MICROBIOLOGY AND TUMORBIOLOGY CENTER KAROLINSKA
INSTITUTE, STOCKHOLM, SWEDEN

CHARACTERIZATION OF A PUTATIVE TUMOR SUPPRESSOR REGION IDENTIFIED BY THE ELIMINATION TEST ON HUMAN 3p21.3

Hajnalka Kiss

Stockholm 2003

Published and printed by Karolinska University Press
Box 200, SE-171 77 Stockholm, Sweden

ISBN 91-7349-558-1

To my Family

CONTENTS

ABSTRACT

An experimental system, called the elimination test (Et), was developed to identify tumor growth antagonizing genes. Chromosome 3 containing microcell hybrids (MCHs) were passaged through SCID mice and regular deletion of the short arm of chromosome 3 (3p) were found by cytogenetic and molecular methods. The Et focused on 3p due to its frequent deletion in 23 tumor types including renal cell carcinoma (RCC), small cell lung carcinoma (SCLC) and breast cancer. By increasing the number of the tumors and by marker enrichment in the region of interest the common eliminated region 1 (CER1 or C3CER1) was reduced gradually to 1.6 cM on 3p21.3. Our aim was to further decrease the size of C3CER1 and identify the gene content.

Additional markers were selected to further study the panel of MCH derived SCID tumors. We reduced C3CER1 to approximately 1 Mb and covered it with a PAC contig. Fiber-FISH was performed with PAC clones to assess the integrity of our contig.

From the telomeric part of C3CER1 two overlapping PACs were selected for sequencing. We identified a novel LIM domain containing gene, which we named to LIM domain containing 1 (*LIMD1*). We have characterized and mapped the mouse ortholog (*Limd1*) of the human gene. They both contain three tandemly arrayed LIM domains at their C-terminal ends.

As the continuation of the gene identification in C3CER1, we constructed a physical and transcriptional map of a 250 kb region, which included four additional genes (*LZTFL1*, *KIAA0851/SAC1*, *XT3*, *CCR9*). We characterized a novel leucine zipper containing gene, the Leucine Zipper Transcription Factor-Like 1 (*LZTFL1*) and its mouse ortholog (*Lztfl1*). The *LZTFL1* gene has two isoforms displaying alternative polyadenylation.

The further detailed description of the transcriptional content of the C3CER1 was possible due to further sequencing of eleven PAC clones and the appearance of new sequences in the public database. We assembled all the available data and detected 19 genes and 3 processed pseudogenes within a 1.4 Mb comprehensive transcriptional map. We identified and characterized four novel genes: FYVE and coiled-coil domain containing 1 (*FYCO1*), Transmembrane protein 7 (*TMEM7*), Leucine-rich repeat-containing 2 (*LRRC2*), Leucine Zipper Protein 3 gene (*LUZP3*). The refined centromeric border of C3CER1 is located inside the *LRRC2* gene. A chemokine receptor cluster was recognized within C3CER1, which consisted of 8 genes.

The availability of the mouse sequence from Celera database made it possible to apply a comparative study to further investigate C3CER1. A 1.32-Mb human genomic sequence was compared with its 907-kb mouse orthologous sequence. The corresponding mouse region was found divided into two blocks, but their gene content and gene positions were highly conserved between human and mouse. We observed that two orthologous mouse genes (*Xtrp3s1* and *Cmkbr1*) were duplicated. We also recognized a large number of conserved elements that were neither exons of known genes, CpG islands, nor repeats. We further identified and characterized five novel orthologous mouse genes (*Kiaa0028*, *Xtrp3s1*, *Fyco1*, *Tmem7*, and *Lrrc2*). The murine/human conservation breakpoint region (CBR) has features that characterize unstable chromosomal regions: deletions in YAC clones, late replication, gene and segment duplications, pseudogene insertions.

The detailed transcriptional map of C3CER1 was prerequisite for the delineation of its role in tumorigenesis. We found 18 genes within C3CER1, which will be studied further. Preliminary results indicate that the *LTF* and the *LIMD1* are candidate tumor antagonizing genes.

LIST OF PUBLICATIONS

This thesis is based on the following publications that will be referred in the text by their roman numerals.

I. Yang Y, **Kiss H**, Kost-Alimova M, Kedra D, Fransson I, Seroussi E, Li J, Szeles A, Kholodnyuk I, Imreh MP, Fodor K, Hadlaczky G, Klein G, Dumanski JP, Imreh S. A 1-Mb PAC contig spanning the common eliminated region 1 (CER1) in microcell hybrid-derived SCID tumors.
Genomics. 1999; 62:147-55.

II. Kiss H, Kedra D, Yang Y, Kost-Alimova M, Kiss C, O'Brien KP, Fransson I, Klein G, Imreh S, Dumanski JP. A novel gene containing LIM domains (LIMD1) is located within the common eliminated region 1 (C3CER1) in 3p21.3.
Human Genetics. 1999; 105:552-9.

III. Kiss H, Kedra D, Kiss C, Kost-Alimova M, Yang Y, Klein G, Imreh S, Dumanski JP. The LZTFL1 gene is a part of a transcriptional map covering 250 kb within the common eliminated region 1 (C3CER1) in 3p21.3.
Genomics. 2001; 73:10-9.

IV. Kiss H, Yang Y, Kiss C, Andersson K, Klein G, Imreh S, Dumanski JP. The transcriptional map of the common eliminated region 1 (C3CER1) in 3p21.3.
European Journal of Human Genetics. 2002; 10:52-61.

V. Kiss H, Darai E, Kiss C, Kost-Alimova M, Klein G, Dumanski JP, Imreh S. Comparative human/murine sequence analysis of the common eliminated region 1 from human 3p21.3.
Mammalian Genome. 2002; 13:646-55.

VI. Kost-Alimova M, **Kiss H**, Fedorova L, Yang Y, Dumanski JP, Klein G, Imreh S. The coincidence of synteny breakpoints with malignancy related deletions on human chromosome 3.
Proc Natl Acad Sci USA. 2003; in press

ABBREVIATIONS

aa	Amino acid
ABI	Applied Biosystems
bp	Base pair
BAC	Bacterial artificial chromosome
CBR	Conservation breakpoint region
CCR	Chemokine receptor
CCS	Conserved chromosomal segment
cDNA	Complementary cDNA
CER	Common eliminated region
CpG island	Cytosine and guanosine rich region
CRR	Common retained region
DNA	Deoxyribonucleic acid
EST	Expressed sequence tag
Et	Elimination test
FER	Frequently eliminated region
FISH	Fluorescence in situ hybridization
FYCO1	FYVE and coiled-coil domain containing gene 1
HD	Homozygous deletion
kb	Kilo base pair
kD	Kilo dalton
LF (*LTF*)	Lactoferrin, Lactotransferrin gene
LIMD1	LIM domains containing gene 1
LOH	Loss of heterozygosity
LRRC2	Leucine-rich repeat-containing gene 2
LUZPP1	Leucine zipper protein pseudogene 1
LZTFL1	Leucine zipper transcription factor-like gene 1
Mb	Million base pairs
MCH	Microcell hybrid
MMCT	Microcell mediated chromosome transfer
mRNA	Messenger ribonucleic acid
ORF	Open reading frame
PAC	P1 artificial chromosome
PCR	Polymerase chain reaction
RACE	Rapid amplification of cDNA ends
RCC	Renal cell carcinoma
RNA	Ribonucleic acid
RT-PCR	Reverse transcriptase PCR
SCID	Severe combined immunodeficiency
SCLC	Small cell lung carcinoma
TMEM7	Transmembrane protein gene 7
UTR	Untranslated region
YAC	Yeast artificial chromosome
WI	Whitehead Institute
Ψ	Pseudogene

1. INTRODUCTION

1.1 Tumorigenesis

Cancer usually arises from a single cell that has accumulated genomic alterations that provide the cell with selective growth advantage. It is generally accepted that multiple genetic changes are necessary for tumor development. Hanahan and Weinberg postulates that the vast range of genetic changes in cancer manifests in six essential alterations in cell physiology that collectively dictate malignant growth: self-sufficiency in growth signals, insensitivity to growth-inhibitory signals, evasion of programmed cell death (apoptosis), limitless replicative potential, sustained angiogenesis and tissue invasion and metastasis (Hanahan et al., 2000). They proposed that these six changes are shared by most, if not all, types of human tumors. All these crucial regulatory circuits are heavily guarded normally; therefore cancer is a relatively rare event during an average human lifetime.

One of the best studied examples for multistage cancer development is the colorectal cancer. Colorectal carcinomas arise through a series of well-characterized histopathological transformations as a result of specific genetic changes (Figure 1) (Kinzler et al., 1996). One oncogene (*K-RAS*) and three tumor suppressor genes (*APC*, *SMAD4*, *TP53*) are the main targets of these alterations (Fodde et al., 2001). There are two main classes of genes that are involved in tumorigenesis: oncogenes and tumor suppressor genes.

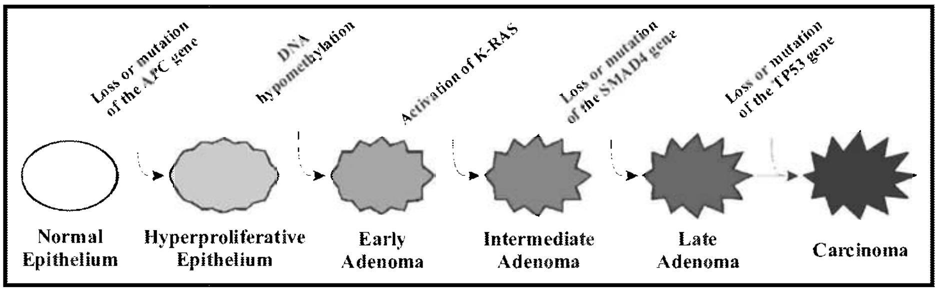

Loss or mutation of the APC gene
DNA hypomethylation
Activation of K-RAS
Loss or mutation of the SMAD4 gene
Loss or mutation of the TP53 gene
Normal Epithelium
Hyperproliferative Epithelium
Early Adenoma
Intermediate Adenoma
Late Adenoma
Carcinoma

Figure 1. Genetic changes associated with colorectal tumorigenesis (Kinzler et al., 1996) modified.

1.1.1. Oncogenes

The first oncogenes were discovered as transforming genes in the tumorigenic viruses (retroviruses) in animals. The oncogene research expanded rapidly, when it was discovered that human tumors contained activated oncogenes orthologous to those found in retroviruses. Oncogenes are altered forms of normal cellular genes called proto-oncogenes. They are mainly involved in the regulation of cell growth: growth factors and their receptors, signal transducers, transcription factors and regulators of programmed cell death (Table 1). Proto-oncogenes are mostly activated by mutation, gene amplification or chromosomal rearrangement. These mechanisms lead to qualitative or quantitative gain-of-function by either an alteration of proto-oncogene structure or an increase in proto-oncogene expression. In the majority of cancers mutations in proto-oncogenes arise somatically in the tumor cells, although germline mutations activating the function of the *RET* gene have been identified in multiple endocrine neoplasia type 2 and familiar medullary thyroid cancer patients. (Santoro et al., 1995). Germline mutations of the *MET* gene have also been found in affected members of families with hereditary papillary renal cell carcinoma (Schmidt et al., 1997; Zhuang et al., 1998).

Table 1. Examples of oncogenes

Oncogene	**Neoplasm**	**Mechanism of activation**	**Protein function**
Growth factors			
KS3	Kaposi's sarcoma	Constitutive expression	Member of FGF family
HST	Stomach carcinoma	Constitutive expression	Member of FGF family
Growth factor receptors			
EGFR	Squamous cell carcinoma	Gene amplification/ increased protein	EGF receptor
TRK	Colon/thyroid carcinomas	DNA rearrangement/ constitutive activation	NGF receptor
RET	Carcinomas of thyroid, MEN2A, MEN2B	DNA rearrangement/ Point mutation	GDNF/NTT/ART/PSP receptor
Signal transducers			
SRC	Colon carcinoma	constitutive activation	Protein tyrosine kinase
ABL	CML	DNA rearrangement	Protein tyrosine kinase
H-RAS	Colon, lung, pancreas carcinomas	Point mutation	GTPase
Transcription factors			
N-MYC	Neuroblastoma, lung carcinoma	Deregulated activity	Transcription factor
L-MYC	Lung carcinoma	Deregulated activity	Transcription factor
Others			
BCL2	B-cell lymphomas		Antiapoptotic protein
MDM2	Sarcomas		Complexes with p53

1.1.2. Tumor suppressor genes

Tumor suppressor genes are involved in the control of cell proliferation. Their loss or inactivation is associated with the development of malignancy. The existence of tumor suppressor genes was suggested by experiments showing the suppression of malignancy in somatic cell hybrids and a consistent loss of selected chromosomal regions in hereditary and sporadic cancers.

The studies of Klein, Harris and Ephrussi provided the evidence that the ability of cells to form a tumor is a recessive trait (Ephrussi et al., 1969; Harris et al., 1969; Klein et al., 1971). They observed that if malignant cells were fused with normal diploid cells, the resulting hybrids were non-tumorigenic (Figure 2). These hybrid cells could not grow in immunocompromised hosts. The suppression of malignancy was dependent on the retention of specific chromosomes. The chromosome loss resulted in the reoccurrence of tumorigenicity, suggesting the existence of tumor suppressor genes. The fusion of two malignant cells also gave rise to non-tumorigenic hybrids suggesting complementation between lesions. Detailed cytogenetic analyses of hybrids identified specific chromosomes responsible for the suppression of the malignant phenotype. For example, when HeLa cervical carcinoma was fused with normal fibroblast, the retention of chromosome 11 from the normal partner was necessary for the suppression (Stanbridge, 1985). When a single chromosome 11 was transferred from normal cells to the HeLa cervical carcinoma by microcell mediated chromosome transfer (MMCT) suppression of the tumorigenic phenotype occured (Saxon et al., 1986). Many studies have demonstrated that the transfer of even very small chromosome fragments can specifically suppress the tumorigenic properties of certain cancer cell lines.

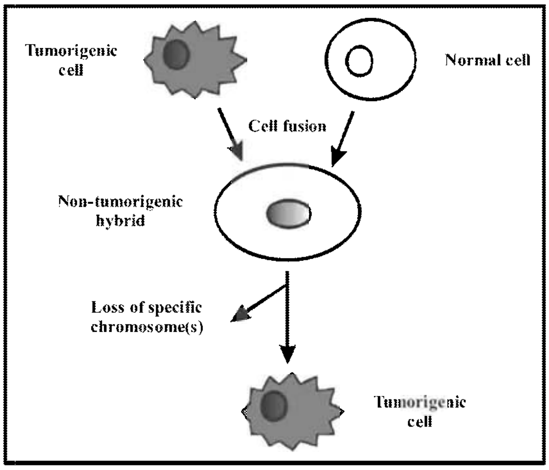


Figure 2. Suppression of malignancy by somatic cell fusion

The idea of the existence of tumor suppressor genes was further supported by Knudson's epidemiological studies of retinoblastoma. Retinoblastoma is a rare childhood tumor of the retina, which occurs sporadically in most cases, but in some families it displays autosomal dominant inheritance. He proposed that two 'hits' or mutagenic events were necessary for retinoblastoma development (Knudson, 1971). Schematic representation of the 'two-hit hypothesis' is shown on Figure 3.

In the familial form of retinoblastoma, the first hit was proposed to be present already in the germline and the second to occur somatically. The cancer usually appears at an early age, since only one mutation is required in the second allele ('second hit'). In the sporadic form of the disease both mutations occur in the somatic tissue. Since the probability of two mutations occurring in the same cell is low, therefore the disease is unilateral and characterized with late onset. Loss of heterozygosity (LOH) was found for polymorphic markers on 13q14 in retinoblastoma, these deletions were interpreted as one of the hit in the 'two hit model' (Cavenee et al., 1983). The gene responsible for retinoblastoma, *RB1*, was identified by positional cloning (Friend et al., 1986). The 'two-hit hypothesis' has been confirmed by molecular genetic analysis of retinoblastoma families and sporadic retinoblastoma tumors.

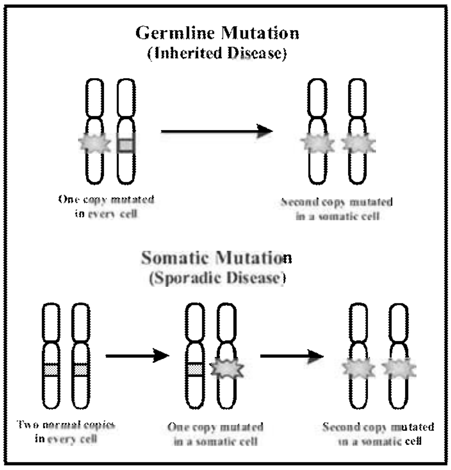

Germline Mutation
(Inherited Disease)
One copy mutated in every cell
Second copy mutated in a somatic cell
Somatic Mutation
(Sporadic Disease)
Two normal copies in every cell
One copy mutated in a somatic cell
Second copy mutated in a somatic cell

Figure 3. Two-hit hypothesis' for inactivation of tumor suppressor genes

Many other familiar cancer syndromes fit the Knudson's 'two hit model' and some of the responsible tumor suppressor genes have been identified (Table 2). Comparison of allele loss in hereditary cancers and sporadic cancers suggested that the fundamental mechanism for carcinogenesis may be the same in both cases. Inactivation of a tumor suppressor gene can happen by deletion, mutation or methylation. Any of these can occur in germline or in somatic cells. Deletions and mutations are essentially irreversible, but methylation changes can be reversible. There is a tendency for some tumor suppressor genes to become inactivated by a specific combination of these mechanisms.

Functionally tumor suppressor genes can be divided into 'gatekeepers' and 'caretakers' (Kinzler et al., 1997). Gatekeepers directly regulate tumorigenesis either by inhibiting cell growth or promoting apoptosis. It has been suggested that each of the cell type has one (or a few) gatekeeper(s), for example *NF2* for Schwann cells, *VHL* for kidney cells and *APC* for colon cells. Inactivation of gatekeepers causes familiar and sporadic cancers according to Knudson's 'two hit model'. Caretakers are responsible for maintaining the integrity of the genome; their inactivation does not promote tumor initiation directly. Inactivation of these genes exposes the cell to a very high mutagenic load that eventually may involve the activation of oncogenes and the inactivation of tumor suppressors. Known caretaker genes include the mismatch repair genes mutated in hereditary non-polyposis colorectal cancer (HNPCC), the *ATM* gene mutated in ataxia telangiectasia and *BRCA1* and *BRCA2* genes mutated in familiar breast and ovarian cancers.

1.2. The short arm of human chromosome 3

The short arm of chromosome 3 is frequently deleted in at least 23 tumor types, including renal cell carcinoma (RCC), small cell lung carcinoma (SCLC) and breast cancer. Earlier work at the Karolinska Institute on sporadic and hereditary RCC indicated the possible involvement of the 3p14 and 3p21 regions as candidate tumor suppressor regions (Boldog et al., 1991; Erlandsson et al., 1990; Erlandsson et al., 1988; Kovacs et al., 1988). The available human tumor LOH data were summarized and it was concluded that tumor suppressor genes can be located anywhere in 3p12-3p23 (Kok et al., 1997). It is very likely that 3p contains multiple tumor suppressor genes. Homozygous deletions are excellent indicators of the locations of tumor suppressor genes and several have been described at regions 3p12, 3p14.2 and 3p21.3. Recent reviews about the candidate tumor suppressor genes on 3p involved in the pathogenesis of breast, lung and other cancers also suggest multiple candidates (Imreh et al., 2003; Yang et al., 2002; Zabarovsky et al., 2002). The characteristics of the candidate tumor suppressor genes are summarized in Table 3. The locations of these genes on 3p are shown on Figure 4.

Table 2. Examples of tumor suppressor genes

Gene	Location	Familiar disorder	Cancer with somatic mutations	Presumed function of protein
APC	5q21	Adenomatous polyposis coli	Colorectal, desmoid tumors	Regulates β-catenin levels in the cytosol, binding to microtubules
BRCA1	17q21	Inherited breast, ovarian cancer	Ovarian, rare in breast cancer	DNA repair, complexes with Rad51 and BRCA2, transcriptional regulation
BRCA2	13q12	Inherited breast, pancreatic cancer	Rare in pancreatic cancer	DNA repair, complexes with Rad51 and BRCA1
DPC4	18q21	Familiar juvenile polyposis syndrome	Pancreatic (~50%), colorectal cancer (~15%)	Transcriptional factor in TGF- β signaling pathway
E-CAD	16q22.1	Familiar diffuse-type gastric cancer, lobular breast cancer	Gastric, lobular breast carcinoma	Cell-cell adhesion molecule
INK4a, p16	9q21	Familiar melanoma, familiar pancreatic carcinoma	Many different cancer types	Cyclin-dependent kinase inhibitor
INK4a, p19ARF	9q21	Familiar melanoma	Many different cancer types	Regulates Mdm-2 protein stability and hence p53 stability
MEN-1	11q13	Multiple endocrine neoplasia type 1	Parathyroid, pituitary adenoma, endocrine tumors	Not known
MSH2 ***MHL1*** ***PMS1*** ***PMS2*** ***MSH6***	2p21 3p22 2q32.2 7p22 2p16	Hereditary non-polyposis colorectal cancer	Colorectal, gastric, endometrial	DNA mismatch repair
NF1	17q11.2	Neurofibromatosis type 1	Melanoma, neuroblastoma	P21ras-GTPase
NF2	22q12	Neurofibromatosis type 2	Schwannoma, meningioma	membrane link to cytoskeleton
TP53	17p13	Li-Fraumeni syndrome	Approx. 50% of all cancers	Transcription factor, regulates cell cycle and apoptosis
RB1	13q14	Familiar retinoblastoma	Retinoblastoma, osteosarcoma, SCLC, breast, others	Transcriptional regulator, E2F binding
TSC1	9q34	Tuberous sclerosis	Not known	Not known
TSC2	16p13.3	Tuberous sclerosis	Not known	Putative GTPase activating protein for Rap1 and rab5
VHL	3p25.3	Von-Hippel Lindau syndrome	Renal, hemanglioblastoma	Regulator of protein stability
WT-1	11p13	WAGR, Denys-Drash syndrome	Wilms' tumor	Transcription factor

Microcell mediated chromosome transfer (MMCT) studies with the entire chromosome 3 and different fragments from chromosome 3 showed tumor suppression in renal cell carcinoma cell lines (Ohmura et al., 1995; Sanchez et al., 1994; Shimizu et al., 1990), in A549 lung adenocarcinoma cells (Satoh et al., 1993), in ovarian carcinoma cell line (Rimessi et al., 1994), in nasopharyngeal carcinoma line (Cheng et al., 1998) and in oral squamous cell carcinoma (Uzawa et al., 1998). Two Mb long fragment from 3p21.3 homozygously deleted region in SCLC and breast cancer suppressed tumorigenicity of mouse fibrosarcoma A9 in athmic nude mice (Killary et al., 1992). Later a 80 kb P1 clone located inside the 2 Mb region, including SEMA3F gene, showed similar suppression of tumor growth of the A9 (Todd et al., 1996).

