An integrated approach in the discovery and characterization of a novel nuclear protein over-expressed in liver and pancreatic tumors

Meng Ling Choong[a,*], Li Kiang Tan[a], Siaw Ling Lo[a], Ee-Chee Ren[b], Keli Ou[a], Shao-En Ong[a], Rosa C.M.Y. Liang[a], Teck Keong Seow[a], Maxey C.M. Chung[a,c]

[a] *Bioprocessing Technology Centre, National University of Singapore, MD 11 Level 5, 10 Medical Drive, Singapore 117597, Singapore*
[b] *Singapore Genomics Program, National University of Singapore, 1 Research Link, Singapore 117604, Singapore*
[c] *Department of Biochemistry, National University of Singapore, 10 Kent Ridge Crescent, Singapore 119260, Singapore*

Received 5 March 2001; accepted 4 April 2001

First published online 24 April 2001

Edited by Gianni Cesareni

Abstract An integrated approach in protein discovery through the use of multidisciplinary tools was reported. A novel protein, Hcc-1, was identified by analysis of the hepatocellular carcinoma (HCC)-M cell proteome. The assembled EST sequence of the 210 amino acid novel protein was subsequently confirmed by rapid amplification of cDNA ends (RACE). A total of 687 bp at the 5′ untranslated region of Hcc-1 was identified. Promoter activity and several upstream open reading frames (uORFs) were demonstrated at this region. Bioinformatics prediction showed that the first 42 amino acids of the protein is a SAP domain with sequence matches to hnRNP from various vertebrate species. The Hcc-1 protein was localized to the cell nucleus while the gene was localized to chromosome 7q22.1. Hcc-1 cDNA level was increased in pancreatic adenocarcinoma. The level was also increased in well-differentiated hepatocellular carcinoma but decreases as the carcinoma progressed to a poorly differentiated stage. © 2001 Published by Elsevier Science B.V. on behalf of the Federation of European Biochemical Societies.

Key words: HCC-M; Nuclear protein; Upstream open reading frame; Intronless gene

1. Introduction

Hepatocellular carcinoma (HCC) is a major type of primary liver cancer. It originated from liver parenchymal cells (hepatocytes) and is one of the major malignant diseases of the world [1]. HCC is responsible for approximately one million deaths per year, mainly in Asia and Africa [2]. Data from epidemiology and experiments have identified several major risk factors associated with HCC including chronic infection with hepatitis B and hepatitis C viruses, and exposure to mycotoxins (aflatoxin B1) [3]. Recent advances in molecular genetics have generated much interest in the role of cellular oncogenes and the many oncogenic pathways leading to the malignant transformation of liver cells [4]. Treatment of HCC is largely palliative, and long-term survival is rare. As such, HCC remains one of the most challenging areas of study in medical oncology.

The reductionist approaches of the 20th century have successfully generated information about individual cellular components and their functions. However, an integrative analysis of the function of multiple gene products has become a critical issue for the future development in biology [5]. The advent of high-throughput techniques, such as genomics and proteomics, is enabling biologists to study cells as systems and to analyze globally the interplay between gene expression, protein levels, pathways and activity. This comprehensive approach is an extremely powerful tool for studying changes in gene expression and the resulting metabolic and phenotypic effects over the course of disease development [6].

Such global approach has not been widely applied in HCC studies. An earlier study on HCC using the proteomics approach was to analyze the global expression of nuclear matrix proteins [7]. Our group is currently looking at the development of HCC by taking global proteome snapshots of the human liver during different stages of disease progression. We have generated a database of the HCC-M cell proteome [8] recently. The large collection of proteins will facilitate proteomic strategies to uncover protein interaction networks [9]. We anticipate that the HCC-M database will be a source for the HCC research community and will foster the collaborative and consortial interactions necessary for global approaches to important biological questions.

In this paper, we demonstrated an integrative approach in protein discovery. A novel protein, Hcc-1, was first discovered by two-dimensional gel electrophoresis (2DE) and mass spectrometry (MS) analysis of HCC-M cells proteome. This was followed by de novo peptide sequencing of selected peptides using the tandem MS. Bioinformatics tools were subsequently used in predictions and expressed sequence tag (EST) assembly of a full-length in silico protein. Finally, a preliminary characterization of the novel protein or gene was carried out through an arsenal of molecular biology techniques.

2. Materials and methods

2.1. Proteome analysis and in silico protein assembly

The detailed procedure for proteome analysis of the HCC-M [10] cell line by 2DE and matrix-assisted laser desorption-time of flight (MALDI-TOF) MS has previously been published [8]. Selected peptide fragments were subjected to de novo sequencing using a PE Sciex OSTAR tandem hybrid quadrupole-TOF (QqTOF) MS system (Concord) equipped with a nanoelectrospray ionization (nESI) source (Pro-

*Corresponding author. Fax: (65)-7754933.
E-mail: btccml@nus.edu.sg

Abbreviations: MS, mass spectrometry; PCR, polymerase chain reaction; RACE, rapid amplification of cDNA ends; SEAP, secreted placental alkaline phosphatase; uORF, upstream open reading frame; 2DE, two-dimensional gel electrophoresis; MALDI-TOF, matrix-assisted laser desorption ionization-time of flight

PII: S0014-5793(01)02409-7

tana). The amino acid sequences were then sent for BLAST EST search. Batch Entrez (NCBI, NIH) was used to get a list of EST in FASTA format before submission to Cap3 Assembly (TigemNET) software for contigs searching. The contig that contained the peptide sequences in the appropriate open reading frame (ORF) was used to check for possible sequence extension with other contigs obtained from other peptide sequences. The sequence extension exercise was reiterated until the largest possible open reading frame was found. The assembled putative protein sequence was double-checked by comparing a theoretical trypsin digestion of the protein with the generated protein peak list of the peptide mass fingerprints to ensure that all significant peaks were identified and the range of listed peaks covered the putative protein.

2.2. Structure and function prediction for the novel protein

The putative structure and function of the novel protein Hcc-1 were predicted using bioinformatics tools available from the internet. The presence of trans-membrane segment and nuclear localization signal was predicted using PredictProtein [11], mitochondrial targeting region using PSORT [12], and secretory signal using SignalP [13]. The 5′-untranslated segment of the gene was searched with ProScan [14] for promoter region. Direct repeats in the gene were searched with GCG Repeat (Wisconsin Package Version 10.0, Genetics Computer Group (GCG), Madison, WI, USA).

