The 630-kb Lung Cancer Homozygous Deletion Region on Human Chromosome 3p21.3: Identification and Evaluation of the Resident Candidate Tumor Suppressor Genes

Michael I. Lerman and John D. Minna

For The International Lung Cancer Chromosome 3p21.3 Tumor Suppressor Gene Consortium

Hamon Center for Therapeutic Oncology Research, University of Texas Southwestern Medical Center, Dallas, Texas 75390-8591 [J. D. M.], and Laboratory of Immunobiology, National Cancer Institute, Frederick Research and Development Center, Frederick, Maryland 21702 [M. I. L.]

ABSTRACT

We overmapped and nested homozygous deletions, contig building, genomic sequencing, and physical and transcript mapping to further define a ∼630-kb lung cancer homozygous deletion region harboring one or more tumor suppressor genes (TSGs) on chromosome 3p21.3. This location was identified through somatic genetic mapping in tumors, cancer cell lines, and premalignant lesions of the lung and breast, including the discovery of several homozygous deletions. The combination of molecular methods and computational predictions permitted us to detect, isolate, characterize, and annotate a set of 25 genes that likely constitute the complete set of protein-coding genes residing in this ∼630-kb sequence. A subset of 19 of these genes was found within the deleted overlap region of ∼370-kb. This region was further subdivided by a nesting 200-kb breast cancer homozygous deletion into two gene sets: 8 genes lying in the proximal ∼120-kb segment and 11 genes lying in the distal ∼250-kb segment. These 19 genes were analyzed extensively by computational methods and were tested by manual methods for loss of expression and mutations in lung cancers to identify candidate TSGs from within this group. Four genes showed loss-of-expression or reduced mRNA levels in non-small cell lung cancer (CACNA2D2a26-2, SEMA3B [formerly SEMA(N), BLU, and HYAL1] or small cell lung cancer (SEMA3B, BLU, and HYAL1) cell lines. We found six of the genes to have two or more amino acid sequence-altering mutations including BLU, NPRL2/Gene21, FUS1, HYAL1, FUS2, and SEMA3B. However, none of the 19 genes tested for mutation showed a frequent (>10%) mutation rate in lung cancer samples. This led us to exclude several of the genes in the region as classical tumor suppressors for sporadic lung cancer. On the other hand, the putative lung cancer TSG in this location may either be inactivated by tumor-acquired promoter hypermethylation or belong to the novel class of haploinsufficient genes that predispose to cancer in a hemizygous (+/−) state but do not show a second mutation in the remaining wild-type allele in the tumor. We discuss the data in the context of novel and classic cancer gene models as applied to lung carcinogenesis. Further functional testing of the critical genes by gene transfer and gene disruption strategies should permit the identification of the putative lung cancer TSG(s), LUCA. Analysis of the ∼630-kb sequence also provides an opportunity to probe and understand the genomic structure, evolution, and functional organization of this relatively gene-rich region.

INTRODUCTION

Lung cancer kills >150,000 patients each year in the United States and many more around the world. These are more deaths than attributable to colon, prostate, and breast cancer combined (1−4). The scale of this epidemic has heightened efforts to understand the molecular pathogenesis of lung cancer, including genes frequently undergoing somatic mutation. The isolation of such “lung cancer genes” should guide the development of new therapeutic interventions and new early detection and prevention strategies. Although tobacco smoking is a well-established environmental etiology in lung carcinogenesis (5), our understanding of the acquired genetic changes leading to lung cancer is still rudimentary and incomplete (6). The lack of (and difficulty of performing) genetic linkage studies of lung cancer in families (4, 7) that have the potential of identifying initiating cancer causing genes has directed the search toward allele loss mapping in tumors, cell lines, and premalignant lesions of the lung and breast (8−13). A convergence of evidence from these allele loss mapping studies, including identification of overlapping homozygous deletions, strongly suggests the presence of a TSG1 in the chromosome 3p21.3 band (6, 9, 10, 14). Allele loss in the 3p21.3 area is the earliest premalignant change thus far detected in lung cancer development (12, 15, 16). Biallelic or monoallelic inactivation of this putative TSG gene(s) likely represents a critical (rate-limiting) step in the development of sporadic lung cancer. As part of our efforts to identify a lung cancer TSG (which we will provisionally call LUCA) on 3p21.3, we physically mapped and cloned the genomic sequence surrounding this locus as defined by homozygous deletions (6, 9, 10, 14, 17−20). Subsequently, the ∼630-kb clone contig was sequenced jointly by The Washington University and The Sanger4 Human Genome Sequencing Centers. In more recent work, we

4 The abbreviations used are: TSG, tumor suppressor gene; SCLC, small cell lung cancer; NSCLC, non-small cell lung cancer; NCI, National Cancer Institute; LOH, loss of heterozygosity; EST, expressed sequence tag; MTN, multiple tissue Northern blots from ClonTech; SSCP, single strand conformation polymorphism; ORF, open reading frame; NLS, nuclear localization signal; FISH, fluorescence in situ hybridization; VWA, Von Willebrand factor type A; PCK, protein kinase C; RT-PCR, reverse transcription-PCR; DAG, diacylglycerol.
5 Internet address: http://genome.wustl.edu/gsc/human/chrom3.shtml.
6 Internet address: http://www.sanger.ac.uk/cgi-bin/hummap?map=3p21.3.
placed the putative 3p21.3 TSG(s) in a ∼120-kb segment that was defined by a homozygous deletion in a breast cancer specimen that was nested within the three small cell lung cancer homozygous deletions (10). In parallel with these genetic and physical studies, we have been constructing a map of transcript sequences with the aim to identify a complete set of all transcripts encoded in the region and to define/annotate the respective genes. Here we report the catalogue of genes we have discovered to be residing in the 630-kb sequence and their experimental and informatics characterization. Of these, only two “G protein” genes, i.e., GNAI2 (21) and GNAT1 (22), had been cloned and characterized previously, and from this catalog we positioned these two genes within the contig DNA sequence. The set of 19 genes found in the overlapping homozygous deletions in SCLC NCI-H740, NCI-H1450, and GLC20 (18), including eight in the smaller critical 120-kb sequence defined by a breast homozygous deletion (10), were analyzed extensively. We used both manual experimental methods to study expression and search for mutations and web-based computational servers to predict possible protein functions. Four of the genes by Northern analysis showed frequent reduced or absent mRNA levels in NSCLC (CACNA2D2, SEMA3B, BLU, and HYAL1) and SCLC (BLU, HYAL1, and SEMA3B) cell lines. We found that six of the genes had mutations, but none of the 19 genes showed a high frequency of mutations (>10%) in the analyzed lung tumor samples. This raises the possibility that the putative TSG, LUA1, may be one of the genes with frequent loss of expression that occurs through acquired tumor promoter hypermethylation (23). Alternatively, it could belong to the class of haploinsufficient TSGs. This novel class of TSGs is predicted to predispose to cancer in a hemizygous (+/-) state but does not show a second hit in the remaining wild-type allele in tumors. Further functional experimental analysis such as growth suppression studies and gene knockout strategies will be required to reveal the identity of the putative 3p21.3 TSG(s). In addition, our study shows that genomic DNA sequencing in combination with high-quality gene annotation is an effective method of gene discovery.

MATERIALS AND METHODS

Cell Lines and DNA Samples

Lung cancer cell lines were started and maintained by us using published methods, and information on the lines are summarized in the NCI-Navy collection (American Type Culture Collection; Ref. 24). DNA from the SCLC line, GLC20 (17), was a gift of C. H. C. Buys (University of Groningen, Groningen, the Netherlands); the SCLC cell line, U2020 (25), was a gift of P. Rabbitts (MRC, Cambridge United Kingdom). Lung cancer normal/tumor paired DNA samples were from the NCI-Navy collection established by B. Johnson (26).

Commercial Reagents

The following materials were purchased from the vendors indicated: PCR tool kits, Perkin-Elmer Cetus; DNA sequencing tool kits, Applied Biosystems (Foster City, CA); fluorescent in situ hybridization reagents, Boehringer Mannheim (Mannheim, Germany); cDNA and cosmid libraries, Clontech Laboratories (Palo Alto CA) and Stratagene (La Jolla, CA); EST clones from public databases, the I.M.A.G.E consortium7 or Research Genetics (Rockville, MD); SSCP tool kits and blotting nylon membranes, Amersham (Arlington Heights, IL); oligonucleotide primers, Life Technologies, Inc. (Rockville, MD); restriction enzymes, Life Technologies, Inc., New England Biolabs (Beverly, MA), and Amersham (Arlington Heights, IL); MTN poly(A)+ RNA blots, Clontech Laboratories (Palo Alto, CA); buffers, blotting solutions, and RNase free water, Quality Biologicals (Gaithersburg, MD); chemicals, Sigma Chemical Co. (St. Louis, MO); and cell culture media, Life Technologies, Inc. (Rockville, MD).

Informatics Tools

The software package, GENSCAN (27), was licensed from Christopher Burge (MIT, Cambridge, MA) and installed at the NCI Advanced Biomedical Computing Center at the Frederick Cancer Research and Development Center. The integrated informatics package, PANORAMA, incorporating BLAST, GENSCAN, GRAIL, and other gene interpretation features, was developed at University of Texas Southwestern Medical Center at Dallas (TX) by H. Garner and run on a Hewlett Packard Exemplar supercomputer. PANORAMA is available for internet use.8 For this analysis, GenBank was downloaded December 1999.

Manual Molecular Procedures

All molecular manipulations (DNA and RNA isolations, screening genomic and cdNA libraries, Northern and Southern blot analyses, and PCR) were performed using standard methods according to Sambrook et al. (28). For DNA sequencing, cdNA clones were sequenced on an Applied Biosystems 373 or 377 DNA sequencer (Stretch) using Taq Dideoxy Terminator Cycle Sequence kits (Applied Biosystems, Foster City, CA) with either vector or clone-specific walking primers. Cosmid and PI phage DNAs were sequenced by the Washington University and Sanger Human Genome Sequencing Centers using the shotgun procedure as described (21, 22). FISH and two-color FISH were used to locate and orient the cosmid contig on chromosome 3p. Normal metaphase chromosomes were hybridized simultaneously with digoxigenin-labeled NotI linking clone NLJ-210, part of cosmid LUCA1, and biotin-labeled cosmid LUCA20 (red) (29). 4',6-Diamidino-2-phenylindole was used as a counterstain. Both the metaphase spreads and interphase nucleus staining confirmed the single-site location of each probe on 3p21.31, establishing the following order: centromere-cosmids LUCA1-LUCA20-telomere. Pulsed-field gel electrophoresis analysis was performed as follows. High molecular weight DNA was prepared in agarose plugs as described (30). Slices containing ∼10³ cells were digested for 16 h with 50 units of enzyme (NotI, NruI, and MluI; Boehringer Mannheim) and resolved on 1% agarose gels using a Bio-Rad CHEF Mapper (Hercules, CA) and electrophoresis profiles, allowing separation in the range 50–1000 kb. For expression analyses, Northern blot hybridization was performed with cdNA probes using commercial MTN poly(A)+ RNA blots Clontech (Palo Alto, CA) from a variety of adult human tissues and tumor cell lines and in-house blots with total or poly(A)+ RNA were prepared from lung cancer cell lines. Radioactive DNA probes were prepared by random priming Rediprime II (Amersham, Arlington Heights, IL). Hybridization was performed in ExpressHyb hybridization solution according to manufacturer’s instructions (Clontech Laboratories, Palo Alto, CA). In addition, the presence of gene transcripts was monitored in situ by BLAST homology searches (31) in public EST databases. Mutational analyses were performed by RT-PCR-SSCP or exon-PCR-SSCP, followed by sequencing of shifted bands as described previously (32, 33). Experimental gene discovery by using conserved and transcribed genomic fragments was performed as detailed previously (18).