Table 3. Candidate tumor suppressor genes located on 3p

Chr 3 region	Gene	Function	Comments
3p24.2	***RARβ*** Retinoic acid receptor β	Receptor for retinoic acid	*RARβ* is not mutated but inactivated by promoter methylation in tumors (Maruyama et al., 2002; Virmani et al., 2000). *RARβ2* has *in vitro* suppression activity in lung cancer cell lines (Toulouse et al., 2000). Transgenic mice expressing antisense *RARβ2* transcripts develop lung tumors (Berard et al., 1996).
	THRβ1 Thyroid hormone receptor β1	Transcriptional activator or silencer	Aberrant expression and/or mutations in *THRβ1* was associated with carcinogenesis (Gittoes et al., 1997; McCabe et al., 1999). Promoter hyper-methylation and a concurrent reduction of *THRβ1* transcripts in breast cancer cell lines was reported (Li et al., 2002).
3p24.1	***TGFβR2*** Transforming growth factor-beta type II receptor	Serine-threonine kinase receptor for TGFβ	The *TGFβR2* is a colon cancer suppressor gene that is inactivated by mutation in 90% of human colon cancers arising via the microsatellite instability (MSI) pathway of carcinogenesis. *TGFβR2* mutation is a late event in MSI adenomas and correlates tightly with progression of adenoma to carcinoma (Grady et al., 1998).
3p21.3T (AP20) HD: Lung, kidney, etc (Daigo et al., 1999)	***DLEC1*** Deleted in lung cancer 1		Mutational analysis of *DLEC1* gene by RT-PCR revealed the lack of functional transcripts and an increase of non-functional RNA transcripts in a significant proportion (33%) of esophageal, renal cell and NSCLC cancer cell lines and primary cancers. Introduction of the cDNA significantly suppressed the growth of four different cancer cell lines (Daigo et al., 1999).
3p21.3 HD: Mesothelioma	***CTNNB1*** Catenin beta 1-cadherin-associated protein	Regulation of cell adhesion, signal transducer in the wnt signaling pathway	The *CTNNB1* has been shown to be genetically mutated or otherwise dysregulated in various human malignancies (Garcia-Rostan et al., 2001; Schlosshauer et al., 2000). Reduced beta-catenin expression in surgically treated NSCLC is clearly associated with lymph node metastasis and an infavourable prognosis, suggesting a functional relation between E-cadherin and beta-catenin (Retera et al., 1998).
3p21.3 (C3CER1)	***LIMD1*** LIM domain containing 1		Tumor cell line panel Northern blot was hybridized with *LIMD1* and several truncated transcripts were detected in most of the tumor cell lines. By Marathon RACE experiment truncated transcripts in HL-60 cell lines were detected, in some cases all the 3 LIM domains were missing. In Marathon-ready liver cDNA library an alternatively spliced *LIMD1* transcript was found, which encodes only two LIM domains (Kiss et al, unpublished). The LIMD1 protein was suggested to be an interaction partner of the wild type pRB, as established by the yeast two-hybrid screening (Sharp et al, unpublished).

3p21.3 (C3CER1)	***LTF*** Lactotransferrin	Prevention of microbial infection, activates NK cells, regulates granulopoiesis, etc.	*LTF* was reported to suppress the growth of a fibrosarcoma cell line and v-ras-transformed NIH3T3 cells, and inhibited experimental metastasis of melanoma cells in mice (Bezault et al., 1994). An alternatively spliced form (delta LTF) of human *LTF* mRNA was expressed at various levels in adult and fetal human tissues, but not in any of 14 diverse tumor-derived cell lines (Siebert et al., 1997). Promoter methylation and/or rearrangement of the insertion site may be responsible for human *LTF* down regulation in mouse fibrosarcoma cells (Yang et al., 2003).
3p21C (LUCA) HD: Lung, breast, etc (Lerman et al., 2000)	***SEMA3F, SEMA3B***	Semaphorins are involved in nerve growth cone migration	Only few mutations of *SEMA3F* and *SEMA3B* were found in lung cancer cell lines. *SEMA3B* inactivated in lung cancer by allele loss and promoter region methylation. *SEMA3B* inhibited lung cancer cell growth and induced apoptosis after reexpression (Tomizawa et al., 2001). HEY cells expressing *SEMA3B* exhibited a diminished tumorigenicity in BALB/c nu/nu mice (Tse et al., 2002). Exogenously expressed *SEMA3F* suppressed mouse fibrosarcoma line A9 xenograft growth and blocked apoptosis induction in A9 cells (Xiang et al., 2002). *SEMA3F/SEMA3B* action in tumorigenesis may involve inhibition of angiogenesis through interference with VEGF function.
	HYAL1, HYAL2	Hyaluronidases	Hyaluronidases catabolize hyaluronic acid to oligosaccharides. Loss of expression of *HYAL1* in many cancer cell lines correlated with promoter methylation (Csoka et al., 2001). Loss of hyaluronidase might provide the cancer cell with the Hyaluronan-rich environment that stimulates growth, movement and metastatic spread.
	RASSF1 Ras association domain family 1	Regulation of cyclin D1	It has at least six different isoforms, *RASSF1A* and *RASSF1C* are the major transcripts. Several studies have shown that loss of *RASSF1A* expression in tumors occurs because of promoter methylation. Missense mutations of *RASSF1A* were also reported. *RASSF1A* can induce cell cycle arrest (Shivakumar et al., 2002). *RASSF1A* is involved in the development or progression of a majority of human tumors, it was suggested that it is an early 'gatekeeper' in lung cancer.
3p21.1-p21.2 (FER) HD: Breast	***DRR1*** Downregulated in renal cell carcinoma		Loss of expression of *DRR1* was found in RCC, cervical, NSCLC and some other cancer cell lines. Transfection of *DRR1* into *DRR1*-negative RCC cell lines resulted in growth retardation (Wang et al., 2000).
3p14 HD: Lung, renal, etc (Huebner et al., 1998)	***FHIT*** Fragile histidine triad	Diadenosine hydrolase	FRA3B and the breakpoint involved in the t(3;8) chromosome translocation in the familiar renal cell carcinomas map within the *FHIT* gene. *FHIT* is inactivated (by deletion, aberrant transcription, loss of expression, DNA methylation) in about 60% of human tumours, therefore *FHIT* is the most commonly altered gene in human cancer (Pekarsky et al., 2002). Fhit knockout mice are healthy, but have an increased susceptibility to spontaneous tumors and are very sensitive to carcinogenes (Fong et al., 2000).
3p12-p13 HD: Lung, breast (Sundaresan et al., 1998)	***DUTT1*** Deleted in-U-Twenty-Twenty	Neural-cell adhesion molecule	Mice homozygous for deletion of exon two frequently die at birth of respiratory failure, because of delayed lung maturation (Xian et al., 2001). Tumor specific promoter region methylation of *DUTT1* has been found in human cancers (Dallol et al., 2002).

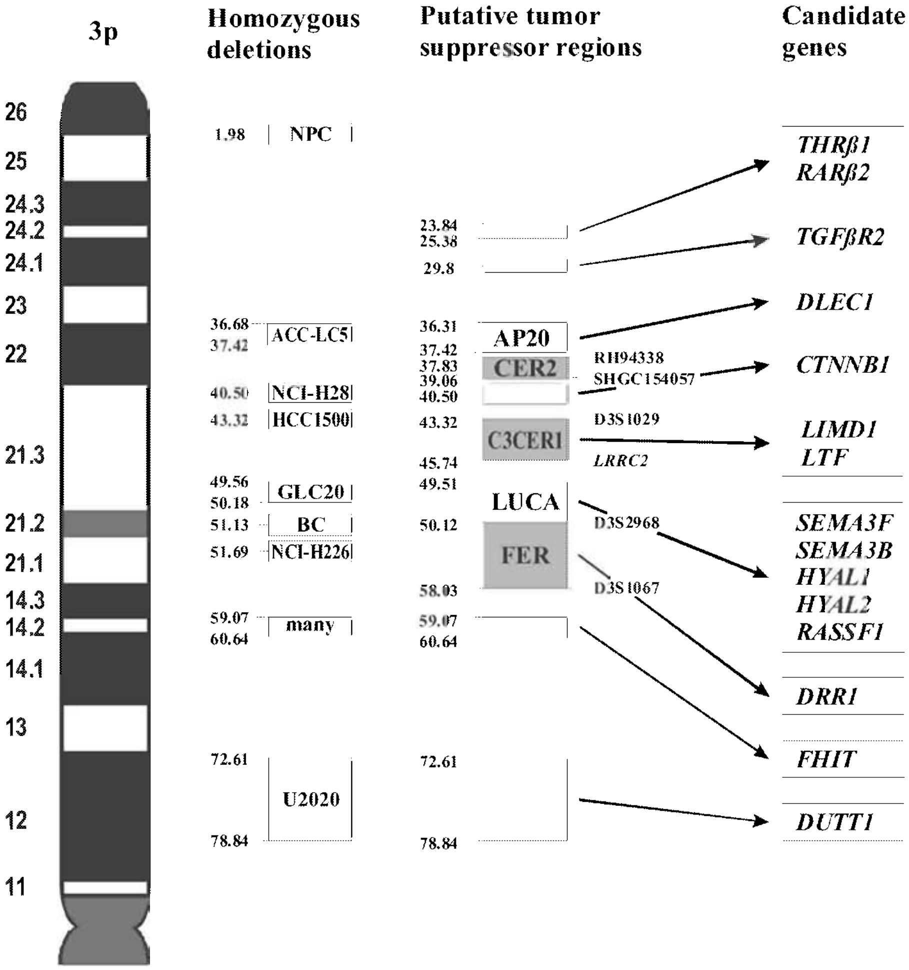


Figure 4. Locations of the homozygous deletions, putative tumor suppressor regions and candidate genes on 3p. Megabase positions are presented accordingly to the Human Genome Project Working Draft (Genome Browser, November 2002 at UCSC), (Imreh et al., 2003), modified

1.3. The elimination test (Et)

The elimination test (Et) was developed by our group with the aim to identify specific chromosomal regions that contain tumor growth antagonizing genes. The Et utilizes microcell hybrids (MCHs) that contain the normal chromosome 3 for inoculation into SCID mice. The obtained tumors are analyzed by cytogenetic and molecular methods in order to find regularly eliminated regions. The hypothesis behind the Et is that the regions, that are regularly lost during tumor development in SCID mice may contain tumor inhibitory genes and the elimination of these regions results in selective growth advantage. Figure 5 shows the experimental design of the Et.

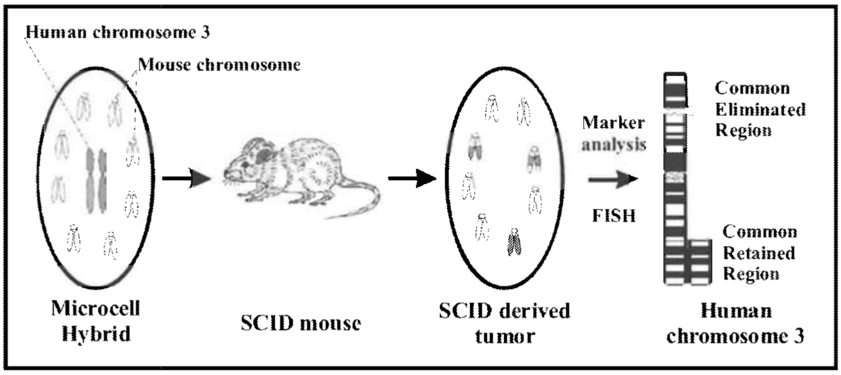


Figure 5. Experimental design of the elimination test

The microcell mediated chromosome transfer (MMCT) is a powerful method for functional testing of suppressor activity. MMCT permits the introduction of a single chromosome into a tumor cell, which results in the microcell hybrid (MCH). In collaboration with Eric Strabridge (Irvine, USA), Carl Barett (Research Triangle Park, USA), Andrew Cuthbert and Rob Newbold (Brunel University, UK) we generated MCHs that contained single normal cell derived human chromosome 3 on A9 mouse fibrosarcoma background. The transgenomic chromosome was retained in the MCH in vitro due to the presence of selective marker that provided Neomycine (G418) or Hygromycine resistance. The tumorigenicity of the MCHs was tested by subcutaneous inoculations into SCID mice. The arising tumors were explanted and expanded in vitro. Retention or elimination of certain regions was detected by cytogenetic (FISH, reverse painting) and molecular (PCR) analysis. The advantage of the Et was the limitless amount of SCID derived tumors and the usage of PCR markers instead of polymorphic markers.

The first experiments using the Et were published in 1994 (Imreh et al., 1994). Five MCHs were used for SCID inoculation; Two of them contained entire chromosome 3 (MCH903.1, MCH906.8), while three contained deleted chromosome 3 (MCH910.7: del(p25-p21), MCH939.2: del(p22-p14), MCH924.4: del(p24-p14)(q21-q26). The MCH901 and MCH904.11 were used as controls since they contained intact chromosome 1 and 13, respectively. FISH analysis of the SCID derived tumors from the intact chromosome 3 carrying MCHs showed chimeric translocations of chromosome 3. The deleted chromosome 3 containing MCHs and the control MCHs were unchanged after SCID passage. By PCR analysis the regular absence of four markers were found in all of the 20 examined tumors: AP20R (3p21.3), D3S1029 (3p21.3), D3S32 (3p21.3) and THRB (3p24). The border markers were GNAI2 (3p21.3) and VHL (3p25) on the centromeric and telomeric site, respectively. The common eliminated region (CER) was located on 3p24-21 with a size estimated to be 40 cM.

In an attempt to narrow down the size of CER, twenty-two new SCID tumors were studied that were derived from five chromosome 3 containing MCHs (Kholodnyuk et al., 1997). FISH, PCR and Southern blot analysis were used to analyze the derived tumors and it was suggested that the CER was 7cM at 3p21.3. The telomeric border marker were D3S1260 and the centromeric D3S643. The CER included 8 PCR markers: AP20R, D3S966, D3S3559, D3S1029, WI-7947, D3S2354, AFMb362wb9 and D3S32.

Four consecutive passages of the microcell hybrids showed a gradual elimination of the transgenomic chromosome and a non-random retainment of 3q26-q29 markers with a minimal region of overlap surrounding GLUT2 (3q26.2). This delimited a common retained region (CRR) (Imreh et al., 1997).

In our next experiment, to further reduce the CER, we used MCH910.6 and MCH906.8 intact chromosome 3 carrying A9 mouse fibrosarcoma cell lines (Szeles et al., 1997). By PCR analysis, we could identify CER1 (CER1 or C3CER1 as it was approved by the HUGO gene nomenclature committee for 'chromosome 3 common eliminated region 1') inside of the previous CER that contained two markers D3S32 and D3S2354. CER1 was bordered distally by D3S1029 and proximally by D3S643. According to the available database information, we concluded that CER1 is not larger than 1.6 Mb.

Detailed analyses of 30 SCID mouse tumors derived from MCH910.6 and MCH906.8 revealed the existence of a second commonly eliminated region (CER2) at 3p22 (Kholodnyuk et al., 2002). The size of CER2 is about 1 Mb, it is flanked distally by RH94338 and proximally by SHGC-154057. CER2 is located about 0.5 Mb centromeric to the known homozygous deletion region, identified in lung cancer. A third region was eliminated from the majority, but not all tumors at 3p21.1-p14.2 called as 'frequently eliminated region' (FER, originally called as eliminated region-2, ER2) (Kholodnyuk et al., 1997; Kholodnyuk et al., 2002; Yang et al., 2001). The location of the C3CER1, CER2 and FER is shown on Figure 4.

Chromosome 3 was transferred by microcell fusion into the human nonpapillary renal cell carcinoma line KH39 that contained uniparentally disomic chromosome 3. The four generated MCHs gave fewer and smaller tumors after longer latency periods in SCID mice, than KH39. The tumors were analyzed in comparison with corresponding MCHs by chr3 arm-specific painting, FISH probes, and polymorphic markers. We concluded that the human/human MCH-based elimination test identified similar eliminated and retained regions on chr3 as the human/murine MCH-based test (Yang et al., 2001).

1.4. Sequencing of the human genome

The human genome holds an extraordinary amount of information about human development, physiology, physiopathology and evolution. In 1990 the Human Genome Project (HGP) was initiated in the United States under the direction of the National Institutes for Health and the U.S. Department of Energy with a 15 year, $ 3 billion plan for decoding the human genome sequence. As a result of this international collaboration, a draft sequence of the human genome and its analysis were reported (Lander et al., 2001). The HGP used a 'clone-by-clone' approach: a set of large insert clones were generated and organized covering the genome and subsequently performed shotgun sequencing on appropriately chosen clones. A draft genome sequence was generated from a physical map covering more then 96% of the euchromatic part of the human genome. During their work the sequencing data was available without restriction and updated daily throughout the project. According to their estimation there are about 30,000-40,000 protein-coding genes in the human genome.

In 1998 Celera Genomics announced their goal to build a unique genome sequencing facility to determine the sequence of the human genome over a 3 year period. The sequencing was performed by a whole-genome random shotgun method. Two assembly strategies were used; a whole genome assembly and a regional chromosome assembly. Celera Genomics combined together their own and the publicly available sequencing data. At the beginning of 2001 Celera Genomics reported to have a 2.91-billion bp consensus sequence of the human genome (Venter et al., 2001). The Celera database is not freely available. Using the obtained sequences, the number of the existing genes was estimated to be 26,000-38,000. Both Celera Genomics' and HGP's estimations were far less than it had been predicted previously (50 000-140 000). The comparison of expressed sequence tags (ESTs) with the human genome sequence indicated that 47% of human genes might be alternatively spliced (Modrek et al., 2001).

The sequencing of the human genome has and will have a major impact on biomedical research, therapeutic and preventive health care (van Ommen, 2002). High throughput 'functional genomics' using DNA-chip and protein-chip approaches and specially designed animal model systems will open great prospects for pharmacological and genetic therapies. Genetically determined differences in drug metabolism could help to design more effective drugs with lower toxicity for individual patiens.

1.5. Functional and positional cloning

There are two main strategies for cloning disease genes: functional and positional cloning. Functional cloning uses the information about the function of a yet unknown disease gene to isolate the gene. However, most of the time no functional information is available about the gene of interest. Positional cloning involves mapping the chromosomal region containing a certain gene by linkage analysis in affected families and then searching the region for the gene itself. The candidate genes should be tested for mRNA and protein expression, mutation analysis and the causative role should be confirmed by functional/animal studies. The experimental design of the positional cloning is shown on Figure 6. Positional cloning is the most commonly used method for identification of disease-related genes, in spite of being extremely tedious before the availability of sequence databases. The human genomic sequence in public databases allows rapid *in silico* identification of candidate genes.

The Online Mendelian Inheritance in Man (OMIM^{TM}) is a continuously updated catalogue of human genes and genetic disorders. OMIM focuses primarily on inherited or heritable genetic diseases. It is also considered to be a phenotypic companion to the human genome project. OMIM contains approximately 5000 monogenic diseases. The majority of these genes are not determined on the molecular level yet.

An important requirement to perform a positional cloning project is the availability of a detailed physical map covering the chromosomal region of interest. Genomic fragments in the size range of million bp (Mbp) can be cloned in yeast artificial chromosomes (YACs), which mimic the normal yeast chromosomes (Burke et al., 1987). They contain yeast telomere and centromere sequences, as well as autonomously replicating sequences that allow their recognition by the yeast replication machinery. YACs are often unstable and can partially lose their cloned genomic fragments. Chimerism is also a very frequent problem. Because of these limitations, more reliable vector systems were developed: bacterial artificial chromosomes (BACs) and bacteriophage P1-derived artificial chromosomes (PACs) (Ioannou et al., 1994; Shizuya et al., 1992). These are circular cloning vectors. BACs can accommodate inserts up to 300 kb in size. YACs, BACs and PACs are low copy number vectors, therefore isolation of large quantities of DNA is more difficult compared with the high copy number cloning systems, for example plasmids and cosmids. Comparison of the most important properties of YACs, BACs, PACs and cosmids are summarized in Table 4.

