKTGP AAAQVQ.... ..VGKQEEDS RSSSEE.SD. SDREALAAMN AAQ
16 - AKPSGKTHQI RAALAPAKES PRKGAAPT.. ...PPGKTGP SAAQAG.... ....KQ.DDS GSSSEE.SD. SDGEAPAAVT SAQ
 7 - VKASEKILQV RAASAPAKGT PGKGATPA.. ...PPGKAGA VASQTK.... ..AGKPEEDS ESSSEESSD. SEEETPAAKA LLQ
10 - VKPLGKSPQV KPASTMGMGP LGKGAGPV.. ...PPGKVGP ATPSAQ.... ..VGKWEEDS ESSSEESSDS SDGEVPTAVA PAQ
13 - AKSVGKGLQV KAASVPVKGS LGQGTAPV.. ...LPGKTGP TVTQVK.... ..AEK.QEDS ESSEEESDSE EAAAS..... PAQ
11 - EKSLGNILQA KPTSSPAKG. .......... ...PPQKAGP VAVQVK.... ..AEKPMDNS ESSEESSDSA DSEEAPAAMT AAQ
15 - VKPPVRNPQN STVLARG... ...PASVP.. ...SVGKAVA TAAQAQ.... ..TG.PEEDS GSSEEESDSE EEAET..... LAQ
12 - AKPALKIPQT KACPKKTNTT ASAKVAPVRV GTQPPRKAGT ATSPAGSSPA VAGGTQRPAE DSSSSEESDS EEEKTGLAVT VGQ
```

Figure 2 Alignment of the most closely related repeated units identified within *Treacle*. The exon onto which each motif maps is indicated at *left*; the area of multiple potential sites for CKII phosphorylation is in boldface type.

and 16 were particularly homologous, with the most closely related being exons 8 and 16 (77.41% identity, 87.10% similarity) and the least closely related being exons 7 and 13 (50% identity, 59.09% similarity) (Fig. 2). The protein sequence was also compared against pattern databases in an attempt to identify functional motifs within the sequence. These comparisons resulted in the identification of multiple motifs for casein kinase II phosphorylation (CKII) and protein kinase C phosphorylation. However, as these elements tend to be overpredicted by the programs, their significance is uncertain. It is, however, striking that multiple sites for CKII phosphorylation were observed in an identical position within each repeated unit (Fig. 2). Weak similarity to *Xenopus laevis* nucleolar phosphoprotein (S57757; $P = 4.2^{e-9}$) (Cairns and McStay 1995) and nucleolar phosphoprotein 140 (M94288; $P = 2.6^{e-8}$) (Meier and Blobel 1992) were demonstrated using the BLAST programs (Altschul et al. 1990). These highly phosphorylated proteins have a role in protein transport between the cytoplasm and the nucleus (Meier and Blobel 1992). Alignment of *Treacle* with these proteins indicated that this was predominantly a result of homology in the regions of the CKII sites within the repeated units of *Treacle* (Fig. 3). As is the case for the nucleolar phosphoproteins, the CKII phosphorylation sites occur within clusters of acidic amino acids, which are separated by stretches of residues that are relatively rich in lysine, proline, and alanine, with few acidic residues and are, therefore, basic in nature (Fig. 2). In addition, both *Treacle* and the nucleolar phosphoproteins display a number of motifs of the type K-K/R-X-R/K, which represents the minimal nuclear localization signal consensus sequence, toward the 3′ end of the coding sequence.