Computational and Bioinformatics Procedures

World Wide Web-based Servers and Databases. World Wide Web-based servers and databases (34) were used to analyze genomic, cdNA, and predicted protein sequences. In addition, the Wisconsin Genetics Computer Group, package 10 (35), and the GENSCAN (27) programs were run at the Advanced Biomedical Computing Center (Frederick Cancer Research and Development Center), whereas the University of Texas Southwestern Medical Center integrated gene analysis software PANORAMA was run at the University of Texas Southwestern Medical Center.

DNA and Protein Sequence Analyses. Global sequence alignments were done using BLAST (31) and Advanced BLAST² (36) programs as provided by National Center for Biotechnology Institute.10 and BLAST2/WU-BLAST.11 Multiple sequence alignments, global and local, were done using the CLUSTAL version W program12 as provided by EMBL, Baylor Computing

7 Internet address: http://www-bio.llnl.gov/bbrp/image/image.html.
8 Internet address: http://pompous.swmed.edu/panorama.htm.
9 Internet address: http://blast.wustl.edu/ and http://www2.ebi.ac.uk/blast2/.
11 Internet addresses: http://blast.wustl.edu/ and http://www2.ebi.ac.uk/blast2/.
12 Internet address: http://www.bork.embl-heidelberg.de/Alignment/.
Center,13 and the Wisconsin Genetics Computer Group, package 10 program (35). Protein structural features were delineated using the EXPASY proteomics tools.14 Protein domains were discovered using Pfam (37) and SMART (38) programs as provided by EMBL.15 Protein subcellular localization was predicted using PSORT25(39). Signal peptides, transmembrane helices, and membrane topologies were predicted by SPLITT(40), TMHMM18(41), and PSORT (39) programs. Protein motifs were found by visually inspecting local alignments or using the protein motifs (42), ProfileScan and Prosite (43) programs. In addition, we used the INTERPRO server (44,19).

**Discovery of Orthologous Genes in Model Organisms.** Stringent criteria for identification of candidate orthologous genes were applied as suggested (45, 46). In the mouse, orthologous pairs were >90% identical on the protein level with >90% alignment of their entire lengths (47). In the fly, worm, and yeast, candidates were identified with 20–50% identity over at least 80% of their nucleotides, Unigene, and EST databases of the model organisms. EST clusters were then built by the EST assembly machine server20 or the EST Assembler21 at Max Delbrück Center.22 The advanced BLAST2 and Orthologue program (48) at EMBL, was used to confirm the putative orthologous relationships and obtain and ascertain phylogenetic trees.

**In Silico Gene Discovery Was Performed following Two Different Protocols.** Genome-wide repeats and low complexity regions in the genomic DNA sequences were identified and masked using the program RepeatMasker.23 They were then used in BLASTN searches against EST, Unigene, and nonredundant nucleotide databases to identify potential transcripts (ESTs and cDNAs) and build EST clusters. Next, genomic sequences assembled from the individual cosmid sequences were subjected to gene prediction programs, i.e., GENSCAN (27) and XGRAIL (49), with default settings to identify coding DNA sequences and corresponding protein sequences. These were then used in BLASTN and TBLASTN searches, respectively, against nonredundant nucleotide and EST databases to identify ESTs and cDNAs. The ESTs were then assembled into clusters (see above). Genomic information (repetitive elements, EST database) was used in the contig sequence. The TBLASTN program was used to search nonredundant nucleotide databases and EST databases assembled into clusters (see above). Genomic information (repetitive elements, coding exons, ESTs, and known and predicted genes) was obtained from EMBL (24) as provided by EMBL. 13 and the Wisconsin Genetics Computer Group, package 10 program (35). Protein structural features were delineated using the EXPASY proteomics tools.14 Protein domains were discovered using Pfam (37) and SMART (38) programs as provided by EMBL.15 Protein subcellular localization was predicted using PSORT25(39). Signal peptides, transmembrane helices, and membrane topologies were predicted by SPLITT(40), TMHMM18(41), and PSORT (39) programs. Protein motifs were found by visually inspecting local alignments or using the protein motifs (42), ProfileScan and Prosite (43) programs. In addition, we used the INTERPRO server (44,19).

RESULTS

**Overlapping and Nested Homozygous Deletions Define 120-kb and 250-kb Regions for a TSG Search and Identification of 25 Resident Genes in the Overall ~630-kb Region.** LOH (allele loss) is an important sign of somatically acquired genetic events in the natural history of many tumors and is a useful tool to discover the location of new TSGs. However, the regions of allelic losses involving 3p in lung cancer (6,9) and premalignant lesions (12,14–16) are multiple and often large, making it difficult to define a small consensus region that would facilitate a realistic positional cloning effort. Fortunately, three overlapping homozygous deletions in lung cancer cell lines (NCI-H740, NCI-H1450, and GLC20; Refs. 9, 17–19, 51) and one in a breast tumor specimen (HCC1500; Ref. 10), in conjunction with functional evidence of suppressor activity (9, 52–54), strongly support the existence of a putative TSG(s) in this location and successively narrowed the critical gene region, first to 370 kb (18), and more recently, to ~120 kb (Ref. 10; Fig. 1). This latest reduction of the critical segment by a nesting deletion in a breast carcinoma assumes that the same gene(s) was targeted in both lung and breast malignancies (10). However, if the targeted gene(s) were different, the lung cancer gene(s) could lie in either the ~120-kb region or in the more telomeric ~250-kb segment of the 370-kb region.

The genomic DNA cosmid/P1 clone contig (Fig. 1) covering the overlapping homozygous deletions in 3p21.3 is now sequenced almost to completion and is ~630 kb long.26 Fig. 1 summarizes the genetic and physical map of the 3p21.3 region on which we have searched for a lung cancer TSG, including contig structure, and successive steps leading to the definition of the critical regions of ~370 kb, ~250 kb, and ~120 kb. We have systematically proceeded to identify and test all of the genes in the region for their candidacy as a TSG (Fig. 1; Table 1). In addition to the genes we detected by manual methods (such as screening cDNA libraries with cosmid DNAs and using the cosmids for exon capture), the genomic sequence allowed us to detect other gene candidates in silico. We used BLAST searches for ESTs, and gene prediction programs such as GRAIL, GENQUEST, and GENSCAN (27, 49) including our own integrated informatics tool, PANORAMA.27 The total number of verified resident genes in the 630-kb region currently is 25, 6 of which were in the centromeric portion of the 630-kb region and outside of the overlap of SCLC or breast cancer homozygous deletions (Fig. 1; Table 1). For these 25 genes, we have reported mutational and expression analysis on 3pK, 1FDR2/SM15, and SEMA3B and laid out the rough outline of some of the gene locations in the contig (18, 55–57). The gene prediction programs, particularly GENSCAN, identified all of the 25 verified genes and fairly accurately (>90%) predicted their ORFs and intron/exon structure (although in some cases distinct genes were combined into one gene). In addition to the 25 genes, by an intensive informatics effort using the genomic DNA sequence, we found several potential genes (Table 1). These included two gene candidates (arbitrary names Gene29, cosmid LUCA2, and Gene19, cosmid LUCA10; Table 1), identified by one or two EST hits that did not have intron/exon structure, were not detected by the GENSCAN prediction program, did not have a significant ORF, and did not have a detectable mRNA signal on MTN blots. We believe these may represent genomic DNA contamination of the EST database. Another set of gene candidates were detected by GENSCAN (29) and GRAIL-EXP (49) gene prediction programs analysis of genomic DNA (arbitrary names Gene22, Gene23, Gene24, Gene25, and Gene27; Table 1). Although programs provide a predicted cDNA sequence, ORF, and intron/exon structure, thus far we have found no EST hits in GenBank, no mRNA signal on MTN blots, and no homology of the predicted protein to proteins in GenBank. Thus, at the present time we cannot verify these GENSCAN predicted candidates as actual genes, and they are noted for future reference. These observations suggest that we have found all, or nearly all, resident genes in the ~630-kb sequence. This contention is likely true for the ~120-kb sequence (Fig. 1) that contains all of seven complete genes and a large part of the CACNA2D2 gene. In this critical region,
the genes reside in the following order with the indicated sizes and intergenic distances: CACNA2D2 (80 kb), 5-kb intron; PL6 (4.5 kb), 1.5-kb intron; 101F6 (2.3 kb), 161-bp intron; NPLR2/21 (3.3 kb), 1.8-kb intron; BLU (4.7 kb), 4.4-kb intron; RASSF1/123F2 (7.6 kb), 1.6-kb intron; FUS1 (3.3 kb), 4.5-kb intron; and HYAL2 (2.7 kb). With a total of ~18 kb of intronic and ~17 kb of intronic sequence, this region is extremely gene dense and contains 20–25% of coding sequence.

The location of the contig on 3p21.3 and its position relative to other 3p genes were determined by the presence in the ~630-kb sequence of several framework genetic markers mapped to this location (e.g., D3S1621 and D3S1568; Fig. 1), multiple radiation hybrid mapped sequence tagged sites (SHGC-11855 shown), and other linked genetic markers (see detailed marker information on genomic sequence contigs NT_002322, NT_000067, and NT-000069). In addition, in situ hybridization [FISH and two-color FISH with DNA from cosmids LUCA1 (Z74618), LUCA8 (Z84495), and LUCA20 (AC004693)] located the contig to 3p21.3 and proved the orientation as shown on Fig. 1 (data not shown). We also performed radiation hybrid mapping with the TNG panel with selected markers for finer resolution of the location (not shown). In addition, using the contig sequence, we have determined intergenic distances, exon-intron structures, the intron sizes of the resident genes, and ascertained the direction of transcription. These are all features that are immediately available by performing a BLAST analysis with our deposited cDNA sequences (Table 1) against the individual cosmids or assembled genomic sequences. In total, the 25 genes occupy ~630-kb of genomic DNA, resulting in an average size of ~25 kb/gene, which agrees well with the size of an average gene (~30 kb) estimated for the whole human genome. 