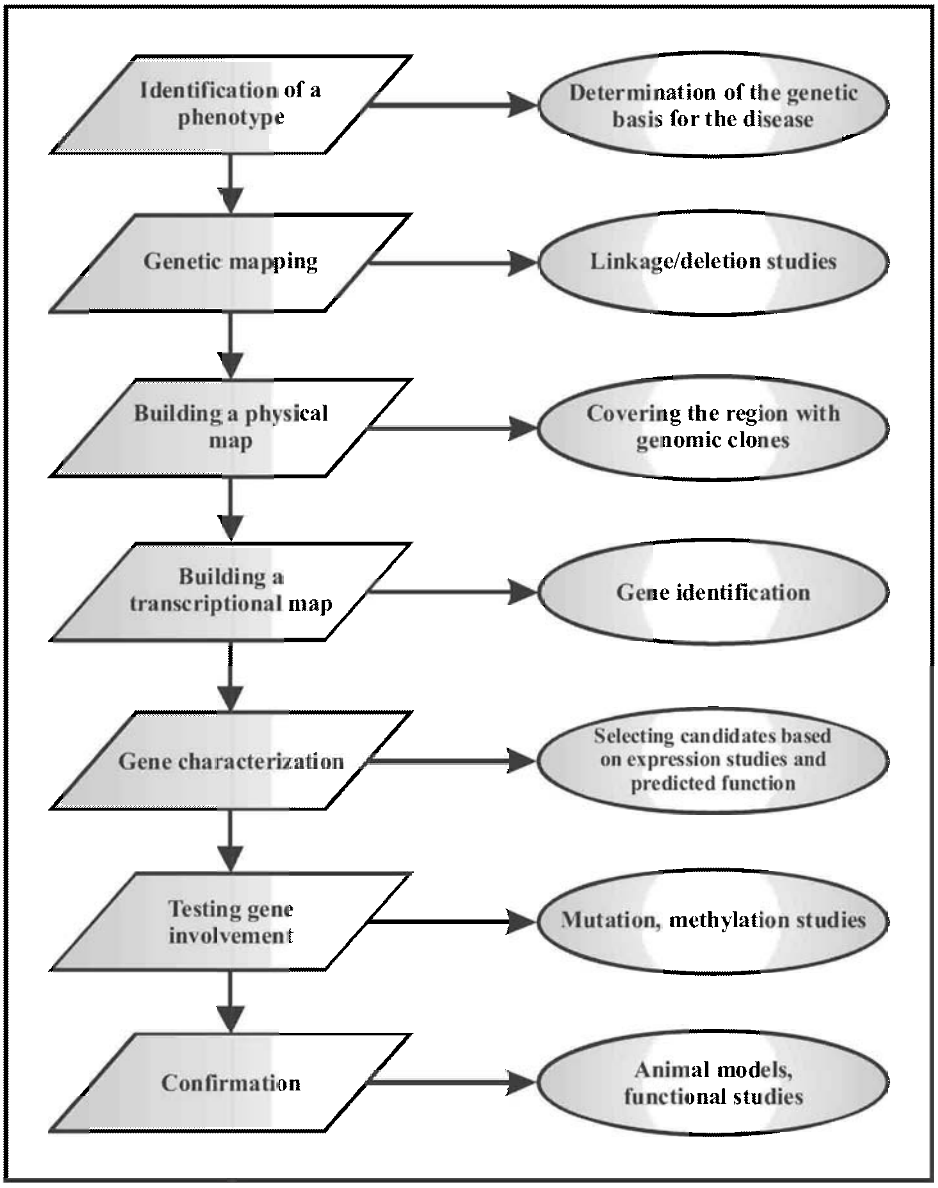

Identification of a phenotype
Determination of the genetic basis for the disease
Genetic mapping
Linkage/deletion studies
Building a physical map
Covering the region with genomic clones
Building a transcriptional map
Gene identification
Gene characterization
Selecting candidates based on expression studies and predicted function
Testing gene involvement
Mutation, methylation studies
Confirmation
Animal models, functional studies

Figure 6. Positional cloning

Table 4. Characteristics of different cloning systems

	Host cell	Based on	Insert size range (bp)	Stability	Copy number	Suitable for sequencing
YAC	Yeast	Yeast artificial chromosome, linear structure	200-2000	Frequent deletions, chimerism	Low	Difficult
BAC	Bacteria	*E. coli* fertility plasmid (F-factor)	100-300	Stable	Low	Yes
PAC	Bacteria	P1 bacteriophage based vector and T4 phage in vitro packaging system	80-150	Stable	Low	Yes
Cosmid	Bacteria	Insertion of cos sequences of lambda phage into a plasmid	30-46	Deletions	High	Yes

1.6. Comparative genomics

The genetic code is more or less universal in all organisms. Different organisms share a larger or smaller number of genes, depending how much time has passed since divergence occurred. The degree of sequence and structural similarity between orthologous genes generally follow this rule. Genes that have vital functions are strongly conserved during evolution. For example, similarity of some crucial metabolic enzymes is preserved among evolutionary distant species. The discovery of mismatch repair genes mutated in human hereditary non-polyposis colon cancer (HNPCC) was facilitated by functional comparison between very distant species, *E. coli* and *S. cerevisiae* (Fishel et al., 1993; Strand et al., 1993). Information can be obtained about the position of a gene, if the studied organisms are closely related. A group of genes located close on a single chromosome in one species is often linked in another closely related species. This phenomenon is called synteny. Depending on the evolutionary distance, synteny may be restricted to very small regions.

Comparative genomics is a large-scale, holistic approach that compares two or more genomes to discover similarities and differences between genomes. The practical applications of comparative genomics are numerous and their scientific impact profound. There are three important areas of comparative genomic analysis: genome structure, coding regions and non-coding regions. The analysis of the global structure of genomes (nucleotide composition, syntenic relationships, gene order) provides information on the organization and the evolution of genomes, and highlights the unique features of individual genomes. The comparative analysis of coding regions between different genomes involves the identification of gene coding regions, comparison of gene content. Non-coding regions of the genome gained a lot of attention recently, because of their predicted role in regulation of transcription, DNA replication, and other biological functions (Hardison, 2000; Meisler, 2001). Comparative genomics were used for the identification of regulatory segments by comparing the genomic non-coding DNA sequences from diverse species to identify conserved regions (Pennacchio et al., 2001). This approach is based on the presumption that selective pressure causes regulatory elements to evolve at a slower rate than that of non-regulatory sequences in the non-coding regions. Regulatory elements involved in the gene regulation of gene expression were successfully identified in many cases, using this approach (Wei et al., 2002). The specificity of regulatory region detection increases significantly when more than two species are used in the comparative analysis. It was found that only half of the human-mouse conserved non-coding sequence were also conserved in a third mammal (Frazer et al., 2001).

2. AIMS OF THIS THESIS

The general aim of my investigation was to study C3CER1 functionally and structurally. In particular the aims were:

1. To reduce further the size of C3CER1 by collecting additional PCR markers and analyzing more SCID derived tumors.
2. To assemble a high resolution full-coverage contig over C3CER1.
3. To sequence the PAC clones by shot gun approach.
4. To identify and characterize the C3CER1 gene content.
5. To compare the human C3CER1 to its orthologous region in mice.
6. To identify and characterize the mouse orthologs of the human C3CER1 genes.
7. To find relationship between conservation breakpoint regions (CBRs) and cancer-associated deletions.
8. To find common characteristics for CBRs.

3. MATERIALS AND METHODS

3.1. Generation of SCID tumors

Human monochromosome/mouse MCHs were generated by MMCT as described (Saxon et al., 1987). The A9 mouse fibrosarcoma line served as recipient. Neomycin-resistant clones of a normal human diploid fibroblast line (HFDC) randomly tagged with the pSV2neo marker were used as chromosome 3 donors. MCH910.6 and MCH906.8 that carried intact chromosome 3 were maintained in growth medium (Iscoves) supplemented with 10% fetal calf serum (FCS) containing 500 µg/ml Geneticin. Cells were injected subcutaneously into SCID mice for tumor formation (105 cells/mouse). SCID mice were observed palpatorily for tumor formation once a week for 20 weeks. Tumors were explanted and cultured in vitro to obtain the necessary cell number for cytogenetic and molecular analysis.

3.2. PCR marker analysis

Genomic DNA was isolated by proteinase K digestion and followed by phenol/chloroform extraction (Sambrook et al., 1989). YAC DNA was prepared according to the published protocol (Chumakov et al., 1992). DNA from PACs was prepared using alkaline lysis mini-preparation method or CsCl ultra-centrifugation (Sambrook et al., 1989). PCR markers were selected from the following databases: the Human Genome (Hudson et al., 1995) from the Whitehead Institute (www.genome.wi.mit.edu/cgi-bin/contig/phys-map), the genome database (GDB, gdbwww.gdb.org), and the CEPH-Genethon integrated map (www.cephb.fr). PCR was performed in the volume of 20 to 30 µl containing 50-100 ng DNA, 200 µM dNTP, 200 nM of each primer and 1.7 units Taq polymerase. The thermal conditions were 95 °C 5 min., 30-35 cycles with 95 °C for 30 sec., on annealing temperature 1 min., 72 °C for 1 min., followed by 72 °C for 7 min.

3.3. PAC library screening

Three types of probes were used for PAC library screening:

- PCR products derived from PCR amplification using DNA markers localized within C3CER1.
- End-fragments of PACs prepared by plasmid rescue procedure (also called delta-cloning). Three restriction endonucleases, that do not cut inside the pCYPAC2 (BamHI, XbaI and HeaI) or pPAC4 (BamHI) vectors, were chosen for cloning the insert ends. PAC DNA was digested to completion, diluted to a concentration of 100 ng/ml, ligated using T4 ligase (BRL) and transformed by electroporation into XL-Blue *E. coli* cells. DNA from plasmids containing PAC ends was isolated by alkaline lysis and ends were released by digesting with NotI+BamHI or NotI+XbaI or NotI+HeaI.
- Specific probes for PAC ends were designed on the basis of PAC DNA direct sequencing, using vector-specific primers.

Probe labeling, hybridization and washing of primary and secondary colony hybridization filters were performed according to standard methods (Feinberg et al., 1984; Sambrook et al., 1989). High-density filters with human PAC libraries were constructed at the Roswell Park Cancer Institute, Buffalo, USA (Ioannou et al., 1994) (Table 5). Anonymous male (RPCI-4, 5) or female (RPCI-6) blood DNA was isolated, partially digested with MboI and cloned between the BamH1 sites of the pCYPAC2 (RPCI-4, 5) or pPAC4 (RPCI-6) vector. The ligation products were transformed into DH10B electrocompetent cells (BRL Life technologies). The libraries have been arrayed into 384-well microtiter dishes and gridded onto 22x22cm nylon high density hybridization filters for screening purposes.

Table 5. The characteristics of the PAC libraries

Library	**Cloning vector**	**DNA source**	**Plate numbers**	**Total clones**	**Empty wells (%)**	**Avarage insert size (kbp)**	**Geno-mic cove-rage**
RPCI - 4	pCYPAC2	Human Male	529-816	105 251	4.8	116	4X
RPCI - 5	pCYPAC2	Human Male	817-1200	142 773	3.2	115	6X
RPCI - 6	pPAC4	Human Female	1-240	87 897	4.6	135	4X

3.4. Fluorescence in situ hybridization (FISH)

DNA from PAC and BAC clones was prepared by CsCl ultracentrifugation (Sambrook et al., 1989) or Qiagen columns (QIAGEN, Inc.). Expand Long Template PCR system (Roche Molecular Biochemicals, Mannheim, Germany) was used according to the recommendations of the supplier to PCR amplify long DNA sequences (>6kb) for FISH. The PACs, BACs, YACs and PCR amplified fragments were labeled using nick-translation with either biotin-dUTP (Bionick labeling system, BRL) or digoxigenin-dUTP (DIG-Nick Translation Mix, Boehringer Mannheim). One, two or three color FISH using labeled probes was performed on metaphase chromosomes and interphase nuclei as described (Fedorova et al., 1997). A fluorescence microscope (Leitz-DMRB, Leica, Heidelberg, Germany) equipped with a Hamamatsu C 4800 cooled CCD camera (Hamamatsu, Herrsching, Germany) and Adobe Photoshop 5.5 image analysis program (Adobe Systems, San Jose, Calif., USA) were used for the analysis of FISH results.

3.5. High resolution mapping by fiber-FISH

Fiber-FISH was performed as described (Fedorova et al., 1997) with minor modifications. Briefly, phytohemagglutinin-stimulated human lymphocytes were cultured for 72 h and harvested without colcemid treatment. Hypotonic treatment was performed in 0.075M KCl for 10 min at 37°C and the cells were fixed in methanol:acetic acid (3:1). To release the chromatin, fixed cells were spread on clean moist slides and before evaporation of the fixative slides were placed in PBS solution for 1 min. The slides were treated thereafter with 70% formamide in 2xSSC, pH7.0, 1 min, rinsed with methanol abs., fixed with methanol: acetic acid (3:1), air-dried, passed through 70%, 95% and 100% ethanol and air dried again. FISH was performed as previously reported (Fedorova et al., 1997). A fluorescent microscope (LEITZ-DMRB, Leica, Heidelberg, Germany) equipped with a Hamamatsu 4800 cooled CCD camera (Hamamatsu, Herrsching, Germany) was used. Image analysis was performed using the Image-Pro Plus (Media Cybernetics, Silver Spring, MD, USA) and Adobe Photoshop 4.0 (Adobe Systems Inc., San Jose, CA) programs.

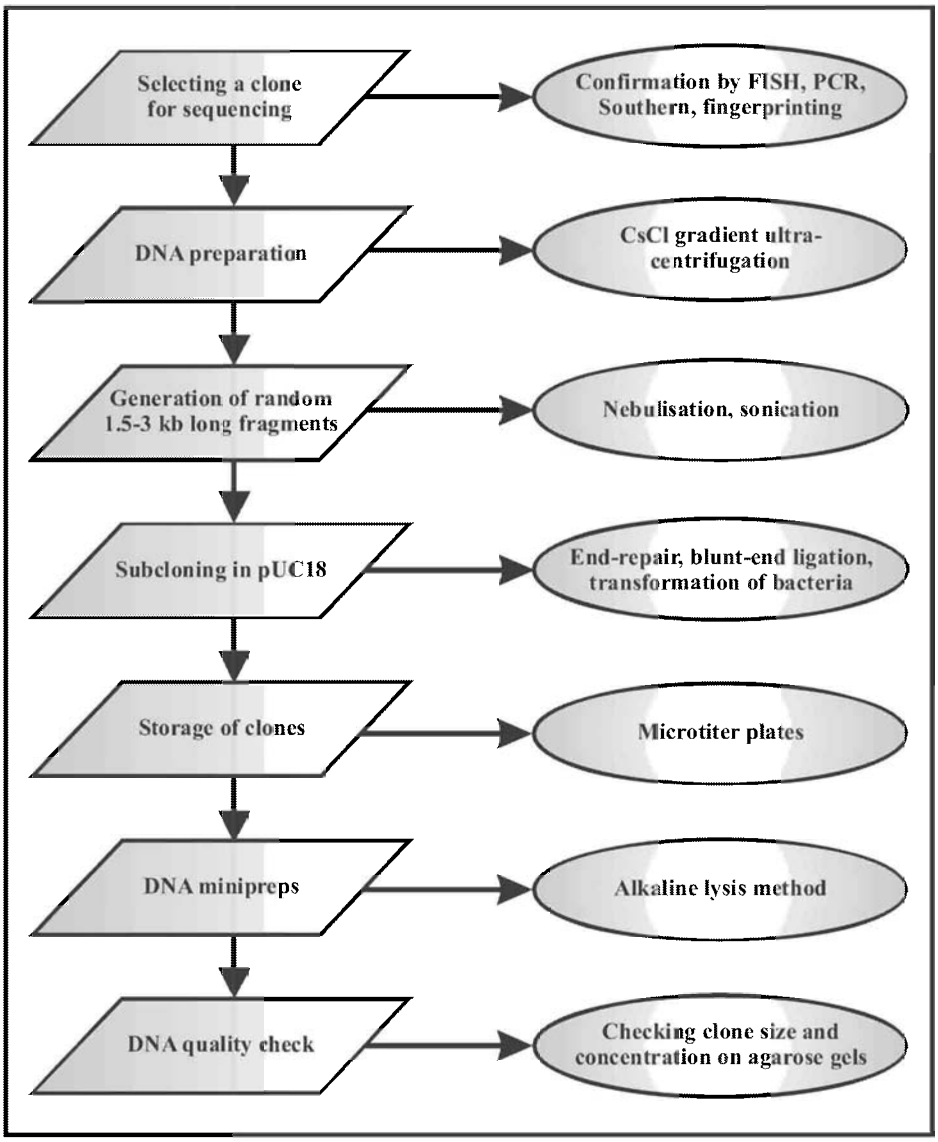
3.6. DNA sequencing

Two basic methods were used in the course of this work, depending on the size of the region to be sequenced. During the gene identification studies, we obtained short PCR fragments that were sequenced with the primers used for PCR amplification. If the fragment could not be sequenced directly, we used primer walking strategy to sequence the entire fragment. The primer walking strategy relies on designing a sequencing primer based on the new sequence to sequence the fragment further. This can be repeated until the whole region is sequenced. For sequencing of PAC clones, we used the random shotgun method. The principle of this method is to generate random subclones covering the whole region. The subclones are sequenced using vector specific primers. Finally all the sequencing data are assembled by the computer using specific programs. During the shotgun clone preparation it is very important to obtain a pure, bacterial DNA free PAC DNA. Therefore we used CsCl gradient ultracentrifugation. The template preparation for shotgun sequencing is shown on Figure 7.

Because of the shotgun sequencing technique used, the sequenced random clones covered 6-10 times the large genomic PACs. This redundancy helps to correct sequencing errors and assemble large contigs. When the number of contigs was reduced to 20 we continued to sequence by primer walking to span the gap between the contigs. Important factors determining sequence quality include: purity of the template DNA, complexity of the templates, base composition, quality of the sequencing primer, chemistry of the sequencing reaction, type of acryl-amide and speed of the electrophoresis.

Figure 8 shows the data flow starting from the ABI automatic sequencer to the final sequence assembly. The collecting of the data starts with the analysis of the sequencing gel files produced by ABI slab gel sequencer. These gel files are color images, containing a number of lanes (32 to 96) corresponding to individual sequence chromatograms. Recognition of the lanes (lane tracking) and subsequent extraction of the data from each lane is performed by Sequence Analysis software (Perkin Elmer). The lane tracking often needs to be corrected manually before the extraction of the data. The extracted ABI chromatograms contain sequence data and numerical values of color intensity for each base along the gel run. The ASP program uses these ABI chromatogram files as input data and processes them through a number of external programs. The ASP program is able to perform:

- Conversion of ABI trace files to SCF format and base calling using phred program.
- Quality clipping: The regions of poor sequence quality are identified and hidden at the start and at the end of each sequence. Sequencing files that contain only 20 bp of quality sequence data are classified as 'failed' and rejected from further analysis.
- Sequencing vector clipping: Sequence of the sequencing vector (pUC18) is identified and hidden. Sequencing files that composed completely of the sequencing vector, are classified as 'failed' and are rejected from further analysis.
- Cloning vector clipping: The sequence of the cloning vector (PAC) is identified using the vector clip program from the Staden package. Sequencing files, which composed completely of cloning vector, are classified as 'failed' and rejected from further analysis.
- Screening against contaminating sequences: Some degree of contamination (<5%) with bacterial DNA during the PAC DNA isolation is impossible to avoid. Contaminants are identified by comparing the vector-clipped reads against the *E. coli* genome sequence using BLAST. Files with positive matches are classified as 'failed' and rejected from the further analysis.



Selecting a clone for sequencing
Confirmation by FISH, PCR, Southern, fingerprinting
DNA preparation
CsCl gradient ultra-centrifugation
Generation of random 1.5-3 kb long fragments
Nebulisation, sonication
Subcloning in pUC18
End-repair, blunt-end ligation, transformation of bacteria
Storage of clones
Microtiter plates
DNA minipreps
Alkaline lysis method
DNA quality check
Checking clone size and concentration on agarose gels

Figure 7. The template preparation for shotgun sequencing

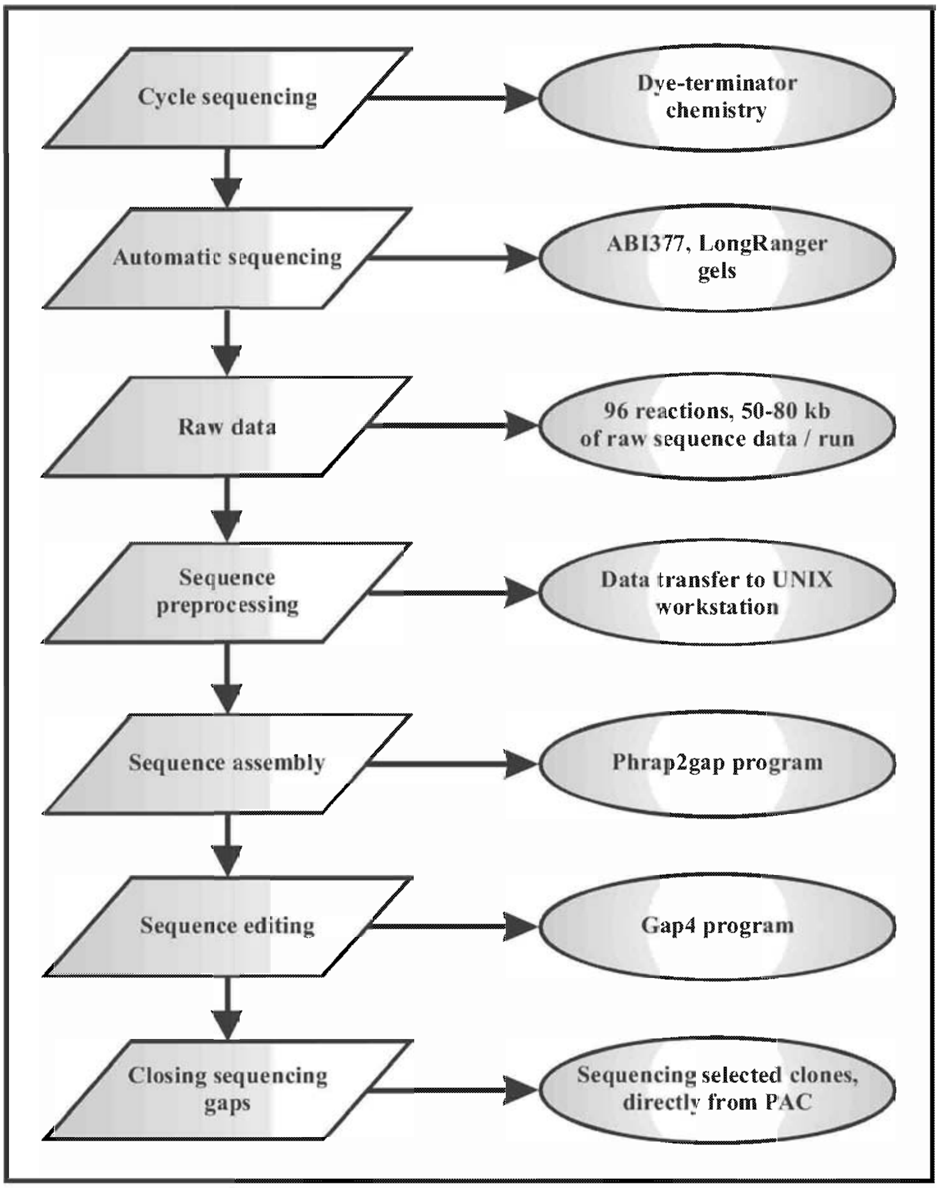


Figure 8. Sequence generation and processing

- Feature marking: Features like repetitive sequences (for example Alu, LINE repeats) are identified and tagged.
- Creation of the experimental (EXP) files. EXP and trace files are transferred to the corresponding project.