DISCUSSION

In this study we report the entire coding sequence of the *TCOF1* gene and its genomic organization. The gene has an ORF of 4233 bp encoded by 25 exons. Exons 4, 5, 6, 15, 18, and 25 have also been recovered by exon amplification strategies (Treacher Collins Syndrome Collaborative Group 1996; Gladwin et al. 1996; this study). All of the splice junctions of the gene conform to the published consensus sequences (Breathnach and Chambon 1981), with the exception of the splice donor site of exon 14, which has GC in place of the expected GT. A number of other genes exhibiting this unusual variation have been published, including the fibrillin gene (Pereira et al. 1993). Although the significance underlying this splice variant is unclear, it seems to be prone to mis-splicing events resulting from apparently minor nucleotide changes in the extended splice consensus sequence (Gladwin et al. 1996). A second unusual variation in the genomic organization of the gene has been observed in the current study in that the 3′ UTR of *TCOF1* is not single exonic as has been reported for the vast majority of genes (Wilcox et al. 1991). Exceptions to this rule include the $p58^{clk-1}$ protein kinase gene (Eipers et al. 1992) and the genes encoding the β1-subunit of the voltage-dependent sodium channel in mouse, rat, and human (Dib-Hajj and Waxman 1995). These genes do not, however, appear to be related. The presence of the intron in the 3′ UTR is validated by the fact that exon 25 was successfully "rescued" by exon amplification using the pSPL3 vector system, which does not rescue the 3′ terminal exon. The sequence and splicing of the exon-amplified product corresponds exactly to the sequence of exon 25.

The nucleotide sequence of *TCOF1* predicts a protein, *Treacle*, of 1411 amino acids, that is relatively rich in alanine and serine residues. Although *Treacle* does not exhibit strong homology to any known proteins, the use of a number of bioinformatics tools has highlighted several interesting features of the gene. There is a series of repeated units of unknown function within the gene that map onto individual exons. The splicing phase of all of the exons in this region is the same (type-0 junctions); therefore, it is possible that this region of the

gene has arisen by exon duplication during evolution. The gene also appears to be rich in potential phosphorylation sites, a number of which map to a similar position within the repeated units. However, as these motifs tend to be overpredicted by the programs, their significance remains unclear. In this regard cloning of the homolog of *TCOF1* in other species will help to identify conserved and, hence, potentially functionally important, domains. Nevertheless, *Treacle* shows weak but significant homology to a small number of phosphorylated proteins, the nucleolar phosphoproteins. Alignments have indicated further that these homologies are detected most strongly in the regions of the CKII sites that exist within the repeated region of *Treacle*. Although the homology between *Treacle* and the nucleolar