Structural organization and splice variants of the *POLE1* gene encoding the catalytic subunit of human DNA polymerase ϵ

Deqi HUANG, Helmut POSPIECH, Tapio KESTI¹ and Juhani E. SYVÄOJA²

Biocenter Oulu and Department of Biochemistry, University of Oulu, FIN-90570 Oulu, Finland

The catalytic subunit of human DNA polymerase ϵ , an enzyme involved in nuclear DNA replication and repair, is encoded by the *POLE1* gene. This gene is composed of 51 exons spanning at least 97 kb of genomic DNA. It was found to encode three alternative mRNA splice variants that differ in their 5'-terminal sequences and in the N-termini of the predicted proteins. A CpG island covers the promoter region for the major transcript in HeLa cells. This promoter is TATA-less and contains several putative binding sites for transcription factors typical of S-phase-up-regulated and serum-responsive promoters. Potential pro-

motor regions were also identified for the two other alternative transcripts. Interestingly, no nuclear polyadenylation signal sequence was detected in the 3'-untranslated region, although a poly(A) tail was present. These results suggest a complicated regulatory machinery for the expression of the human *POLE1* gene, including three alternative transcripts expressed from three promoters.

Key words: alternative splicing, gene structure, non-conforming splice site.

INTRODUCTION

DNA polymerases (Pols) are enzymes of central importance in the maintenance and replication of DNA. In mammalian cells, six distinct Pols, designated α , β , γ , δ , ϵ and ζ , have been characterized [1,2]. In *Saccharomyces cerevisiae*, Pols α , δ and ϵ are all implicated in replication and are essential for cell viability [3]. Consistent with this concept, the same Pols are also cross-linked to nascent nuclear DNA in replicating mammalian cells [4]. In contrast, Pols α and δ were found to be the only Pols involved in the replication of simian virus 40 viral DNA *in vivo* and *in vitro* [4,5]. Thus replication of simian virus 40 DNA, the most common model system for mammalian replication, appears to be a simplification of the cellular process.

Pol ϵ has also been implicated in the repair of UV-damaged DNA in human fibroblasts [6], and in nucleotide-excision repair [7]. Furthermore, Pol ϵ seems to play a role in meiotic recombination and/or in recombinational repair [8–10].

Human Pol ϵ is composed of two subunits, with molecular masses of 261 and 59 kDa [11–14]. The catalytic subunit of Pol ϵ is highly conserved among eukaryotic species and contains both the DNA polymerase and 3'-exonuclease activities in its N-terminal half [12,15,16]. The function of the C-terminal domain is still unclear, although in *S. cerevisiae* it has been implicated in cell-cycle regulation during S-phase [17]. Like many genes encoding enzymes involved either in nucleotide metabolism or in DNA replication, the genes for mammalian Pols α , δ and ϵ belong to the group of late-responsive genes whose mRNAs are induced in cells that are stimulated to proliferate from a quiescent state by serum or growth factors [18–20].

As a step towards better understanding of the regulation of the human *POLE1* gene, which encodes the catalytic subunit of Pol ϵ , we have revealed the structural organization of the entire > 97 kb gene, provide evidence for three alternative splice variants of mRNA which differ in their 5'-ends, and present the sequences of the putative promoter regions for the three transcripts.

EXPERIMENTAL

Escherichia coli strains and cell cultivation

E. coli XL1 Blue was used for the propagation of plasmids and phages, unless mentioned otherwise. HeLa S3 cells were cultivated in suspension as 'spinner cells' in Joklik's modification of minimum essential medium (Flow Laboratories) supplemented with glutamine, penicillin, streptomycin, sodium bicarbonate and 5% (v/v) calf serum.

Rapid amplification of cDNA ends (RACE)

Human fetal thymus Marathon-Ready[®] cDNA (7421-1; Clontech) was used for cloning of the 3'-untranslated region (3'-UTR) by 3'-RACE with the specific sense primer hd39 and the inner primer hd38 (all primers utilized in this study are listed in Table 1) for primary and nested PCR amplifications respectively. The PCR products were subcloned and sequenced. The sequence of the 3'-UTR was confirmed by 3'-RACE from a human CEM T-cell cDNA library (HL1063f; Clontech) and a human HeLa S3 cDNA library (HL1100a; Clontech).

The 5'-terminal sequence was obtained by 5'-RACE using human bone marrow Marathon-Ready[®] cDNA (7416-1; Clontech), human fetal thymus Marathon-Ready[®] cDNA (7421-1; Clontech) and human testis Marathon-Ready[®] cDNA (7414-1; Clontech) as templates. The specific antisense primer hd34 and the inner primer hd35 were used for primary and nested PCR respectively. The presence of alternative 5'-terminal sequences was verified by PCR using a sense primer (seq222, seq229 or seq223) and the antisense primer seq194.

Northern hybridization and ribonuclease protection assay

Total RNA from HeLa S3 cells was isolated with TRIzol[®] LS reagent (Gibco-BRL). Poly(A)⁺ RNA was isolated from the total RNA using oligo(dT)-cellulose (Pharmacia). A 10 μ g portion of

Abbreviations used: Pol, DNA polymerase; RACE, rapid amplification of cDNA ends; UTR, untranslated region; PAC, P1 artificial chromosome; RT-PCR, reverse transcriptase-PCR; NF, nuclear factor; AP, activator protein.

¹ Present address: Department of Molecular Biology, MB-3, The Scripps Research Institute, 10550 N. Torrey Pines Road, La Jolla, CA 92037, U.S.A.