The assembly of the sequences is performed by phrap2gap. It assembles a set of raw reads (EXP files) into a GAP database using phrap program. During the database preparation the program clips and edits reads automatically, finally it creates annotated sequence assemblies (sequence contigs). The obtained GAP database can be viewed using gap4. The gap4 was our main sequence assembly and editing program (Bonfield et al., 1995). It contains an excellent contig editor and it contains a number of additional, useful features. These include contig joining, assembly checking, repeat searching, read-pair analysis, sequence comparisons, restriction analysis. It has also graphical views of contigs, templates, readings and traces.

All DNA sequencing reactions were performed using the Sanger method (Sanger et al., 1977). Fluorescent dideoxynucleotides were used in the PCR-sequencing reactions. Unincorporated dyes were removed by Sephadex columns. The cleaned sequencing products were separated on 0.2 mm thick denaturing polyacrylamide on ABI 377 automatic sequencers. Fluorochromes incorporated in the synthesized DNA fragments were excited by a laser beam and emit light of a certain wavelength corresponding to a product ending base. PCR-sequencing and electrophoresis were performed according to standard protocols recommended by Perkin Elmer.

3.7. Gene identification

Gene identification is a process in which the cDNA sequence of a gene is determined. Two types of gene identifications were performed during this work: finding genes within C3CER1 and finding orthologs in mice. Two basic approaches have been established for computational gene-finding: the sequence similarity search and the integrated compositional and signal search methods (Fickett, 1996). Figure 9 shows the design of our gene identification scheme.

3.7.1. Gene identification by similarity search

A DNA sequence and its translated aminoacid sequence in all six possible translational reading frames can be compared against all available DNA and protein sequences by homology search. The largest nucleotide sequence databases are the EMBL (http://www.ebi.ac.uk) and GenBank (http://www.ncbi.nlm.nih.gov) and protein sequence databases are PIR, TrEMBL and SWISS-PROT. The most popular algorithms that are used in sequence searching are the BLAST (basic local sequence alignment tool) programs (Altschul et al., 1990). The BLAST family of programs is listed in Table 6. Any significant matching between the test sequence and a sequence of a known gene, cDNA or protein, whether human or non-human origin, indicates a gene associated sequence. Identification of a new gene based on the use of information from the database of ESTs (dbest) is the most useful and efficient approach. EST sequences are generated at a very high rate and the dbest is the most rapidly growing section of the sequencing databases. A large number of tissues at various developmental stages have been used to construct cDNA libraries. Clones from these libraries are used for the EST sequencing project. ESTs from other organism (*M. musculus*, *R. norvegicus*, *D. melanogaster*, *C. elegans*, etc.) are also a valuable asset for gene identification. These EST data can be used for *in silico* cloning. EST sequences together with their trace files can be assembled in the gap4 gatabase. The obtained EST sequences can be compared to its genomic sequence to find the exon/intron structure of the putative gene.

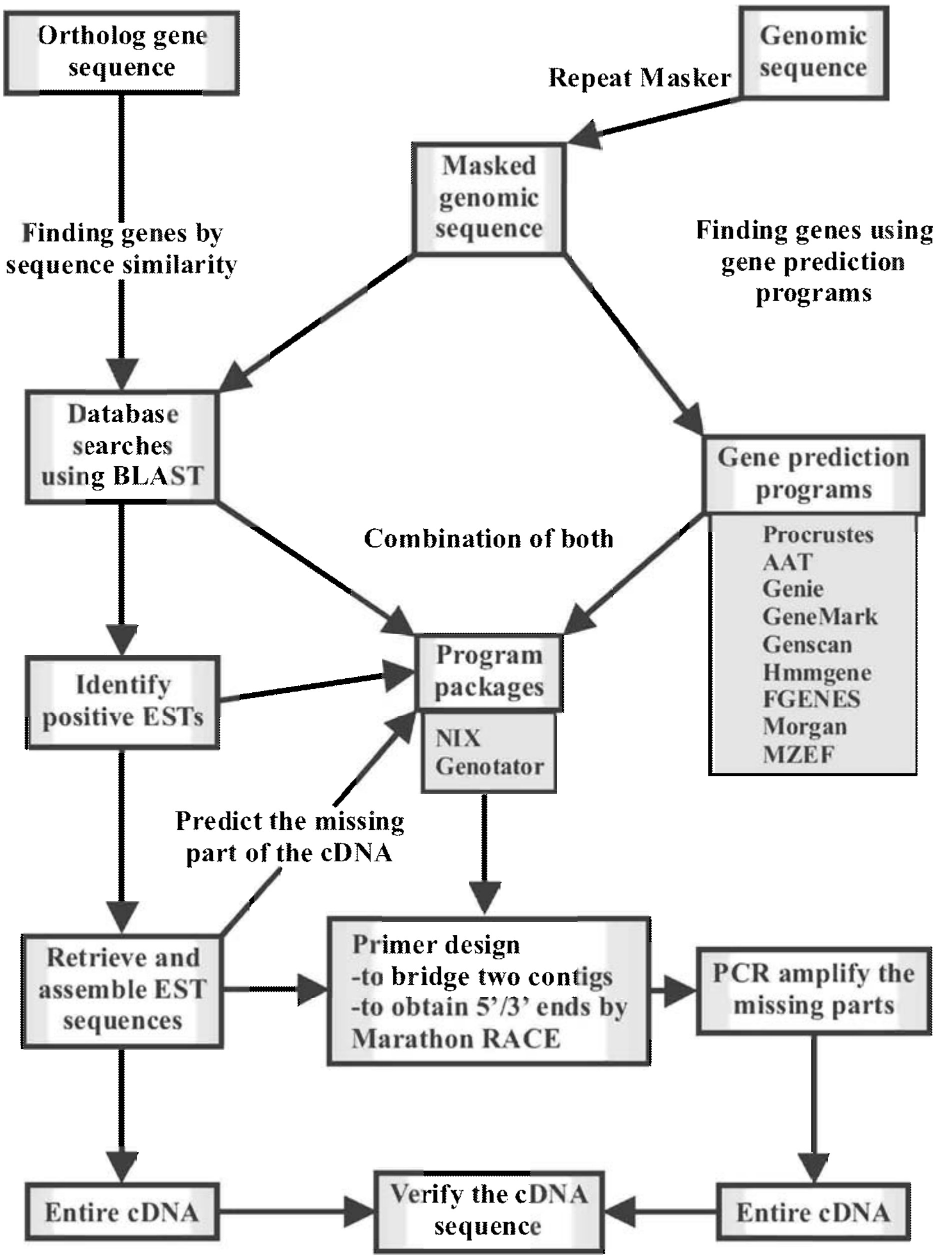


Figure 9. Experimental design of gene identification

Table 6. BLAST programs for sequence comparisons

Program	Comparison
BLASTN	Compares a nucleotide query sequence against a nucleotide sequence database.
BLASTP	Compares an amino acid query sequence against a protein sequence database.
BLASTX	Compares a nucleotide query sequence translated in all reading frames against a protein sequence database.
TBLASTN	Compares a protein query sequence against a nucleotide sequence database dynamically translated in all six reading frames
TBLASTX	Compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database

During my work, in several instances, the EST clones did not cover the entire gene. Gaps between EST contigs can be closed by resequencing the EST clones or by designing primers for the end of the contigs and perform PCR from cDNA libraries. Missing 3' and 5' end fragments of the gene can be obtained by Marathon RACE (rapid amplification of cDNA ends) technique (Frohman et al., 1988; Siebert et al., 1995) (Clontech) (Figure 10). Marathon-ready cDNA libraries are adaptor-ligated double stranded cDNAs ready for use as templates in Marathon RACE. The principle is to amplify sequences between a known sequence and an adapter sequence which is coupled to the 3' or 5' end of the gene. One primer should be designed from the known sequence (gene specific primer- GSP) of the gene and the other from the adaptor sequence (adaptor primer-AP). Nested primers can improve the efficiency of Marathon RACE if the gene expression is low. The PCR product of the Marathon RACE should be sequenced to determine the missing part of a gene.

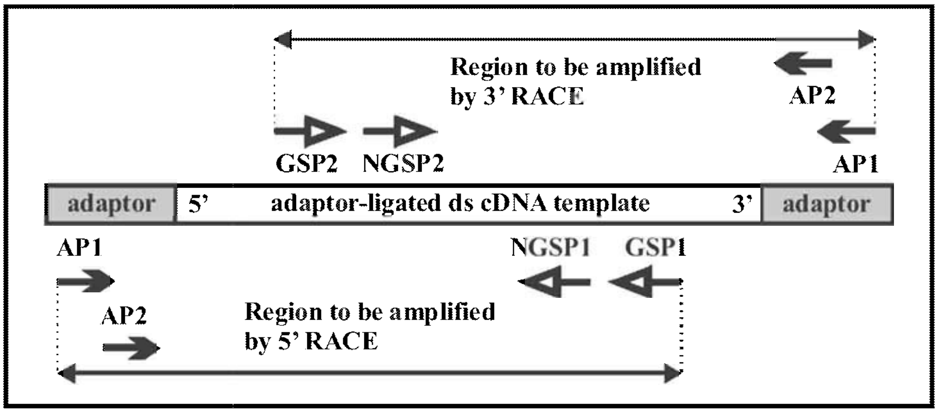


Figure 10. Marathon RACE. AP1-Adaptor primer, AP2-Nested adaptor primer, GSP-Gene specific primer, NGSP-Nested gene specific primer

The pitfalls of the EST-based cloning should also be mentioned: Artifacts like chimeric cDNA clones, cDNA clones containing intronic sequences and incorrectly spliced clones may seriously hinder the efforts to clone a gene. Sequencing errors (inaccurate base-calling, wrong tracking) can also be a problem. Therefore it is not enough to rely on *in silico* cloning; the sequence of the novel gene should be verified by RT-PCR, followed by sequencing.

3.7.2. Gene prediction programs

Gene prediction programs are becoming increasingly important as the sequence of the human genome is unveiled. The most frequently used programs are listed in Table 7. Computational approaches for the prediction of gene structures in the genomic DNA sequences integrate coding statistics with signal detection into one framework. Coding statistics behave differently on coding and non-coding regions and they are measures indicative of protein coding functions. A number of these measures have been evaluated (Fickett et al., 1992). Signal sensors are usually just several nucleotides-long subsequences, which are recognized by cell machinery and are initiators of certain processes. The signals that are usually modeled by gene-finding programs are promoter elements, start and stop codons, splice sites, and poly-A sites. Both codon statistics and signal models are 'learned' from a training set. There is also a group of programs that integrate a third component in their systems: similarity with an annotated sequence. Examples of such programs are Procrustes and AAT. Older programs were trained to identify just one gene in a sequence, rarely predicting any promoter elements. Recently developed programs are capable of identifying more complex genomic structures: any number of genes with either complete or partial structure. This is the case with Genie, GeneMark, Genscan and HMMgene. Seven recently developed gene-finding programs (FGENES, GeneMark, Genie, Genscan, HMMgene, Morgan, and MZEF) were evaluated (Rogic et al., 2001). Among them only Genscan and HMMgene have reliable scores for exon prediction according to the authors. The major problems associated with software-based de novo exon prediction are over-prediction and under-prediction. Over-prediction stands for predicting false-positive exons, while under-prediction is a failure of the program to recognize a real exon. Distinguishing pseudogenes from true genes is a major problem in gene prediction. Regulatory regions and poly-A sites usually remain unidentified, 5' and 3' untranslated regions are not specified, alternative splice variants are not considered, and overlapping or nested genes are not detected. Nevertheless, the prediction of the coding sequence of typical genes is an important first step in deciphering the content of any genome.

Table 7. Gene prediction programs

Program	Reference	URL
Procrustes	(Gelfand et al., 1996)	http://hto-13.usc.edu/software/procrustes/index.html
AAT	(Huang et al., 1997)	http://genome.cs.mtu.edu/aat/aat.html
Genie	(Kulp et al., 1996)	http://www.fruitfly.org/seq_tools/genie.html
GeneMark	Borodovsky M. and Lukashin A. (unpublished)	http://opal.biology.gatech.edu/GeneMark/eukhmm.cgi
Genscan	(Burge et al., 1997)	http://genes.mit.edu/GENSCAN.html
HMMgene	(Krogh, 1997)	http://www.cbs.dtu.dk/services/HMMgene/
FGENES	(Solovyev, 2002)	http://www.softberry.com/berry.phtml?topic=gfind
Morgan	(Salzberg et al., 1998)	http://www.tigr.org/~salzberg/morgan.html
MZEF	(Zhang, 1997)	http://argon.cshl.edu/genefinder/human.htm
NIX		http://www.hgmp.mrc.ac.uk/Registered/Webapp/nix/
Genotator	(Harris, 1997)	http://www.fruitfly.org/~nomi/genotator/

Recently a variety of software packages have been released that use general sequence homology-based database searching programs together with programs designed to identify gene-associated motifs and exons. These programs produce the output in a graphical format where a picture of a genomic sequence with database hits, putative exons predicted by several programs, regions containing repetitive sequences, ORFs, putative promoters, polyadenilation signals and other important features are combined. Two popular packages are NIX (nucleotide identification) from the UK Human Genome Mapping Project Resource Centre (Figure 11) and Genotator from the US Lawrence Berkeley National Laboratory.

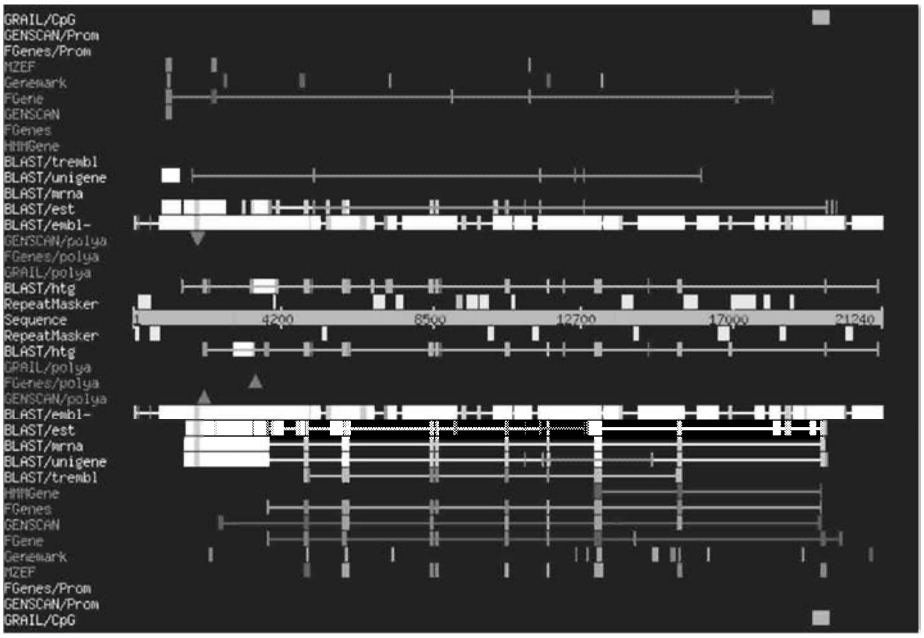


Figure 11. NIX output of the genomic region containing *LZTFL1* gene. The length of the region is illustrated by the green bar in the middle. Analysis include the use of programs to scan for gene-associated motifs such as promoter sequences, polyadenilation sites and various exon prediction programs. Significant homologies to other sequences at the nucleotide and protein levels are indicated by the boxes for the various BLAST programs

3.8. Gene characterization

3.8.1. Analysis at DNA level

A complete genomic sequence of the gene provides a basis for further examination of the gene structure and its regulatory elements. Direct comparison of the cDNA with the genomic sequence gives a picture about the exon/intron organization of the gene. The gap4 program can be used to identify the splice-donor and splice-acceptor sites and annotate the exon borders. The computerized detection of gene regulatory elements includes: promoter prediction, detection of potential CpG islands and finding transcriptional binding sites (Figure 11, 12). The programs, useful for these purposes are listed in Table 8. The function of the eukaryotic promoter is the initiation of the transcription. It has been shown that multiple functional sites in the primary DNA are involved in the polymerase binding process. These elements, such as TATA-box, GC-box, CAAT-box and the transcription start site are known to function as binding sites for transcription factors and other proteins, which are involved in the initiation process. These promoter elements are present in various combinations separated by various distances in the sequence. The location of a gene promoter can be predicted by Neural Network Promoter Prediction (NNPP), Promoter 2.0, PROMOTER SCAN II, CONsensus PROmoter predictor (CONPRO). Some gene prediction programs such as Genescan visualize potential promoters, as well. The positions of the CpG islands located within the promoter region or elsewhere can be detected by using CpG island detection programs, such as CpGPlot/CpGReport and CpG Island Promoter Detection (CpGProD). Transcription factor binding sites can be localized by AliBaba 2.1, PatSearch 1.1 and TFSEARCH.

Table 8. Programs for detection of gene regulatory elements

Program	Reference	URL
NNPP		http://www.fruitfly.org/seq_tools/promoter.html
Promoter 2.0	(Knudsen, 1999)	http://www.cbs.dtu.dk/services/promoter/
PROMOTER SCAN II	(Prestridge, 1995)	http://www.molbiol.ox.ac.uk/promoterscan.htm
CONPRO	(Liu et al., 2002)	http://stl.bioinformatics.med.umich.edu/conpro/
CpGPlot/CpGReport		http://www.ebi.ac.uk/emboss/cpgplot/
CpGProD	(Ponger et al., 2002)	http://pbil.univ-lyon1.fr/software/cpgprod_query.html
AliBaba 2.1	(Grabe, 2002)	http://www.alibaba2.com/
PatSearch 1.1	(Pesole et al., 2000)	http://transfac.gbf.de/cgi-bin/patSearch/patsearch.pl
TFSEARCH	(Heinemeyer et al., 1998)	http://molsun1.cbrc.aist.go.jp/research/db/TFSEARCH.html

3.8.2. Analysis at protein level

The first step in analysing new protein sequences is traditionally to search the protein databases for similar sequences. If the similarity is significant to a protein, information about this protein could be obtained depending on the quality of the annotation in the database and the availability of experimental results in the scientific literature. In many cases the similarity is restricted to a domain sequence. Depending how much is known about this domain, it is possible to have a clue about the function of a novel protein. There are a number of well known signature databases publicly available that can produce diagnostic signatures for protein families, domains, repeats, active sites and post-translational modifications (Figure 12). These include PROSITE, Pfam, SMART, BLOCKS and PRINTS (Table 9). There are also several databases that identify protein families or domains using sequence clustering and alignment methods; these include ProDom, DOMO and ClusSTr (Table 9).