Secondary structures of the protein were predicted by PHDsec [15]. The presence of domains in the protein was predicted by PredictProtein coiled-coil prediction, low-complexity regions (SEG) [16] and COILS [17]. The predicted domains were then used to search for conserved domains by the reverse position-specific (RPS)-BLAST [18]. Sequence and domain homology searches were carried out with position-specific iterative (PSI)-BLAST on a non-redundant database [18]. Hydropathicity of the protein was analyzed by the Kyte–Doolittle plot [19].

2.3. Molecular characterizations of Hcc-1

The novel protein Hcc-1 was subjected to wet laboratory characterizations to confirm its existence and its putative role in the cell. Unless otherwise stated for experiments listed below, all DNA primers were synthesized by Genset Singapore, restriction enzymes, dNTPs and Taq polymerase were obtained from Promega, agarose from FMC Bioproducts, polyacrylamide gel from Bio-Rad, and other chemical reagents from Sigma. All polymerase chain reaction (PCR) amplifications were carried out in a PTC-200 thermal cycler (MJ Research).

2.4. HCC-M RNA extraction and rapid amplification of cDNA ends (RACE)

RNA was extracted from the HCC-M cells using the Trizol reagent (Gibco). The extraction procedure was as recommended by the manufacturer. The amount and purity of RNA extracted was measured by absorbance reading at 260 and 280 nm. Typical total RNA yield from 3×10^7 cells is about 700 μg and 260/280 ratio is 2.0.

RACE was performed using the SMART RACE cDNA amplification kit (Clontech). Gene-specific primers for both the 5′ and 3′ RACE were designed from EST sequences deduced from the peptide fragments identified by tandem MS. The gene-specific primers were designed such that there would be a region of overlap between the region to be amplified by 5′ RACE and the region to be amplified by 3′ RACE. The overlapping region permitted the use of the primers together in a control PCR to eliminate non-specific amplification. The DNA sequences for the 5′ and 3′ RACE gene-specific primers for Hcc-1 are 5′-CTTTGCCTCTGTATCCTCTGTGGTTCC-3′ and 5′-TTCAATGTACCTGTGAGCTTGGAGAGTAAG-3′ respectively. RACE was performed with protocols from the kit manufacturer.

2.5. Chromosome walking

After the DNA sequence of the 5′ RACE product was identified, chromosome walking was performed to explore the 5′-untranslated region of the Hcc-1 gene. Four libraries of uncloned, adaptor-ligated genomic DNA fragments were purchased (GenomeWalker, Clontech). Chromosome walking was performed by nested PCR on the four libraries using primers designed from the 5′-end of the gene. The first PCR was performed using an outer adaptor primer and a gene-specific primer from the 5′-end of Hcc-1 (5′-CTTATGGAGCTCCACCGTCTCGGTC-3′). The PCR product was subjected to a second PCR using an inner adaptor primer (located 3′ to the outer adaptor primer) and a gene-specific primer located upstream of the first gene-specific primer (5′-ATCTTGTTACCCCTCACTCCACTCCCC-3′). The nested PCR mixture and conditions were as recommended by the kit manufacturer. PCR products were analyzed on a 1% agarose gel and sequenced (ABI377 automated DNA sequencer, Applied Biosystems).

Genomic DNA from HCC-M cells was extracted using a DNA extraction kit (Roche) and used to identify the genomic sequence of the Hcc-1 gene. Gene-specific primers were designed from the 5′-untranslated region (5′-AGGGGTAACAAGATGGCGACCGAGAC-3′) and 3′-untranslated region (5′-GTACCTGGGGTACATGCTCCCTCATTG-3′) of the gene. PCR was performed with 50 ng of DNA, 0.2 μM each of the primers, 0.2 mM of dNTPs and AdvanTaq Genomic DNA polymerase (Clontech) in a reaction buffer of the same manufacturer. The cycling parameters were 94°C for 30 s, 28 cycles of 94°C for 30 s and 68°C for 2 min, and a final extension at 68°C for 5 min. The PCR product was examined by electrophoresis on a 1% agarose gel and sequenced.

2.6. Hcc-1 promoter study

A promoter study was carried out to investigate whether the 5′-untranslated region of the Hcc-1 gene identified through chromosome walking contained any promoter function. The DNA from the 5′-untranslated region, identified through chromosome walking, was cloned into the pSEAP2 vector containing a human secreted placental alkaline phosphatase (SEAP) gene downstream from the cloning site (Clontech). *Eco*RI restriction sites were generated on to the 5′-untranslated region of the gene by PCR. The restricted fragment was ligated into the pSEAP2 vector that was similarly restricted. The clones were plated and selected by ampicillin resistance. The orientation of the inserted sequence was determined by DNA sequencing. The pSEAP2 vectors containing either a SV40 early enhancer sequence downstream of the SEAP or a SV40 early promoter sequence upstream of the SEAP were also constructed.

Successfully constructed vectors were transfected into Huh-7 cells [20] by the SuperFect reagent method (Qiagen). Normalization and measurement of transfection efficiency were performed by co-transfection with a pZeoSV2/lacZ vector (Invitrogen), which constitutively expressed the LacZ gene for β-galactosidase assay. Average transfection efficiency of 2% was routinely achieved at 24 h post-transfection. All transfection experiments were performed in triplicate. Quantitation of the SEAP was performed 48 h post-transfection using the SEAP Fluorescent Detection kit (Clontech) with a fluorescence reader (Spectrafluor, Tecan).

2.7. Chromosome localization of Hcc-1

The location of Hcc-1 on the chromosome was determined by radiation hybrid (RH) panel screening. Two RH panels, Genebridge 4 and Stanford G3, were purchased (Research Genetics). DNA primers flanking a region of 255 bp were generated from the 3′-untranslated region of the gene. The sequences of the primers were 5′-TGAAAAAGAGGAAGGAGCGA-3′ and 5′-TTCATGGATGTACCTGGGGT-3′. DNA (25 ng) from each of the cell lines was used as PCR template for the primers and each cell line was screened a minimum of two times. Using standard PCR conditions and protocols, the finished reaction was assayed by electrophoresis on a 3% agarose gel. The results were scored for the presence (score = 1) or absence (score = 0) of a PCR product of the expected size. Results which were ambiguous among the multiple runs were scored as a 2. These were then submitted to the Whitehead/MIT RH server (http://www-genome.wi.mit.edu) for data generated from Genebridge 4, and to the Stanford Human Genome Center RH server (http://shgc-www.stanford.edu) for data from Stanford G3 panel, where it was tested against the framework maps in the databases.