phosphoproteins is weak, the proteins do have several features in common. In all cases, alanine, serine, lysine, glutamic acid, and proline make up the majority of the amino acids. These low complexity proteins do, however, contain repeating units that consist of sites for CKII phosphorylation, which are embedded in clusters of acidic amino acids separated by stretches of basic residues. In the case of *Treacle*, we have identified 10 repeat units, although 1 appears to be atypical. Interestingly, rat nucleolar phosphoprotein and its human homolog also contain 10 repeats (Meier and Blobel 1992; Pai et al. 1995), although the *Xenopus* homolog contains 17 (Cairns and McStay 1995). In the case of *TCOF1*, these repeating units map onto individual exons; however, the genomic organization of the nucleolar phosphoproteins has not been determined, to our knowledge. Finally, all of the proteins possess nuclear localization signals toward their carboxyl termini. Although the precise role of the nucleolar phosphoproteins has not been elucidated, they have been shown to shuttle along curvilinear tracks from the nucleolus to the cytoplasm and have, as such, been implicated as a chaperone in nucleolar–cytoplasmic transport. In this regard, it has been suggested that the alternating acidic and basic domains could function to cover and neutralize highly charged domains of preribosomal particles

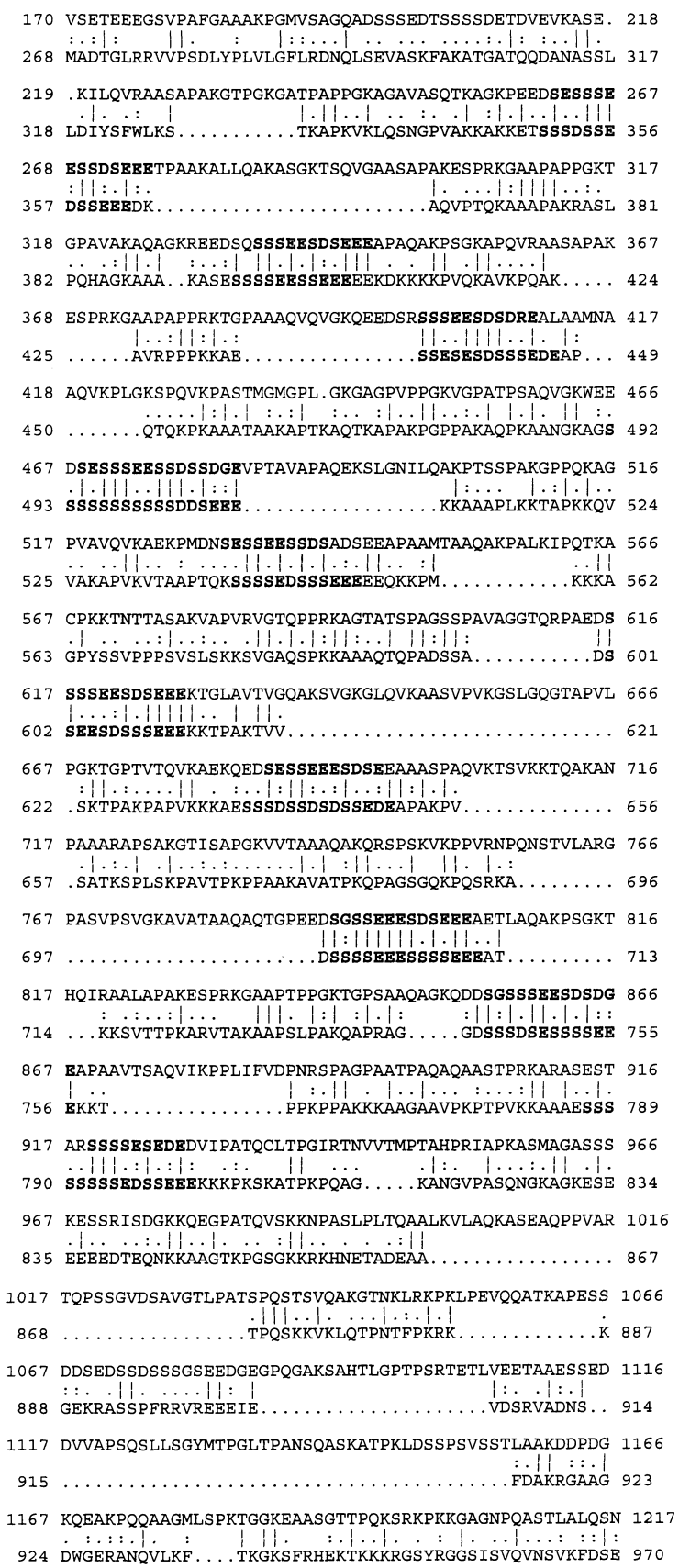

Figure 3 (*See facing page for legend.*)

(export) or ribosomal proteins (import) (Meier and Blobel 1992). The role of phosphorylation may be to increase the net negative charge in the acidic domain, thus increasing the affinity of the protein for oppositely charged species such as basic ribosomal proteins (Cairns and McStay 1995). Interestingly, recent data suggest that the complex events of epithelial–mesenchymal transformation in the neural crest cell system can be triggered by epigenetic events involving differential protein phosphorylation (Newgreen and Minichiello 1995). Nevertheless, although these interesting similarities provide potential insight into the function of the protein, additional experiments are required to confirm its role during facial development.

On the basis of the TCS phenotype, the gene must play a fundamental role in early embryonic development, possibly via an effect on neural crest cell migration and/or differentiation (Poswillo 1975; Wiley et al. 1983). In this regard, determination of the precise spatiotemporal expression patterns of the gene, and its protein product *Treacle* during development will be important in the investigation of the function of the gene. Here, the use of animal models, where accurately staged embryos from a wide variety of developmental stages are readily obtainable, will be extremely valuable. Furthermore, the isolation of the complete coding sequence of *TCOF1* and the determination of its genomic organization reported here will prove to be helpful in the isolation and characterization of its murine homolog, which is essential in the creation of a mouse model of TCS by gene targeting.