² To whom correspondence should be addressed (e-mail syvaoja@csc.fi).

The nucleotide sequence data reported will appear in DDBJ, EMBL and GenBank Nucleotide Sequence Databases under the accession numbers AF127937–127975 and AF128541–128544.

Table 1 Primers used in the study

For the cDNA numbering, the first nucleotide of the translation initiation codon (ATG) of the major transcript in HeLa S3 cells (GenBank accession no. L09561) is designated position +1.

Primer	Sequence	Sense (+)/antisense (—)	Position	
			cDNA	Gene
form4-40	5'-CGCTCCACCCAGACTCGCGAGAGCCGGAATGCCGCGT-3'	—	— 45 to — 84	Exon 1b
hd34	5'-CTTCTCGCTCACAACCTTTCTGGTCGC-3'	—	373–346	Exons 3–4
hd35	5'-TGCTTCTCACCAGGCTCCTTCAGCCGC-3'	—	207–180	Exon 2
hd38	5'-GCACTACGGCATGTCGTACCTCCTGGA-3'	+	6789–6815	Exon 49
hd39	5'-CTGCAGGACCTGGTCTGCCTGAAGTGC-3'	+	6646–6672	Exons 47–48
seq30	5'-ACAGTCTCCACTTTTGC-3'	—	530–514	Exon 5
seq54	5'-CGGAGTCAGTGGACGGA-3'	+	109–125	Exon 2
seq55	5'-CACCTCCGGCTTCTCAC-3'	—		Intron 2
seq56	5'-GTCCACAGGACCTGT-3'	+		Intron 8–exon 9
seq57	5'-GGCAGATGCTGCTGTAG-3'	—		Intron 9
seq103	5'-CCGTAGTGTGGCAATGTTCC-3'	—	6797–6776	Exon 49
seq104	5'-TGTCCTGGACACCAACATCAC-3'	+	6335–6356	Exon 46
seq107	5'-CTGCTGGTATGACTCCACGT-3'	—	5678–5657	Exon 41
seq114	5'-GAGCTATTCGAGCTCATCTCTG-3'	+	3082–3103	Exon 26
seq118	5'-TGCCCGCTTCATGAGCCGACTG-3'	—		Intron 20
seq119	5'-GTGTGTGATGAGATTAAGAGC-3'	+	1795–1815	Exon 17
seq121	5'-TTCAAAGTTGGTGACTTGCTCC-3'	—	1791–1770	Exon 16
seq122	5'-CCCTGAAGCAGAGGTGGTAGCC-3'	+	— 31 to — 10	Exon 1b
seq127	5'-GGGGTCATCAGTCATCTACGCC-3'	+	5577–5598	Exon 41
seq129	5'-GTACCACAACATCTATGCAGAC-3'	+	5424–5445	Exon 40
seq134	5'-CTGTCTCCTGCCCTCCAGGGTG-3'	—	3582–3561	Exon 29
seq135	5'-AAATCGGGTCCTTCTCGTCC-3'	+	4152–4173	Exon 33
seq136	5'-ATCCTTGACCATCTGGTCTCC-3'	—	3213–3192	Exon 26
seq137	5'-TTTCCAAGCAGAGCCACGGTG-3'	+	3264–3285	Exon 27
seq142	5'-CACTACCTACGTACCCGCTC-3'	+	2753–2774	Exon 24
seq145	5'-CCCTTGCGCATGACATAGCCAT-3'	—	2468–2447	Exon 21
seq146	5'-CCTGTTCTCGCGGGACAGTTC-3'	—	2146–2125	Exon 19
seq150	5'-TGCCTTCTCCTTGGAGGCTGG-3'	—	2848–2827	Exon 24
seq156	5'-GGAGGCTGTGGATCAGTCG-3'	—	5295–5274	Exon 39
seq164	5'-GATGCCAGCCTTGGCGCCCTG-3'	+	4566–4587	Exon 36
seq169	5'-TGCTCGAATATCAAAGTCTGA-3'	—	3378–3357	Exon 27
seq170	5'-AGGCCCTTAGAGCTGGACACAG-3'	+	2563–2584	Exon 23
seq181	5'-GAGCTCAGGGCCACCTTCTCC-3'	—	4635–4614	Exon 36
seq182	5'-TAGGCTATGATCCCGTGGAGCT-3'	+	1295–1316	Exon 13
seq183	5'-CCAGTGCCAGCCAGTCGACGG-3'	+		Intron 18–exon 19
seq194	5'-CGCAAATCCATCTTATCCGTC-3'	—	140–119	Exon 2
seq207	5'-ATTTCTACATGATCGATGGCC-3'	+	886–907	Exon 9
seq208	5'-GCATGCTCAGACCGTGGACTGC-3'	—	1150–1129	Exon 12
seq209	5'-CGAGTTCAAAGGGCTCCACAAG-3'	+	2298–2319	Exon 20
seq210	5'-GAAGGTCTCCTTTGAATGGATG-3'	—	5700–5679	Exon 42
seq222	5'-GCATCTACTGATCACTGACCAG-3'	+		Exon 1a
seq223	5'-AAGGCTGCTGCTGGGATAAGTA-3'	+		Exon 1c
seq229	5'-CGCGAGTCTGGTGGGAGCG-3'	+	— 64 to — 45	Exon 1b
seq326	5'-GGCCCAATTGGAGCGCCTCGG-3'	+	— 140 to — 120	Exon 1b
seq327	5'-CGCCGCGCGCGCTACACACACT-3'	+	— 205 to — 182	5'-Flanking region of exon 1b
seq329	5'-CGCTCCACCCAGACTCGCGAG-3'	—	— 45 to — 66	Exon 1b

poly(A)⁺ RNA was separated by agarose-gel electrophoresis and blotted on to a Hybond-N membrane (Amersham). The membrane was cut into several pieces, which were hybridized to different ³²P-labelled RNA or DNA probes. The ribonuclease protection assay was performed with an RPA II[®] ribonuclease protection assay kit (Ambion).