2.8. Hcc-1 recombinant protein expression

Full-length cDNA of Hcc-1 was generated by PCR from the HCC-M total cDNA pool. The primers used were 5′-GGATCCAACAAGATGGCGACC-3′ which contained a *Bam*HI restriction site, and 5′-CAGCTGGTACCTGGGGTACAT-3′ which contained a *Sal*I restriction site. The amplified region spanned the entire expressed 630 bp of the gene. PCR amplification was as described above. The PCR product was analyzed by electrophoresis on a 2% agarose gel, purified, restricted, and cloned into the pQE-30 expression vector (Qiagen). The gene was cloned 3′ to a 6×His tag provided by the vector.

The vector was transformed into *Escherichia coli* strain DH5α (Gibco). Expression and purification of the protein were done according to the QiaExpressionist (Qiagen) protocol. The purity of the expressed recombinant protein was examined by electrophoresis on a 10% sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS–PAGE) and stained with Coomassie blue dye. The gel was also Western blotted and the protein detected with anti-6×His monoclonal antibody (Qiagen). Quantitation of the expressed protein was performed using the DC Protein Assay kit (Bio-Rad). The expressed recombinant protein was confirmed by both N- and C-terminal sequencing.

2.9. Subcellular localization of Hcc-1

Polyclonal antibody against the Hcc-1 protein was raised in rabbits as described [21]. Pre-immunized rabbit serum was collected as control. The specificity and sensitivity of the antiserum was determined by Western blot analysis of HCC-M lysate after 2DE. It was also compared to the anti-6×His antibody in detecting the His-tagged recombinant protein in SDS–PAGE.

The subcellular localization of the Hcc-1 protein was determined by immunofluorescence staining [21] of HCC-M and Huh-7 cells with the diluted (1:100) antiserum. Co-localization was performed with mouse anti-human Golgi or mitochondria antibodies (LabVision). The antiserum was detected with diluted (1:50) fluorescein isothiocyanate (FITC)-tagged sheep anti-rabbit antibody (Dako) while the anti-Golgi or anti-mitochondria antibodies were detected with similarly diluted rhodamine-tagged goat anti-mouse antibody (Dako). Images of the cells were generated by overlaying the fluorescent microscopy (Zeiss Laser Scanning Confocal Microscope 410) images of each dye together with an image generated from the light microscopy using the same microscope.

2.10. Hcc-1 cDNA distributions in tissues

The cDNA distribution of Hcc-1 gene in different tissues was performed by PCR screening of commercially available multiple tissue cDNA panels (Clontech) using the manufacturer's protocol. The expression in four liver cell lines was also examined. The commercial panels have been normalized using four housekeeping genes (α-tubulin, β-actin, glyceraldehyde-3-phosphate dehydrogenase and phospholipase A2). Paired liver samples, one from the hepatocellular carcinoma tissue and the other from the non-tumor part of the liver, were obtained from informed consent subjects. All procedures with the study subjects were in accordance with the Helsinki Declaration of 1975, as revised in 1983. The same pair of primers used for chromosome walking was used for detection of Hcc-1 cDNA. A region of 778 bp spanning from the 5′-untranslated region to the 3′-untranslated region of the gene was covered. The PCR product (5 μl) was examined by electrophoresis on a 2% agarose gel. The experiment was performed in duplicate to ensure reproducibility.

3. Results and discussions

We report the molecular cloning and initial characterization of a novel nuclear protein discovered through concerted efforts in proteomics and bioinformatics. The protein is one of several novel proteins discovered in a proteomics initiative in creating a database of expressed proteins in the HCC-M cell line [8]. This discovery demonstrated an integrated approach through combined disciplines in proteomics, bioinformatics and functional genomics.

A protein spot which was not identified in SwissProt or NCBI databases by peptide mass fingerprinting from MALDI-TOF data was subjected to de novo sequencing of three of its in-gel trypsin-digested peptide fragments (Fig. 1). The amino acid sequences from the fragments were then used to search the EST database by using BLAST (tblastn). The Cap3 Assembly (TigemNET) software was used to assemble some 40 ESTs into a putative in silico ORF. Based on the consensus DNA sequence derived from EST assembly, DNA primers were designed from two of the peptide fragments (Fig. 2). RACE was performed to confirm the assembled putative protein and to obtain more information on the 5′- and 3′-untranslated regions of the gene. A total of 873 bp was identified. The RACE results were found to agree with the EST assembled sequence of the protein. We named the protein Hcc-1. The start codon of the gene was in agreement with the Kozak consensus sequence for eukaryotic genes [22].

With the exception of the first 42 amino acids, Hcc-1 protein has no significant match in SwissProt or NCBI non-redundant databases. Bioinformatics tools were not able to determine the structure of Hcc-1. The current protein structure prediction is usually done via homology modeling of known