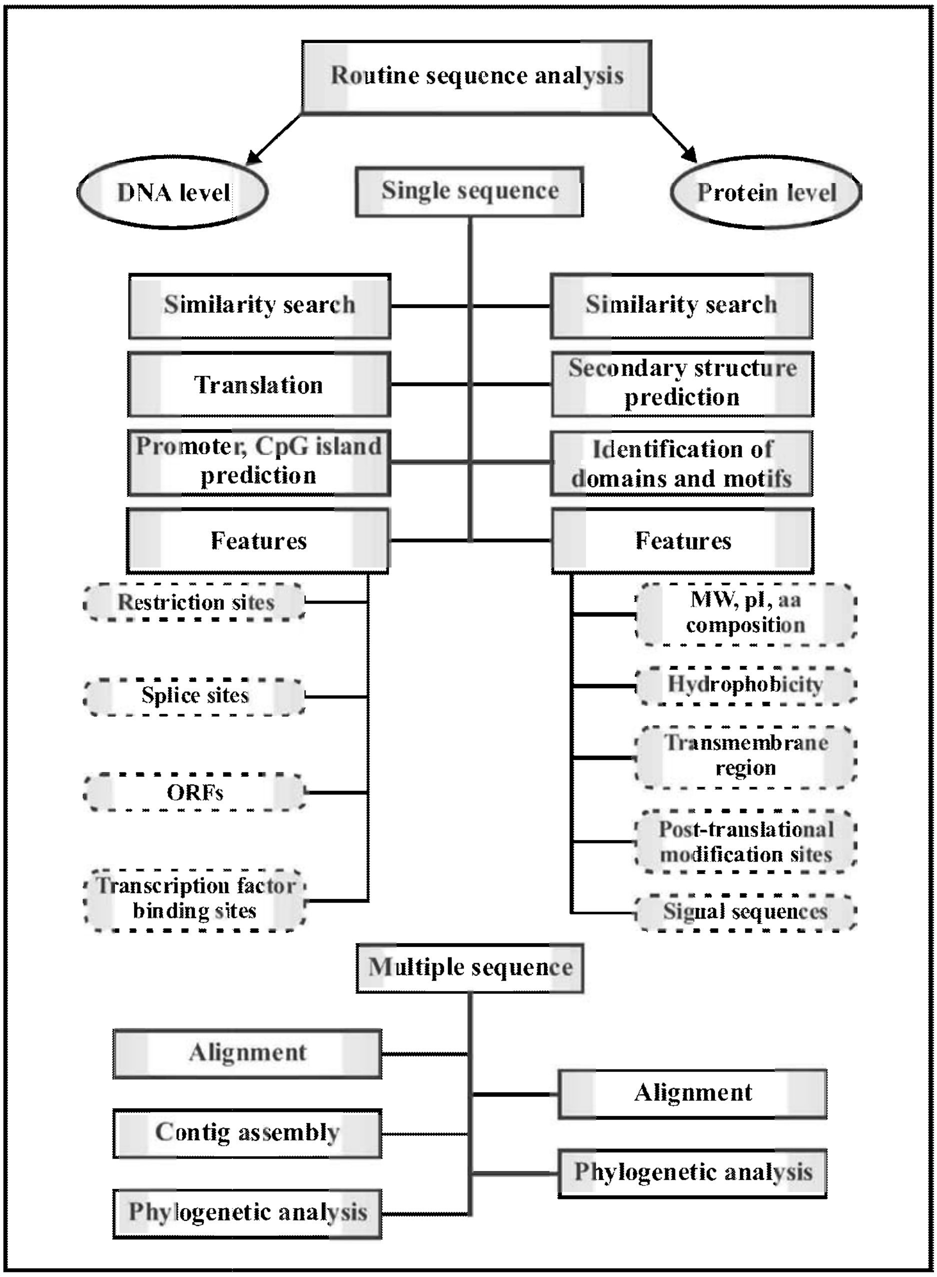


Figure 12. The scheme of gene characterization

PROSITE is a database for both patterns and profiles. Pfam emphasizes extracellular domains. There are two parts of the Pfam database: PfamA, a set of manually evaluated and annotated models and PfamB, which has a higher coverage but is fully automated. The Simple Modular Architecture Research Tool (SMART) facilitates the identification and annotation of genetically mobile domains and the analysis of domain architectures. The database is highly populated with models for domains found in signaling, extracellular and chromatin-associated proteins. The models rely on hand-curated multiple sequence alignments of representative family members. BLOCKS and PRINTS are two motif databases that represent protein or domain families by several short, ungapped multiple alignment fragments. The integrated resources for protein family and domain signature databases such as InterPro have several uses for the scientific community and for the member databases. The integration reduces duplication of effort of the member databases and facilitates the communication between the disparate resources. The PSORT program predicts the subcellular localization sites of proteins from their amino acid sequences. It uses many subprograms which calculate various scores. In the course of this work we also have used Coils and Multicoils servers for prediction of coiled-coil domains and SOSUI system and TopPred 2 programs for transmembrane domain prediction (Table 9). To perform amino acid sequence alignment we used Clustalw program. It calculates the best match for the selected sequences, and lines them up so that the identities, similarities and differences can be seen. The output of the Clustalw program was visualized by the Boxshade server. Phylogenetic tree was calculated with the TreeTop-Phylogenetic Tree Prediction program.

Table 9. Programs for protein analysis

Program	**Reference**	**URL**
PROSITE	(Falquet et al., 2002)	http://www.expasy.org/prosite/
Pfam	(Bateman et al., 2002)	http://www.sanger.ac.uk/Software/Pfam/
SMART	(Schultz et al., 2000)	http://SMART.embl-heidelberg.de
Blocks	(Henikoff et al., 2000)	http://blocks.fhcrc.org
Prints	(Attwood, 2002)	http://www.biochem.ucl.ac.uk/bsm/dbbrowser/PRINTS/
ProDom	(Corpet et al., 2000)	http://www.toulouse. inra.fr/prodomCG.html
DOMO	(Gracy et al., 1998a; Gracy et al., 1998b)	http://www.infobiogen.fr/services/domo/
ClusSTr	(Kriventseva et al., 2003)	http://www.ebi.ac.uk/clustr/)
InterPro	(Mulder et al., 2003)	http://www.ebi.ac.uk/interpro
PSOPT II	(Nakai et al., 1999)	http://psort.nibb.ac.jp/form2.html
Coils	(Lupas et al., 1991)	http://www.ch.embnet.org/software/COILS_form.html
Multicoils	(Wolf et al., 1997)	http://nightingale.lcs.mit.edu/cgi-bin/multicoil
SOSUI system	(Hirokawa et al., 1998)	http://sosui.proteome.bio.tuat.ac.jp/cgi-bin/sosui.cgi?/sosui_submit.html
TopPred 2	(von Heijne, 1992)	http://bioweb.pasteur.fr/seqanal/interfaces/toppred.html
Clustalw	(Higgins et al., 1996)	http://www.ebi.ac.uk/clustalw/
Boxshade		http://www.ch.embnet.org/software/BOX_form.html
TreeTop	(Brodskii et al., 1995)	http://www.genebee.msu.su/services/phtree_reduced.html

3.9. Comparative genomics

Sequence similarity search and gene prediction might fail to identify an active gene within the genomic sequence. An alternative method to handle this problem is to perform cross-species comparisons between genomic sequences in the syntenic chromosomal regions. Evolutionary comparisons between species could be informative in the detection of putative genes and other elements that are important in the control of gene regulation.

We used two programs for comparative genomics analysis of the human C3CER1 and its mouse orthologous region. PipMaker (http://bio.cse.psu.edu) is a World-Wide Web site for comparing two long DNA sequences to identify conserved segments and for producing informative, high-resolution displays of the resulting alignments (Schwartz et al., 2000). PipMaker returns the alignments generated by BlastZ in any or all of four different formats: a pip, a dot plot, a conventional textual alignment, and a compact listing of the coordinates of the aligning segments. In the pip output, the program plots the position of the first sequence and percent identity of each gap-free segment of the alignments (Figure 13A). The top horizontal axis automatically shows the positions of repeats, CpG islands and exons. Dot-plot displays the positions of alignments in both sequences as diagonal lines. PipMaker is appropriate for comparing genomic sequences from any two related species, although the types of information that can be inferred (e.g., protein-coding regions and cis-regulatory elements) depend on the level of conservation and the time and divergence rate since the separation of the species. Gene regulatory elements are often detectable as similar, non-coding sequences in species that diverged as much as 100-300 million years ago, such as *C. elegans* and *C. briggsae*, or *E. coli* and *Salmonella spp.* PipMaker supports analysis of unfinished or 'working draft' sequences by permitting one of the two sequences to be in unoriented and unordered contigs.

VISTA (http://www-gsd.lbl.gov/vista) is a program for visualizing global DNA sequence alignments of arbitrary length (Mayor et al., 2000). The VISTA plot is based on moving a user-specified window over the entire alignment and calculating the percent identity over the window of each base pair (Figure 13B). The x-axis and the y-axis represent the base sequence and the percent identity, respectively. If the user supplies an annotation file, genes and exons can be marked above the plot. The direction of genes is indicated by an arrow, while the coding exons and UTRs are marked with rectangles of different color. Conserved regions (defined with percentage and length of cutoffs) are highlighted under the curve.

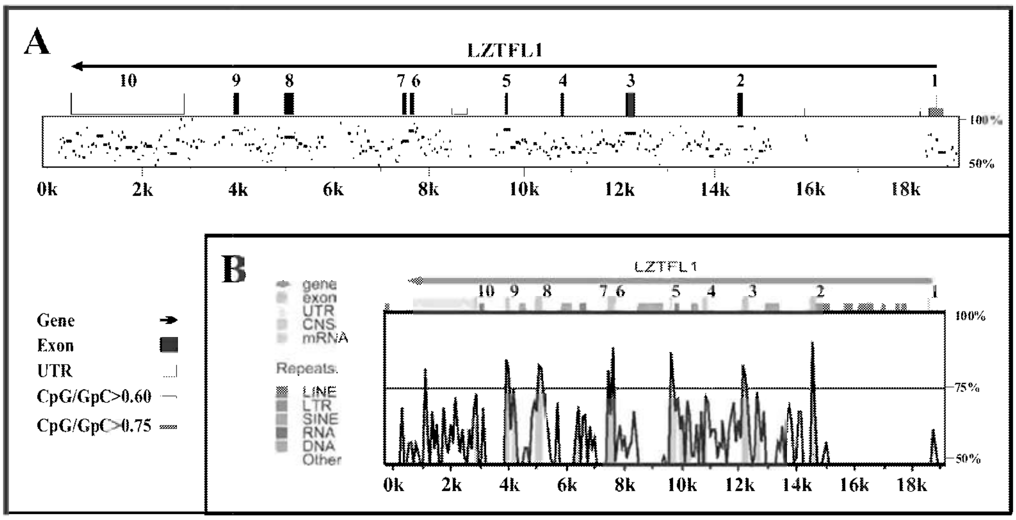


Figure 13. Comparative analysis of the *LZTFL1* gene. A, The pip output from the PipMaker. B, The VISTA plot. UTR-untranslated region, CNS-Conserved non-coding sequence

4. RESULTS AND DISCUSSION

4.1. Narrowing down C3CER1 and assembly of a PAC contig over C3CER1 (paper I)

Previously C3CER1 contained two markers (D3S2354 and D3S32) and the available linkage map indicated that it spanned 1.6 cM (Szeles et al., 1997). To construct a physical map, we tested additional markers from the Whitehead Institute (WI) and the GEPH-Geneton integrated maps. By constructing a tentative YAC-contig between D3S2354 and D3S32, we were able to select 10 additional markers from the WI-map. We localized three genes inside C3CER1 (*LTF*, *CCR1* and *CCR3*) and we used them as markers. Altogether 30 markers from 3p24-p21 were tested in 19 SCID derived tumors from MCH906.8 and MCH910.6. Among them, fourteen markers were deleted in all tumors. We established a new telomeric border of C3CER1 as D3S3582. The centromeric border remained D3S643.

The first two PAC clones (86:16E and 94k13, containing D3S2354 and D3S32 markers, respectively) were purchased from Genome Systems. We used markers from the C3CER1 region for PAC library screening. We screened RPCI-4, RPCI-5 and RPCI-6 libraries from Roswell Park Cancer Institute (Ioannou et al., 1994). To close the gaps between the selected PAC clones, we used additional probes for screening. We prepared probes from the end sequences of the existing PACs, which we obtained by direct sequencing of PAC ends and PAC-ends prepared by plasmid rescue. As a result of multiple rounds of screenings, we obtained 94 PAC clones. Among them, 47 were tested with markers and we concluded that twelve PAC clones formed a minimal tiling path and fully covered C3CER1.

The order of PACs was established by two color FISH on metaphase chromosomes, followed by statistical analysis. Fiber-FISH was performed with 11 PACs to verify the integrity of the contig and to measure its length. Various combinations of PAC clones were labeled differentially with two colors and hybridized pair-wise to stretched chromatin fibers. The length of the PAC signals, gaps and overlaps were measured in pixels on Adobe Photoshop images. PAC 86:16E was selected as the standard 'ruler' of the measurement. Therefore the size of the C3CER1 was estimated to be ~1 Mb.

BLAST searches using C3CER1 marker sequences revealed a fully sequenced BAC clone (BAC110P12, Acc. U95626, 143,068 bp). Lactotransferrin (*LTF*) and three chemokine receptor genes (*CCR2*, *CCR5* and *CCR1L2*) were identified on this BAC clone.

In this paper, we narrowed down C3CER1 by using additional PCR markers and analyzing more SCID derived tumors. We constructed a PAC contig over C3CER1 that was verified and measured by fiber-FISH. We localized several genes (*LTF*, *CCR1*, *CCR3*, *CCR2*, *CCR5* and *CCR1L2*) within C3CER1

.

4.2. Identification of *LIMD1* gene (paper II)

Previously we constructed a PAC contig over C3CER1. PAC86:16E was the first clone that we selected for sequencing. Analysis of sequence similarity by Blast programs using the obtained 87 665 bp of genomic sequence revealed the presence of two genes. The last 2 exons of *KIAA0028* gene located within PAC 86:16E. The human *KIAA0028* cDNA encodes the precursor of the mitochondrial leucyl-tRNA synthetase. The 5' end of a novel gene was also present on the PAC clone. In order to fully cover this new gene at the genomic level, we selected and partially sequenced PAC965c9, which overlaps with PAC86:16E on the centromeric side. We identified and assembled multiple human and mouse ESTs corresponding to this novel gene. In order to fill the gaps between the EST sequences, we designed PCR primers. The PCR fragments were amplified from placenta Marathon cDNA library and sequenced. The predicted protein sequence revealed the presence of three tandemly arranged LIM domains. We therefore named this novel gene as LIM domain-containing 1 gene (*LIMD1*).

The RPCI-21 mouse PAC library was screened by hybridization with a part of the mouse Limd1 cDNA, which resulted in 7 positive clones. The chromosomal localization of mouse Limd1 gene was determined by FISH using PAC415k6. The DAPI banding determined the localization of the *Limd1* gene to the mouse chromosome 9F telomeric region.

The LIM domain defines a unique double zinc finger structure found in a class of proteins involved in cell identity, differentiation, and growth control (Dawid et al., 1998; Sanchez-Garcia et al., 1993). The LIM domain is characterized by the cysteine-rich consensus sequence: CX2CX16-23HX2CX2CX2CX16-21CX2(C/D/H), (C=Cysteine, H=Histidine, D=Aspartic acid, and X=any aa). The LIM motif was initially identified in three developmentally important transcription factors, C. elegans Lin-11, rat Isl-1, and C. elegans mec-3, from which the acronym LIM is derived (Freyd et al., 1990; Karlsson et al., 1990; Way et al., 1988). LIM domains are highly conserved among proteins present in organisms representing a wide range of evolution. They are thought to function as versatile protein modules, capable of acting within diverse cellular contexts and in multiple subcellular compartments. Many have been shown to participate in direct protein-protein interactions, and they may also have the capacity to bind DNA directly (Beckerle, 1997; Gill, 1995; Schmeichel et al., 1997).

LIM domain-containing proteins have been classified according to the sequence similarities among the LIM domains and the overall structure of the protein (Dawid et al., 1998). Group 1 proteins contain LIM domains linked to a homeodomain and a potential transcription activation domain (e.g., Lin-11, Isl-1 and mec-3). They are nuclear transcription factors involved in cell fate determination and differentiation. Group 2 proteins are LIM-only (LMO) proteins consisting of one to five LIM domains without additional structural or functional motifs. These proteins can be nuclear (e.g., the erythrocyte protein rhombotin 2 or LMO2), cytosolic (e.g., muscle cell cysteine-rich protein), or both (e.g., myogenic LIM protein). Group 3 proteins contain three to four tandem LIM domains at the C terminus in association with distinct N-terminal domains. Members of this group include zyxin (Crawford et al., 1994), Enigma (Wu et al., 1994), paxillin (Turner et al., 1994), lipoma partner protein (*LPP*) (Petit et al., 1996), Trip6 (Lee et al., 1995), and the protozoal proteins AvL3-1 and OvL3-1 (Oberlander et al., 1995). Proteins not conforming to definitions of groups 1 to 3 constitute a fourth group.

The majority of group three LIM proteins are cytosolic. The LIM domains of these proteins have been shown to interact with cell surface proteins (e.g., Enigma) (Durick et al., 1996; Wu et al., 1994), cytoskeletal proteins at sites of cell adhesion (e.g., zyxin and paxillin) (Beckerle, 1986; Turner et al., 1990), or other LIM proteins (e.g., zyxin) (Sadler et al., 1992). In addition to their LIM domains, group 3 proteins contain extensive N-terminal non-LIM, or pre-LIM, domains that are relatively divergent in sequence. However, all of them are rich in proline residues, with some proline-rich stretches conforming to consensus SH3 recognition sites (Alexandropoulos et al., 1995). Indeed, some have been shown to interact with the SH3 domains of various cytosolic proteins in vitro (Hobert et al., 1996; Weng et al., 1993). However, the functional significance of these interactions *in vivo* has not been demonstrated yet. In addition, the pre-LIM domain of zyxin also mediates an interaction with actinin and members of the VASP protein family that are important for the assembly and maintenance of the actin cytoskeleton (Reinhard et al., 1995).

The human and mouse LIMD1/Limd1 proteins belong to the third group, which contain a variable number of LIM domains at the C-terminus (Dawid et al. 1998). As in the other proteins of this group, the non-LIM domain part of the LIMD1/Limd1 proteins has a high content of proline.

Upon blastp analysis without filtering of low complexity sequences, ZYXIN, ESP-2 and LPP proteins display similarity with LIMD1/Limd1, which is not only restricted to the LIM domains.

In summary, we have sequenced two PAC clones from C3CER1, which revealed the presence of the *KIAA0028* gene and a novel LIM domain containing gene. We have identified and characterized the human and the mouse *LIMD1* genes. The mouse *Limd1* gene was localized to the mouse chromosome 9F.

4.3. Identification of *LZTFL1* gene (paper III)

As a continuation of our large-scale sequencing project, we have sequenced the RP5-965C9 and the RP6-123I13 clones. We identified a fully sequenced BAC clone (RP11-165I16) in the public database that partially overlaps with the two newly sequenced PAC clones. We assembled all these sequencing data and obtained a 251 078 bp sequence contig. Blast searches with this sequence revealed four genes previously fully or partially characterized in humans or other species (*LIMD1*, *KIAA0851/SAC1*, *XT3*, *CCR9*). We detected also EST clones that belonged to a novel gene.

The ESTs from human and mouse were assembled separately and the obtained cDNA was verified by PCR amplification and sequencing from human testis, mouse kidney and mouse breast cDNA libraries. The name of this gene, Leucine Zipper Transcripion Factor Like-1 (*LZTFL1*), was based on the predicted presence of a leucine zipper pattern and coiled-coil domains within the LZTFL1 protein. The leucine zipper pattern is present in many proteins that regulates gene expression and it consists of a periodic repetition of leucine at every seventh position, covering the distance of eight helical turns. This segment forms an alpha-helix that can interact with another similar alpha-helix, facilitating protein dimerization. The structure formed by co-operation of these two regions forms a coiled-coil. The COILS program confirmed the existence of a coiled-coil domain. Furthermore, the PROSITE profile of bZIP transcription factor was detected. The bZIP superfamily of eukaryotic DNA-binding transcription factors group together proteins that contain a basic region mediating sequence-specific DNA-binding followed by a leucine zipper required for dimerization (Hurst, 1995). Two human and mouse *LZTFL1* transcripts were detected on Northern blots. Assembling of ESTs and confirmatory cDNA sequencing revealed that each of these two transcript forms contain separate polyA-tails. This is likely due to the alternative usage of polyadenylation signals. The localization of mouse *Lztfl1* gene to the chromosome 9F telomeric region was determined by FISH using a PCR amplified genomic fragment (>6 kb) from the mouse gene.

The *KIAA0851/SAC1* cDNA sequence was identified from size-fractionated human brain cDNA library in the course of a large-scale identification of human transcripts (Nagase et al., 1998). The KIAA0851/SAC1 protein shows 34% identity and 52% similarity to the recessive suppressor of secretory defect gene in yeast (*RSD1* also called as *SAC1*, suppressor of actin). Using the sequence of the human *KIAA0851/SAC1* gene as bait we found several mouse EST clones, which were assembled into contigs. The missing part of the cDNA was obtained by PCR. Both the human and the mouse proteins contain two transmembrane regions and a leucine zipper pattern. Expression of the mouse *Sac1* gene was studied. The mouse *Sac1* gene was localized to the mouse chromosome 9F telomeric region, in the close vicinity of the *Limd1* gene.

The *S. cerevisiae* gene SAC1 was isolated having ability to suppress the effects of ACT1 and sec14p mutations (Cleves et al., 1989; Novick et al., 1989). Sac1p is an integral membrane polyphosphoinositide phosphatase. Phosphorylated derivates of phosphatidylinositols (PtdIns) play an important role in lipid-based signal transduction. These molecules participate in the regulation of various cellular functions including membrane trafficking, cytoskeletal organization, cell proliferation and metabolism (Simonsen et al., 2001). Phosphoinositide phosphatases are consequently important mediators of Phosphoinositide signaling events. Phosphatases belonging to a subclass share a specific domain, which was first identified in the yeast Sac1 protein (Hughes et al., 2000a). *In vitro* studies have shown that the Sac1 homology domain is capable of dephosphorylating PtdIns(3)P, PtdIns(4)P and phosphatidyl-inositol 3,5-biphosphate (Hughes et al., 2000b; Stolz et al., 1998). The yeast *SAC1* gene product is an integral membrane protein of the endoplasmic reticulum (ER) and the Golgi complex. Genetic and biochemical analysis defined that Sac1p is an important regulator of ATP uptake into the ER lumen (Kochendorfer et al., 1999; Mayinger et al., 1995). Sac1p regulates a pool of PtdIns(4)P in the Golgi that is important to forward trafficking to the cell periphery (Schorr et al., 2001). The rat ortholog of the yeast SAC1 gene was identified and functionally characterized (Nemoto et al., 2000). The rat Sac1 localizes predominantly to the ER. Rat Sac1 exhibits intrinsic phosphoinositide phosphatase activity towards the same substrates, like the yeast Sac1p. Both the yeast and the rat proteins contain a C-terminal transmembrane domain. Sac1 ortholog proteins play evolutionary conserved role in eukaryotic cell physiology.

Blast searches with human genomic sequence identified a partial cDNA of the *XT3* gene. The *XT3* gene belongs to the plasma membrane neurotransmitter transporter superfamily. This family has a common structure of 12 transmembrane helices. Comparison of the human genomic sequence with the mouse and rat cDNA sequences indicated that at least half of the human *XT3* cDNA sequence was missing. We assembled all available ESTs and verified the human cDNA sequence by sequencing overlapping PCR fragments that were amplified from human brain and pancreas cDNA libraries. During the cDNA characterization, we discovered a differently spliced form of the gene in brain, designated as XT3a. This mRNA isoform is predicted to miss one transmembrane domain from the XT3 protein.

The *CCR9* gene was first identified and named as *GPR-9-6* as a G-protein-coupled receptor gene. Recently, this gene was renamed to CC-chemokine receptor 9 (*CCR9*), based on the fact that it is a receptor for a TECK (thymus-expressed chemokine) (Zaballos et al., 1999).