Isolation and characterization of genomic clones

A human genomic λ library (HL1067J; Clontech) was screened with ³²P-labelled DNA probes derived from the 5'-terminus of the cDNA (GenBank accession no. L09561) and screened for genomic clones representing exon 1a. A positive human P1 artificial chromosome (PAC) clone (PAC-261-24B; control no. 4713) was identified by PCR screening with two primer pairs

(seq54/55 and seq56/57) from a human PAC library (PAC-6539; Genome Systems, Inc). Cosmid clones were isolated from a human cosmid library (951202; Stratagene) using a probe (110 bp) derived from the 5'-end of the PAC clone.

Determination of exon–intron boundaries

The exon–intron boundaries for exons 1a, 1b and 2–15 were sequenced from λ clones. The remaining exons were sequenced directly from the PAC clone. The sizes of the introns were determined by sequencing, PCR amplification and/or restriction mapping. DNA sequencing was performed by the dideoxynucleotide chain-termination method using T7 DNA polymerase (Pharmacia Biotech) or the 71001 Delta TAQ[®] Cycle Sequencing Kit (Amersham).

© 1999 Biochemical Society

Figure 3 Ribonuclease protection analysis of the three transcripts encoded by the human *POLE1* gene

Samples of 5–10 μg of HeLa mRNA (h), and of 5–10 μg of yeast RNA (y) as a control, were hybridized to RNA probes derived from cDNA containing 87 nt of exon 2 sequence and sequences specific for exon 1b (left) or exon 1c (right). Transcripts with exons 1b and 1c are expected to generate full-length protection fragments of 213 nt (left) and 179 nt (right) respectively. A 87 bp fragment corresponding to exon 2 gives a stronger signal than the full-length fragment for exon 1c (right), indicating that this transcript is not predominant. See the Results section for details.

nuclear poly(A) signal sequence (AAUAAA) could be found at the 3'-UTR region. Instead, a putative cytoplasmic polyadenylation element UUUUUAU [22] was identified 21 nt upstream of the poly(A) tail.

Three alternatively spliced 5'-termini of human *POLE1* transcripts

5'-RACE from Marathon-Ready[®] cDNAs revealed the existence of three transcripts that differed in their first exons, designated exons 1a, 1b and 1c (Figure 1, top and middle panels). The cDNA sequences representing the three mRNA variants were also present in cDNAs from several tissues and cancer cell lines.

as well as in HeLa mRNA, as determined by RT-PCR analysis (results not shown).

The fact that the majority of human 5'-end cDNA clones contained exon 1b sequence rather than the other two alternative exon forms (results not shown) suggested that exon 1b encodes the 5'-terminus of the predominant transcript, and that exons 1a and 1c represent minor transcripts. Results from Northern hybridization and ribonuclease protection analysis of HeLa mRNA were consistent with this suggestion. When exon 1b was used as a DNA probe in Northern hybridization, a relatively strong signal was detected (Figure 2). The size of this signal was identical with that obtained by an RNA probe recognizing all three transcripts. No signal was detected with RNA probes representing exons 1a or 1c (Figure 2) or with the corresponding DNA probes (results not shown).

Ribonuclease protection analysis revealed no specific protection fragments with a probe derived from exon 1a (545 bp of exon 1a and 38 bp of exon 2), suggesting a low abundance of the corresponding mRNA (results not shown). Two protection fragments were detected with a 179 nt RNA probe representing exon 1c and 87 nt of exon 2 (Figure 3). The protection fragment corresponding to the entire RNA probe gave a very weak signal when compared with a shorter protection fragment corresponding to 87 bp of exon 2, thus indicating a low abundance of the transcript containing exon 1c-encoded sequence. An RNA probe containing sequence encoded by exon 1b (126 nt of exon 1b and 87 nt of exon 2) gave three specific signals that were 20, 27 and 38 nt shorter than the probe. Only a weak signal was detected in the area corresponding to the 87 bp of exon 2 sequence (Figure 3). The 5'-UTR sequence transcribed from exon 1b has a G+C content of 74% and a potentially stable secondary structure (results not shown), which may explain the incomplete protection by the probe representing exon 1b.

The putative translation initiation sites, and hence the N-termini of the deduced translation products from the three transcripts, are different (Figure 1, bottom panel). Whereas both exons 1a and 1c contain putative translation start sites, a stop codon is present in the correct reading frame of exon 1c. Therefore translation of this form is presumably initiated from exon 2. The unique N-termini encoded by exons 1a and 1b share no apparent

Figure 4 Physical map of the human *POLE1* gene

The cDNA of the major transcript in HeLa cells and important functional domains of the encoded protein are shown on the top. The exon/intron organization of the gene is shown in the middle. Genomic clones used to analyse the structure of the *POLE1* gene and respective restriction sites are shown at the bottom. The enzymes used for restriction mapping were *Bss*HI (B), *Eag*I (E), *Hind*III (H), *Kpn*I (K), *Mlu*I (M), *Not*I (N), *Sac*I (Sc) and *Sal*I (Sl).