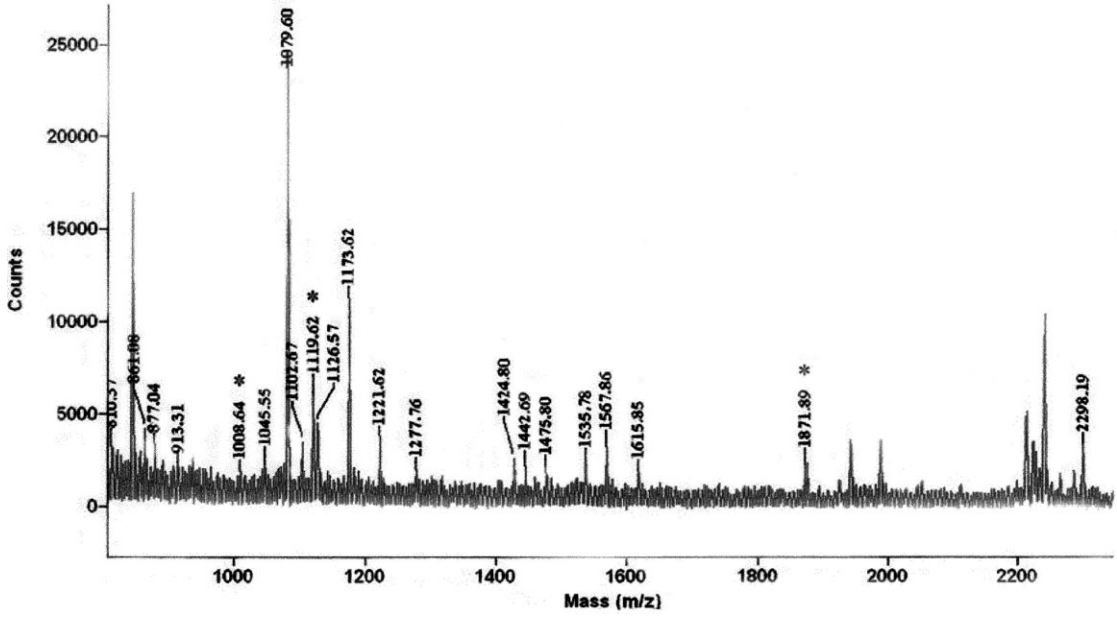

Fig. 1. Peptide mass fingerprint of Hcc-1 from MALDI-TOF analysis. Three peptide fragments (asterisked) were randomly chosen for de novo sequencing using tandem MS. The amino acid sequences of the peptides were used for EST assembly into an in silico protein.

```
   1  CAGGGGCAGC AGTGATTATC TGAACTCGGA TCTTTAAAAT TGTGGTAGCT CTAAAGCTGA
  61  TGATGTCTGG TTAGGAAGTG GCTCTTGCCC GCCCCAGCCC CACCGCCAGT TCCTTAAGCC
 121  CGCCCCATGC CCCTCCCAGC TTCCTCCTCA TGTTCATCGG TTTTTTCAGG GCTCCCTTCA
 181  ACGCTCCCCT CTCAGTATTT AGGTCACCAC TCCCTCGGCG CCCCTTTCGC CTCCCACCAT
 241  TTTTCCTCAG CAACCCTTAC AGTCTTTGCA GCTCCTACCT GCCAGCTCAG ATCCCCGTCC
 301  GGCTATGGGC GCGGCGCCGG CTACCACACC TGAAGTCTCC AGGAAGTAAC GCCTCTCCTT
 361  CTGCCCCTTT CCTGTTGGAG GAACAGAATC AGCGCTGCCA CCACCCATTG GTTGGTGGTC
 421  TGTAATGCAG AAGCACAGTT GGTTGCCATT TCTGTCGTTC GCAAGATACA GTGCCCGCCC
 481  CTCTCCCAGT TCCACCTTTT GAAAGAGGTG GGGCAAGCTG CCTAGAGAAG TGAGAGCGAC
 541  GTCAGCTATT GACCAATGGG AAGAGCTGAT GGTATGGCGT GGGAGCAAGA GTGACAACGA
 601  TTGGTCAGCC TTGCATCTCT ACGCCTAAGG CGGGAACTCC TGGAGGCGGA GGCCGCGGGT
                                   M  A  T  E  T  V  E  L  H  K  L
 661  GGGGGGAGTG GAGTGAGGGG TAACAAGATG GCGACCGAGA CGGTGGAGCT CCATAAGCTA
       K  L  A  E  L  K  Q  E  C  L  A  R  G  L  E  T  K  G  I  K
 721  AAGCTTGCCG AACTAAAGCA AGAATGTCTT GCTCGTGGTT TGGAGACCAA GGGAATAAAG
       Q  D  L  I  H  R  L  Q  A  Y  L  E  E  H  A  E  E  E  A  N
 781  CAAGATCTTA TCCACAGACT CCAGGCATAT CTTGAAGAAC ATGCTGAAGA GGAGGCAAAT
       E  E  D  V  L  G  D  E  T  E  E  E  E  T  K  P  I  E  L  P
 841  GAAGAAGATG TACTGGGAGA TGAAACAGAG GAAGAAGAAA CAAAGCCCAT TGAGCTCCCT
       V  K  E  E  E  P  P  E  K  T  V  D  V  A  A  E  K  K  V  V
 901  GTCAAAGAGG AAGAACCCCC TGAAAAAACT GTTGATGTGG CAGCAGAGAA GAAAGTGGTG
       K  I  T  S  E  I  P  Q  T  E  R  M  Q  K  R  A  E  R  F  N
 961  AAAATTACAT CTGAAATACC ACAGACTGAG AGAATGCAGA AGAGGGCTGA ACGATTCAAT
       V  P  V  S  L  E  S  K  K  A  A  R  A  A  R  F  G  I  S  S
1021  GTACCTGTGA GCTTGGAGAG TAAGAAAGCT GCTCGGGCAG CTAGGTTTGG GATTTCTTCA
       V  P  T  K  G  L  S  S  D  N  K  P  M  V  N  L  D  K  L  K
1081  GTTCCAACAA AAGGTCTGTC ATCTGATAAC AAACCTATGG TTAACTTGGA TAAGCTGAAG
       E  R  A  Q  R  F  G  L  N  V  S  S  I  S  R  K  S  E  D  D
1141  GAAAGAGCTC AAAGATTTGG TTTGAATGTC TCTTCAATCT CCAGAAAGTC TGAAGATGAT
       E  K  L  K  K  R  K  E  R  F  G  I  V  T  S  S  A  G  T  G
1201  GAGAAACTGA AAAAGAGGAA GGAGCGATTT GGGATTGTCA CAAGTTCAGC TGGAACTGGA
       T  T  E  D  T  E  A  K  K  R  K  R  A  E  R  F  G  I  A
1261  ACCACAGAGG ATACAGAGGC AAAGAAGAGG AAAAGAGCAG AGCGCTTTGG GATTGCCTGA
1321  TGAAAAGTTC CTGATACTTT CTGTTCTCCA GTGTTTTCCA TTTCTCTCCT TCTTCTTGGT
1381  CACATATATG CCTAAATGCA CAGTCATGTG CCTACGTCCT GCCTCGCAAT GAGGGAGCAT
1441  GTACCCCAGG TACATCCATG AACTGCGGCA GCAGTTTGAC TTATTGCTGT TTCAGCTTTA
1501  AGGTTGTTGT GTTTTTGTTT TTGATTATGT TGCTTGTTAA Taaaaaaaaa tagaaaa
```

Fig. 2. The complete Hcc-1 gene (AJ409089). Underlined fragments are amino acid sequences obtained by tandem MS analysis. Primers used for RACE are highlighted in bold. Extra bases from EST assembly are in lowercase. The translated product has 210 amino acids (in single-letter code). Boxed sequences are the direct repeats at both ends of the gene. Dotted sequences are uORFs before the start of the Hcc-1 coding region. A total of 687 nucleotides were identified at the 5′-untranslated region. Sequence with wavy line is the 246 nucleotides used in promoter study to circumvent the uORFs.

protein structures with sequence homologies higher than 30% [23]. Homology modeling could not be carried out as all the sequences were having less than 30% homology with Hcc-1 and do not have any established structures. Even though possible secondary structures of Hcc-1 can be predicted using software such as the PHD, we are unable to predict tertiary structures of the protein.