Fig. 1. Genetic and physical characterization of the human tumor nested homozygous deletion region on 3p21.3 showing the steps leading to the identification of the cloned resident candidate tumor suppressor genes. Top, ideogram of the banding pattern of chromosome 3p with some genetic framework markers found in and flanking the deleted region. Next is shown a physical pulsed-field gel electrophoresis map of the homozygously deleted region and its flanking sites. The positions of the rare cutter restriction sites are indicated above the line and the sizes (in kb) of the corresponding NotI restriction fragments are given below the line. The physical map is followed by a diagrammatic representation of the overlapping homozygous deletions discovered in SCLC cell lines (NCI-H1450 in blue, NCI-H1740 in magenta, and GLC20 in green) and a nested deletion in a breast cancer tumor sample and corresponding cell line (H1500 in brick). The sizes of the homozygous deletions, deletion overlaps, and the precisely mapped breakpoints are indicated. The minimum tiling cosmid contig covering the deleted region is shown below the deletions. The position of the framework genetic markers is shown by downward arrows; the GenBank accession nos. for cosmid sequences are given under the lines, and the abbreviated cosmid/P1 clone numbering system given to the Genome Centers (e.g., Luca01, Luca02, . . . , Luca22, and P1 clone p3938) is given above the lines. The sequencing was done jointly by The Washington University (Luca12 to P3938) and The Sanger (Luca01 to Luca11) Genome Sequencing Centers using the high-redundancy shotgun procedure and is available from GenBank or the centers ftp sites (see: ftp.sanger.ac.uk/pub/human/sequences/Chr_3/ and ftp://genome.wustl.edu/pub/gsc1/sequence/st.louis/human/). Finally, a line representing the ~630-kb assembled genomic sequence with the positioned resident gene and breakpoints defining critical regions is shown at the bottom. The genes are represented by pointed rectangles, indicating the orientations of transcription. The gene names are given above and the GenBank accession nos. below the rectangles. The * with downward arrows indicates the position of two genes (temporary names LUCA1.2 and LUCA2.5) on cosmids Luca01 and Luca02, respectively, whose sequencing is not yet completed. The 6 genes outside of the critical region in the centromeric portion are in magenta, the 8 genes in the combined breast and lung 120-kb critical region are in blue, whereas the remaining 11 genes still within the rest of the lung cancer nested homozygous deletions (and additional 250 kb) are all in green. The 6 genes exhibiting at least some point or small mutations have an “M” in their box. Probable pseudogenes and other in silico predicted genes for which we currently cannot confirm their status as a bona fide gene are not shown.
and found no abnormalities. Xiang et al. (63) reported studying the entire ORF in 28 small cell lung cancers and found no abnormalities.

However, distribution of these genes along the sequence is rather uneven, with the highest gene density (genes in green in Fig. 1) in the 120-kb region (average gene size, ~15-kb). Gene sizes varied dramatically, with the smallest gene size of 2.3 kb (g101F6) and the largest of ~140 kb (CACNA2D2). Actually, this large gene is interrupted by the centromeric breakpoints of both the HCC1500 breast cancer homologous deletion (in cosmid LUCA6) and in the SLC GLC20 homologous deletion in LUCA10 (Fig. 1). The intergenic distances also differed enormously, from 161 bp (between g101F6 and NPR2/g21) to ~60 kb (between genes g20 and CACNA2D2). The intron sizes also varied dramatically, the smaller ones ranging from 50–1000 bp and the largest being 40–50 kb (in CACNA2D2).

It is worthwhile to examine what type of genes (if any) would have been missed by our manual and informatics prediction analyses. These omissions could include genes that lack GenBank matches, genes whose sequences would not be recognized by the available gene prediction programs, such as non-protein coding genes, and genes whose mRNAs have a very restricted pattern of tissue expression. Another approach to this gene saturation problem will be to compare the human sequence with the corresponding mouse genomic sequence. Because functional sequences such as transcriptional enhancers and mRNA-like noncoding RNAs are highly conserved in mammalian genomes, they might be readily detected in this comparison (58, 59).

The three lung cancer homologous deletions identified an ~370-kb region extending from cosmid LUCA10 (Z75742; defined by the centromeric end of the GLC20 homologous deletion) to P3938 (AC004814; defined by the telomeric end of the NCI-H440 homologous deletion (Fig. 1). The nested homologous deletion in breast cancer HCC1500 (10), which covered part of cosmid LUCA6 (Z84493) to part of cosmid LUCA13 (AC002455; Fig. 1), divided the 19 genes in the 370-kb region into two critical gene sets: 8 genes in a 120-kb segment extending from part of cosmid LUCA10 to part of cosmid LUCA13, and 11 genes in the telomeric portion (~250 kb, extending from cosmid LUCA13 to P1 clone P3938, AC004814). These deletions eliminated the 3pk/MAPKAP3 (U95758), CISH (AF132297, temporarily called Gen18 in our GenBank deposit), HEMK isoforms I and II (AF131220 and AF172244), Gen19 (AF188706), and two partially characterized genes [Gen22 (Luc2a.2) and Gen19 (Luca10)], have ESTs but appear to be nonexpressed, intronless with no good ORF, and thus probable pseudogenes. Several candidates were predicted by GENSCAN but have no EST hits, no detectable mRNA on multiple tissue Northern blots, and no obvious protein homologies. These include Gen18 (cosmid LUCA14), Gen23 (cosmid LUCA14), Gen25 (cosmid LUCA20), and Gen27 (cosmids LUCA19–22). No EST/mRNAs were found, no MTN blot mRNA signals are seen, and the predicted amino acid sequences have no known protein homologies.

We tested 456 of 753 amino acids of the SEMA3F open reading frame and found no abnormalities. Xiang et al. (63) reported studying the entire ORF in 28 small cell lung cancers and found no abnormalities.
sizes of the respective transcripts and patterns of expression in normal human tissues and lung cancers (Figs. 2 and 3; Table 2). Expression in Northern blots prepared from total or poly(A)+ RNA of 20–30 RNA samples from lung cancer cell lines, representing both SCLC and NSCLC, revealed, for many of the genes, levels of expression similar to normal lung. No abnormal transcript sizes suggestive of mutations were found. However, several genes showed reduced expression in the lung cancer lines, i.e., the CACNA2D2 gene not expressed in 50% of the lines, the BLU gene expressed only in 30%, the HYAL1 gene expressed in <30% (Fig. 3), and SEMA3B and SEMA3F (19, 57, 63) expressed in <50%. Thus, expression analysis in lung cancers identified these five genes as potential TSG candidates based on loss of expression in a sizable number (but not all lung cancers). Because a possible mutation mechanism is tumor-acquired promoter hypermethylation (23), the methylation status of the CpG islands associated with the genes showing reduced or absent expression is currently under investigation.

Mutation Analysis of the Candidate Tumor Suppressor Genes. Initially, we performed Southern blot analysis on ~100 genomic DNAs from our large panel of lung cancer cell lines representing both SCLC and NSCLCs (24) using cDNA or genomic probes representing each of the 25 genes in our search for other homozygous deletions or genomic DNA rearrangements (data not shown). We found only the homozygous deletions listed in Fig. 1 and ~30-kb homozygous deletion in SCLC NCI-H524 involving most of the genomic sequence in cosmid LUCA13-interrupting gene FUS1 to gene HYAL1 (data not shown, and see discussion below). Mutational analyses of the resident genes (summarized in Tables 1 and 3) were performed on lung cancer cDNAs or genomic DNAs by RT-PCR-SSCP or PCR-SSCP, respectively, followed by DNA sequencing of any altered bands detected. A

Table 2  mRNA expression of 3p21.3 genes in Multiple Tissue Northern (MTN) blots

<table>
<thead>
<tr>
<th>Gene locus (mRNA kb)</th>
<th>GenBank no.</th>
<th>Heart</th>
<th>Brain</th>
<th>Placenta</th>
<th>Lung</th>
<th>Liver</th>
<th>SkMus</th>
<th>Kidney</th>
<th>Pancreas</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAPKAP2 (2.5)</td>
<td>U09578</td>
<td>3+</td>
<td>Trace</td>
<td>Trace</td>
<td>2+</td>
<td>1+</td>
<td>3+</td>
<td>3+</td>
<td>2+</td>
</tr>
<tr>
<td>CISH/Gen18 (2.4)</td>
<td>AF132297</td>
<td>3+</td>
<td>Neg</td>
<td>Trace</td>
<td>Trace</td>
<td>Trace</td>
<td>Trace</td>
<td>Trace</td>
<td>Trace</td>
</tr>
<tr>
<td>HEMK (1.7, 7, 7.5)</td>
<td>AF131220</td>
<td>3+</td>
<td>Trace</td>
<td>Trace</td>
<td>Trace</td>
<td>Trace</td>
<td>Trace</td>
<td>Trace</td>
<td>Trace</td>
</tr>
<tr>
<td>Gene20 (2.4, 2.6)</td>
<td>AF188706</td>
<td>3+</td>
<td>2+</td>
<td>1-</td>
<td>3+</td>
<td>Neg</td>
<td>Neg</td>
<td>2+</td>
<td>1+</td>
</tr>
<tr>
<td>CACNA2D2 (5.5–5.7)</td>
<td>AF040709</td>
<td>2+</td>
<td>2+</td>
<td>Neg</td>
<td>Trace</td>
<td>Trace</td>
<td>Trace</td>
<td>Trace</td>
<td>Trace</td>
</tr>
<tr>
<td>Gene19 (none)</td>
<td>AF173154</td>
<td>2+</td>
<td>2+</td>
<td>Neg</td>
<td>Trace</td>
<td>Trace</td>
<td>Trace</td>
<td>Trace</td>
<td>Trace</td>
</tr>
<tr>
<td>PL6 (2.2)</td>
<td>U09584.1</td>
<td>1+</td>
<td>Trace</td>
<td>Trace</td>
<td>2+</td>
<td>1-</td>
<td>2+</td>
<td>3+</td>
<td>1+</td>
</tr>
<tr>
<td>101F6 (1.5)</td>
<td>AF040706.1</td>
<td>1+</td>
<td>Trace</td>
<td>Trace</td>
<td>2+</td>
<td>1-</td>
<td>2+</td>
<td>3+</td>
<td>1+</td>
</tr>
<tr>
<td>NPR2/Gene21 (1.5)</td>
<td>AF040707</td>
<td>2+</td>
<td>1+</td>
<td>1-</td>
<td>2+</td>
<td>Trace</td>
<td>Trace</td>
<td>Trace</td>
<td>Trace</td>
</tr>
<tr>
<td>BLU (2.0)</td>
<td>U70824, U70880</td>
<td>Trace</td>
<td>Trace</td>
<td>Trace</td>
<td>2+</td>
<td>1-</td>
<td>2+</td>
<td>3+</td>
<td>1+</td>
</tr>
<tr>
<td>RASSF1/123F2 (2.0)</td>
<td>AF040703</td>
<td>3+</td>
<td>1+</td>
<td>2+</td>
<td>1+</td>
<td>1+</td>
<td>2+</td>
<td>1+</td>
<td>3+</td>
</tr>
<tr>
<td>FUS1 (1.8)</td>
<td>AF055479.1</td>
<td>3+</td>
<td>2+</td>
<td>1+</td>
<td>2+</td>
<td>3+</td>
<td>3+</td>
<td>3+</td>
<td>3+</td>
</tr>
<tr>
<td>HYAL2 (2.0)</td>
<td>U09577</td>
<td>2+</td>
<td>Trace</td>
<td>Trace</td>
<td>3+</td>
<td>3+</td>
<td>2+</td>
<td>3+</td>
<td>1+</td>
</tr>
<tr>
<td>HYAL2 (2.6)</td>
<td>U03056, AF173154</td>
<td>Neg</td>
<td>Neg</td>
<td>Trace</td>
<td>Neg</td>
<td>Trace</td>
<td>Trace</td>
<td>Trace</td>
<td>Trace</td>
</tr>
<tr>
<td>FUS2 (1.9)</td>
<td>AF040705</td>
<td>3+</td>
<td>2+</td>
<td>Neg</td>
<td>1+</td>
<td>3+</td>
<td>1+</td>
<td>2+</td>
<td>3+</td>
</tr>
<tr>
<td>HYAL3 (2)</td>
<td>AF040710</td>
<td>Neg</td>
<td>1+</td>
<td>Neg</td>
<td>Neg</td>
<td>Neg</td>
<td>Neg</td>
<td>Neg</td>
<td>Neg</td>
</tr>
<tr>
<td>IFDR2/SKMC15 (4)</td>
<td>U09585</td>
<td>3+</td>
<td>Trace</td>
<td>Trace</td>
<td>1-</td>
<td>1+</td>
<td>3+</td>
<td>3+</td>
<td>1+</td>
</tr>
<tr>
<td>SEMA3B (3.4)</td>
<td>U28369</td>
<td>1+</td>
<td>1+</td>
<td>3+</td>
<td>Trace</td>
<td>Neg</td>
<td>1+</td>
<td>1+</td>
<td>1+</td>
</tr>
<tr>
<td>G17 (3.0)</td>
<td>U49082.1</td>
<td>2+</td>
<td>2+</td>
<td>Neg</td>
<td>Neg</td>
<td>3+</td>
<td>3+</td>
<td>1+</td>
<td>3+</td>
</tr>
<tr>
<td>SEMA3F (3.9, 2.9)</td>
<td>U38276, U33920</td>
<td>2+</td>
<td>1+</td>
<td>2+</td>
<td>2+</td>
<td>Trace</td>
<td>Trace</td>
<td>Trace</td>
<td>Trace</td>
</tr>
<tr>
<td>G15/RBM5 (4, 2, 1.5)</td>
<td>U29346</td>
<td>3+</td>
<td>1+</td>
<td>1-</td>
<td>Neg</td>
<td>3+</td>
<td>1+</td>
<td>3+</td>
<td>3+</td>
</tr>
<tr>
<td>G16/RBM6 (4.0)</td>
<td>U50839</td>
<td>1+</td>
<td>Trace</td>
<td>Trace</td>
<td>2+</td>
<td>1+</td>
<td>2+</td>
<td>3+</td>
<td>3+</td>
</tr>
</tbody>
</table>