Table 2 Exon-intron boundaries of the human *POLE1* gene

Exon sequence is in upper case, and intron sequence is in lower case. The positioning of introns between codons is indicated by phase 0, interruption after the first nucleotide by phase I, and interruption after the second nucleotide by phase II. For nucleotide position numbering, the first nucleotide of the translation initiation codon (ATG) for the major transcript in HeLa S3 cells is designated position + 1.

Exon	Splice acceptor site	Exon size (bp)	Nucleotide positions	Amino acid position	Splice donor site	Intron size (bp)	Intron phase
1a		≥ 745			AGGCG/gtgagtctag	≥ 50kb	II
1b		217	—155–62	1–21	AGCAG/gtgagggcg	7.5kb	II
1c		≥ 206			GGGAG/gtgagtgtgc	6 kb	
2	tttactcag/GGATG	142	63–204	22–68	ATCCT/gtaagcacta	448	0
3	tcctttag/ACCGA	81	205–285	69–95	TTAAG/gtaagccct	384	0
4	ctcttaacag/GTGGC	45	286–330	96–110	GAAAG/gtgagtgtgt	131	0
5	tcctgattag/GTTTG	93	331–423	111–141	ACTTG/gtgagtatca	240	0
6	cgcatatag/CCAAA	155	424–578	142–193	TCCAG/gtaaaccttc	1730	II
7	taattccag/TGTTC	142	579–720	194–240	ACGTG/gtgagtgggg	134	0
8	tgtttgcag/GCTCA	81	721–801	241–267	GACCT/gtaagttttg	670	0
9	gtgtccacag/GACCC	108	802–909	268–303	GCCAG/gtgagcaggt	290	0
10	acttggtcag/GGCTA	111	910–1020	304–340	ATGAG/gtaagaggt	230	0
11	tgggttcag/GCTCA	86	1021–1106	341–369	GAAGT/gtgagtctgt	170	II
12	tccaaccag/GCCAT	120	1107–1226	370–409	CTCAG/gtagggcca	1660	II
13	tccttcacag/GTGGG	133	1227–1359	410–453	CCCAG/*gcacgtaag	280	0
14	ctctctcag/ACTCT	114	1360–1473	454–491	ACGAG/gtgagcatta	300	0
15	tccttcgag/GTGCT	213	1474–1686	492–562	GGATG/gtgtctct	270	0
16	gttgtcag/AATCC	108	1687–1794	563–598	AAGAG/gtaaccatca	2960	0
17	ttctctcag/GTGTG	129	1795–1923	599–641	TGCAG/gtgagacttt	73	0
18	ccctccacag/CCCTC	103	1924–2026	642–675	GTTCA/gtgagtgctg	132	I
19	tcctccacag/TGCCA	147	2027–2173	676–724	GGCGG/gtgagtgcct	670	I
20	ctgcgtcag/ATTAC	146	2174–2319	725–773	ACAAG/gtgagctctg	2100	0
21	tgctctcag/GTGTG	149	2320–2468	774–823	AAGGG/gtaggtgggc	800	II
22	gtccctcag/GGCTC	93	2469–2561	824–854	ATTGG/gtgagtgccg	240	II
23	ttctctgaag/GAGGC	145	2562–2706	855–902	TCAAG/gtgagtctgg	2210	0
24	tgtttcag/GAAGG	158	2707–2864	903–955	AAGAG/gtgggactct	330	II
25	ttttttcag/GTATG	196	2865–3060	956–1020	GCAAG/gtaagggcct	1410	0
26	ctttctcag/GCAGC	215	3061–3275	1021–1092	GAGAG/gtgagggctc	1210	II
27	cctaccacag/GGCCA	103	3276–3378	1093–1126	GAGCA/gtgagtgcga	480	0
28	gcacatcag/ATTCT	81	3379–3459	1127–1153	AGCAG/gtaagggcgg	130	0
29	cctcacacag/GTAAA	123	3460–3582	1154–1194	GACAG/gtgacggctg	1560	0
30	gtctttcag/GTCAC	213	3583–3795	1195–1265	GCCAG/gtaaggcaca	150	0
31	atccatccag/GAGGA	206	3796–4001	1266–1334	TGCAG/gtatagacct	210	0
32	ggccttcag/ATCAG	148	4002–4149	1335–1383	GCAAG/gtagggagga	4000	0
33	atctgccag/GTAAA	141	4150–4290	1384–1430	CTCAG/gtgactcccc	260	0
34	attctccag/GTTCC	154	4291–4444	1431–1481	ACCAG/gtatggcctg	76	I
35	ccttctcag/GGAGT	107	4445–4551	1482–1517	AACTT/gtgagtcttg	270	0
36	gttgtctcag/GTGCG	177	4552–4728	1518–1576	ACAAG/gtgagagtg	90	0
37	cttgaccacag/GAGGA	224	4729–4952	1577–1651	AGCAG/gtgagcacag	106	II
38	cctgggcag/GTACT	221	4953–5173	1652–1725	CACAG/gtaagtgggc	320	I
39	tcctgtcag/GTGTG	205	5174–5378	1726–1793	TTCAG/gtgccacct	2350	II
40	tcctttcag/GATCC	174	5379–5552	1794–1851	CTGCA/gtgagtgctc	950	II
41	tgctccacag/GCTCA	126	5553–5678	1852–1893	AGCAG/gtgcggctgg	2050	II
42	tcctctcag/CATCC	133	5679–5811	1894–1937	GACTG/gtaagcgagc	1550	0
43	tgtgttacag/CAAGA	193	5812–6004	1938–2001	TTCAG/gtaagcccag	1440	I
44	ctctccacag/CGTAC	132	6005–6136	2002–2045	TCCCG/gtgagcagtc	155	I
45	atactccag/GAATG	194	6137–6330	2046–2110	GCAAG/gtgagagtg	3650	0
46	gtctctcag/GTGCT	201	6331–6531	2111–2177	CAGAG/gtgcggctga	320	0
47	ctctgcctcag/GATGG	126	6532–6657	2178–2219	ACCTG/gtgagtcgtg	620	0
48	ctctttcag/GTCTG	90	6658–6747	2220–2249	CCCAG/gtgggctctc	94	0
49	attctgcag/GTCTT	1049	6748–7796	2250–2286	GTTTC		