The protein was found to have no predicted trans-membrane segment (PredictProtein), no mitochondrial targeting sequence (PSORT), and no secretory signal (SignalP). It was also predicted to be a globular protein with predominantly α-helix secondary structure (PHDsec). The protein may have three domains based from searches using PredictProtein (coiled-coil region at amino acid positions 30–51 and 146–160), COILS (at amino acid positions 25–64 and 145–172), and SEG low-complexity region (at amino acids 42–79 and 165–179). The novel protein was then divided into three domains for subsequent searches.

The first domain (consisting of the first 42 amino acids) was searched with the Conserved Domain database using RPS-BLAST. The domain was found to have homology with the SAP domain (e value: 5e-04), which is a putative bi-helical DNA binding motif predicted to be involved in chromosomal organization and transcriptional regulations [24]. The SAP (after SAF-A/B, ACINUS and PIAS) motif is found in diverse nuclear proteins. SAF-A/B or heterogeneous nuclear ribonucleoprotein U (scaffold attachment factor A/B) binds RNA through RGG box [25] and cleavage of the protein by enzyme caspase-3 will result in apoptosis [26]. ACINUS or apoptotic chromatin condensation inducer in the nucleus is a caspase-3-activated protein required for apoptotic chromatin condensation [27]. PIAS are protein inhibitors of activated STAT (signal transducer and activator of transcription). STAT proteins are latent cytoplasmic transcription factors that become activated in response to stimulation by various cytokines [28]. PIAS have been identified as potentially important down-regulators of this pathway.

Using PSI-BLAST on non-redundant database, amino acid sequence 1–42 of Hcc-1 was matched to vertebrate heterogeneous nuclear ribonucleoprotein (hnRNP) with identities match of above 45%: hnRNP U scaffold attachment factor A (SAF-A) (Q00839) of *Homo sapiens*, hnRNP U protein (X65488) of *H. sapiens*, hnRNP U (AF073992) of *Mus musculus*, SP120 (D14048) (nuclear scaffold protein that binds the matrix attachment region DNA) of *Rattus norvegicus*, and SAF A (AF068847) of *Xenopus laevis*. The other two domains (amino acids 43–170 and 171–210) were not found to have any known homologous domain.

Chromosome walking was carried out to determine the 5′-untranscribed region of the Hcc-1 gene. A total of 687 bp upstream of the start codon ATG was determined (Fig. 2). A number of mini-cistrons or upstream open reading frames (uORFs) were noted. The 5′-untranslated region of Hcc-1 gene was found to have a number of start and stop codons.

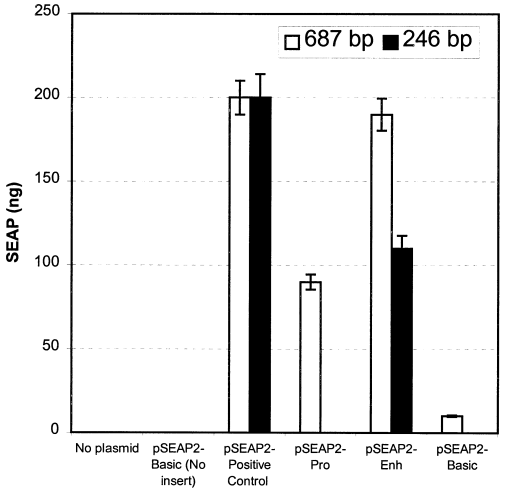

Fig. 3. Promoter study on the 5′-untranslated region of the Hcc-1 gene. The activity of the whole fragment (687 bp) was compared to a shorter 246 bp fragment located at the 5′-portion of the untranslated region. pSEAP2-Basic: the basic pSEAP2 vector; pSEAP2-Positive Control: a control pSEAP2 vector with a SV40 promoter upstream, and a SV40 enhancer downstream, of the SEAP gene; pSEAP2-Pro: a pSEAP2 vector with the SV40 promoter upstream of the multiple cloning site; pSEAP2-Enh: a pSEAP2 vector with the SV40 enhancer downstream of the SEAP gene.

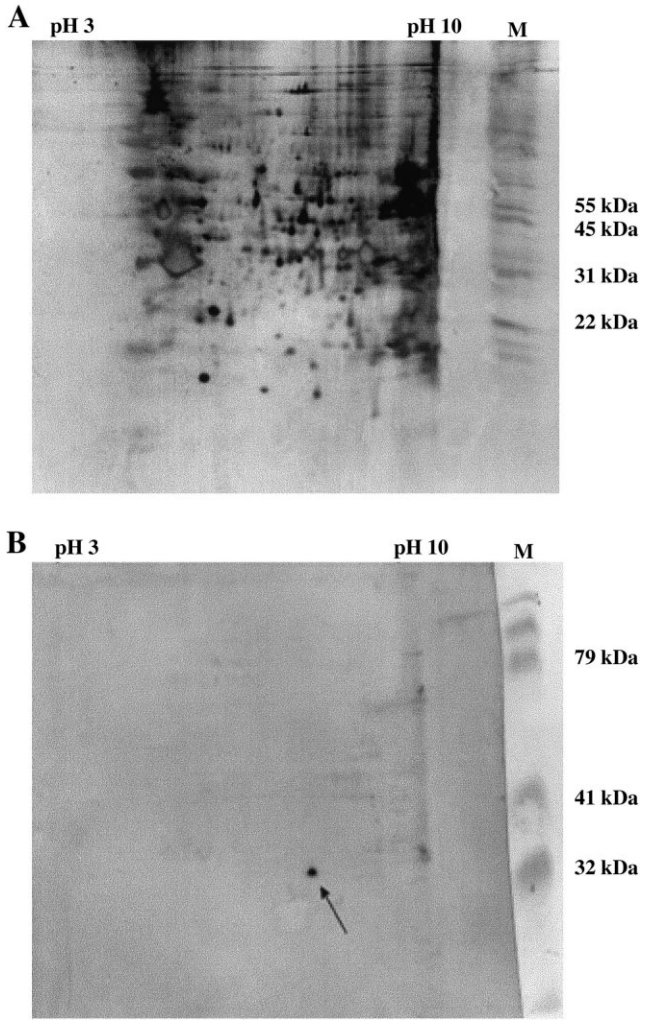

Fig. 4. A: Silver stained 2D gel image of HCC-M total lysate. B: Western blot detection of Hcc-1 in HCC-M 2D gel. Western blot determination of the sensitivity and specificity of the rabbit-raised polyclonal antibody against Hcc-1. 2DE was carried out in duplicate with HCC-M total protein lysate loaded at 40 g each. One of the gels was silver stained (A) while another was used for Western blot detection using the polyclonal antibody with calorimetric detection (B). The location of the Hcc-1 protein was indicated by an arrow.