* Northern blot data for the following genes available in these references: MAPKAP2, Sitananadam et al. (55); IFDR2/SKMC15, Latif et al. (56); SEMA3B, SEMA3F, Sekido et al. (57); and HYAL2, this report and Coka et al. (79). The size of the most prominent transcripts in kb are given in parentheses next to the gene locus.
* Neg, negative.
* Gene19 is given as an example of one of the six genes detected by either a few ESTs or predicted by GENSCAN, which do not give mRNA signals on MTN blots (see Table 2 and “Discussion”).
large number of both SCLC and NSCLC cell lines and for some genes, paired normal and tumor DNA samples, were used. In addition, the coding sequences of the FUS1 and the BLU genes were sequenced completely in a large number of lung cancer samples, including paired tumor and normal tissue samples. The results for the eight-gene set in the 120-kb region show that the genes either had no mutations at all (CACNA2D2, PL6, 101F6, 123F2, and HYAL2), or the mutation rate was in the range of 5% (NPR2/g21, BLU, and FUS1; Table 1). The same absence or low frequency of mutations was detected in the 11 genes lying in the more telomeric ~250-kb portion of the 630-kb region with HYAL1, FUS2, and SEMA3B, each exhibiting a few mutations (Table 1). Many examples of the mutations detected in lung cancer cell lines are given in Table 3. Thus, this extensive mutational analysis (involving 1102 separate tumor sample/gene mutation tests; Table 1) did not pinpoint a strong candidate gene with a high frequency of mutation among either of the critical gene sets. This finding was unexpected and disappointing. Because it is possible that genes not involved in tumorigenesis may show a low frequency of mutations in common tumors related to a “mutator phenotype” expressed by tumors (64, 65), the finding of a low mutation rate in a candidate TSG(s) must be regarded with caution. Accordingly, other lines of evidence besides finding frequent mutations are needed to rule out positionally defined candidate TSGs. These include analysis for tumor-acquired promoter hypermethylation, gene transfer into tumor cells with tests for suppression of the malignant phenotype, and disruption of the candidate genes in “knock-out” mice. In addition, we now also have to consider the newly recognized class of haploinsufficient TSGs, where the presence of only one wild-type allele facilitates tumorigenesis (see “Discussion”).

**Informatics Analysis for Predicted Protein Functions and Orthologue Identification in Model Organisms.** Computational predictions of biochemical functions could also provide clues as to whether a particular gene could serve a tumor suppressor function by affecting cell growth and/or survival. Therefore, we next studied the biochemical functions and subcellular localization of the proteins encoded by the genes using a variety of computational tools. Information about protein function is a continuum that begins with the finding of homology between proteins within and between species, domain composition, functional motifs, and subcellular localization signals, extending ultimately to demonstration of function by biochemical analysis. Finding orthologous genes in model organisms (mouse, fly, worm, and yeast) permits the transfer of any functional annotation to the human gene in question; therefore, we performed an extensive search for orthologues of the resident candidate TSGs in these model organisms. The results of these computations are summarized and discussed in the annotations for each of the genes (see...
below). As expected, we found that all 19 genes have true murine orthologues discovered in mouse EST databases. These genes showed nearly 90–100% identity/similarity on the protein sequence level and about 80–90% on the mRNA sequence level (GenBank accession nos. are provided in the annotations). In the worm, highly likely orthologous pairs were identified for 14 of the 19 genes using stringent criteria of orthology, i.e., 30–50% amino acid sequence identity or similarity with >80% alignment of their entire amino acid lengths (45, 46). In the fly, highly likely orthologues were found for 10 of the 19 genes among the complete set (~14,000) of fly genes (67) (66). We noticed that orthologous pairs fall into three categories: common for both worm and fly, only present in the worm, and only present in the fly. Yeast genes sharing ~50% similarity in common domains/features were found for two of the genes, i.e., PL6 and NPR2/L2/Gene21, probably only the Schizosaccharomyces pombe counterpart (NPR2) of NPR2/g21 should be considered a candidate orthologous gene. The availability of the complete DNA sequences of yeast (50), worm (51), and fly (66) genomes makes it unlikely that we have missed any of the orthologous gene pairs for our TSG candidates in these three model genomes. We now provide annotations for each of the 19 genes found in the ~370-kb segment starting with the genes in the smaller ~120-kb sequence (Fig. 1).

The α2δ-2 calcium channel subunit gene, CACNA2D2, was discovered in silico by both finding EST matches with fragments of genomic sequence and by exons predicted by GENSCAN. The gene occupies ~140 kb of genomic space and is composed of at least 40 exons. It is expressed as a 5.5–5.7 kb mRNA. Three mRNA splice forms have been detected that code for two protein isoforms in several normal tissues. GenBank deposits AF040709 (mRNA isoform 3) and AF042792 (mRNA isoform 1) differ in the 5′ untranslated region and encode the same amino acid sequence (protein isoform I), whereas AF042793 (mRNA isoform 2) differs in the 5′ translated region and has a slightly different amino terminal amino acid sequence (protein isoform II). The expression of CACNA2D2 is reduced or absent in >50% of lung cancer cell lines, particularly NSCLCs. However, no mutations were detected in analysis of 60 lung cancer cell lines and 40 paired normal/SCLC tumor samples. The nucleotide sequence suggests that the gene encodes an auxiliary regulatory α2-δ subunit of calcium channels and joins the α2-δ-1 (previously A subunit) gene (67) as a new and second member of the α2-δ gene family. Three putative transmembrane helices predicted previously in the α2-δ-1 protein (67) were also predicted in both protein isoforms of the CACNA2D2 gene with the SPLITT 35 program (40). In addition, protein isoform I of the CACNA2D2 gene has another membrane helix at the very amino terminus. Using the TMHMM program (41), all three α2δ proteins were predicted to span the membrane only once at the amino (α2δ-2 isoform I) or at the carboxy termini (α2δ-1 and α2δ-2 isoform II), favoring the single-transmembrane model for the α2δ subunit proteins, which was verified experimentally for the α2δ-1 protein (67). A protein binding a VWA-like domain was discovered by the PFAM (37) program in the extracellular part at similar positions in all three α2δ subunit proteins (amino acid residues: 291–469 and 222–400 for α2δ-2 protein isoforms I and II, respectively, and residues 253–430 for the α2δ-1 protein). The VWA-like domain may facilitate the binding of the α2δ complex with the calcium channel α-1 pore forming subunit protein (67). The almost identical membrane topologies, similar domain structures, and posttranslational modifications of all three α2δ subunit proteins strongly support the identity of the new α2δ-2 gene as a member of the α2δ gene family. To provide experimental confirmation of this predicted function, through injection of CACNA2D2 cRNA into Xenopus oocytes, we have confirmed recently that CACNA2D2 acts as a regulatory subunit of voltage-gated calcium channels able to augment the function of all three pore-forming units (68). BLAST (36) searches in the mouse EST database detected two different nonoverlapping EST clones (accession nos. AA000341 and AA008996), which showed 91 and 85% cDNA sequence identity with CACNA2D2 (residues 2925–3421 and 4989–5391), respectively. These EST sequences showed only limited homologies to the murine α2δ-1 gene splice forms, indicating that they represent true orthologous sequences of the human CACNA2D2 gene. This was further corroborated by protein alignment of the 86-amino acid ORF encoded by mouse EST AA000341, which was 96% identical to the CACNA2D2 isoform I protein (amino acid residues 922–1005). The worm genome also contains two α2δ genes; by stringent criteria of orthology (i.e., ~50% identity/similarity with >80% alignment of their entire amino acid sequence) the worm gene, T24F1.6, appears to be the orthologue of CACNA2D2, whereas the second worm α2δ gene, UNC-36 (accession no. P34374), is the orthologue of the α2δ-1 gene (67). The UNC-36 phenotypes do not affect growth, and no phenotypes were yet reported for the T24F1.6 locus. The fly proteome (~14,000 proteins; Ref. 66) contains three α2δ proteins (accession nos. AAF53505, AAF53476, and AAF58335) of which the first contains a likely orthologue of α2δ-1 (44), the second is a likely orthologue of the α2δ-2 gene, and the third is a still-not-cloned human gene; all three have the VWA_DOMAINT. The yeast proteome (50) contains only one ion channel gene and appears to have no orthologues for either of the α2δ genes. Despite the lack of mutations, the absence of CACNA2D2 expression in many but not all NSCLCs with high CACNA2D2 expression in normal lung makes CACNA2D2 an excellent candidate TSG with the need for testing of function in tumor cells and study of acquired promoter hypermethylation as a method of inactivation of gene expression.