In conclusion, we localized additionally *KIAA0851/SAC1*, *XT3*, *CCR9* and a novel gene (*LZTFL1*) within C3CER1. We identified and characterized the human and mouse *LZTFL1* genes, the mouse *Sac1* gene and further characterized the human *XT3* gene. The mouse *Lztfl1* and *Sac1* genes were localized to the mouse chromosome 9F telomeric region.

4.4. Transcriptional map of the C3CER1 (paper IV)

As the continuation of this project, we have sequenced and analyzed 8 additional PAC clones from C3CER1. During this work and as a result of the Human Genome Project, the sequences of additional fully or partially sequenced genomic clones were released in the public databases. We assembled all the sequencing data for C3CER1 and reported a physical map of 21 BAC/PAC clones as well as a comprehensive transcriptional map of 1.4 Mb. We detected 18 active genes and three pseudogenes within C3CER1. The characteristics of the active genes are summarized in Table 10.

4.4.1. Identification of *FYCO1*, *TMEM7*, *LRRC2*, *LUZP3* genes

Blastn search against dbest database revealed similarities with multiple ESTs and with mouse *Mem2* partial cDNA sequence. The *Mem2* cDNA clone had been identified by differential display analysis of cDNA libraries prepared from unfertilized eggs and preimplantation embryos and named as Maternal Embryonic Message 2 (Heyer et al., 1997). The human ESTs were assembled and compared with the genomic sequences. The cDNA sequence of the novel gene contained several gaps; therefore we used the GENSCAN prediction program to detect the missing exons. Several primer pairs were designed to verify the already available cDNA sequences and fill in the missing parts. The sequence of the 5'end of the gene was verified by Marathon RACE. Northern hybridization with the human cDNA probe revealed an 8.5 kb transcript, which is expressed mainly in heart and skeletal muscle. The presence of FYVE zinc finger domain and a coiled-coil domain was predicted, therefore it was named to FYVE and Coiled-coil domain containing 1 (*FYCO1*).

FYVE zinc finger is a cystein-rich domain, which can bind two Zn2+ ions. The FYVE finger can bind with high specificity to the membrane lipid phosphatidyl-inositol-3-phosphate (PtdIns(3)P) (Gaullier et al., 1998). PtdIns(3)P have been shown to play a role in signal transduction, membrane trafficking, cytoskeletal regulation and apoptosis (Leevers et al., 1999; Rameh et al., 1999). Mammalian cells express more than 25 different FYVE-domain containing proteins. Little is known about the functions of the majority of these proteins, they comprise a group with a wide range of different structures and functions. Some of them (EEA1, Hrs and Vac1p) are involved in membrane trafficking, others (FGD1-3 and Frabin) regulate the cytoskeleton. SARA is involved in signal transduction and it is responsible for the recruitment of Smad2 and Smad3 to the TGFβ receptor upon receptor stimulation (Tsukazaki et al., 1998). It was suggested that FYVE finger proteins typically regulate 'housekeeping' cellular functions rather than agonist-induced processes (Stenmark et al., 2002).

We have identified the *TMEM7* gene by Marathon RACE using primary and nested gene specific primers for the detected EST cluster, both in 5' and 3' direction. Northern hybridization with the *TMEM7* cDNA revealed a transcript exclusively in liver. The TMEM7 protein contains a single transmembrane domain located near the C-terminus. We also noticed a KKXX-like motif (VKTA) at the C-terminus predicted to function as endoplasmic reticulum (ER) membrane retention signal.

During the cloning of the *LRRC2* gene, we assembled the identified ESTs into 5 contigs. Using Marathon-RACE PCR, we were able to identify the entire gene. The *LRRC2* transcript was detected only in heart, skeletal muscle and kidney. The LRRC2 protein contains 7 Leucine-rich repeats (LRRs) which are relatively short motifs (22-28 aa) found in a variety of cytoplasmic, membrane and extracellular proteins 40. LLRs proteins are associated with widely different functions, with a common characteristic involving protein-protein interaction. The closest relative of LRRC2 is RSP-1 (Ras Suppressor Protein 1) that plays a role in the ras signal transduction pathway. RSP-1 is capable of suppressing v-ras transformation *in vitro* (Cutler et al., 1992). The PSORT II program detected two putative nuclear localization signals.

Cloning of the *LUZP3* was initiated by a match between the genomic sequence and a cluster of 6 ESTs. The RACE PCR allowed us to obtain the entire gene. The only domain/motif that could be detected in the protein sequence is a leucine zipper pattern. It consists of one exon that is located on the opposite strand and within the first intron of *LRRC2*. By RT-PCR we detected PCR products in IB-4 cell line, kidney, trachea and skeletal muscle. *LUZP3* was later renamed to leucine zipper protein pseudogene 1 (*LUZPP1*).

We identified 3 processed pseudogenes (NRBF-2Ψ, UQCRC2Ψ and FLT1Ψ) within 290 kb of the centromeric part of C3CER1. The centromeric part of C3CER1 is very dense in active genes, which has prompted us to redefine its centromeric border. We, therefore, tested the MHC906.8 microcell hybrid-derived panel of SCID tumors that was used in our previous study (Yang et al., 1999). We have used 6 new STSes that are located in RP6-91P17 clone. We concluded that the centromeric border of C3CER1 is positioned within the *LRRC2* gene.

4.4.2. Chemokine receptor cluster in C3CER1

We detected the presence of a large cluster of chemokine receptors (CCRs) in C3CER1 that include 8 genes; *CCR9*, *STRL33* (also named *TYMSTR* or Bonzo), *CCXCR1*, *CCR1*, *CCR3*, *CCR2*, *CCR5* and a chemokine receptors like *CCRL2* (also named *CRAM-B*). CCRs belong to the superfamily of G-protein-coupled receptors that possess seven transmembrane domains. Families of chemokine genes (over 45 members) and chemokine receptor genes (19 members) occur in clusters on chromosome 4 and 17 and chromosome 2 and 3, respectively. Interestingly, two CCRs, *CCR8* and *CX3CR1*, are located within CER2.

Chemokines and their receptors mediate signals that are critical for the leukocyte recruitment and activation in processes that require active cell migration, such as inflammatory responses, bacterial or viral infections and wound healing. Other functions of chemokines have been described; they are mediators of cell growth and differentiation, cellular trafficking, blood vessel growth. Some CXC chemokines have the ability to regulate angiogenesis directly (Gupta et al., 1998), through interaction with their receptors on endothelial cells, or indirectly by attracting inflammatory cells which release angiogenic factors such as basic fibroblast growth factor and vascular endothelial growth factor (Goede et al., 1999; Polverini, 1997). CXC chemokines can be subdivided based on the presence of the ELR (Glu-Leu-Arg) motif. ELR+ chemokines are generally potent neutrophil chemo-attractants with pro-angiogenic properties, whereas ELR-chemokines are generally monocyte and T-cell chemo-attractants with potent angiostatic properties (Baggiolini et al., 1997; Belperio et al., 2000).

The role of chemokines in human disease has been suggested, although their function in cancer is not entirely clear. Some chemokines are expressed by tumor cells or the surrounding stroma and appear to exert a growth potentiating effect on cancers. The ability to induce cell migration and angiogenesis has suggested the possibility that chemokines may aid in the metastatic spread of tumors. Indeed, it has been shown that chemokines and their receptors have a critical role in determining the metastatic destination of tumor cells. Signaling through *CXCR4* or *CCR7*, chemokines mediate actin polymerization and pseudopodia formation in breast cancer cells, and induce chemotactic and invasive responses. Furthermore, organs representing the main sites of breast cancer metastasis are the most abundant sources of ligands for these tumor-associated receptors (Muller et al., 2001).

Table 10. The characteristics of the C3CER1 genes

Genes in C3CER1	Gen. size (kb)	Size of prot (aa)	Domain predicted (SMART)	Expression detected	Predicted protein function, comments
KIAA0028 also called *LARS2*	151	903	tRNA-synthetase class 1	Widely expressed (UniGene)	This gene encodes a class 1 aminoacyl-tRNA synthetase, mitochondrial leucyl-tRNA synthetase.
LIMD1 LIM domains containing 1	86	676	three tandemly arrayed LIM domains at the C-termini	Ubiquitosly expressed in the tested tissues (Kiss et al., 1999)	Plays a role in DNA-dependent transcription, cell communication, signal transduction, embryogenesis and morphogenesis (predicted by Celera) See Table 3 for more information.
KIAA0851/ SACM1L suppressor of actin 1	56	587	SacI homology domain, two transmembrane domains at the C-termini	Heart, brain, lung, liver, kidney, pancreas, testis, ovary, spleen (Nagase et al., 1998)	Phospho-inositide phosphatase, ortholog of the yeast SAC1 protein. Plays a role in intracellular protein traffic, cell growth and maintenance (predicted by Celera)
XT3 Isoform 1,2	41	592 555	neurotransmitter symporter family domain (twelve transmembrane domains)	Kidney, small intestine (Nash et al., 1998)	Na+/Cl- neurotransmitter transporter
LZTFL1 leucine zipper transcription factor-like 1	18	299	Coiled coil, Ribosomal protein S20, Leucine zipper pattern	Heart, brain, placentla, lung, sk. muscle, kidney, pancreas, thymus, testis (Kiss et al., 2001)	Plays a role in cytoskeleton organization and biogenesis, control of mitosis, peptidoglycan catabolism, cell cycle checkpoint, non-selective vesicle transport (predicted by Celera)
CCR9A ***CCR9B*** chemokine (C-C motif) receptor 9	15	369 357	Seven transmembrane receptor domain	Pooled clear cell type tumors, adenocarcinoma (UniGene)	This receptor appears to bind the majority of beta-chemokine family members, however, its specific function remains to be unknown.
FYCO1 FYVE and coiled-coil domain containing 1	78	1478	Coiled coil, RUN domain, FYVE domain	Heart, sk. muscle (Kiss et al., 2002)	Plays a role in organelle organization and biogenesis, embryogenesis and morphogenesis, protein modification, skeletal development, cell proliferation (predicted by Celera)
STRL33 Also called TYMSTR, chemokine (C-X-C motif) receptor 6, CXCR6	5	342	Seven transmembrane receptor domain	Liver and spleen; placenta; pooled pancreas and spleen; uterus; adenocarcinoma; osteosarcoma, etc. (UniGene)	STRL33 probably functions in interactions between dendritic cells and T cells and in regulating T-cell migration in the splenic red pulp (Matloubian et al., 2000). Potential co-receptors for HIV-1 virus.
CCXCR1 Chemokine XC receptor 1	1	333	Seven transmembrane receptor domain	Placenta, spleen, tymus (Yoshida et al., 1998)	This receptor is closely related to RBS11 and the MIP1 alpha/ RANTES receptor. The viral macrophage inflammatory protein -II is an antagonist of this receptor and blocks signaling.

Genes in C3CER1	**Gen. size (kb)**	**Size of prot (aa)**	**Domain predicted (SMART)**	**Expression detected**	**Predicted protein function, comments**
CCR1 chemokine (C-C motif) receptor 1	6	355	Seven transmembrane receptor domain	Kidney, prostate, colon, placenta uterus, lung, liver, spleen, breast, etc. (UniGene)	Mice lacking the chemokine receptor *CCR1* have defects in neutrophil trafficking and proliferation and increased susceptibility to Toxoplasma gondii infection (Khan et al., 2001).
CCR3 chemokine (C-C motif) receptor 3	1.2	355	Seven transmembrane receptor domain	Pooled brain, lung, testis; leukocyte (UniGene)	The presence of *CCR3* and its ligands in epithelial cells may contribute to the accumulation and activation of eosinophils and other inflammatory cells in the allergic airway (Stellato et al., 2001). Potential co-receptor for HIV-1 virus.
CCR2A ***CCR2B*** chemokine (C-C motif) receptor 2	5.3	374 360	Seven transmembrane receptor domain	Liver and Spleen; stem cell 34+/38+; Blood; prostate (UniGene)	*CCR2* is required for proper trafficking of antigen-presenting cells capable of inducing IFNG production by T cells (Peters et al., 2000) *CCR2B* is a potential co-receptor for HIV-1 virus.
CCR5 chemokine (C-C motif) receptor 5	6	352	Seven transmembrane receptor domain	pooled germ cell tumors; prostate; pooled colon, kidney, stomach; germinal center B cell, etc. (UniGene)	*CCR5* is an important co-receptor for macrophage-tropic virus, including HIV, to enter host cells. Defective alleles of this gene have been associated with the HIV infection resistance.
CCRL2 chemokine (C-C motif) receptor-like 2	1.6	344	Seven transmembrane receptor domain	colon tumor; adenocarcinoma; pooled brain, lung, testis ;pooled pancreas and spleen, etc. (UniGene)	G-protein coupled receptor
LTF Lactotrans-ferrin	29	711	Two transferrin domains	Wildly expressed (UniGene)	*LTF* belongs to a family of iron-binding proteins that modulate iron metabolism, hemopoiesis, and immunologic reactions. See Table 3 for more information.
TMEM7 Transmemb-rane protein 7	3	232	Transmembrane domain	Liver (Kiss et al., 2002)	
LRRC2 leucine-rich repeat-containing 2	51	371	Leucine-rich repeats	Heart, skeletal muscle (Kiss et al., 2002)	DNA-dependent transcription, RNA processing, nurse cell/oocyte transport, cell growth and maintenance, ectoderm development (predicted by Celera)
LUZPP1 leucine zipper protein pseudogene 1	2.3		Leucine zipper pattern	pooled germ cell tumors (UniGene)	

4.5. C3CER1 orthologous region in mice (paper V)

As a result of our large scale sequencing project, we have obtained the sequence of the C3CER1 (Kiss et al., 2002) and used the 1.32 Mb sequence for comparative analysis. We searched the Celera mouse database with genes from this region and it revealed that C3CER1 corresponded to two distinct conserved blocks on mouse chromosome 9F. A 907 kb long mouse sequence was selected and used for the PipMaker analysis.

High conservation within the coding regions of 17 genes identified within the human C3CER1 additionally supports that the described transcriptional map of the human C3CER1 was correct (Kiss et al., 2002). Lack of conservation of the detected human pseudogenes confirms their functional insignificance. We have detected and statistically analyzed a number of non-coding conserved elements.

The GC content of the entire human and mouse region was similar, 43.54% and 42.40%, respectively. RepeatMasker analysis identified a striking discrepancy in the number of repetitive elements. We detected 13 and 12 CpG islands in the human and in the mouse sequence, respectively. In the human region, 8 genes (out of the 17) have CpG islands at their 5' end. Interestingly, at the 5' end of *LIMD1*, three CpG islands were detected. Among the 13 CpG islands that were detected in the human sequence, 7 were conserved in the mouse. The remaining 5 CpG islands in the mouse sequence were located in intragenic regions.

We used EST based approach combined with Marathon RACE method to identify five novel mouse genes: *Kiaa0028*, *Xtrp3s1*, *Fyco1*, *Tmem7* and *Lrrc2*. The mouse *Tmem7* gene has two transcripts of different sizes, which were expressed mainly in the liver. The human and the mouse TMEM7 proteins do not show conservation within their C-termini. However, both proteins contain predicted transmembrane domains at their C-terminal parts.

The order and the structure of the genes in the studied human and mouse regions are well conserved with the exception of pseudogene insertions and generation of two mouse genes by duplication. The occurrence of gene duplications is a pronounced feature of the studied regions. In the mouse sequence, we could identify two new genes that were generated by duplication of the genomic sequence. From protein similarity analysis, we suggest that *Xtrp3* and *Cmkbr1l1* were a result of local duplications of *Xtrp3s1* and *Cmkbr1* genes, respectively, which occurred after human/mouse divergence. In C3CER1 we noticed the presence of a large cluster of chemokine receptor genes, which include eight genes: *CCR9*, *STRL33*, *CCXCR1*, *CCR1*, *CCR3*, *CCR2*, *CCR5*, and a chemokine receptor-like *CCRL2*. We constructed a phylogenetic tree of the human and mouse chemokine receptors that lie in the C3CER1 and its mouse othologous region. It can be hypothesized that all the chemokine receptor genes arose as a result of sequential intrachromosomal duplications. This notion is supported by the comparisons of three pairs of genes (*CCR9* and *STRL33*; *CCR1* and *CCR3*; *CCR2* and *CCR5*) in both species. All three pairs are located next to each other within the human and mouse loci and are the closest relatives.

In this study we applied a comparative analysis to further examine C3CER1. The orthologous region in mice was divided into two parts, but the gene content and gene positions were highly conserved between species. We found two mouse genes (*Xtrp3s1* and *Cmkbr1*) duplicated. Five novel mouse genes (*Kiaa0028*, *Xtrp3s1*, *Fyco1*, *Tmem7* and *Lrrc2*) were identified and characterized.

4.6. Study of human/mouse conservation breakpoints on human 3p12-22 (paper VI)

We compared our FISH-derived map to the human and mouse genome sequence-based maps, available from the Celera and UCSC databases. Our results were in agreement with these sequence-based maps throughout the entire analyzed region. In conclusion, the human C3CER1 belongs to two conserved chromosomal segments (CCSs). The distances between genes and their order in each of these two CCSs are similar in man and mouse and a murine/human conservation breakpoint region (CBR) lies between the *CCR5* and the *LTF* genes. There are several breakpoints of tumor-related chr 3 rearrangements in the region of 300 kb surrounding the CBR and we got further indications that this region is genetically unstable:

- YACs show often instability in this region.
- The presence of a cluster of chemokine receptors suggest that gene duplication events occurred during evolution. According to the phylogenetic tree of this family, the latest evolutionary duplications could be those two forming *CCR1-CCR3* and *CCR2-CCR5* pairs. These genes are very similar and are located within C3CER1 immediately adjacent to CBR. Human specific processed pseudogene insertions nearby CBR and enrichment of CBR with LTRs reflect increased transposition capacity, which is also characteristic for unstable regions. Moreover, comparison of mouse and human C3CER1 sequences revealed additional duplications in mouse involving one chemokine receptor gene and an *XT3* ortholog.
- We found that the CBR region replicated latest, as compared to the other studied sites, suggesting that a mechanism of instability within CBR may be similar to that in fragile sites.
- A TATAGA repeat capable to form hairpin-like secondary structure co-localizes with CBR. Hairpin formation may play a destabilizing role in eukaryotic genomes (Sinden, 2001).

We tested our hypothesis about the role of regional instability in the cancer-associated chromosomal rearrangements extending our analysis to a larger chromosome 3 segment. Seven regions were identified on the chromosome 3 as preferentially involved in tumor growth associated deletions located within 3p12-22. In order to analyze murine/human conservation over this chromosomal segment, we used the Celera computation data about mouse orthologs of 303 human genes located between Mb positions 40 and 85. We found that 278 genes match into 7 CCSs. Analysis of the rest of the genes indicated gene duplication events and additional chromosomal rearrangements or transpositions surrounding the CBRs. This may reflect the evolutionary instability of these chromosomal regions.

We found that these evolutionarily unstable regions are non-randomly localized within the 7 mentioned tumor associated deletions. Four regions, HD1, C3CER1, FER and HD4 contain CBRs, while two other (CER2, HD2) do not, but contain conservation mismatches and gene duplications. In particular within CER2 closely related paralogous genes *CCR8* and *CX3CR1* are most likely derived from the same locus as their closest relatives, i.e. *CCR1*, *CCR2*, *CCR3* and *CCR5* on C3CER1. This suggests that CER2 was involved in evolutionary rearrangements prior to the murine/human divergence. The HD2 contains 3 groups of paralogous genes and we detected the duplication of a large mouse segment corresponding to its telomeric part. Tandem Repeat Finder program at the UCSC server defined 8 sites of TATAGA repeats in the 3p12-22 segment. These sites were located non-randomly within the sites of tumor-associated deletions and in the vicinity of CBRs. Moreover the repeats, identified with the highest scores were located in two of our common eliminated regions (C3CER1 and CER2).

Our findings suggest that regional instabilities seem to play an important role in both, cancer-associated and evolutionary chromosomal rearrangements. The presence of TATAGA repeat is associated, but not necessarily causally, with regional instability.

5. CONCLUDING REMARKS AND FUTURE PERSPECTIVES

Our group has developed a functional assay, called elimination test, for the identification of tumor growth antagonizing regions. Using this system, we identified the C3CER1, which was commonly eliminated when chromosome 3 containing MCHs were xenografted in SCID mice. C3CER was fine mapped by PCR marker analysis and was fully covered by a PAC contig. We initiated a large scale sequencing project to uncover the gene content of C3CER1. Sequencing of the PAC clones revealed six novel genes: LIM domains containing gene 1 (*LIMD1*), Leucine zipper transcription factor-like gene 1 (*LZTFL1*), FYVE and coiled-coil domain containing gene 1 (*FYCO1*), Transmembrane protein gene 7 (TMEM), Leucine-rich repeat-containing gene 2 (*LRRC2*) and Leucine zipper protein pseudogene 1 (*LUZPP1*). Their mouse orthologs were also identified and characterized and selected examples were localized by FISH to the mouse chromosome 9F telomeric region. Comparative analysis of the human C3CER1 with its mouse orthologous region showed that the corresponding mouse region was divided into two blocks. The gene content and the order of the genes were highly conserved. The human/mouse conservation breakpoint region (CBR) lies between *CCR5* and *LTF* genes. A chemokine receptor cluster of 8 genes was recognized within C3CER1. We hypothesized that they were generated by duplication events and according to their family tree the latest duplications (*CCR1-CCR3*, *CCR2-CCR5*) occurred around the CBR.

The detailed transcriptional map of C3CER1 was prerequisite for the delineation of its role in tumorigenesis. We found 18 genes within C3CER1, which should be further studied. Preliminary results suggest that the *LTF* and the *LIMD1* genes are the best candidates. Mutation and methylation analysis, microarray technology can be applied for the further study of the C3CER1 genes before functional testing.