The occurrence of a long 5′-untranslated region with minicistrons or uORFs is not uncommon. It is found in a number of proto-oncogenes and growth factors [29]. It is a structure used in transcriptional regulation and translational control of genes whose products are important for cell growth [30,31].

Chromosome walking into the Hcc-1 gene revealed no intron. The amplified genomic DNA product was sequenced and confirmed to have the same sequence with the expressed cDNA of Hcc-1. Direct repeats of eight nucleotides long were detected at the 5′- and 3′-untranslated regions of the gene (Fig. 2). A processed mRNA can sometimes re-integrate back into the genome and expressed on occasions. Processed genes that are re-expressed as mRNA are sometimes called retrogenes [32]. Several common characteristics of a retrogene are lack of intron, direct eight nucleotides repeats, and retaining the functionality and ORF of their parental gene [33,34]. Except from finding a parental gene with introns, the Hcc-1 gene satisfied most criteria as a retrogene.

Promoter region was predicted from nucleotides 294–544 at the 5′-untranslated region of the gene by ProScan (Fig. 2). The prediction was based on regions of DNA that contain a significant number and type of transcriptional elements that are usually associated with eukaryotic polymerase II promoter sequences. No TATA box and transcriptional start site were predicted.

To confirm the bioinformatics prediction, the 687 bp fragment was cloned into the pSEAP2 vector upstream of the SEAP gene. The amount of the expressed SEAP was assayed 48 h post-transfection of the vector into Huh-7 cells. This provided an indirect measurement of the strength of the promoter activity of the gene. Low promoter activity was observed (10 ng SEAP expressed per 5 μg DNA) with the 687 bp fragment (Fig. 3). For the pSEAP2 vector construct with a SV40 early promoter, an increase in SEAP transcription was observed (90 ng SEAP expressed). However, high transcription activity was obtained when the 687 bp fragment was constructed into a vector containing SV40 early enhancer sequence (190 ng SEAP expressed). This showed that an enhancer element is needed for the transcriptional activity of the promoter.

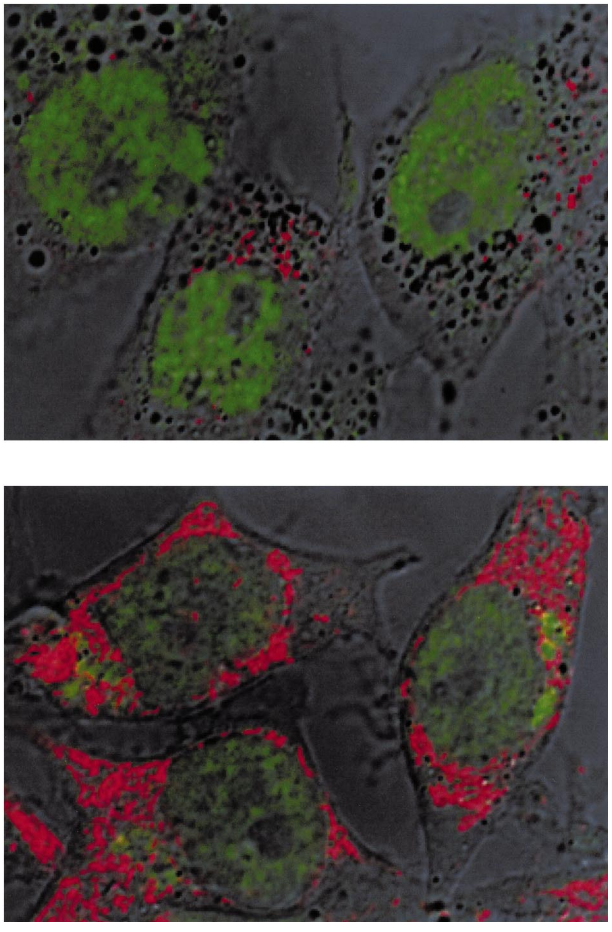

Fig. 5. Upper panel: Huh-7 cells stained with anti-Hcc-1 (green) and anti-Golgi (red) antibodies. Lower panel: HCC-M cells stained with anti-Hcc-1 (green) and anti-mitochondria (red) antibodies. Subcellular localization of the Hcc-1 protein in two liver cell lines. The Hcc-1 antibody was detected with FITC labeled anti-rabbit antibody. Co-localization was performed with rhodamine labeled anti-mouse antibodies against the Golgi (upper panel) and mitochondria (lower panel).

To circumvent the mini-cistrons, 246 bp (Fig. 2) from the 5′-end of the 687 bp fragment was amplified and inserted into the pSEAP2 vector. No activity was observed when the pSEAP2 vector was constructed without SV40 early enhancer or promoter sequences. Transcriptional activity was observed at half (110 ng of SEAP expressed) of that from 687 bp fragment when the SV40 early enhancer sequence was included in the construct (Fig. 3). The results showed that the promoter region is located primarily at the middle of the identified 5′-untranslated region of the gene. The enhancer sequence is probably further upstream from the 687 bp sequence.

The placement of the Hcc-1 gene was the best possible when scored against the framework markers at the time of writing. The gene was assigned to chromosome 7 at position 7q22.1, or 3.36 cR from a framework microsatellite marker, D75651 (GDB accession: AFM249za5). A search in the Mitelman Database of Chromosome Aberrations in Cancer from the National Cancer Center [35] revealed that there were no chromosome aberrations (insertions, deletions, translocations, etc.) reported for pancreatic tumors. There have been several reported cases of hepatic adenocarcinomas associated with aberrations in chromosomes 1, 8, 14 and 16, hepatoblastoma is associated with chromosomes 1, 4 and 8, and mesenchymal tumor of the liver with chromosomes 11 and 19. A search in the above database also shown that chromosome 7q22 is an area implicated in a large number of human tumors (e.g. non-Hogkins lymphoma, acute myeloid leukemia, germ cell tumor). It is yet to be determined whether Hcc-1 plays a role in chromosome aberrations leading to carcinoma.