The PL6 gene was discovered manually by probing Northern blots with genomic fragments. The gene occupies 4.5 kb of genomic space, is composed of two exons, and expressed as a 2.2-kb mRNA in many normal human tissues including lung. The expression of PL6 is slightly reduced in some SCLC lines and abundantly represented in the human and mouse EST databases. No mutations were detected in 38 cell lines and 40 paired normal/SCLC tumor samples. By sequence analysis, PL6 encodes an integral plasma membrane protein [PSORT program (39)] with six [SPLIT program (40)] or seven to eight [TMHMM program (41)] transmembrane helices. The predicted cytoplasmic portion of the protein (the last 103 amino acids, residues 249–351) contains an OMPdecase domain [residues 274–298; PFAM, (39)] that may involve PL6 in protein-protein interactions and a bipartite NLS (residues 282–299; Ref. 39) that may guide it to the nucleus. The mouse orthologue was discovered in an EST (accession no. W96860), sequenced (our accession no. AF134238), and shown to be 92% identical on protein and 87% on cDNA levels. The worm F11A10.3 gene encoding a multidomain protein that aligns with PL6 and the aligned region contains the NLS and the OMPdecase domains, suggesting that it is the orthologue of PL6. The fly gene CG9536 product (450 residues; accession no. AAFO32388) is a likely orthologue of PL6 (48% similarity over the first 306 residues of PL6) and is also an integral membrane protein with seven to eight transmembrane helices. It has two HMW kinogen domains (residues 325–349 and 353–3760) but no NLS and OMPdecase domains. The yeast gene Yol107w product has substantial homology with PL6 but does not have the NLS and the OMPdecase domain. The absence of mutations and robust expression of PL6 in most lung cancers suggest that PL6 is an unlikely candidate TSG.

The 101F6 gene was discovered manually by screening arrayed...
cDNA libraries with cosmid LUCA12 DNA. The gene space of 3.2 kb contains four exons encoding a 1.5-kb mRNA. The gene is expressed in many normal tissues including lung, is highly expressed in SCLC and NSCLC cell lines, and is abundantly represented in the EST data bases. No mutations were detected in 40 cell lines and 40 paired normal/SCLC tumor samples. By sequence analysis, 101F6 encodes an integral plasma membrane protein [PSORT program (39)] with six [SPLIT program (40) and TMHMM program (41)] transmembrane helices with both termini in the cytoplasm. No other known domains or significant motifs were detected. The mouse orthologue was discovered in the mouse EST database (accession nos. AA285935, AA198541, and AA198960), sequenced (our accession no. AF131206), and shown to be 95% identical on the protein and 88% on the cDNA sequence level. No orthologous pairs were detected in the fly (66), worm, and yeast proteomes (50). The absence of mutations and robust expression of 101F6 in most lung cancers suggest that 101F6 is an unlikely candidate TSG.

The NPRL2/Gene21 gene was discovered in silico by finding both ESTs matches and GENSCAN predicted exons. The gene space of 3.3 kb contains 11 exons coding for a 1.5-kb mRNA with multiple splice isoforms that are expressed in many normal tissues including lung and testis and is abundantly represented in the EST databases. NPRL2/Gene21 is well expressed in SCLC and NSCLC lines except for the SCLC line NCI H1514. A frameshift mutation producing a stop codon was detected in 1 of 40 lung cancer cell lines. Sequence analysis shows NPRL2/Gene21 encodes a soluble protein that has a bipartite NLS (residues 62–79) and a protein binding domain, granulin (residues 86–98), predicted by PFAM (35, 37). The mouse orthologue was discovered in mouse EST databases (accession nos. AI037102, AA764527, AA709972, and W64225), sequenced (our accession no. AF131206), and shown to be 97% identical on protein and 90% on cDNA sequence levels. True orthologues were identified: in yeast (the Saccharomyces cerevisiae, GenBank accession no. P39923, and the hypothetical M, 47,000 protein in S. pombe, accession no. Z99163); in the fly (66), the CG9104 gene product (accession no. AAFA48677) with 65% similarity over the whole length of the NPRL2/g21; and in the worm (accession no. U61949) proteome databases. However, only the mouse orthologue contains the bipartite NLS (residues 62–79) and the granulin domain (residues 86–98). NPRL2/Gene21 mRNA is expressed in most lung cancers. The mutations in NPRL2/Gene21, particularly the stop mutations, indicate the need for further study of this gene as a candidate TSG.

The BLU gene was discovered manually (and serendipitously) using PCR primers (kindly provided by B. Vogelstein, Johns Hopkins, Baltimore, MD) to screen for the presence of the β-catenin gene (at the time recently assigned to chromosome region 3p21) in our cosmid contig. Although a PCR product was identified, DNA sequence analysis showed no sequence relationship of the product to β-catenin. This PCR product was used as a probe that identified a mRNA on Northern blot analysis, which then led to the subsequent isolation of the full BLU cDNA by library screening. The gene space of ~4.5 kb contains 11 (testis version) or 12 (lung version) exons coding for a 2-kb, alternatively spliced mRNA, well expressed in lung and testis but not expressed in all other tested human tissues. The EST databases contain a moderate number of hits, mostly from lung and testis cDNA libraries. The testis isoform contains 11 exons because of a complex selection of an alternative acceptor site. The testis-specific protein isoform contains a different amino acid sequence between residues 199 and 234 as compared with the lung-specific isoform; this change results in the loss of one of three PKC phosphorylation sites (residues 229–231). The expression in SCLC and NSCLC cell lines is reduced or virtually undetectable in 70% of tested lines. Three missense mutations were discovered in a sample of 61 lung cancer cell lines. The BLU protein is likely a soluble cytoplasmic protein and shares 30–32% identity over a stretch of 100–312 amino acids (residues 334–437 or 318–430) with proteins of the MTG/ETO family of transcription factors (69) and the suppressors (70) that may regulate entry into the cell cycle and suppress growth of colon carcinoma cells. The “Zn knuckle” motif involved in specific protein-protein interactions is part of this domain and is present in many proteins. No orthologous pairs were found in the worm and yeast proteomes (50). However, the fly genome (66) contains a true orthologue of BLU: the CG11253 gene product (accession no. AAF98580) is of similar size (451 residues), has 49% amino acid sequence similarity over the whole length of BLU, and also has a MYND finger domain (residues 412–448). Several other fly, worm, and S. pombe proteins share 35% identity with the MTG/ETO domain. The mouse orthologue was discovered in mouse EST databases (accession nos. A1595515 and AI429164), sequenced (our accession no. AF123386), and shown to be 89% identical on protein and 87% on cDNA levels. The loss of expression in most lung cancers and the occurrence of a few mutations make BLU an attractive TSG candidate requiring further functional and promoters methylation status studies.

The RASSF1/123F2 gene was discovered manually by screening gridded cDNA libraries with cosmid LUCA12 DNA. The gene space of 7.6 kb contains 5 exons coding for 2-kb, alternatively spliced mRNAs (“short” and “long” forms, 123F2SF and 123F2FL, that should now be referred to by Human Genome Organization-approved nomenclature as RASSF1C and RASSF1A, respectively) that are well expressed in all analyzed human tissues including lung. The RASSF1C/123FSF but not the RASSF1A/123F2LF mRNAs are well expressed in most lung cancer cell lines. The mRNA is well represented in EST databases from normal and tumor tissues. Using GENSCAN prediction programs, the RASSF1A/123F2LF splicing form was discovered using RT-PCR on mRNA with a difference in amino acid sequence in the NH2 terminus, giving a total amino acid sequence of 340 amino acids compared with 270 amino acids for the RASSF1C/123F2SF. The amino acid sequence of RASSF1A/123F2LF contains a predicted DAG binding domain also found in the related gene Nore1 but not found in the RASSF1C/123F2SF cDNA sequence. RASSF1A/123F2LF mRNAs also come in multiple tissue-related splicing forms, with slight differences in amino acid sequence, including forms for lung (RASSF1A, AF102770), heart (RASSF1D, AF102771), and pancreas (RASSF1E, AF102772). No mutations were detected in 40 paired normal tumor (SCLC/NSCLC) DNA samples (studied for RASSF1C/123F2SF and RASSF1/123F2 common region) and in 38 lung cancer cell lines (RASSF1C/123F2SF and RASSF1A/123F2LF, all regions). The RASSF1/123F2 protein is a soluble cytoplasmic protein that contains a Ras association domain (residues 124–218) discovered by the SMART (38) and PFAM (37) programs. Although not all Ras association domains bind RasGTP, the Ras association domain in the mouse parologue of RASSF1/123F2, Nore1 was found to bind RasGTP (71). The Nore1 protein also contains the PKC-C1 and DAG/PE domains, which are found in the RASSF1A/123F2LF predicted protein but not in the RASSF1C/123F2SF protein. Recently, the Kastan group has identified RASSF1/123F2 amino acid sequence (common to both the RASSF1C/123F2SF and RASSF1A/123F2LF proteins) as a potential phosphorylation target for ataxia telangiectasia mutated (72). The mouse orthologue of RASSF1/123F2 was discovered in mouse EST databases (accession nos. AA538990, AA161846, and AA466998), sequenced (our accession no. AF132851), and shown to be 97% identical on protein and 88% on cDNA sequence levels. In contrast, the human orthologue of the...
malignant phenotype, manuscript in preparation. Lerman, E. Zabarovsky, M. White, and J. D. Minna. RASSF1A in the 3p21.3 homozygous D. Randle, A. Virmani, S. Bader, Y. Sekido, F. Latif, S. Milchgrub, A. F. Gazdar, M. I. NORE1 and the rat MAXP1 (accession no. AF002521) genes is present in a single human EST (accession no. AA362184). Thus, RASSF1A/123F2 is part of the same gene family as (but not the orthologue of) NORE1 and the rat gene MAXP1. The worm gene, T24F1.3 (accession no. Z49912), encodes a 615-amino acid hypothetical protein that shares 33% identity and 53% similarity over 95% of the length of the RASSF1/123F2 protein. The T24F1.3 protein contains in the shared portion with RASSF1/123F2 the Ras association domain (residues 396–496), and in addition a PH domain (residues 1–53), and PKC-C1, DAG/PE binding domains (residues: 164–214), which are found in the RASSF1A/123F2LF predicted protein. The fly (66) and yeast (50) proteomes do not contain a gene with substantial homology to 123F2. The absence of expression of RASSF1A/123F2LF in many lung cancers makes this isoform an attractive candidate for further promoter hypermethylation and tumor-suppressing functional studies. In fact, recent studies by us have shown that RASSF1A/123F2LF promoter region CpG islands undergo tumor-acquired hypermethylation associated with loss of expression, and that forced re-expression of RASSF1A/123F2LF leads to suppression of the malignant phenotype.33