* Non-conforming splice signal sequence.

identity (Figure 1, bottom panel). The translation product from the transcript containing exon 1b is identical with that deduced from the previously published cDNA (GenBank accession no. L09561)

Organization of the human *POLE1* gene

To elucidate the exon/intron structure of *POLE1*, we isolated and characterized two genomic λ clones, H11-16 and H11-17.

H11-16 contained exon 1b, and H11-17 contained a region from exon 2 to exon 15 (Figure 4). A subsequently characterized human PAC genomic clone (PAC-261-24B) contained exons 1b and 1c and all downstream exons of the gene. Exon 1a was not present in this PAC clone. We could not locate it either in four human cosmid clones (Cos1, Cos4, Cos5 and Cos6) that extended 40 kb upstream of exon 1b (Figure 4). Therefore two human genomic λ clones, H5-23 and H5-24, were isolated by screening with a cDNA probe representing exon 1a to determine the

Figure 5 Identification of the transcription-initiation site for the major transcript of the human *POLE1* gene in HeLa cells

Upper panel: primer extension analysis. Primers used for extensions are shown above the lanes. Their sequences and locations with respect to cDNA are shown in Table 1. HeLa mRNA (h; 5.4 μ g) was used as an extension template, and yeast RNA (y; 5.0 μ g) was used as a negative control. The position of the extension products with respect to the genomic sequence is indicated by an arrowhead. Lower panel: RT-PCR analysis. The location of the primers is presented in the scheme below the gel. The samples from left to right are: *Hae*III-digested ϕ X174 DNA size standards; specific PCR amplification products from HeLa mRNA (4 lanes); single-primer controls (5 lanes); and no-template controls (4 lanes). The specific amplification products are indicated by arrowheads.

chromosomal localization of this exon. The λ clones were mapped to the *POLE1* locus at chromosome 12q24.3 [23] by fluorescence *in situ* hybridization (FISH) (results not shown). Exon 1a was also successfully amplified by PCR using genomic DNA from the human/mouse hybrid cell line GM12072C containing only human chromosome 12 (Coriell Cell Repositories), but could not be amplified from control genomic mouse DNA, thereby verifying the localization to this chromosome (results not shown). The evidence provided above strongly suggests that exon 1a is located at least 50 kb upstream of exon 2 (Figure 4).

The *POLE1* gene consists of 48 common exons and three alternative first exons, designated exons 1a, 1b and 1c, and of 47

Figure 6 CpG plot of the 5'-end of the human *POLE1* gene

CpG islands were identified with the CpGPlot program in the EGC8.1 package [24]. The value for the ratio of observed/expected CpG dinucleotides is shown by the solid line, and the G + C content (%) is shown by the broken line. The locations of the two CpG islands relative to the positions of exons 1b and 1c in the genomic sequence are presented below the plot.

common introns and the alternative introns 1a, 1b and 1c (Figure 4 and Table 2). The gene spans at least 97 kb of DNA. The known lengths of exons range from 45 to 1049 bp. The accurate lengths of exons 1a and 1c remain unknown. The intron sizes range from 76 bp to 7.5 kb, except for intron 1a, which spans over at least 50 kb. Intron-exon junctions match well with the reported consensus sequences (GT/AG), with the exception of the exon 13-intron 13 junction, where GT has been replaced by non-conforming GC (Table 2).

Transcription-initiation site of the major transcript in HeLa cells, and analysis of the 5'-flanking regions of the three transcripts

In order to identify the transcription-initiation site for the major transcript in HeLa cells, two primers of different lengths were used for primer extension. Identical extension products were generated with the two primers (Figure 5, upper panel), suggesting that transcription is initiated from a single site 155 bp upstream of the translation start site (Figure 1, middle panel). The transcription initiation site was also mapped by RT-PCR analysis. PCR products could be amplified only with sense primers from exon 1b representing the region expected to be transcribed, but not with a sense primer from the 5'-flanking, non-transcribed region (Figure 5, lower panel). Thus the results from primer extension analysis were verified and the presence of introns in the 5'-UTR of the gene was ruled out.

The 5'-flanking genomic region of exon 1b (832 bp) is very G + C-rich and lacks a consensus TATA box upstream of the transcription initiation site. The promoter region for exon 1b is covered by a CpG island of 1437 bp (Figure 6). More than 20 potential transcription factor-binding sites were identified in the region encompassing 500 bp upstream and 150 bp downstream of the transcription-initiation site (Table 3).

Attempts to identify the transcription start sites of the two minor transcripts in HeLa cells by primer extension failed, probably due to the very low abundance of these transcripts. Computational analysis of the 5'-flanking region of exon 1c by

Table 3 Putative regulatory elements of the human *POLE1* gene in the region encompassing the transcription initiation site for the major transcript in HeLa S3 cells

A region of 500 bp upstream and 150 bp downstream from the transcription initiation site (accession no. *to be added*) was analysed for potential transcription factor binding sites using the MatInspector V2.2 program [25] in conjunction with the TRANSFAC MATRIX TABLE V3.3 database. Only core sequences showing a perfect match to a consensus transcription factor binding site are listed (upper-case letters). (—) denotes the presence of the element on the cDNA strand.