The specificity and sensitivity of the rabbit-raised polyclonal antibody against Hcc-1 was tested by Western blot analysis after HCC-M 2DE. Two similar 2D gels of HCC-M were created; each loaded with 40 μg of total protein lysate. One of the gels was silver stained (Fig. 4A), and the other Western blotted. The polyclonal antibody at 1000× dilution was able to detect the presence of the Hcc-1 protein in a colorimetric assay with minimal non-specificity (Fig. 4B). The antibody was then used in subcellular localization experiments whereby Hcc-1 was found to localize to the nucleus of liver cells (Fig. 5). PredictProtein further supported the possibility of a nuclear protein with the prediction of a nuclear localization signal at amino acid positions 197–203.

The Hcc-1 protein migrated consistently at 35 kDa in 2DE of the HCC-M lysate despite that the protein has only 210 amino acids with a theoretical molecular mass of 24 kDa (Fig. 4B). The theoretical molecular mass of the recombinant protein is 26 kDa due to 13 additional amino acids at the N-terminal. The extra amino acids were the 6×His tag and amino acids from the multiple cloning site. The recombinant Hcc-1 protein has also migrated as a 35 kDa band in SDS–PAGE (results not shown). Both N- and C-terminal amino acid sequencing results showed that the band at 35 kDa contained the full-length Hcc-1 protein. The deviation in migration may be due to the fact that Hcc-1 is a very hydrophilic

Table 1
A semi-quantitative analysis of Hcc-1 cDNA levels in human tissues, liver cell lines and paired liver samples

	Samples	Hcc-1 cDNA levels
Healthy tissues	colon	−
	ovary	−
	peripheral blood leukocyte	−
	prostate	+/−
	small intestine	−
	spleen	+
	testis	++
	thymus	+
	brain	−
	heart	+
	kidney	+
	liver	+/−
	lung	−
	pancreas	+
	placenta	−
	skeletal muscle	−
Diseased tissues	breast carcinoma (GI-101)	+/−
	lung carcinoma (LX-1)	−
	colon adenocarcinoma (CX-1)	−
	lung carcinoma (GI-117)	−
	prostatic adenocarcinoma (PC3)	−
	colon adenocarcinoma (GI-112)	−
	ovarian carcinoma (GI-102)	−
	pancreatic adenocarcinoma (GI-103)	+++
Liver cell lines	HCC-M	+
	HepG2	+
	Hep3B	+
	HuH4	+
	HuH7	+
Paired liver samples[a]	non-tumor (subject A)	−
	tumor (subject A)	+++
	non-tumor (subject B)	++
	tumor (subject B)	+

The semi-quantitative analysis was performed at 22 cycles of PCR. Saturation of PCR products began at 26 cycles. Expression level: +++ > ++ > + > +/−; −, no expression.

[a]Both subjects were positive for hepatitis B virus infection. Subject A had well-differentiated tumor while subject B has poorly differentiated tumor.

protein (with only 29.5% hydrophobic amino acids by Kyte–Doolittle plot) that interferes with the SDS binding of the protein in SDS–PAGE.

The PCR screening method was chosen over the conventional Northern blot method because it permits detection of mRNA of all abundance levels and it is much faster than hybridization analyses. A trial PCR experiment was performed at 20–32 cycles with two-cycle intervals. An aliquot of the PCR mixture at each interval was electrophoresed on a 2% agarose gel. Exponential phase of the PCR was determined visually and found to range from 20 to 24 cycles. Saturation was found to begin from cycle 26. We have chosen to stop the PCR at cycle 22 for the semi-quantitative analysis of Hcc-1 cDNA level in tissues.

The Hcc-1 cDNA distributions in various human tissues are shown in Table 1. Experiments were performed in duplicate and identical results were obtained. The Hcc-1 cDNA is found at varying levels in several healthy tissues. The level is raised markedly in the pancreatic adenocarcinoma tissue when compared with healthy pancreas. Hcc-1 cDNA levels in the paired liver samples are as tabulated (Table 1). Both subjects were positive for hepatitis B virus infection. Subject A had well-differentiated tumor while subject B had poorly differentiated tumor. From the table, Hcc-1 seems to be differentially expressed. It is increased in well-differentiated hepatocellular carcinoma but its level decreased as the tumor progressed to poorly differentiated hepatocellular carcinoma. The pancreas and liver have the same embryonic developmental origin [36]. Thus it is not surprising to find the increase of Hcc-1 cDNA in both types of tumor.

Expression of a human retrogene in several organs is not common, though there are several reports on human retrogenes having single organ expression [37]. Most retrogenes are transcriptionally inactive and contain a number of mutations that render them non-functional. However, there are some reported functional retrogenes [33,34]. Hcc-1 is atypical as it is expressed in several organs. Our non-saturating PCR on normalized cDNA gives an accurate comparison of relative mRNA copy numbers [38,39]. Hcc-1 cDNA is apparently not expressed to high level except in pancreatic adenocarcinoma and hepatocellular carcinoma. Whether the translated product has any biological activity remains to be determined.

During the course of this work, a human CD34+ stem cell partial mRNA sequence was deposited with GenBank (AF161434, February 2000). The sequence was found to have 95.2% DNA sequence identity at the coding and 3′-untranslated regions, and 70.2% amino acid sequence identity with the Hcc-1 protein. We are reasonably convinced that our sequence is authentic as it was obtained independently through RACE and matched the assembled consensus EST sequence derived from some 40 ESTs. Its existence was further confirmed by chromosome localization using the 3′-untranslated region, chromosome walking into both the coding sequence and the 5′-untranslated region, and demonstrated promoter activity within the 5′-untranslated region. The minor differences between the two genes were probably caused by inaccuracies in EST sequencing of the CD34+ stem cell cDNA library. Another human pre-pro-B cell genomic clone (AP001207, June 2000) from chromosome 8q23 was found to have 53.3% DNA sequence identity with the Hcc-1 gene. The region that has sequence identity with the Hcc-1 gene occurred at 49–50 kb of the 154 kb genomic clone. The fact that there were no suitable ORFs within that region of the genomic clone and that Hcc-1 gene was localized to chromosome 7q22.1 showed that the Hcc-1 gene is a separate and independent entity.