The FUS1 gene was discovered manually by screening cDNA libraries with a genomic fragment from the region of cosmids LUCA12 and LUCA13 showing sequence conservation by Southern blot hybridization and isolated as the fusion (FUS = ‘fusion’) junction of the ends part of a ~30-kb homozygous deletion in SCLC NCI-H524 linking LUCA12 with LUCA13 sequences. The gene space of 3.3 kb contains three exons coding for a 1.8-kb mRNA that is well expressed in all analyzed human tissues including lung and in 20 lung cancer cell lines. The mRNA is well represented in EST databases from normal and tumor cells. Three mutations were discovered in 79 lung cancer cell line DNAs leading to truncated products. The FUS1 protein (110 amino acids) is probably a soluble cytoplasmic protein with a high pI of 9.69; no domains or known motifs were detected by SMART (38) or PFAM (37) programs. The mouse orthologue was discovered in mouse EST databases (AA867009, AA473614, and AA672013), sequenced (our accession no. AF123387), and shown to be 93% identical on the protein level and 87% on the cDNA level. The fly proteome (66) does not contain a gene with substantial homology to FUS1. However, the worm gene, C09E9.1, shows 41% identity on global alignment and 43% identity over 83% of the FUS1 protein length and should be considered a candidate orthologue of FUS1. This small worm protein (123 amino acids) is predicted to have a bipartite NLS (residues 84–101) and weak similarity with DNA-directed RNA polymerase subunit A’ (accession no. P31813). The mutations found in FUS1 make it an attractive candidate for further functional TSG studies.

The HYAL2 gene along with HYAL1 was discovered manually by screening cDNA libraries with a genomic fragment from LUCA13 conserved across species in Southern blotting. Because these were the first two genes we isolated in our positional cloning effort, they were initially given the working names LUCA1 and LUCA2 (see GenBank deposits). With the discovery of their function as hyaluronidases, they should now be referred to as HYAL1 and HYAL2 and the LUCA1, 2 names reserved for future discovery of a lung cancer functional TSG. The gene space of 2.8 kb contains three exons that encode a 2.6-kb mRNA well expressed in all analyzed human tissues including lung, and in 20 lung cancer cell lines. The mRNA is well expressed in lung cancer cell lines except SCLC line NCI-H524 because of a small (~30 kb) homozygous deletion/rearrangement.34 HYAL2 is abundantly represented in EST databases from normal and tumor tissues. No mutations were detected in 40 lung cancer cell lines tested. The HYAL2 protein is a member of a large family of hyaluronidases (EC3.2.1.35), and in fact the expressed recombinant protein was shown to have enzymatic activity (73). PSORT predicts a signal peptide and cell surface and lysosomal sublocalizations (37, 39). Similarly, SMART predicts a signal peptide and a Ca2+ binding epidermal growth factor-like domain (residues 365–440; Ref. 38). Recently, the mouse orthologue (A3000059) was cloned and mapped to the syntenic region of mouse chromosome 9 between the microsatellite markers D9Mit183 and D9Mit17 (74). The worm gene, T22C8.2, encodes a similar size protein of 458 amino acids and shows 32% global identity and 50% similarity with the HYAL2 protein and is predicted to be an orthologous protein by the Orthologue program (48). The fly (66) and yeast (50) proteome databases do not have any members with homology to the hyaluronidase family of proteins.

HYAL1 was discovered along with HYAL2, manually by screening cDNA libraries with a conserved genomic fragment from cosmid LUCA13. It is another hyaluronidase with amino acid sequence homology to HYAL2 and HYAL3 (see below). The gene space of ~3.5 kb contains three exons coding for a 2.6-kb mRNA well expressed in all analyzed human tissues, including lung, and is abundantly represented in EST databases from normal and tumor tissues. However, it is not expressed in 18 of 20 lung cancer cell lines. Two missense mutations were detected in 40 lung cancer cell lines. The HYAL1 protein is a member of a large family of hyaluronidases (EC3.2.1.35) and in fact was shown to have enzymatic activity (accession nos. U03056 and U96078.1). Triggs-Raine et al. (75) identified two mutations in the HYAL1 alleles of a patient with newly described lysosomal disorder, mucopolysaccharidosis IX, a mutation that introduces a nonconservative amino acid substitution (Glu268Lys) in a putative active site residue and a complex intragenic rearrangement, 136del37ins14, which results in a premature termination codon. They reasoned that the mild phenotype engendered by these mutations was the result of redundancy resulting from the three tandemly located hyaluronidases HYAL1, HYAL2, and HYAL3 (discussed below). Thus far, no increased incidence of cancer has been reported in these kindreds. PSORT predicts a signal peptide and a cell surface and lysosomal sublocalizations (39). Similarly, SMART predicts a signal peptide and a visible Ca2+ binding epidermal growth factor-like domain (residues 357–430; Ref. 38). Recently, the mouse orthologue was cloned and shown to map to the syntenic mouse chromosome 9 region (accession no. AF011567; Ref. 76). The worm gene, T22C8.2, encodes a similar size protein of 458 amino acids and shows 31% global identity and 46% similarity with the HYAL1 protein and contains the same domains. It is predicted to be an orthologous protein by the Orthologue program (48). The fly (66) and yeast (50) proteome databases do not have a member homologous to the hyaluronidase family of proteins. The absent expression and occurrence of mutations make HYAL1 an attractive candidate for future promoter methylation and TSG functional studies.

The FUS2 gene was discovered manually by screening cDNA libraries with a genomic fragment occurring in cosmid LUCA14 that showed conservation in Southern blot cross species hybridizations. FUS2 was also present in the fusion junction genomic DNA clone isolated from the NCI-H524 30-kb homozygous deletion but was not involved or rearranged in this deletion but was given the “FUS” working name at the time of its isolation. The gene space of ~3.5 kb

---


contains an intronless, single-copy gene (accession no. AF040705) coding for a 1.9-kb mRNA expressed in normal human tissues including lung. However, an alternatively spliced form (accession no. AF040706) with one intron exists that results in the same predicted amino acid sequence. The alternatively spliced form contains an intron in the 5′ untranslated region, whereas the other form is intronless. The mRNA is well represented in EST databases from normal and tumor tissues. Four FUS2 missense mutations were detected in 78 lung cancer cell lines. The FUS2 protein was predicted to be a soluble nuclear protein [predicted by PSORT (39)] with interesting domains and motifs. SMART (38) and PFAM (37) programs predicted an acetyltransferase (GNAT) domain (residues 66–189) and a proline-rich domain (residues 239–262) that overlaps (residues 234–249) with the Wilms’ tumor protein signature. A Src homology 2 domain (residues 240–250) was detected by the BLOKS (77) program, which may explain the suggested weak similarity to furin-like proteases (accession no. AAC02732). The presence of these domains raises the intriguing possibility that FUS2 may be directly related like proteases (accession no. AAC02732). The presence of these domains raises the intriguing possibility that FUS2 may be directly involved in nuclear activities. However, Zegerman et al. (78) demonstrated recently that FUS2 functions as an N-acetyltransferase using a ping-pong mechanism with a specificity for substrates and is a soluble cytoplasmic protein. The worm protein, C56G2.15, shows 32% identity and 65% similarity on global alignment and should be considered a true orthologue of the FUS2 gene. As expected, it also contains all but the ZP predicted protein domains and is predicted to be a nuclear protein. Interestingly, the worm gene contains three small introns in contrast to the one or no introns of the human FUS2 gene. The mouse orthologue was discovered in mouse EST databases (accession nos. AA051756, AA051686, AI425576, and AA833145), sequenced and shown to be 69% identical on protein level and 87% on cDNA level (accession no. AF172275). The mouse mFUS2 protein contains an additional 28-amino acid stretch, and similar to the human FUS2 protein belongs to the hyaluronidase family of enzymes (EC3.2.1.35) and its human orthologue, PC4, is 39% identical and 60% similar to the human FUS2 protein and is predicted to have an acetyltransferase (GNAT) domain (-residues 92–217) and a proline-rich domain (residues 267–291). The Flybase (66) contains several ESTs (accession nos. AI064351, AI109425, and AI404849), which could be assembled into a partial cDNA coding for a 168-residue protein that is 39% identical and 60% similar to the human FUS2 protein and is predicted to have an acetyltransferase (GNAT) domain (residues 13–138). However, the fly proteome (66) does not have a true orthologue of FUS2, only several proteins with a GNAT domain. The occurrence of mutations and the demonstration of its biochemical activity make FUS2 and attractive candidate for future TSG functional studies.

The HYAL3 gene was discovered in silico by finding EST matches and sequence relationship to the HYAL1 and HYAL2 genes. It occupies 5.5 kb of genomic space and codes for a ~2.0-kb mRNA composed of two or three coding exons, expressed in several human tissues including lung and testis and well represented in EST databases. The protein belongs to the hyaluronidase family of enzymes (EC3.2.1.35) and thus represents the third member of this family in the region. Phylogenetically, it is closer to the worm hyaluronidase gene, T22C8.2, than the other members of the human family. No mutations were found in 40 lung cancer cell lines. No mouse orthologous sequences were found in the databases as of April 2000. HYAL3 RNA was not expressed in any lung cancer cell lines (data not shown); however, it has a very restricted pattern of expression in normal tissues and was not found by Csoka et al. to be expressed in normal lung (79). The lack of mutations and absent expression in normal lung make HYAL3 a less attractive TSG candidate.

The IFRD2/SKMC15/SM15 gene was discovered experimentally by screening cDNA libraries with conserved genomic fragments (56). GenBank refers to SKMC15/SM15 as IFRD2, and thus we will use the terminology IFRD2/SM15. The gene is ~6 kb and codes for a ~4-kb mRNA composed of 12 exons, expressed in several human tissues including lung (56). It is well represented in EST databases from normal and tumor cells. The IFRD2/SM15 protein is a soluble nuclear protein as predicted by PSORT (39) and contains a bipartite NLS at residues 115–132 predicted by ProfileScan (43). PFAM (37) discovered one Armadillo-β catenin-like repeat (residues 249–288), which suggests the possibility of involvement in APC signaling. No mutations were found in 63 lung cancer cell lines (56). The mouse orthologue was discovered in ESTs (accession no. W65790), sequenced and shown to be 93% identical on the protein level and 87% on the cDNA level. This true mouse orthologue of IFRD2/SM15 is different from the mouse gene mIFRD1/PC4, which has its own human orthologue located on chromosome 7q22–31 (80). Interestingly, mIFRD1/PC4 and its human orthologue, IFRD1, are not localized in the nucleus and probably are membrane proteins. IFRD2/SM15 and IFRD1/PC4 have different patterns of expression during mouse development (47, 80). The relation of IFRD2/SM15 and probably other members of the family (PC4 and IFRD1) to the IFNs is not really supported. The slightly shorter worm protein, F58B3.6 (accession no. Z73427), shows on global alignment 36% identity and 52% similarity to the SM15 protein and should be considered a potential orthologue. The fly gene CG3098 product (accession no. AAF51186) is shorter (324 residues) and has 43% similarity to 277 residues of SM15, has a NLS signal (residues 93–110), and should be considered a potential orthologue. The expression of IFRD2/SM15 and lack of mutations make it a less attractive TSG candidate.