Element	Position	Sequence	Refs.
NF-1	—6 to —23	cgtTGGCtaccacctgtg (—)	[26]
	—310 to —293	ggcTGGCcgggcgcgag	
E2F	—38 to —45	TTTGgcgc (—)	[27]
Sp1	—110 to —98	gaggGGCGgagcg	
NF-Y	—121 to —109	ggggGGCGggggga	[28,29]
	—221 to —233	accgGGCGgcgct (—)	
	—509 to —497	ccggGGCGgcgctc	
	—127 to —137	gctCCAAttgg (—)	
	—140 to —130	ggcCCAAttgg	
AP-2	—137 to —148	ggCCCGgcgccc (—)	[30]
	—144 to —155	cgCCCGgcgggg (—)	
	—156 to —145	ccCCCGccgggc	
	—327 to —338	ccCCCGgcgcca (—)	
	—384 to —373	ggCCCGgcgcct	
AP-4	—416 to —405	cgCCCGccgcgg	[31]
	—236 to —227	cgCAGCgccc	
	—348 to —349	ccCAGCgccc	
	—531 to —522	cgCAGCggcg (—)	
	—624 to —615	ttCAGCgccc	
AP-1	—263 to —253	ccTGACatttc	[32]
	—588 to —578	tcTGACccggg	

Sequence analysis reveals the presence of a putative TATA box, but no CCAAT box could be detected. The sequence preceding the putative TATA box has a high G + C content, revealing the presence of several potential transcription factor binding sites, including multiple Sp1 consensus sites. The other possible promoter, located approx. 300 bp upstream of exon 1c, contains a putative TATA box, but no apparent CCAAT box or consensus GC boxes can be identified in this region.

A potential promoter, located approx. 300 bp upstream of exon 1a and covered by a short CpG island of 271 bp (results not shown), was predicted by scanning of 2.7 kb of 5'-flanking sequence. Sequence analysis further revealed the presence of a candidate TATA box and, among other putative transcription factor-binding sites, consensus sites for Sp1 and nuclear factor-1 (NF-1) [26,27].

Copy number of the human *POLE1* gene

To study the possible presence of other *POLE1*-related genes or pseudogenes in the human genome, primer pairs covering different regions of the *POLE1* gene were used for comparative PCR amplification from human placental genomic DNA and from the genomic clone PAC-261-24B. Each primer pair resulted in amplification products of the same size from the two templates (Figure 7), suggesting that no other *POLE1*-related genes or pseudogenes are present in the human genome. Furthermore, the hybridization patterns from restriction-digested human placental genomic DNA were identical with those from control PAC-261-24B DNA using probes that covered almost the entire human Pol ϵ cDNA (results not shown).

Figure 7 Search for human *POLE1*-related genes

Products were amplified by PCR from the PAC-261-24B clone (g) and human placental genomic DNA (p) with primer pairs representing specific exons/introns as shown above the lanes. The left-most and right-most lanes (M) contain DNA size standards. The single-primer controls and no-template controls are not included in the figure. The specific amplification products are indicated by arrowheads. In some instances, additional small PCR products were amplified from PAC and/or genomic DNA. These products were shown to be unspecific by PCR under more stringent conditions or by subcloning and sequence analysis.

the methods of Prestridge [33] and of Reese et al. [34] predicted the presence of two potential promoters. One is located approx. 650 bp upstream of exon 1c in a CpG island of 522 bp (Figure 6).

DISCUSSION

In this paper we (i) present the exon/intron organization of the human *POLE1* gene, (ii) elucidate the 3'- and 5'-termini of the corresponding cDNA, (iii) provide evidence for the presence of three alternative splice variants, resulting apparently in three different N-termini of the protein, and (iv) report the sequence and putative transcription factor-binding sites of the promoter regions of the transcripts. The three transcripts encoding the Pol ϵ catalytic subunit differ in their first exons. Considering the physical distance between these exons, it is probable that transcription is under the control of three distinct promoters. This could reflect complicated regulation of transcription, possibly in a tissue-specific manner or as a response to different cellular processes, such as DNA replication and repair. To our knowledge, no alternative splicing of the mRNAs for the replicative Pols α and δ has been reported. Several alternative splice variants of the mammalian Pol β transcript have been described, but their functions remain unclear [35–37]. Furthermore, a single promoter controls the expression of the mammalian Pol β gene [38]. The alternative Pol ϵ transcripts would encode proteins with different N-termini. Since no function has been assigned to the very N-terminus of the catalytic subunit of Pol ϵ , it is possible that the differences in the amino acid sequence do not influence the activity of the protein, but rather are a consequence of the use of three different promoters.