4. Conclusions

The bioinformatics predictions and experimental data suggested that Hcc-1 is a nuclear protein with potential nucleic acid binding capability that may participate in important transcriptional or translational control of cell growth, metabolism, and possibly carcinogenesis. Further functional characterizations are needed to fully understand the biology of the novel nuclear protein and its roles in carcinogenesis. We are currently in the process of finding interacting partners for Hcc-1 by protein–protein interaction study. A DNA protein binding study will also be carried out to determine the ability of Hcc-1 in binding RNA or DNA. It is hoped that a potential use of this gene product in the diagnosis or therapeutics of an important class of tumors can be discovered in the near future.

Acknowledgements: The authors wish to thank Ms. Y. Mao for raising the polyclonal antibody, Ms. M.S. Teo for animal cell culture work, Ms. S.S. Ng and C.L. Leaw for excellent technical help in molecular characterizations of the gene. The paired human liver samples were provided by Prof. C.K. Leow, Department of Surgery, National University of Singapore. The work was supported by a Core Competencies grant from the National Science and Technology Board, Singapore.

References

[1] Schafer, D.F. and Sorrell, M.F. (1999) Lancet 353, 1253–1257.
[2] Bosch, F.X. (1997) in: Liver Cancer (Okuda, K. and Tabor, E., Eds.), pp. 13–28, Churchill Livingstone, New York.
[3] Miyazaki, M. and Namba, M. (1994) in: Atlas of Human Tumor Cell Lines (Hay, R.J., Park, J-G. and Gazdar, A., Eds.), pp. 185–212, Academic Press, San Diego, CA.
[4] Buendia, M.A. (2000) Sem. Cancer Biol. 10, 185–200.
[5] Palsson, B. (2000) Nat. Biotechnol. 18, 1147–1150.
[6] Glassbrook, N., Beecher, C. and Ryal, J. (2000) Nat. Biotechnol. 18, 1142–1143.
[7] Chew, E.C., Liew, C.T., Wu, S., Yang, L., Yam, H.F., Wang, S.W., Lee, S.M., Wang, Z.H. and Chew-Cheng, S.B. (1997) Anticancer Res. 17, 3581–3586.
[8] Seow, T.K., Ong, S-E., Liang, R.C.M.Y., Ren, E-C., Chan, L., Ou, K. and Chung, M.C.M. (2000) Electrophoresis 21, 1787–1813.
[9] Ito, T., Tashiro, K., Muta, S., Ozawa, R., Chiba, T., Nishizawa, M., Yamamoto, K., Kuhara, S. and Sakaki, Y. (2000) Proc. Natl. Acad. Sci. USA 97, 1143–1147.
[10] Watanabe, T., Morizane, T., Tsuchimoto, K., Inagaki, Y., Munakata, Y., Nakamura, T., Kumagai, N. and Tsuchiya, M. (1983) Int. J. Cancer 32, 141–146.
[11] Rost, B. (1996) Methods Enzymol. 266, 525–539.
[12] Nakai, K. and Kanehisa, M. (1992) Genomics 14, 897–911.
[13] Ni, H., Engelbrecht, J., Brunak, S. and von Heijne, G. (1997) Protein Eng. 10, 1–6.
[14] Prestridge, D.S. (1995) J. Mol. Biol. 249, 923–932.
[15] Rost, B. and Sander, C. (1993) J. Mol. Biol. 232, 584–599.
[16] Wootton, J.C. and Federhen, S. (1996) Methods Enzymol. 266, 554–571.
[17] Lupas, A. (1996) Methods Enzymol. 266, 513–525.
[18] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Nucleic Acids Res. 25, 3389–3402.
[19] Kyte, J. and Doolittle, R.F. (1982) J. Mol. Biol. 157, 105–132.

[20] Nakabayashi, H., Taketa, K., Miyano, K., Yamane, T. and Sato, J. (1982) Cancer Res. 42, 3858–3863.
[21] Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Seidman, J.G., Smith, J.A. and Struhl, K., Eds. (1998) Current Protocols in Molecular Biology, Wiley, New York.
[22] Kozak, M. (1987) Nucleic Acids Res. 15, 8125–8148.
[23] Brenner, S.E., Chothia, C. and Hubbard, T.J.P. (1998) Proc. Natl. Acad. Sci. USA 95, 6073–6078.
[24] Massari, M.E. and Murre, C. (2000) Mol. Cell. Biol. 20, 429–440.
[25] Kiledjian, M. and Dreyfuss, G. (1992) EMBO J. 11, 2655–2664.
[26] Kipp, M., Schwab, B.L., Przybylski, M., Nicotera, P. and Fackelmayer, F.O. (2000) J. Biol. Chem. 275, 5031–5036.
[27] Sahara, S., Aoto, M., Eguchi, Y., Imamoto, N., Yoneda, Y. and Tsujimoto, Y. (1999) Nature 401, 168–173.
[28] Sturm, S., Koch, M. and White, F.A. (2000) J. Mol. Neurosci. 14, 107–121.
[29] Willis, A.E. (1999) Int. J. Biochem. Cell Biol. 31, 73–86.
[30] Brown, E.J. and Schreiber, S.L. (1996) Cell 86, 517–520.
[31] Clemens, M.J. and Bommer, U-A. (1999) Int. J. Biochem. Cell Biol. 31, 1–23.
[32] Dahl, H.H., Brown, R.M., Hutchison, W.M., Maragos, C. and Brown, G.K. (1990) Genomics 8, 225–232.
[33] McCarrey, J.R. and Thomas, K. (1987) Nature 326, 501–505.
[34] Gebara, M.M. and McCarrey, J.R. (1992) Mol. Cell. Biol. 12, 1422–1431.
[35] Mitelman, F., Mertens, F. and Johansson, B. (1997) Nat. Genet. 15, 417–474.
[36] Bock, P., Abdel-Moneim, M. and Egerbacher, M. (1997) Microsc. Res. Tech. 37, 374–383.
[37] Kuittinen, T., Eggert, A., Lindholm, P., Horelli-Kuitunen, M., Palotie, A., Maris, J.M. and Saarma, M. (2000) FEBS Lett. 473, 233–236.
[38] Spanakis, E. and Brouty-Boyé, D. (1994) Nucleic Acids Res. 22, 799–806.
[39] Savonet, V., Mainhaut, C., Miot, F. and Pirson, I. (1997) Anal. Biochem. 247, 165–167.