The SEMA3B/SEMA A(V) gene was discovered experimentally by using DNA fragments from cosmids LUC14 to screen cDNA libraries and for capture of exons (57). The correct nomenclature for this member of the semaphorin family is SEMA3B [previously referred to as SEMA-A(V)]. It is composed of 17 exons spread over 8–10 kb of genomic space coding for a 3.4-kb mRNA expressed in several normal tissues including lung and testis and not expressed at all in 12 SCLC lines (Fig. 2; Ref. 57). It is well represented in EST databases from normal and tumor tissues. Three missense mutations were found in 39 lung cancer cell lines; all mutations were in NSCLCs. The mouse semaphorin A gene (accession no. X85990) is most likely the mouse orthologue of the SEMA3B gene (86% identity and 89% similarity on the protein level on global alignment; Ref. 48). Several mouse EST clones (accession nos. AI533114, AA518074, and AA466386) when translated showed 80–94% identity. The worm genome contains three semaphorin genes, of which the CeSema gene (accession no. U15667) shows 33% identity and 49% similarity over the whole length of the CeSema protein and could be considered an orthologous gene (48). The fly proteome (66) contains seven semaphorin proteins, of which the product of the Sema-2a gene (accession no. AA537990) is of similar size, predicted to be secreted, has a similar domain structure, and should be considered a potential candidate orthologue of SEMA3B. The SEMA3B protein is predicted by PSORT (39) to be an extracellular secreted protein. SMART (38) and PFAM (37) programs identify a signal peptide (residues 1–25), a PFAM:SEMA domain (residues 55–497), and one IGC2 domain (residues 587–646). Interestingly, the PFAM:SEMA domain is also present in the extracellular part of the MET and RON oncproteins belonging to the MET family of receptor tyrosine kinases, as discovered by the PFAM program (37). Thus, it will be reasonable to test the hypothesis that interaction of SEMA3B and SEMA3F (see below) proteins with these oncocenes may disrupt the activation of MET and RON and therefore convey a negative growth signal. The lack of
expression and mutation make SEMA3B an attractive candidate for methylation and TSG functional analysis.

The GNAI2 gene was discovered and cloned 12 years ago as part of studies on G proteins (21). GNAI2 was mapped to 3p21.3 by us and others and located to the central part of the 370-kb region (Fig. 1; Refs. 9, 17, 18, and 20). It is composed of 8 exons spread over ~22 kb of genomic space. The ~2.5-kb mRNA is well expressed in normal tissues and lung cancer cell lines and is well represented in EST databases. No mutations were found in 34 lung cancer cell lines. The product is a G protein localized to the endoplasmic reticulum. PFAM (37) predicts a G-o domain (residues 6–354) and an arf domain (residues 157–307). The GTPase function was established experimentally. The mouse orthologue (accession nos. RGM512 and P08752) is 98% identical and 99% similar on protein and 96% identical on cDNA levels. The worm orthologue (accession no. P51875) is 67% identical, and the fly orthologue (accession no. P20353) is 76% identical on the protein level. The newly predicted fly gene product G-oec47A (accession no. AAF58790) is identical in size, 72% similar in amino acid sequence, and should be considered a potential orthologue of GNAI2. The lack of mutations and continued expression of GNAI2 in lung cancers suggest it is an unlikely TSG candidate.

The G17 gene was discovered experimentally by cDNA selection onto cosmid LUCA17, which was then used to screen cDNA libraries. The gene space of 17 kb encodes 3-kb mRNA composed of 18 exons. The mRNA is expressed in several human tissues including lung, well represented in EST databases from normal and tumor tissues. No mutations were found in 38 lung cancer cell lines, and Gene17 was expressed in many lung cancers. The product is predicted to function as a plasma membrane amino acid transporter by homology to ABC transporters. It contains 10–11 transmembrane helices to function as a plasma membrane amino acid transporter by homology to ABC transporters. It contains 10–11 transmembrane helices. The mouse orthologue (accession no. 67–455) predicted by ProfileScan and PFAM servers. The mouse orthologue was discovered in several ESTs clones (accession nos. AI098786, AI048261, and AI466351), sequenced (our accession no. AI098786), and shown to be 90% identical on cDNA and 97% on protein levels. The yeast (50), worm, and fly (66) proteomes contain several amino acid transporter genes of similar size. The lack of mutations and continued expression make Gene17 an unlikely TSG candidate.

The GNAT1 gene was cloned 10 years ago (Ref. 22; accession no. X15088) and encodes the transducin protein isolated from the eye. We and others positioned the gene in 3p21.3, i.e., in the homozygous deletion overlap region close to the GNAI2 gene (Fig. 1; Refs. 18 and 20). The gene space of ~3.5 kb contains seven exons and encodes a 1.5-kb mRNA expressed abundantly in the retina and fetal heart tissues and T-cell lines. GNAT1 was not expressed in lung and lung cancer cell lines. No mutations were found in analysis of genomic DNA in 35 lung cancer cell lines. The mouse orthologue (48) was cloned (accession no. P20612) and is 100% identical on global protein alignment and is probably a true orthologue of GNAT1. The transducin protein is an α1 G-protein subunit localized in the endoplasmic reticulum. PFAM (37) predicts a G-α domain (residues 2–349) and an arf domain (residues 161–342). The restricted tissue distribution of expression of GNAT1 and lack of mutations make GNAT1 an unlikely TSG candidate.

The SEMA3F/SEMA-IV/SEEM IIIF gene, the second semaphorin gene in the region, was identified experimentally and cloned independently by several groups (Refs. 19, 57, and 63; accession nos. U38276 and U33920). The current correct nomenclature for this gene is SEMA3F (previously referred to as SEMAIIV and SEM IIIF). The gene space of ~28 kb encodes a 3- and/or 4-kb mRNA composed of 18 exons. The gene is well expressed in several normal tissues including lung (19, 57, 63) and is represented in EST databases from normal and tumor tissues and cell lines. Comparison of the cDNA sequences of Xiang et al. (63) and Roche et al. (63; U38276 and U33920) indicate that there are alternatively spliced forms of SEMA3F. SEMA3F was expressed in several but not all lung cancer cell lines (15 of 19; Ref. 57). No mutations were found in 30 lung cancer cell lines by Sekido et al. (57) testing 456 of the total 753 amino acids for mutations and in tests of the full ORF in 28 SCLC cell lines by Xiang et al. (63). ProfileScan (43) predicts an extracellular (secreted) protein containing several recognized domains. SMART (38) and PFAM (37) predict a signal peptide (residues 1–18), a single immunoglobulin-like domain (residues 619–680), two PFAM:SEMA domains (residues 57–153 and 184–529), and a proline-rich region (residues 715–726). As discussed for SEMA3B, the SEMA3F gene in the region raise the possibility that the SEMA3F protein interrupts signaling by the MET and RON receptor tyrosine kinase receptors. The mouse orthologue isoforms a and b were cloned (accession nos. AF80090 and AF080091) and also exist in several EST clones (accession nos. AA216900, W75532, and AA216899). The fly proteome (66) contains seven semaphorin proteins, of which the product of the CG4383 gene (accession no. AAF57999) is predicted to be secreted, has a similar domain structure, and could be considered a potential candidate orthologue of SEMA3F. This fly protein is similar to the worm gene CeSema (accession no. U15667), which is 29% identical and 48% similar to SEMA3F, and both are predicted to be secreted proteins. The loss of expression of SEMA3F and results from the Naylor laboratory (52), showing suppression of tumorigenicity of mouse A9 fibrosarcoma cells by a P1 clone in the SEMA3F region means this gene has to be considered in functional and methylation analysis.

The G15/RBM5 gene was discovered experimentally by screening arrayed cDNA libraries with cosmid LUCA22 (U73168) DNA. The same gene (called RBM5) was cloned by others (Ref. 81; accession no. AAD04159). The gene space of 18.5 kb encodes 18 exons expressed as a main 4-kb mRNA in several human tissues (a minor 7.5-kb species, and in some tissues, 1.5-kb and 2-kb species are also expressed). The mRNA is well expressed in several lung cancer cell lines and is represented in human and mouse EST databases from normal and tumor tissues. No mutations were found in 18 lung cancer cell lines. According to ProfileScan (43), PFAM (37) and SMART (38), the G15/RBM5 protein is a nuclear RNA binding protein featuring two bipartite NLSs (residues 34–51 and 708–725), two RNA-binding domains (residues 98–178 and 231–315), two zinc finger motifs, i.e., ZF RANBP (residues 181–210), and ZINC FINGER 2H2 2 (residues 647–677), and a recently recognized D111/G-patch DOMAIN (residues 749–786). G15/RBM5 has amino acid sequence homology with its immediate telomeric neighbor, G16/RBM6/NY-LU-12/DEF-3 (see below), indicating that they are part of the same gene family and arose through gene duplication. Studies by Drabkin et al. (82) have shown that both G15/RBM5 and G16/RBM6/NY-LU-12/ DEF-3 recombinant proteins can specifically bind poly(G) RNA homopolymers in vitro. The mouse orthologue was discovered in several EST clones (accession nos. AA574979, AA139814, and AI197332) that are 90% identical on cDNA and 97% on the protein level. The worm gene, T08B2.5 (accession no. AF000263), is 28% identical and 45% similar on global protein alignment and is probably a true orthologue (48). The fly proteome (66) contains three similar proteins, of which the CG4887 gene product (accession no. AAF51398) shows
43% similarity over 90% of the G15 length, has a similar domain structure, and is a likely orthologue of G15. The rat RNA-binding protein S1–1 (accession no. P70501) has a similar domain structure and is close in amino acid sequence to G15/RBM5; however, a rat EST clone (accession no. AI121056) on translation showed 97% identity to the last 37 amino acids of G15/RBM5 (outside the recognized domains) and probably represents the true rat orthologue of G15/RBM5. The lack of mutations and continued expression in most lung cancers means that G15/RBM5 is an unlikely TSG candidate.