The CpG island covering the putative promoter region of the major transcript is typical of 'housekeeping' and widely expressed genes [39], such as genes encoding DNA-metabolizing proteins. CpG islands are believed to represent 'active chromatin' that is readily bound by factors that initiate transcription [40,41]. The transcript levels of enzymes that are involved in DNA synthesis

increase at the G₁/S boundary in actively cycling cells, and upon stimulation of quiescent cells to proliferate by serum or growth factors. This has also been found to apply for Pol ϵ [20]. The absence of a TATA box in the promoter is also typical of genes encoding several other S-phase-specific enzymes, including those for Pols α and δ [42–44]. The transcription factors Sp1 and E2F have been found to play central roles in the modulation of such genes [27], as appears to be the case for Pols α and δ [43,45]. In addition, binding sites for NF-Y (CCAAT box) and AP-1 (activator protein-1) in the Pol α gene are protected from DNase I digestion [43]. The presence of consensus binding sites for transcription factors E2F, Sp1, NF-Y and AP-1 in the putative major *POLE1* promoter in HeLa cells is consistent with S-phase-specific expression, and suggests similar transcriptional regulation of the replicative Pols α , δ and ϵ in human cells.

Transcription of the major form of Pol ϵ is initiated at a single site, as is the case for Pol α . In contrast, transcription of the human Pol δ gene was found to initiate from multiple sites [43,44].

Non-conforming mRNA splice sites, like the one at the exon 13–intron 13 boundary of the human *POLE1* gene, have been described in several other genes of higher eukaryotes [46]. Among these, the 5' splice variant GC commonly replaces the 'invariant' GT at the first two positions of the intron, as was found in the *POLE1* gene. It has been demonstrated that replacement of GT by GC results in normal splicing *in vitro* and *in vivo*, albeit with reduced efficiency [47,48]. The positions surrounding the variant splice sites of the *POLE1* gene (CAG GCACGT) show a very good match to the 5' splice site consensus sequence [A/C]AG GTAAGT [49], as has been reported for other 5' GC splice variants. This can be explained by the requirement of the splice site to pair with the U1 RNA in the spliceosome. The absence of the central U:A pair coded by the invariant GT is compensated by the better complementarity in the surrounding sequence [46]. The non-conforming splice site (GC instead of GT) of the *POLE1* gene is conserved from mouse to human (D. Huang, R. Knuuti, H. Palosaari, H. Pospiech and J. E. Syväoja, unpublished work), suggesting that it plays an important role in the regulation of Pol ϵ expression.

The absence of the nuclear polyadenylation signal (AAUAAA) in the 3'-UTR is unexpected, since this sequence is almost invariant in metazoans [50]. However, some transcripts are polyadenylated in the absence of a recognizable AAUAAA element. These include an mRNA from the mouse dihydrofolate reductase gene [51] and one out of several human cyclin D transcripts [52]. In many cases, only one out of several alternatively polyadenylated transcripts lacks the nuclear polyadenylation signal [53], whereas we found only one form of polyadenylation in the human *POLE1* transcripts. The significance of the putative cytoplasmic polyadenylation consensus element UUUUUAU 21 bp upstream of the poly(A) tail remains unclear, since it is believed to function only in conjunction with the nuclear polyadenylation sequence [54]. Messages harbouring this element undergo cytoplasmic polyadenylation and translation during oocyte maturation.

In conclusion, the human *POLE1* gene possesses a number of intriguing features that suggest complex regulation of transcription and translation. This could be the requisite for the different functions assigned to Pol ϵ .

We thank M. Mäkinen for careful reading of and valuable comments on the manuscript. This work was supported by grants from the National Research Council for the Environment and Natural Resources, Academy of Finland, and from the Cancer Foundation of Finland, and by a scholarship from the Center for International Mobility (CIMO), Finland, to D.H.