As the vast majority of the mutations that have been reported in *TCOF1* to date are unique to the family in which they were identified and are spread throughout the gene (Gladwin et al. 1996; Treacher Collins Syndrome Collaborative Group 1996; M.J. Dixon and S.J. Edwards, unpubl.), elucidation of the complete genomic organization of the gene will facilitate continued mutation screening, which may provide further information on functionally important domains within the gene and confirm the hypothesis that TCS results from haploinsufficiency. With the exception of exon 23 of *TCOF1*, all of the exons are <250 bp in size and are therefore of an appropriate size to be analyzed as a single fragment by single-stranded conformation polymorphism (SSCP) analysis (Orita et al. 1989). Although this technique is not sufficiently sensitive to identify all potential mutations, it does provide an excellent balance between simplicity and sensitivity of detection. In addition to providing information on the mutational spectrum underlying TCS, continued mutation analysis will also prove to be important for postnatal diagnosis of TCS, particularly in cases where there is some doubt as to diagnosis of either parent of a child with obvious clinical signs of TCS. Molecular diagnosis will also be important in providing prenatal diagnostic predictions for "at-risk" families (Edwards et al. 1996) and in assessing whether patients with conditions in which the facial gestalt has some similarities to the TCS facies, such as Nager and Miller syndrome (Gorlin et al. 1990), are also attributable to mutations in *TCOF1*.

METHODS

5′ and 3′ RACE

First-strand cDNA synthesis was performed on 500 ng of poly(A)$^+$ RNA isolated from skeletal muscle (Clontech) and a lymphoblastoid cell line using the 5′ or 3′ RACE kit (BRL) according to the manufacturer's instructions. In the case of 5′ RACE, cDNA synthesis was initiated from the gene-specific primer 5′-TTCCCAGTCTTGCCAG-3′. The original mRNA template was then removed by treatment with RNase H. In the case of 5′ RACE products, a homopolymeric tail was added to the 3′ end of the cDNA using terminal deoxynucleotide transferase (Tdt) and dCTP. PCR amplification of the target cDNA was performed using the universal amplification primer and a 5′ RACE gene-specific primer, 5′-TTTCGGCTTCTGCTTCTTCC-3′, or a 3′ RACE gene-specific primer, 5′-GGCCTCAACCAAAGATTC-3′. After electrophoretic separation on a 1% agarose gel, the PCR product was diluted and subject to a second round of PCR using the abridged universal amplification primer and a nested 5′ RACE gene-specific primer, 5′-CTGAATTCAGATGTCCAGAAGGGTTACG-3′ or a nested 3′ RACE gene-specific primer, 5′-AAGAATTCTGAGTCACCGTCCCAG-3′. The resulting PCR products were gel-purified, digested with *Eco*RI/*Sal*I, cloned into M13mp18/19, and sequenced via the dideoxy chain termination method (Sanger et al. 1977) using the Sequenase version 2.0 kit (U.S. Biochemical Corp.).

RT–PCR

One microgram of RNA extracted from skeletal muscle or a lymphoblastoid cell line was incubated with 100 ng of random primer at 70°C for 10 min. The samples were chilled on ice, and Moloney murine leukemia virus (MMLV) reverse

Figure 3 Comparison of the predicted amino acid sequences of the human *Treacle* protein and the rat nucleolar phosphoprotein using BESTFIT alignment. The *top* line represents the single-letter amino acid sequence of human *Treacle*; and the *bottom* line is the rat nucleolar phosphoprotein sequence. The sequences display 32.43% identity and 46.86% similarity. The regions containing multiple potential sites for CKII phosphorylation are in boldface type.

transcriptase buffer, 10 mM DTT, 1 mM dNTPs (all BRL), and 0.5 units of RNAsin (Promega) were added. The reactions were equilibrated at 37°C for 2 min, 100 units of MMLV reverse transcriptase was added, and the samples incubated at 37°C for 1 hr. The samples were then heated to 95°C for 5 min, and 3 µl of cDNA was used in the PCR with the primers 5′-TTGGATCCAAGTGGGGCGCGCGAGGT-3′ and 5′-TCGAATTCTGGTAGATCAGGGGAAGTAG-3′. Control reactions included those performed in the absence of RNA or in the absence of reverse transcriptase.