6. ACKNOWLEDGEMENTS

It is impossible to mention everybody, who helped me during my PhD studies. Thank you ALL for your support! Special thanks goes to

George Klein, who accepted me as a student immediately after I got my Bioengineer degree. I was very proud that George has taken his precious time and had separate meetings with me at the beginning of my work. The scientific conversations we had during our correspondence will always remain a nice memory for me.

Stefan Imreh, who introduced me to cell culture and the elimination test. He was always standing up and fighting for me and for the group. He made a good atmosphere in the lab and provided me everything I needed for a successful research. I knew that I could always ask his help even for personal problems. Thank you, without you, I could not achieve my PhD degree.

Jan Dumanski, who let me work in his laboratory at CMM for two years. I collected there all the basic knowledge necessary for my studies. I learned a lot from him during the transformation of my manuscripts from 'Hungarian English' to scientific English. He supported my ideas and helped me to build up my self-confidence calling me a 'good student' and 'expert'. This thesis would not exist without him.

Darek Kedra, who introduced me to UNIX and the Staden package. He helped me in the identification and characterization of my first novel gene.

My group members in MTC. **Éva Darai**, who helped to me to finish experiments, when I already expected Peter. Thanks for being my best girlfriend ever, to whom I can turn when I need support. **Maria Kost-Alimova**, who gave me a lot of encouragement and advised me concerning science and life. **Katalin Benedek**, who was always ready for a little chat and correcting my Swedish homework. **Anna Szeles**, who was always enthusiastic about my work and my family. **Ying Yang**, who started her studentship with me. Thanks for your help in the lab and your friendship. **Irina Kholodnyuk**, who taught me how to make PCR and thanks for the discussions about children. **Luda Fedorova**, who was always ready to help me with everything. **Agneta Sandlund**, who encouraged me to speak Swedish.

Group members of Jan Dumanski's group. **Ingegerd Fransson**, who helped me with all the practical things during my stay in that lab. **Isabell Tapia**, with whom I had great discussions about science, life and cats. **Myriam Peyrard,** her attitude towards work and life was a great example for me. **Giedre Grigelioniene,** who shared all my fears during my pregnancy with Robert. I also want to thank to **Cajsa Hansson**, **Kevin O'Brien** and **Carl Bruder** for creating a nice environment during my stay at CMM.

I would like to thank to my **mother** and my sister **Ilonka** all the support and encouragement I got even being far away from home. Special thanks to **Csaba's mother**, who thinks of me as her own daughter and shared with me all her experience about LIFE. I could not accomplish all this work without the support I got from **Csaba**. Thank you for your love, tolerance and understanding. Last but not least, thanks to my beloved sons, **Róbert** and **Péter** for making my life complete.

7. BIBLIOGRAPHY

Alexandropoulos K., Cheng G. and Baltimore D. (1995) Proline-rich sequences that bind to Src homology 3 domains with individual specificities. *Proc Natl Acad Sci U S A* **92**, 3110-4.

Altschul S. F., Gish W., Miller W., Myers E. W. and Lipman D. J. (1990) Basic local alignment search tool. *J Mol Biol* **215**, 403-10.

Attwood T. K. (2002) The PRINTS database: a resource for identification of protein families. *Brief Bioinform* **3**, 252-63.

Baggiolini M., Dewald B. and Moser B. (1997) Human chemokines: an update. *Annu Rev Immunol* **15**, 675-705.

Bateman A., Birney E., Cerruti L., Durbin R., Etwiller L., Eddy S. R., Griffiths-Jones S., Howe K. L., Marshall M. and Sonnhammer E. L. (2002) The Pfam protein families database. *Nucleic Acids Res* **30**, 276-80.

Beckerle M. C. (1986) Identification of a new protein localized at sites of cell-substrate adhesion. *J Cell Biol* **103**, 1679-87.

Beckerle M. C. (1997) Zyxin: zinc fingers at sites of cell adhesion. *Bioessays* **19**, 949-57.

Belperio J. A., Keane M. P., Arenberg D. A., Addison C. L., Ehlert J. E., Burdick M. D. and Strieter R. M. (2000) CXC chemokines in angiogenesis. *J Leukoc Biol* **68**, 1-8.

Berard J., Laboune F., Mukuna M., Masse S., Kothary R. and Bradley W. E. (1996) Lung tumors in mice expressing an antisense RARbeta2 transgene. *Faseb J* **10**, 1091-7.

Bezault J., Bhimani R., Wiprovnick J. and Furmanski P. (1994) Human lactoferrin inhibits growth of solid tumors and development of experimental metastases in mice. *Cancer Res* **54**, 2310-2.

Boldog F., Arheden K., Imreh S., Strombeck B., Szekely L., Erlandsson R., Marcsek Z., Sumegi J., Mitelman F. and Klein G. (1991) Involvement of 3p deletions in sporadic and hereditary forms of renal cell carcinoma. *Genes Chromosomes Cancer* **3**, 403-6.

Bonfield J. K., Smith K. and Staden R. (1995) A new DNA sequence assembly program. *Nucleic Acids Res* **23**, 4992-9.

Brodskii L. I., Ivanov V. V., Kalaidzidis Ia L., Leontovich A. M., Nikolaev V. K., Feranchuk S. I. and Drachev V. A. (1995) [GeneBee-NET: An Internet based server for biopolymer structure analysis]. *Biokhimiia* **60**, 1221-30.

Burge C. and Karlin S. (1997) Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **268**, 78-94.

Burke D. T., Carle G. F. and Olson M. V. (1987) Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. *Science* **236**, 806-12.

Cavenee W. K., Dryja T. P., Phillips R. A., Benedict W. F., Godbout R., Gallie B. L., Murphree A. L., Strong L. C. and White R. L. (1983) Expression of recessive alleles by chromosomal mechanisms in retinoblastoma. *Nature* **305**, 779-84.

Cheng Y., Poulos N. E., Lung M. L., Hampton G., Ou B., Lerman M. I. and Stanbridge E. J. (1998) Functional evidence for a nasopharyngeal carcinoma tumor suppressor gene that maps at chromosome 3p21.3. *Proc Natl Acad Sci U S A* **95**, 3042-7.

Chumakov I. M., Le Gall I., Billault A., Ougen P., Soularue P., Guillou S., Rigault P., Bui H., De Tand M. F., Barillot E., et al. (1992) Isolation of chromosome 21-specific yeast artificial chromosomes from a total human genome library. *Nat Genet* **1**, 222-5.

Cleves A. E., Novick P. J. and Bankaitis V. A. (1989) Mutations in the SAC1 gene suppress defects in yeast Golgi and yeast actin function. *J Cell Biol* **109**, 2939-50.

Corpet F., Servant F., Gouzy J. and Kahn D. (2000) ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res* **28**, 267-9.

Crawford A. W., Pino J. D. and Beckerle M. C. (1994) Biochemical and molecular characterization of the chicken cysteine-rich protein, a developmentally regulated LIM-domain protein that is associated with the actin cytoskeleton. *J Cell Biol* **124**, 117-27.

Csoka A. B., Frost G. I. and Stern R. (2001) The six hyaluronidase-like genes in the human and mouse genomes. *Matrix Biol* **20**, 499-508.

Cutler M. L., Bassin R. H., Zanoni L. and Talbot N. (1992) Isolation of rsp-1, a novel cDNA capable of suppressing v-Ras transformation. *Mol Cell Biol* **12**, 3750-6.

Daigo Y., Nishiwaki T., Kawasoe T., Tamari M., Tsuchiya E. and Nakamura Y. (1999) Molecular cloning of a candidate tumor suppressor gene, DLC1, from chromosome 3p21.3. *Cancer Res* **59**, 1966-72.

Dallol A., Forgacs E., Martinez A., Sekido Y., Walker R., Kishida T., Rabbitts P., Maher E. R., Minna J. D. and Latif F. (2002) Tumour specific promoter region methylation of the human homologue of the Drosophila Roundabout gene DUTT1 (ROBO1) in human cancers. *Oncogene* **21**, 3020-8.

Dawid I. B., Breen J. J. and Toyama R. (1998) LIM domains: multiple roles as adapters and functional modifiers in protein interactions. *Trends Genet* **14**, 156-62.

Durick K., Wu R. Y., Gill G. N. and Taylor S. S. (1996) Mitogenic signaling by Ret/ptc2 requires association with enigma via a LIM domain. *J Biol Chem* **271**, 12691-4.

Ephrussi B., Davidson R. L., Weiss M. C., Harris H. and Klein G. (1969) Malignancy of somatic cell hybrids. *Nature* **224**, 1314-6.

Erlandsson R., Bergerheim U. S., Boldog F., Marcsek Z., Kunimi K., Lin B. Y., Ingvarsson S., Castresana J. S., Lee W. H., Lee E., et al. (1990) A gene near the D3F15S2 site on 3p is expressed in normal human kidney but not or only at a severely reduced level in 11 of 15 primary renal cell carcinomas (RCC). *Oncogene* **5**, 1207-11.

Erlandsson R., Boldog F., Sumegi J. and Klein G. (1988) Do human renal cell carcinomas arise by a double-loss mechanism? *Cancer Genet Cytogenet* **36**, 197-202.

Falquet L., Pagni M., Bucher P., Hulo N., Sigrist C. J., Hofmann K. and Bairoch A. (2002) The PROSITE database, its status in 2002. *Nucleic Acids Res* **30**, 235-8.

Fedorova L., Kost-Alimova M., Gizatullin R. Z., Alimov A., Zabarovska V. I., Szeles A., Protopopov A. I., Vorobieva N. V., Kashuba V. I., Klein G., et al. (1997) Assignment and ordering of twenty-three unique NotI-linking clones containing expressed genes including the guanosine 5'-monophosphate synthetase gene to human chromosome 3. *Eur J Hum Genet* **5**, 110-6.

Feinberg A. P. and Vogelstein B. (1984) "A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity". Addendum. *Anal Biochem* **137**, 266-7.

Fickett J. W. (1996) Finding genes by computer: the state of the art. *Trends Genet* **12**, 316-20.

Fickett J. W. and Tung C. S. (1992) Assessment of protein coding measures. *Nucleic Acids Res* **20**, 6441-50.

Fishel R., Lescoe M. K., Rao M. R., Copeland N. G., Jenkins N. A., Garber J., Kane M. and Kolodner R. (1993) The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. *Cell* **75**, 1027-38.

Fodde R., Smits R. and Clevers H. (2001) APC, signal transduction and genetic instability in colorectal cancer. *Nat Rev Cancer* **1**, 55-67.

Fong L. Y., Fidanza V., Zanesi N., Lock L. F., Siracusa L. D., Mancini R., Siprashvili Z., Ottey M., Martin S. E., Druck T., et al. (2000) Muir-Torre-like syndrome in Fhit-deficient mice. *Proc Natl Acad Sci U S A* **97**, 4742-7.

Frazer K. A., Sheehan J. B., Stokowski R. P., Chen X., Hosseini R., Cheng J. F., Fodor S. P., Cox D. R. and Patil N. (2001) Evolutionarily conserved sequences on human chromosome 21. *Genome Res* **11**, 1651-9.

Freyd G., Kim S. K. and Horvitz H. R. (1990) Novel cysteine-rich motif and homeodomain in the product of the Caenorhabditis elegans cell lineage gene lin-11. *Nature* **344**, 876-9.

Friend S. H., Bernards R., Rogelj S., Weinberg R. A., Rapaport J. M., Albert D. M. and Dryja T. P. (1986) A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma. *Nature* **323**, 643-6.

Frohman M. A., Dush M. K. and Martin G. R. (1988) Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. *Proc Natl Acad Sci U S A* **85**, 8998-9002.

Garcia-Rostan G., Camp R. L., Herrero A., Carcangiu M. L., Rimm D. L. and Tallini G. (2001) Beta-catenin dysregulation in thyroid neoplasms: down-regulation, aberrant nuclear expression, and CTNNB1 exon 3 mutations are markers for aggressive tumor phenotypes and poor prognosis. *Am J Pathol* **158**, 987-96.

Gaullier J. M., Simonsen A., D'Arrigo A., Bremnes B., Stenmark H. and Aasland R. (1998) FYVE fingers bind PtdIns(3)P. *Nature* **394**, 432-3.

Gelfand M. S., Mironov A. A. and Pevzner P. A. (1996) Gene recognition via spliced sequence alignment. *Proc Natl Acad Sci U S A* **93**, 9061-6.

Gill G. N. (1995) The enigma of LIM domains. *Structure* **3**, 1285-9.

Gittoes N. J., McCabe C. J., Verhaeg J., Sheppard M. C. and Franklyn J. A. (1997) Thyroid hormone and estrogen receptor expression in normal pituitary and nonfunctioning tumors of the anterior pituitary. *J Clin Endocrinol Metab* **82**, 1960-7.

Goede V., Brogelli L., Ziche M. and Augustin H. G. (1999) Induction of inflammatory angiogenesis by monocyte chemoattractant protein-1. *Int J Cancer* **82**, 765-70.

Grabe N. (2002) AliBaba2: context specific identification of transcription factor binding sites. *In Silico Biol* **2**, S1-15.

Gracy J. and Argos P. (1998a) Automated protein sequence database classification. I. Integration of compositional similarity search, local similarity search, and multiple sequence alignment. *Bioinformatics* **14**, 164-73.

Gracy J. and Argos P. (1998b) Automated protein sequence database classification. II. Delineation Of domain boundaries from sequence similarities. *Bioinformatics* **14**, 174-87.

Grady W. M., Rajput A., Myeroff L., Liu D. F., Kwon K., Willis J. and Markowitz S. (1998) Mutation of the type II transforming growth factor-beta receptor is coincident with the transformation of human colon adenomas to malignant carcinomas. *Cancer Res* **58**, 3101-4.

Gupta S. K., Lysko P. G., Pillarisetti K., Ohlstein E. and Stadel J. M. (1998) Chemokine receptors in human endothelial cells. Functional expression of CXCR4 and its transcriptional regulation by inflammatory cytokines. *J Biol Chem* **273**, 4282-7.

Hanahan D. and Weinberg R. A. (2000) The hallmarks of cancer. *Cell* **100**, 57-70.

Hardison R. C. (2000) Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet* **16**, 369-72.

Harris H., Miller O. J., Klein G., Worst P. and Tachibana T. (1969) Suppression of malignancy by cell fusion. *Nature* **223**, 363-8.

Harris N. L. (1997) Genotator: a workbench for sequence annotation. *Genome Res* **7**, 754-62.

Heinemeyer T., Wingender E., Reuter I., Hermjakob H., Kel A. E., Kel O. V., Ignatieva E. V., Ananko E. A., Podkolodnaya O. A., Kolpakov F. A., et al. (1998) Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL. *Nucleic Acids Res* **26**, 362-7.

Henikoff J. G., Greene E. A., Pietrokovski S. and Henikoff S. (2000) Increased coverage of protein families with the blocks database servers. *Nucleic Acids Res* **28**, 228-30.

Heyer B. S., Warsowe J., Solter D., Knowles B. B. and Ackerman S. L. (1997) New member of the Snf1/AMPK kinase family, Melk, is expressed in the mouse egg and preimplantation embryo. *Mol Reprod Dev* **47**, 148-56.

Higgins D. G., Thompson J. D. and Gibson T. J. (1996) Using CLUSTAL for multiple sequence alignments. *Methods Enzymol* **266**, 383-402.

Hirokawa T., Boon-Chieng S. and Mitaku S. (1998) SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics* **14**, 378-9.

Hobert O., Schilling J. W., Beckerle M. C., Ullrich A. and Jallal B. (1996) SH3 domain-dependent interaction of the proto-oncogene product Vav with the focal contact protein zyxin. *Oncogene* **12**, 1577-81.

Huang X., Adams M. D., Zhou H. and Kerlavage A. R. (1997) A tool for analyzing and annotating genomic sequences. *Genomics* **46**, 37-45.

Hudson T. J., Stein L. D., Gerety S. S., Ma J., Castle A. B., Silva J., Slonim D. K., Baptista R., Kruglyak L., Xu S. H., et al. (1995) An STS-based map of the human genome. *Science* **270**, 1945-54.

Huebner K., Druck T., Siprashvili Z., Croce C. M., Kovatich A. and McCue P. A. (1998) The role of deletions at the FRA3B/FHIT locus in carcinogenesis. *Recent Results Cancer Res* **154**, 200-15.

Hughes W. E., Cooke F. T. and Parker P. J. (2000a) Sac phosphatase domain proteins. *Biochem J* **350 Pt 2**, 337-52.

Hughes W. E., Woscholski R., Cooke F. T., Patrick R. S., Dove S. K., McDonald N. Q. and Parker P. J. (2000b) SAC1 encodes a regulated lipid phosphoinositide phosphatase, defects in which can be suppressed by the homologous Inp52p and Inp53p phosphatases. *J Biol Chem* **275**, 801-8.

Hurst H. C. (1995) Transcription factors 1: bZIP proteins. *Protein Profile* **2**, 101-68.

Imreh S., Kholodnyuk I., Allikmetts R., Stanbridge E. J., Zabarovsky E. R. and Klein G. (1994) Nonrandom loss of human chromosome 3 fragments from mouse-human microcell hybrids following progressive growth in SCID mice. *Genes Chromosomes Cancer* **11**, 237-45.

Imreh S., Klein G. and Zabarovsky E. R. (2003) Search for unknown tumor antagonizing genes. *Genes Chromosomes and Cancer* In press.

Imreh S., Kost-Alimova M., Kholodnyuk I., Yang Y., Szeles A., Kiss H., Liu Y., Foster K., Zabarovsky E., Stanbridge E., et al. (1997) Differential elimination of 3p and retention of 3q segments in human/mouse microcell hybrids during tumor growth. *Genes Chromosomes Cancer* **20**, 224-33.

Ioannou P. A., Amemiya C. T., Garnes J., Kroisel P. M., Shizuya H., Chen C., Batzer M. A. and de Jong P. J. (1994) A new bacteriophage P1-derived vector for the propagation of large human DNA fragments. *Nature Genetics* **6**, 84-9.

Karlsson O., Thor S., Norberg T., Ohlsson H. and Edlund T. (1990) Insulin gene enhancer binding protein Isl-1 is a member of a novel class of proteins containing both a homeo- and a Cys-His domain. *Nature* **344**, 879-82.

Khan I. A., Murphy P. M., Casciotti L., Schwartzman J. D., Collins J., Gao J. L. and Yeaman G. R. (2001) Mice lacking the chemokine receptor CCR1 show increased susceptibility to Toxoplasma gondii infection. *J Immunol* **166**, 1930-7.

Kholodnyuk I., Kost-Alimova M., Kashuba V., Gizatulin R., Szeles A., Stanbridge E. J., Zabarovsky E. R., Klein G. and Imreh S. (1997) A 3p21.3 region is preferentially eliminated from human chromosome 3/mouse microcell hybrids during tumor growth in SCID mice. *Genes Chromosomes Cancer* **18**, 200-11.

Kholodnyuk I. D., Kost-Alimova M., Yang Y., Kiss H., Fedorova L., Klein G. and Imreh S. (2002) The microcell hybrid-based "elimination test" identifies a 1-Mb putative tumor-suppressor region at 3p22.2-p22.1 centromeric to the homozygous deletion region detected in lung cancer. *Genes Chromosomes Cancer* **34**, 341-4.

Killary A. M., Wolf M. E., Giambernardi T. A. and Naylor S. L. (1992) Definition of a tumor suppressor locus within human chromosome 3p21-p22. *Proc Natl Acad Sci U S A* **89**, 10877-81.

Kinzler K. W. and Vogelstein B. (1996) Lessons from hereditary colorectal cancer. *Cell* **87**, 159-70.

Kinzler K. W. and Vogelstein B. (1997) Cancer-susceptibility genes. Gatekeepers and caretakers. *Nature* **386**, 761, 763.

Kiss H., Kedra D., Kiss C., Kost-Alimova M., Yang Y., Klein G., Imreh S. and Dumanski J. P. (2001) The LZTFL1 Gene Is a Part of a Transcriptional Map Covering 250 kb within the Common Eliminated Region 1 (C3CER1) in 3p21.3. *Genomics* **73**, 10-9.

Kiss H., Kedra D., Yang Y., Kost-Alimova M., Kiss C., O'Brien K. P., Fransson I., Klein G., Imreh S. and Dumanski J. P. (1999) A novel gene containing LIM domains (LIMD1) is located within the common eliminated region 1 (C3CER1) in 3p21.3. *Hum Genet* **105**, 552-9.

Kiss H., Yang Y., Kiss C., Andersson K., Klein G., Imreh S. and Dumanski J. P. (2002) The transcriptional map of the common eliminated region 1 (C3CER1) in 3p21.3. *Eur J Hum Genet* **10**, 52-61.

Klein G., Bregula U., Wiener F. and Harris H. (1971) The analysis of malignancy by cell fusion. I. Hybrids between tumour cells and L cell derivatives. *J Cell Sci* **8**, 659-72.

Knudsen S. (1999) Promoter2.0: for the recognition of PolII promoter sequences. *Bioinformatics* **15**, 356-61.

Knudson A. G. (1971) Mutation and cancer: Statistical study of retinoblastoma. *Proc Natl Acad Sci U S A* **68**, 820-823.

Kochendorfer K. U., Then A. R., Kearns B. G., Bankaitis V. A. and Mayinger P. (1999) Sac1p plays a crucial role in microsomal ATP transport, which is distinct from its function in Golgi phospholipid metabolism. *Embo J* **18**, 1506-15.

Kok K., Naylor S. L. and Buys C. H. (1997) Deletions of the short arm of chromosome 3 in solid tumors and the search for suppressor genes. *Adv Cancer Res* **71**, 27-92.

Kovacs G., Erlandsson R., Boldog F., Ingvarsson S., Muller-Brechlin R., Klein G. and Sumegi J. (1988) Consistent chromosome 3p deletion and loss of heterozygosity in renal cell carcinoma. *Proc Natl Acad Sci U S A* **85**, 1571-5.

Kriventseva E. V., Servant F. and Apweiler R. (2003) Improvements to CluSTr: the database of SWISS-PROT+TrEMBL protein clusters. *Nucleic Acids Res* **31**, 388-9.