REFERENCES

- Wang, T. S.-F. (1996) DNA Replication in Eukaryotic Cells (DePamphilis, M. L., ed.), pp. 461–493, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
- Gibbs, P. E., McGregor, W. G., Maher, V. M., Nisson, P. and Lawrence, C. W. (1998) Proc. Natl. Acad. Sci. U.S.A. **95**, 6876–6880.
- Sugino, A. (1995) Trends Biochem. Sci. **20**, 319–323.
- Zlotkin, T., Kaufmann, G., Jiang, Y., Lee, M. Y. W., Uitto, L., Syväoja, J. E., Dornreiter, I., Fanning, E. and Nathaniel, T. (1996) EMBO J. **15**, 2298–2305.
- Waga, S. and Stillman, B. (1994) Nature (London) **369**, 207–212.
- Nishida, C., Reinhard, P. and Linn, S. (1988) J. Biol. Chem. **263**, 501–510.
- Aboussekhra, A., Biggerstaff, M., Shivji, M. K., Vilpo, J. A., Moncollin, V., Podust, V. N., Protic, M., Hübscher, U., Egly, J. M. and Wood, R. D. (1995) Cell **80**, 859–868.
- Jessberger, R., Podust, V., Hübscher, U. and Berg, P. (1993) J. Biol. Chem. **268**, 15070–15079.
- Jessberger, R., Chui, G., Linn, S. and Kemper, B. (1996) Mutat. Res. **350**, 217–227.
- Kamel, D., Mackey, Z. B., Sjöblom, T., Walter, C. A., McCarrey, J., Uitto, L., Palosaari, H., Lähdele, J., Tomkinson, A. and Syväoja, J. E. (1997) Biol. Reprod. **57**, 1367–1374.
- Syväoja, J. E. and Linn, S. (1989) J. Biol. Chem. **264**, 2489–2497.
- Kesti, T., Frantti, H. and Syväoja, J. E. (1993) J. Biol. Chem. **268**, 10238–10245.
- Li, Y., Asahara, H., Patel, V. S., Zhou, S. and Linn, S. (1997) J. Biol. Chem. **272**, 32337–32344.
- Jokela, M., Mäkinen, M., Lehtonen, S., Szpirer, C., Hellman, U. and Syväoja, J. E. (1998) Nucleic Acids Res. **26**, 730–734.
- Kesti, T. and Syväoja, J. E. (1991) J. Biol. Chem. **266**, 6336–6341.
- Uitto, L., Hallee, J., Hentunen, T., Höyhty, M. and Syväoja, J. E. (1995) Nucleic Acids Res. **23**, 244–247.
- Navas, T. A., Zhou, Z. and Elledge, S. J. (1995) Cell **80**, 29–39.
- Wahl, A. F., Geis, A. M., Spain, B. H., Wong, S. W., Korn, D. and Wang, T. S.-F. (1988) Mol. Cell. Biol. **8**, 5016–5025.
- Hao, H., Jiang, Y., Zhang, S. J., Zhang, P., Zeng, R. X. and Lee, M. Y. W. T. (1992) Chromosoma **102**, S121–S127.
- Tuusa, J., Uitto, L. and Syväoja, J. E. (1995) Nucleic Acids Res. **23**, 2178–2183.
- Ausubel, F. M., Brent, R., Kingston, R. E., Moore, D. D., Seidman, J. G., Smith, J. A. and Struhl, K. (1989) Current Protocols in Molecular Biology, Wiley Interscience, New York.
- Salles, F. J., Darrow, A. L., O'Connell, M. L. and Strickland, S. (1992) Genes Dev. **3**, 803–815.
- Szpirer, J., Pedoutour, F., Kesti, T., Syväoja, J. E., Turc-Carel, C. and Szpirer, C. (1994) Genomics **20**, 223–226.
- Gardiner-Garden, M. (1987) J. Mol. Biol. **196**, 261–282.
- Quandt, K., Frech, K., Karas, H., Wingender, E. and Werner, T. (1995) Nucleic Acids Res. **23**, 4878–4884.
- Gronostajski, R. M., Adhya, S., Nagata, K., Guggenheimer, R. A. and Hurwitz, J. (1985) Mol. Cell. Biol. **5**, 964–971.
- Slansky, J. E. and Farnham, P. J. (1996) BioEssays **18**, 55–62.
- Chang, Z. F. and Liu, C. J. (1994) J. Biol. Chem. **269**, 17893–17898.
- Filatov, D. and Thelander, L. (1995) J. Biol. Chem. **270**, 25239–25243.
- Zeng, Y. X., Somasundaram, K. and el-Deiry, W. S. (1997) Nat. Genet. **15**, 78–82.
- Hu, Y. F., Luscher, B., Admon, A., Mermod, N. and Tjian, R. (1990) Genes Dev. **4**, 1741–1752.
- Foletta, V. J. (1996) Immunol. Cell. Biol. **74**, 121–133.
- Prestridge, D. S. (1995) J. Mol. Biol. **249**, 923–932.
- Reese, M. G., Harris, N. L. and Eeckman, F. H. (1996) Pac. Symp. Biocomput., Abstract.
- Chyan, Y. J., Ackerman, S., Shepherd, N. S., McBride, O. W., Widen, S. G., Wilson, S. H. and Wood, T. G. (1994) Nucleic Acids Res. **22**, 2719–2725.
- Chyan, Y. J., Strauss, P. R., Wood, T. G. and Wilson, S. H. (1996) DNA Cell Biol. **15**, 653–659.
- Nowak, R., Bieganski, P., Konopinski, R. and Siedlecki, J. A. (1996) Int. J. Cancer **68**, 199–202.
- He, F., Narayan, S. and Wilson, S. H. (1996) Biochemistry **35**, 1775–1782.
- Larsen, F., Gundersen, G., Lopez, R. and Prydz, H. (1992) Genomics **13**, 1095–1107.
- Cross, S. H. and Bird, A. P. (1995) Curr. Opin. Genet. Dev. **5**, 309–314.
- Antequera, F., Boyes, J. and Bird, A. (1990) Cell **62**, 503–514.
- Johnson, L. F. (1992) Curr. Opin. Cell Biol. **4**, 149–154.
- Pearson, B. E., Nasheuer, H.-P. and Wang, T. S.-F. (1991) Mol. Cell. Biol. **11**, 2081–2095.
- Chang, L.-S., Zhao, L., Zhu, L., Chen, M.-L. and Lee, M. Y. W. T. (1995) Genomics **28**, 411–419.
- Zhao, L. and Chang, L.-S. (1996) J. Biol. Chem. **272**, 4869–4882.
- Jackson, I. J. (1991) Nucleic Acids Res. **19**, 3795–3798.

-
- 47 Aebi, M., Hornig, H. and Weissmann, C. (1987) *Cell* **50**, 237–246
- 48 Chen, I.-T. and Chasin, L. A. (1993) *Mol. Cell. Biol.* **13**, 289–300
- 49 Shapiro, M. B. and Senapathy, P. (1987) *Nucleic Acids Res.* **15**, 7155–7174
- 50 Wahle, E. and Keller, W. (1996) *Trends Biochem. Sci.* **21**, 247–250
- 51 Hook, A. G. and Kellems, R. E. (1988) *J. Biol. Chem.* **263**, 2337–2343
- 52 Xiong, Y., Connolly, T., Futcher, B. and Beach, D. (1991) *Cell* **65**, 691–699
- 53 Wahle, E. and Keller, W. (1992) *Annu. Rev. Biochem.* **61**, 419–440
- 54 Richter, J. D. (1996) in *Translational Control* (Hershey, J. W. B., Mathews, M. and Sonenberg, N., eds.), pp. 481–503, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY
-

Received 16 September 1998/13 January 1999; accepted 10 February 1999