PCR Conditions

PCR assays were performed in 25-µl volumes containing 50 pmoles of each primer; 200 µM each of dCTP, dGTP, dTTP, and dATP; 10 mM Tris-HCl at pH 8.3, 50 mM KCl, 1.5 mM $MgCl_2$, and 0.01% gelatin. The samples were overlaid with mineral oil, heated to 96°C for 10 min and cooled to 55°C. After addition of 0.75 units of *Taq* DNA polymerase, the samples were processed through 35 amplification cycles of 92°C for 30 sec, 55°C for 30 sec, and 72°C for 30 sec using a Hybaid thermal cycler. The final extension step was lengthened to 10 min. Positive and negative controls were established for all reactions. The PCR products were analyzed on 2%–3% agarose gels.

Screening of cDNA Libraries

Bacteriophage from muscle, placental (Stratagene Cloning Systems), and fetal brain (Clontech) cDNA libraries were plated at 5×10^4 PFU/140-mm petri dish. Approximately 5×10^5 plaques were screened with restriction fragments of the original *TCOF1* cDNA (Treacher Collins Syndrome Collaborative Group 1996) or RACE products using standard procedures. Positive primary clones were purified by two additional rounds of screening and subcloned into pBluescript. The resulting plasmids were restriction mapped, and suitable restriction fragments were subcloned into M13mp18/19 and sequenced.

Exon Amplification

Genomic DNA from cosmids 17-1 and 18-3 were digested to completion with either *Pst*I, or double-digested with *Bam*HI and *Bgl*II. The restriction fragments were ligated into corresponding sites of the pSPL3 vector. The exon amplification protocol of Church et al. (1994) was followed with modifications reported previously (Treacher Collins Syndrome Collaborative Group 1996). Exon amplification clones were sequenced as above.

Determination of Intron/Exon Boundaries

Cosmids 17-1 and 18-3 were digested with *Sau*3A1, *Alu*I, *Pst*I, and *Sst*I and shotgun-cloned into M13. Recombinant plaques were screened with restriction fragments of the *TCOF1* cDNA or RACE products. Sequence data generated from the positive clones were compared with the cDNA sequence, and intron–exon boundaries were identified by comparison with the published consensus sequences (Breathnach and Chambon 1981).

Bioinformatics Analysis

DNA and derived protein sequences were used to query the GenBank, NBRF, Swissprot, and TREMBL databases using the BLAST suite of programs; BLASTP was used to compare *Treacle* to the OWL protein database (Altschul et al. 1990; Bleasby et al. 1994). A dot plot of *Treacle* versus *Treacle* was created using the GCG programs COMPARE (window size, 30; stringency, 18) and DotPlot (Devereux et al. 1984). Multiple alignments of the repeated units of *Treacle* were produced using GCG PILEUP and edited using LINEUP. PROSITE, PRINTS, and BLOCKS databases were searched to identify protein motifs within *Treacle* (Bairoch 1991; Attwood et al. 1994; Henikoff and Henikoff 1994). The above programs were accessed using either SEQNET or the Human Genome Mapping Resource Centre (Hinxton, UK).

ACKNOWLEDGMENTS

The financial support of the Wellcome Trust [grants 044684/Z/95/Z (M.J.D.) and 044327/Z/95/Z (M.J.D.)] and the Hearing Research Trust [grant 150:MAN:MD (M.J.D.)] is gratefully acknowledged. J.D. was supported, in part, by a HUGO travel award. This work benefitted from the use of the SEQNET and Human Genome Mapping Resource Centre computing facilities.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

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

Attwood, T.K., M.E. Beck, A.J. Bleasby, and D.J. Parry-Smith. 1994. PRINTS—A database of protein motif fingerprints. *Nucleic Acids Res.* **22:** 3590–3586.

Bairoch, A. 1991. PROSITE: A dictionary of sites and patterns in proteins. *Nucleic Acids Res.* **19:** 2241–2245.

Bleasby, A.J., D. Akrigg, and T.K. Attwood. 1994. OWL—A non-redundant composite protein sequence database. *Nucleic Acids Res.* **22:** 3574–3577.

Breathnach, R. and P. Chambon. 1981. Organization and expression of eukaryotic split genes coding for proteins. *Annu. Rev. Biochem.* **50:** 349–383.