Krogh A. (1997) Two methods for improving performance of an HMM and their application for gene finding. *Proc Int Conf Intell Syst Mol Biol* **5**, 179-86.

Kulp D., Haussler D., Reese M. G. and Eeckman F. H. (1996) A generalized hidden Markov model for the recognition of human genes in DNA. *Proc Int Conf Intell Syst Mol Biol* **4**, 134-42.

Lander E. S., Linton L. M., Birren B., Nusbaum C., Zody M. C., Baldwin J., Devon K., Dewar K., Doyle M., FitzHugh W., et al. (2001) Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921.

Lee J. W., Choi H. S., Gyuris J., Brent R. and Moore D. D. (1995) Two classes of proteins dependent on either the presence or absence of thyroid hormone for interaction with the thyroid hormone receptor. *Mol Endocrinol* **9**, 243-54.

Leevers S. J., Vanhaesebroeck B. and Waterfield M. D. (1999) Signalling through phosphoinositide 3-kinases: the lipids take centre stage. *Curr Opin Cell Biol* **11**, 219-25.

Lerman M. I. and Minna J. D. (2000) The 630-kb lung cancer homozygous deletion region on human chromosome 3p21.3: identification and evaluation of the resident candidate tumor suppressor genes. The International Lung Cancer Chromosome 3p21.3 Tumor Suppressor Gene Consortium. *Cancer Res* **60**, 6116-33.

Li Z., Meng Z. H., Chandrasekaran R., Kuo W. L., Collins C. C., Gray J. W. and Dairkee S. H. (2002) Biallelic inactivation of the thyroid hormone receptor beta1 gene in early stage breast cancer. *Cancer Res* **62**, 1939-43.

Liu R. and States D. J. (2002) Consensus promoter identification in the human genome utilizing expressed gene markers and gene modeling. *Genome Res* **12**, 462-9.

Lupas A., Van Dyke M. and Stock J. (1991) Predicting coiled coils from protein sequences. *Science* **252**, 1162-4.

Maruyama R., Toyooka S., Toyooka K. O., Virmani A. K., Zochbauer-Muller S., Farinas A. J., Minna J. D., McConnell J., Frenkel E. P. and Gazdar A. F. (2002) Aberrant promoter methylation profile of prostate cancers and its relationship to clinicopathological features. *Clin Cancer Res* **8**, 514-9.

Matloubian M., David A., Engel S., Ryan J. E. and Cyster J. G. (2000) A transmembrane CXC chemokine is a ligand for HIV-coreceptor Bonzo. *Nat Immunol* **1**, 298-304.

Mayinger P., Bankaitis V. A. and Meyer D. I. (1995) Sac1p mediates the adenosine triphosphate transport into yeast endoplasmic reticulum that is required for protein translocation. *J Cell Biol* **131**, 1377-86.

Mayor C., Brudno M., Schwartz J. R., Poliakov A., Rubin E. M., Frazer K. A., Pachter L. S. and Dubchak I. (2000) VISTA : visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* **16**, 1046-7.

McCabe C. J., Gittoes N. J., Sheppard M. C. and Franklyn J. A. (1999) Thyroid receptor alpha1 and alpha2 mutations in nonfunctioning pituitary tumors. *J Clin Endocrinol Metab* **84**, 649-53.

Meisler M. H. (2001) Evolutionarily conserved noncoding DNA in the human genome: how much and what for? *Genome Res* **11**, 1617-8.

Modrek B., Resch A., Grasso C. and Lee C. (2001) Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res* **29**, 2850-9.

Mulder N. J., Apweiler R., Attwood T. K., Bairoch A., Barrell D., Bateman A., Binns D., Biswas M., Bradley P., Bork P., et al. (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res* **31**, 315-8.

Muller A., Homey B., Soto H., Ge N., Catron D., Buchanan M. E., McClanahan T., Murphy E., Yuan W., Wagner S. N., et al. (2001) Involvement of chemokine receptors in breast cancer metastasis. *Nature* **410**, 50-6.

Nagase T., Ishikawa K., Suyama M., Kikuno R., Hirosawa M., Miyajima N., Tanaka A., Kotani H., Nomura N. and Ohara O. (1998) Prediction of the coding sequences of unidentified human genes. XII. The complete sequences of 100 new cDNA clones from brain which code for large proteins in vitro. *DNA Res* **5**, 355-64.

Nakai K. and Horton P. (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem Sci* **24**, 34-6.

Nash S. R., Giros B., Kingsmore S. F., Kim K. M., el-Mestikawy S., Dong Q., Fumagalli F., Seldin M. F. and Caron M. G. (1998) Cloning, gene structure and genomic localization of an orphan transporter from mouse kidney with six alternatively-spliced isoforms. *Receptors Channels* **6**, 113-28.

Nemoto Y., Kearns B. G., Wenk M. R., Chen H., Mori K., Alb J., De Camilli P. and Bankaitis V. A. (2000) Functional characterization of a mammalian Sac1 and mutants exhibiting substrate specific defects in phosphoinositide phosphatase activity. *J Biol Chem.*

Novick P., Osmond B. C. and Botstein D. (1989) Suppressors of yeast actin mutations. *Genetics* **121**, 659-74.

Oberlander U., Adam R., Berg K., Seeber F. and Lucius R. (1995) Molecular cloning and characterization of the filarial LIM domain proteins AvL3-1 and OvL3-1. *Experimental Parasitology* **81**, 592-9.

Ohmura H., Tahara H., Suzuki M., Ide T., Shimizu M., Yoshida M. A., Tahara E., Shay J. W., Barrett J. C. and Oshimura M. (1995) Restoration of the cellular senescence program and repression of telomerase by human chromosome 3. *Jpn J Cancer Res* **86**, 899-904.

Pekarsky Y., Zanesi N., Palamarchuk A., Huebner K. and Croce C. M. (2002) FHIT: from gene discovery to cancer treatment and prevention. *Lancet Oncol* **3**, 748-54.

Pennacchio L. A. and Rubin E. M. (2001) Genomic strategies to identify mammalian regulatory sequences. *Nat Rev Genet* **2**, 100-9.

Pesole G., Liuni S. and D'Souza M. (2000) PatSearch: a pattern matcher software that finds functional elements in nucleotide and protein sequences and assesses their statistical significance. *Bioinformatics* **16**, 439-50.

Peters W., Dupuis M. and Charo I. F. (2000) A mechanism for the impaired IFN-gamma production in C-C chemokine receptor 2 (CCR2) knockout mice: role of CCR2 in linking the innate and adaptive immune responses. *J Immunol* **165**, 7072-7.

Petit M. M. R., Mols R., Schoenmakers E. F., Mandahl N. and Van de Ven W. J. (1996) LPP, the preferred fusion partner gene of HMGIC in lipomas, is a novel member of the LIM protein gene family. *Genomics* **36**, 118-29.

Polverini P. J. (1997) Role of the macrophage in angiogenesis-dependent diseases. *Exs* **79**, 11-28.

Ponger L. and Mouchiroud D. (2002) CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. *Bioinformatics* **18**, 631-3.

Prestridge D. S. (1995) Predicting Pol II promoter sequences using transcription factor binding sites. *J Mol Biol* **249**, 923-32.

Rameh L. E. and Cantley L. C. (1999) The role of phosphoinositide 3-kinase lipid products in cell function. *J Biol Chem* **274**, 8347-50.

Reinhard M., Jouvenal K., Tripier D. and Walter U. (1995) Identification, purification, and characterization of a zyxin-related protein that binds the focal adhesion and microfilament protein VASP (vasodilator-stimulated phosphoprotein). *Proc Natl Acad Sci U S A* **92**, 7956-60.

Retera J. M., Leers M. P., Sulzer M. A. and Theunissen P. H. (1998) The expression of beta-catenin in non-small-cell lung cancer: a clinicopathological study. *J Clin Pathol* **51**, 891-4.

Rimessi P., Gualandi F., Morelli C., Trabanelli C., Wu Q., Possati L., Montesi M., Barrett J. C. and Barbanti-Brodano G. (1994) Transfer of human chromosome 3 to an ovarian carcinoma cell line identifies three regions on 3p involved in ovarian cancer. *Oncogene* **9**, 3467-74.

Rogic S., Mackworth A. K. and Ouellette F. B. (2001) Evaluation of gene-finding programs on mammalian sequences. *Genome Res* **11**, 817-32.

Sadler I., Crawford A. W., Michelsen J. W. and Beckerle M. C. (1992) Zyxin and cCRP: two interactive LIM domain proteins associated with the cytoskeleton. *J Cell Biol* **119**, 1573-87.

Salzberg S., Delcher A. L., Fasman K. H. and Henderson J. (1998) A decision tree system for finding genes in DNA. *J Comput Biol* **5**, 667-80.

Sambrook J., Fritsch E. F. and Maniatis T. (1989) *Molecular cloning; a laboratory manual.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Sanchez Y., el-Naggar A., Pathak S. and Killary A. M. (1994) A tumor suppressor locus within 3p14-p12 mediates rapid cell death of renal cell carcinoma in vivo. *Proc Natl Acad Sci U S A* **91**, 3383-7.

Sanchez-Garcia I. and Rabbitts T. H. (1993) LIM domain proteins in leukaemia and development. *Semin Cancer Biol* **4**, 349-58.

Sanger F., Nicklen S. and Coulson A. R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**, 5463-7.

Santoro M., Carlomagno F., Romano A., Bottaro D. P., Dathan N. A., Grieco M., Fusco A., Vecchio G., Matoskova B., Kraus M. H., et al. (1995) Activation of RET as a dominant transfotming gene by germline mutations of MEN2A and MEN2B. *Science* **267**, 381-3.

Satoh H., Lamb P. W., Dong J. T., Everitt J., Boreiko C., Oshimura M. and Barrett J. C. (1993) Suppression of tumorigenicity of A549 lung adenocarcinoma cells by human chromosomes 3 and 11 introduced via microcell-mediated chromosome transfer. *Mol Carcinog* **7**, 157-64.

Saxon P. J., Srivatsan E. S. and Stanbridge E. J. (1986) Introduction of human chromosome 11 via microcell transfer controls tumorigenic expression of HeLa cells. *Embo J* **5**, 3461-6.

Saxon P. J. and Stanbridge E. J. (1987) Transfer and selective retention of single specific human chromosomes via microcell-mediated chromosome transfer. *Methods Enzymol* **151**, 313-25.

Schlosshauer P. W., Pirog E. C., Levine R. L. and Ellenson L. H. (2000) Mutational analysis of the CTNNB1 and APC genes in uterine endometrioid carcinoma. *Mod Pathol* **13**, 1066-71.

Schmeichel K. L. and Beckerle M. C. (1997) Molecular dissection of a LIM domain. *Mol Biol Cell* **8**, 219-30.

Schmidt L., Duh F. M., Chen F., Kishida T., Glenn G., Choyke P., Scherer S. W., Zhuang Z., Lubensky I., Dean M., et al. (1997) Germline and somatic mutations in the tyrosine kinase domain of the MET proto-oncogene in papillary renal carcinomas. *Nat Genet* **16**, 68-73.

Schorr M., Then A., Tahirovic S., Hug N. and Mayinger P. (2001) The phosphoinositide phosphatase Sac1p controls trafficking of the yeast Chs3p chitin synthase. *Curr Biol* **11**, 1421-6.

Schultz J., Copley R. R., Doerks T., Ponting C. P. and Bork P. (2000) SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res* **28**, 231-4.

Schwartz S., Zhang Z., Frazer K. A., Smit A., Riemer C., Bouck J., Gibbs R., Hardison R. and Miller W. (2000) PipMaker--a web server for aligning two genomic DNA sequences. *Genome Res* **10**, 577-86.

Shimizu M., Yokota J., Mori N., Shuin T., Shinoda M., Terada M. and Oshimura M. (1990) Introduction of normal chromosome 3p modulates the tumorigenicity of a human renal cell carcinoma cell line YCR. *Oncogene* **5**, 185-94.

Shivakumar L., Minna J., Sakamaki T., Pestell R. and White M. A. (2002) The RASSF1A tumor suppressor blocks cell cycle progression and inhibits cyclin D1 accumulation. *Mol Cell Biol* **22**, 4309-18.

Shizuya H., Birren B., Kim U. J., Mancino V., Slepak T., Tachiiri Y. and Simon M. (1992) Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. *Proc Natl Acad Sci U S A* **89**, 8794-7.

Siebert P. D., Chenchik A., Kellogg D. E., Lukyanov K. A. and Lukyanov S. A. (1995) An improved PCR method for walking in uncloned genomic DNA. *Nucleic Acids Res* **23**, 1087-8.

Siebert P. D. and Huang B. C. (1997) Identification of an alternative form of human lactoferrin mRNA that is expressed differentially in normal tissues and tumor-derived cell lines. *Proc Natl Acad Sci U S A* **94**, 2198-203.

Simonsen A., Wurmser A. E., Emr S. D. and Stenmark H. (2001) The role of phosphoinositides in membrane transport. *Curr Opin Cell Biol* **13**, 485-92.

Sinden R. R. (2001) Neurodegenerative diseases. Origins of instability. *Nature* **411**, 757-8.

Solovyev V. V. (2002) *Finding genes by computer: Probabilistic and discriminative approaches.* MIT Press.

Stanbridge E. J. (1985) A case for human tumor-suppressor genes. *Bioessays* **3**, 252-5.

Stellato C., Brummet M. E., Plitt J. R., Shahabuddin S., Baroody F. M., Liu M. C., Ponath P. D. and Beck L. A. (2001) Expression of the C-C chemokine receptor CCR3 in human airway epithelial cells. *J Immunol* **166**, 1457-61.

Stenmark H., Aasland R. and Driscoll P. C. (2002) The phosphatidylinositol 3-phosphate-binding FYVE finger. *FEBS Lett* **513**, 77-84.

Stolz L. E., Huynh C. V., Thorner J. and York J. D. (1998) Identification and characterization of an essential family of inositol polyphosphate 5-phosphatases (INP51, INP52 and INP53 gene products) in the yeast Saccharomyces cerevisiae. *Genetics* **148**, 1715-29.

Strand M., Prolla T. A., Liskay R. M. and Petes T. D. (1993) Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. *Nature* **365**, 274-6.

Sundaresan V., Chung G., Heppell-Parton A., Xiong J., Grundy C., Roberts I., James L., Cahn A., Bench A., Douglas J., et al. (1998) Homozygous deletions at 3p12 in breast and lung cancer. *Oncogene* **17**, 1723-9.

Szeles A., Yang Y., Sandlund A. M., Kholodnyuk I., Kiss H., Kost-Alimova M., Zabarovsky E. R., Stanbridge E., Klein G. and Imreh S. (1997) Human/mouse microcell hybrid based elimination test reduces the putative tumor suppressor region at 3p21.3 to 1.6 cM. *Genes Chromosomes Cancer* **20**, 329-36.

Todd M. C., Xiang R. H., Garcia D. K., Kerbacher K. E., Moore S. L., Hensel C. H., Liu P., Siciliano M. J., Kok K., van den Berg A., et al. (1996) An 80 Kb P1 clone from chromosome 3p21.3 suppresses tumor growth in vivo. *Oncogene* **13**, 2387-96.

Tomizawa Y., Sekido Y., Kondo M., Gao B., Yokota J., Roche J., Drabkin H., Lerman M. I., Gazdar A. F. and Minna J. D. (2001) Inhibition of lung cancer cell growth and induction of apoptosis after reexpression of 3p21.3 candidate tumor suppressor gene SEMA3B. *Proc Natl Acad Sci U S A* **98**, 13954-9.

Toulouse A., Morin J., Dion P. A., Houle B. and Bradley W. E. (2000) RARbeta2 specificity in mediating RA inhibition of growth of lung cancer-derived cells. *Lung Cancer* **28**, 127-37.

Tse C., Xiang R. H., Bracht T. and Naylor S. L. (2002) Human Semaphorin 3B (SEMA3B) located at chromosome 3p21.3 suppresses tumor formation in an adenocarcinoma cell line. *Cancer Res* **62**, 542-6.

Tsukazaki T., Chiang T. A., Davison A. F., Attisano L. and Wrana J. L. (1998) SARA, a FYVE domain protein that recruits Smad2 to the TGFbeta receptor. *Cell* **95**, 779-91.

Turner C. E., Glenney J. R., Jr. and Burridge K. (1990) Paxillin: a new vinculin-binding protein present in focal adhesions. *J Cell Biol* **111**, 1059-68.

Turner C. E. and Miller J. T. (1994) Primary sequence of paxillin contains putative SH2 and SH3 domain binding motifs and multiple LIM domains: identification of a vinculin and pp125Fak-binding region. *J Cell Sci* **107**, 1583-91.

Uzawa N., Yoshida M. A., Hosoe S., Oshimura M., Amagasa T. and Ikeuchi T. (1998) Functional evidence for involvement of multiple putative tumor suppressor genes on the short arm of chromosome 3 in human oral squamous cell carcinogenesis. *Cancer Genet Cytogenet* **107**, 125-31.

van Ommen G. J. (2002) The Human Genome Project and the future of diagnostics, treatment and prevention. *J Inherit Metab Dis* **25**, 183-8.

Venter J. C., Adams M. D., Myers E. W., Li P. W., Mural R. J., Sutton G. G., Smith H. O., Yandell M., Evans C. A., Holt R. A., et al. (2001) The sequence of the human genome. *Science* **291**, 1304-51.

Virmani A. K., Rathi A., Zochbauer-Muller S., Sacchi N., Fukuyama Y., Bryant D., Maitra A., Heda S., Fong K. M., Thunnissen F., et al. (2000) Promoter methylation and silencing of the retinoic acid receptor-beta gene in lung carcinomas. *J Natl Cancer Inst* **92**, 1303-7.

von Heijne G. (1992) Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *J Mol Biol* **225**, 487-94.

Wang L., Darling J., Zhang J. S., Liu W., Qian J., Bostwick D., Hartmann L., Jenkins R., Bardenhauer W., Schutte J., et al. (2000) Loss of expression of the DRR 1 gene at chromosomal segment 3p21.1 in renal cell carcinoma. *Genes Chromosomes Cancer* **27**, 1-10.

Way J. C. and Chalfie M. (1988) mec-3, a homeobox-containing gene that specifies differentiation of the touch receptor neurons in C. elegans. *Cell* **54**, 5-16.

Wei L., Liu Y., Dubchak I., Shon J. and Park J. (2002) Comparative genomics approaches to study organism similarities and differences. *J Biomed Inform* **35**, 142-50.

Weng Z., Taylor J. A., Turner C. E., Brugge J. S. and Seidel-Dugan C. (1993) Detection of Src homology 3-binding proteins, including paxillin, in notmal and v-Src-transformed Balb/c 3T3 cells. *J Biol Chem* **268**, 14956-63.

Wolf E., Kim P. S. and Berger B. (1997) MultiCoil: a program for predicting two- and three-stranded coiled coils. *Protein Sci* **6**, 1179-89.

Wu R. Y. and Gill G. N. (1994) LIM domain recognition of a tyrosine-containing tight turn. *J Biol Chem* **269**, 25085-90.

Xian J., Clark K. J., Fordham R., Pannell R., Rabbitts T. H. and Rabbitts P. H. (2001) Inadequate lung development and bronchial hyperplasia in mice with a targeted deletion in the Dutt1/Robo1 gene. *Proc Natl Acad Sci U S A* **98**, 15062-6.

Xiang R., Davalos A. R., Hensel C. H., Zhou X. J., Tse C. and Naylor S. L. (2002) Semaphorin 3F gene from human 3p21.3 suppresses tumor formation in nude mice. *Cancer Res* **62**, 2637-43.

Yang Q., Yoshimura G., Mori I., Sakurai T. and Kakudo K. (2002) Chromosome 3p and breast cancer. *J Hum Genet* **47**, 453-9.

Yang Y., Kiss H., Kost-Alimova M., Kedra D., Fransson I., Seroussi E., Li J., Szeles A., Kholodnyuk I., Imreh M. P., et al. (1999) A 1-Mb PAC contig spanning the common eliminated region 1 (CER1) in microcell hybrid-derived SCID tumors. *Genomics* **62**, 147-55.

Yang Y., Kost-Alimova M., Ingvarsson S., Qianhui Q., Kiss H., Szeles A., Kholodnyuk I., Cuthbert A., Klein G. and Imreh S. (2001) Similar regions of human chromosome 3 are eliminated from or retained in human/human and human/mouse microcell hybrids during tumor growth in severe combined immunodeficient (SCID) mice. *Proc Natl Acad Sci U S A* **98**, 1136-41.

Yang Y., Li J., Szeles A., Imreh M. P., Kost-Alimova M., Kiss H., Kholodnyuk I., Fedorova L., Darai E., Klein G., et al. (2003) Consistent downregulation of human lactoferrin gene, in the common eliminated region 1 on 3p21.3, following tumor growth in severe combined immunodeficient (SCID) mice. *Cancer Lett* **191**, 155-64.

Yoshida T., Imai T., Kakizaki M., Nishimura M., Takagi S. and Yoshie O. (1998) Identification of single C motif-1/lymphotactin receptor XCR1. *J Biol Chem* **273**, 16551-4.

Zaballos A., Gutierrez J., Varona R., Ardavin C. and Marquez G. (1999) Cutting edge: identification of the orphan chemokine receptor GPR-9-6 as CCR9, the receptor for the chemokine TECK. *J Immunol* **162**, 5671-5.

Zabarovsky E. R., Lerman M. I. and Minna J. D. (2002) Tumor suppressor genes on chromosome 3p involved in the pathogenesis of lung and other cancers. *Oncogene* **21**, 6915-35.

Zhang M. Q. (1997) Identification of protein coding regions in the human genome by quadratic discriminant analysis. *Proc Natl Acad Sci U S A* **94**, 565-8.

Zhuang Z., Park W. S., Pack S., Schmidt L., Vortmeyer A. O., Pak E., Pham T., Weil R. J., Candidus S., Lubensky I. A., et al. (1998) Trisomy 7-harbouring non-random duplication of the mutant MET allele in hereditary papillary renal carcinomas. *Nat Genet* **20**, 66-9.