Cairns, C. and B. McStay. 1995. Identification and cDNA cloning of a Xenopus nucleolar phosphoprotein, xNopp180, that is the homolog of the rat nucleolar protein Nopp140. *J. Cell. Sci.* **108:** 3339–3347.

Church, D.M., C.J. Stotler, J.L. Rutter, J.R. Murrel, J.A. Trofatter, and A.J. Buckler. 1994. Isolation of genes from complex sources of mammalian genomic DNA using exon amplification. *Nature Genet.* **6:** 98–105.

Devereux, J., P. Haeberli, and O. Smithies. 1984. A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res.* **12:** 387–395.

Dib-Hajj, S.D. and S.G. Waxman. 1995. Genes encoding the β1 subunit of voltage-dependent Na^+ channel in rat, mouse and human contain conserved introns. *FEBS Lett.* **377:** 485–488.

Dixon, M.J., A.P. Read, D. Donnai, A. Colley, J. Dixon, and R. Williamson. 1991. The gene for Treacher Collins syndrome maps to the long arm of chromosome 5. *Am. J. Hum. Genet.* **49:** 17–22.

Dixon, M.J., J. Dixon, T. Houseal, M. Bhatt, D.C. Ward, K. Klinger, and G.M. Landes. 1993. Narrowing the position of the Treacher Collins syndrome locus to a small interval between three new microsatellite markers at 5q32-33.1. *Am. J. Hum. Genet.* **52:** 907–914.

Dixon, M.J., H.A.M. Marres, S.J. Edwards, J. Dixon, and C.W.R.J. Cremers. 1994a. Treacher Collins syndrome: Correlation between clinical and genetic linkage studies. *Clin. Dysmorphol.* **3:** 96–103.

Dixon, J., A.J. Gladwin, S.K. Loftus, J. Riley, R. Perveen, J.J. Wasmuth, R. Anand, and M.J. Dixon. 1994b. A yeast artificial chromosome contig encompassing the Treacher Collins syndrome critical region at 5q31.3-32. *Am. J. Hum. Genet.* **55:** 372–378.

Edwards, S.J., A. Fowlie, M.P. Cust, D.T.Y Liu, I.D. Young, and M.J. Dixon. 1996. Prenatal diagnosis in Treacher Collins syndrome using combined linkage analysis and ultrasound imaging. *J. Med. Genet.* **33:** 603–606.

Eipers, P.G., J.M. Lahti, and V.J. Kidd. 1992. Structure and expression of the human p58clk-1 protein kinase chromosomal gene. *Genomics* **13:** 613–621.

Fazen, L.E., J. Elmore, and H.L. Nadler. 1967. Mandibulo-facial dysostosis (Treacher Collins syndrome). *Am. J. Dis. Child.* **113:** 406–410.

Gladwin, A.J., J. Dixon, S.K. Loftus, S. Edwards, J.J. Wasmuth, R.C.M. Hennekam, and M.J. Dixon. 1996. Treacher Collins syndrome may result from insertions, deletions or splicing mutations, which introduce a termination codon into the gene. *Hum. Mol. Genet.* **5:** 1533–1538.

Gorlin, R.J., M.M. Cohen, and L.S. Levin. 1990. *Syndromes of the head and neck*. Oxford University Press, Oxford, UK.

Henikoff, S. and J.G. Henikoff. 1994. Protein family classification based on searching a database of blocks. *Genomics* **19:** 97–107.

Jabs, E.W., X. Li, C.A. Coss, E.W. Taylor, D.A. Meyers, and J.L. Weber. 1991. Mapping the Treacher Collins syndrome locus to 5q31.3-q33.3. *Genomics* **11:** 193–198.

Jabs, E.W., X. Li, M. Lovett, L.H. Yamaoka, E. Taylor, M.C. Speer, C. Coss, R. Cadle, B. Hall, K. Brown, K.K. Kidd, G. Dolganov, M.H. Polymeropoulos, and D. Meyers. 1993. Genetic and physical mapping of the Treacher Collins syndrome locus with respect to loci in the chromosome 5q3 region. *Genomics* **18:** 7–13.

Jones, K.L., D.W. Smith, M.A. Harvey, B.D. Hall, and L. Quan. 1975. Older paternal age and fresh gene mutation: Data on additional disorders. *J. Pediatr.* **86:** 84–88.

Kay, E.D. and C.N. Kay. 1989. Dysmorphogenesis of the mandible, zygoma and middle ear ossicles in hemifacial microsomia and mandibulofacial dysostosis. *Am. J. Med. Genet.* **32:** 27–31.

Kozak, M. 1987a. An analysis of 5′-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res.* **15:** 8125–8148.

———. 1987b. At least six nucleotides preceding the AUG initiator codon enhance translation in mammalian cells. *J. Mol. Biol.* **196:** 947–950.

Loftus, S.K., S.E. Edwards, T. Scherpbier-Heddema, K.H. Buetow, J.J. Wasmuth, and M.J. Dixon. 1993. A combined genetic and radiation hybrid map surrounding the Treacher Collins syndrome locus on chromosome 5q. *Hum. Mol. Genet.* **11:** 1785–1792.

Loftus, S.K., J. Dixon, K. Koprivnikar, M.J. Dixon, and J.J. Wasmuth. 1996. Transcriptional map of the Treacher Collins candidate gene region. *Genome Res.* **6:** 26–34.

Meier, U.T. and G. Blobel. 1992. Nopp140 shuttles on tracks between nucleolus and cytoplasm. *Cell* **70:** 127–138.

Newgreen, D.F. and J. Minichiello. 1995. Control of epitheliomesenchymal transformation. Events in the onset of neural crest cell migration are separable and inducible by protein kinase inhibitors. *Dev. Biol.* **170:** 91–101.

Orita, M., H. Iwahana, H. Kanazawa, K. Hayashi, and T. Sekiya. 1989. Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms. *Proc. Natl. Acad. Sci.* **86:** 2766–2770.

Pai, C.Y., H.K. Chen, H.L. Sheu, and N.H. Yeh. 1995. Cell cycle-dependent alterations of a highly phosphorylated nucleolar protein p130 are associated with nucleologenesis. *J. Cell Sci.* **108:** 1911–1920.

Pereira, L., M. D'Alessio, F. Ramirez, J.R. Lynch, B. Sykes, T. Pangilinan, and J. Bonadio. 1993. Genomic organization of the sequence coding for fibrillin, the defective gene product in Marfan syndrome. *Hum. Mol. Genet.* **7:** 961–968.

Phelps, P.D., D. Poswillo, and G.A.S. Lloyd. 1981. The ear deformities in mandibulofacial dysostosis (Treacher Collins syndrome). *Clin. Otolaryngol.* **6:** 15–28.

Poswillo, D. 1975. The pathogenesis of Treacher Collins syndrome (mandibulofacial dysostosis). *Br. J. Oral Surg.* **13:** 1–26.

Sanger, F., S. Nicklen, and A.R. Coulson. 1977. DNA sequencing with chain terminating inhibitors. *Proc. Natl. Acad. Sci.* **74:** 5463–5467.

Thompson, A. 1846. Notice of several cases of malformation

of the external ear, together with experiments on the state of hearing in such persons. *Monthly J. Med. Sci.* **7:** 420.

Treacher Collins, E. 1900. Cases with symmetrical congenital notches in the outer part of each lid and defective development of the malar bones. *Trans. Ophthalmol. Soc. U.K.* **20:** 190–192.

Treacher Collins Syndrome Collaborative Group. 1996. Positional cloning of a gene involved in the pathogenesis of Treacher Collins syndrome. *Nature Genet.* **12:** 130–136.

Wilcox, A.S., A.S Khan, J.A. Hopkins, and J.M. Sikela. 1991. Use of 3′ untranslated sequences of human cDNAs for rapid chromosome assignment and conversion to STSs: Implications for an expression map of the genome. *Nucleic Acids Res.* **19:** 1837–1843.

Wiley, M.J., P. Cauwenbergs, and I.M. Taylor. 1983. Effects of retinoic acid on the development of the facial skeleton in hamsters: Early changes involving cranial neural crest cells. *Acta Anat.* **116:** 180–192.

Received October 9, 1996; accepted in revised form January 9, 1997.