The G16/RBM6/DEF-3/NYLU12 gene, the second gene encoding an RNA-binding protein in the homologous deletion region, was discovered experimentally by screening cDNA libraries with a conserved breakpoint genomic fragment from P1 clone 3938. Others have also cloned this gene (referred to as RBM6, NY-LU-12, and DEF-3). The identification of the NY-LU-12 clone through expression cloning using antibodies to detect the protein from a patient with lung cancer by means of the SERAX technology, and the DEF-3 clone as a differentially expressed gene during myelopoiensis, are of particular interest (81–83). The gene spans the distal breakpoint of the NCI-740 deletion (Ref. 18; Fig. 1), occupies more than 60 kb of genomic DNA, and encodes more than 16 exons expressed as a 4-kb mRNA in several human and mouse tissues, including lung. Alternatively spliced forms have been described (83). This mRNA is well expressed in several lung cancer cell lines and is represented in human and mouse EST databases from normal and tumor tissues. One mutation was found in 39 lung cancer cell lines. According to ProfileScan (43), PFAM (37), and SMART (38), the G16/RBM6 protein is a nuclear RNA binding protein. The following motifs and domains were predicted: a bipartite NLS (residues 1016–1033), one RNA-biding domain (residues 456–536), two zinc fingers, ZF-RANBP (residues 535–565), and ZF-C2H2 (residues 955–980), and a D111 DOMAIN (residues 1057–1094). Recombinant proteins containing the RNA recognition motifs of DEF-3(g16/NY-LU-12) and LUCA15 specifically bound poly(G) RNA homopolymers in vitro (82). The mouse orthologue was cloned (accession no. A100648) and also exists in several EST clones (accession nos. AA607276 and AA549397) with 90% identity on the cDNA level and 97% on the protein level. The fly proteome (66) contains three similar proteins, of which the CG4887 gene product (accession no. AAF51400) shows 43% similarity over 85% of the G16 length, has a similar domain structure, and is an orthologous orthologue of G16. The lack of mutations and continued expression in lung cancers make G16/RBM6 an unlikely TSG candidate.

**DISCUSSION**

We initiated these studies to gain insight into the molecular mechanisms of lung cancer pathogenesis with the goal of identifying and cloning a lung cancer TSG(s) residing in chromosome region 3p21.3. Our interest in this region began with our initial cytogenetic description of a chromosome 3p deletion abnormality in SCLC and subsequently in NSCLCs >17 years ago (84–86). This was followed by our discovery of 3p allele loss in SCLC (87–90). During this period, many groups also described 3p allele loss in lung cancer and other common cancers and found 3p cytogenetic abnormalities (9). Karyotyping, comparative genomic hybridization, interphase FISH (summarized in Ref. 91), and identification of homologous deletions, in conjunction with numerous molecular allelotyping studies, have identified a minimum of four separate consensus regions of 3p allele loss, i.e., 3p25-p26, 3p21-p22, 3p14.2 (FHH region), and 3p12 (U2020 homozygous deletion region; Refs. 6, 8, 9, 11, 17, 19, 51, and 92–95). A more recent study with very high density allelotyping, as well as results from multiple groups, have identified at least eight different 3p LOH sites (14). To aid in this TSG positional cloning effort, the homozygous deletions were precisely mapped at 3p12 (U2020 region; Refs. 9, 11, and 96–98), 3p14.2 (FHH region; Refs. 93 and 95), 3p21.3 (6, 9, 10, 17–19, 51), and 3p21.3–3p22 (94, 99). Allelotyping of histological subtypes of lung cancer showed differences in allele loss patterns at several chromosomal sites in addition to 3p between SCLC versus NSCLC histological types (8, 14). In addition, both SCLCs and squamous cell lung cancers usually undergo allele loss of a large part of the 3p arm. In contrast, adenocarcinomas of the lung tend to have much smaller regions of 3p allele loss (6, 8, 9, 14, 16, 91).

The 3p21.3 location was chosen for the current effort based on allele loss mapping studies showing that this is perhaps the first chromosomal region with allele loss in preneoplastic lesions and even in the histologically normal bronchial epithelium of current and former smokers (12, 14, 16, 100). Thus, TSG(s) residing in this region are likely to play a causative role in the earliest steps of lung cancer pathogenesis. In addition, the presence of overlapping/nested homologous deletions in lung and breast cancer and functional studies demonstrating the presence of a tumor phenotype-suppressing function gave us both positional and functional reasons to pursue this search (6, 9, 10, 17, 19, 20, 52–54). This exact region has also been identified as a site frequently undergoing loss in peripheral blood lymphocytes of lung cancer patients treated *in vitro* with the tobacco smoke carcinogen benzo(a)pyrene diol epoxide (101). We first obtained a physical map of the homozygously deleted area, which we then covered with a cosmid and P1 clone contig (Fig. 1 and Ref. 18). Through the auspices of the NCI, we were able to have this ~630-kb clone contig sequenced by the Sanger Genome Center (Hinxton, United Kingdom; cosmids LUCA1-LUCA11) and the St. Louis Genome Center (LUCA12-LUCA22, P3938). The cosmid DNA sequence accession nos. are given in Fig. 1, and the DNA sequences are available from GenBank. These DNA sequence have been condensed to several contigs (NT_002322, NT_000067, and NT_000069).29

We identified genes in the 630-kb region by experimental and informatics methods (discussed above), bringing the total number of verified resident genes to 25. In addition, several candidate “genes” were predicted by BLAST EST hits, GENSCAN and XGRAIL. Gene prediction software that we could not yet confirm as being real genes as defined by either detection of a mRNA on Northern blots or the presence of a cDNA with exon/intron structure and a reasonable ORF. With the complete genomic DNA sequence, all of the genes could, in retrospect, be identified by informatics methods “in silico.” The 19 genes residing in the largest homozygous deletion overlap region (from cosmid LUCA6 to part of P1 clone P3938) were analyzed by experimental (expression and mutational) analyses and computational methods (for predicted function using web-based servers) for information supporting their candidacy as a TSG. The extensive search for mutations in the 19-gene set did not identify a single gene with a high frequency of mutation in the analyzed lung tumors but did identify *NPRI2/Gene21*, *BLU*, *FUS1*, *HYAL1*, *FUS2*, and *SEMA3B* as candidates for further study because of the occurrence of a few mutations. In addition, a significant frequency of absent mRNA expression in lung cancer cell lines was found for the *CACA2D2*, *RASSF1A*, *HYAL1*, *BLU*, and *SEMA3B* genes, also indicating their possible involvement in lung carcinogenesis as candidate TSGs with homologous inactivation in tumors. The combination of mutations and absent expression in *BLU*, *HYAL1*, *SEMA3B*, and for the *RASSF1A*/*SEMA3B* mRNA isoform suggests special attention be paid to these genes. With the recognition of tumor-acquired promoter hypermethylation as a frequent mechanism of gene inactivation, those genes with loss of expression need to be studied for such acquired promoter changes (23). In addition, functional information documenting tumor suppressor or other relevant activity (e.g., DNA repair) will be required to

---

**3p21.3 TUMOR SUPPRESSOR GENE POSITIONAL CLONING**

---
provide a convincing argument for TSG identification. Computational analyses predicting function for the encoded proteins among the 19-gene set did not reveal obvious homology to a known TSG, nor did it exclude genes from having a possible TSG function.

There are several obvious features of this 3p21.3 region that provide information on the evolution of this part of the genome. There are four examples of gene duplication: the G proteins GNA12 and GNA17 (40); the hyaluronidases HYAL1, HYAL2, and HYAL3; the semaphorins, SEMA3B and SEMA3F; and the RNA binding proteins encoded by Gene15/RBM5 and Gene16/RBM6/DEF-3/Ny-Lu-12 genes. Each of these duplications (or triplications in the case of the hyaluronidase genes) occurred within 50–100-kb regions. In the case of the G-protein and semaphorin gene regions, a complex or sequential duplication must have occurred to account for the orientation of these four genes. In the case of Gene16 and Gene15 and the hyaluronidase genes, it appears tandem duplications occurred. Other genes (e.g., FUS2 and Gene17) then arose between these gene clusters. These duplicated genes pairs (or triplets) all have about the same amino acid sequence homology (~40–50%) to their duplicated partners, indicating that these duplications arose at approximately the same time. Also with the exception of the G-protein genes (showing 67% homology), the hyaluronidase, semaphorin, and RNA binding protein encoding genes all have ~30% amino acid sequence homology with their orthologous/homologous worm or fly genes. In each case, the worm or fly genomes have only one copy of the gene and not two, as are present in the human and mouse genomes.

The absence of frequent mutations in the resident candidate TSGs suggests we have to consider several other possibilities:

(a) We need to consider the possibility that the homoygous deletions only represent random genetic damage from tobacco smoke carcinogen exposure or the presence of a previously unknown fragile site, and thus, this specific 3p21.3 region does not contain a TSG. There are no obvious ways to prove or disprove that the three overlapping homoygous deletions in the SCLCs are random deletions. However, the recurrent presence of allele loss in this very specific region in multiple lung cancers and respiratory preneoplastic lesions and the occurrence of a homoygous deletion in a breast cancer not exposed to tobacco smoke carcinogens, as well as the microcell hybrid data indicating a tumor suppressor function in this region, all argue against this possibility (10, 14, 16, 19, 52–54, 100).

Homoygous deletions have also been crucial in the identification of several TSGs in other tumor types such as p16 (102), SMAD4 (103), BRCA2 (104), and p53 (105). Also, homoygous deletions have been found in lung cancer for these and other TSGs including RB, p53, p16, FHIT, and PTEN (33, 93, 95, 106–108). Finally, in recently completed genome-wide scans of lung cancers using >600 polymorphic markers, we have found only a few chromosomal sites with recurrent homoygous deletions that include 3p12 (U2020 region), 3p14.2 (FHIT region), 3p21.3 (the current region under study), 9p21 (p16/INK4 region), 10q (PTEN region), and a few other regions [Girard et al. (109) and Refs. 8, 11, 14, 33, and 95]. These studies include over 40 markers distributed over chromosome arm 3p. Thus, in our experience, the occurrence of recurring homoygous deletions in the same chromosomal region are very uncommon and appear to target TSGs.

(b) It is possible that we have not identified all of the genes residing in the critical region. Although this remains a possibility, the combination of extensive computational and experimental approaches (screening multiple cDNA libraries with cosmid clones and using the cosmid clones for exon capture) would argue against this possibility. No matches in EST databases suggestive of other genes have been detected in regular searches using the intergenic and intronic sequences. In addition, the gene identification programs GENSCAN (27) and XGRAIL (49) predicted only five candidates that we could not confirm as real genes (i.e., that did not give positive signals on Northern blots or show several matches in searches of the EST databases). In contrast, GENSCAN accurately predicted the location and nearly all of the intron/exon structures of the genes we did identify. Related to the gene detection question is the possibility that there is a tissue-specific splicing isoform that we have not detected, and it is this isoform that is mutated or has altered expression. Because of this possibility, we have particularly scrutinized GENSCAN exon predictions, ESTs mapping to intragenic regions not covered by our current cDNA sequences, and poly(A) Northern blots containing normal lung RNA for other possible mRNA forms. In fact, this methodology has allowed us to detect many different mRNA splicing variants, including those for at least 8 of the 19 genes including CACNA2D2, NPRL2/Gene21, BLU, RASSF1/123F2, HYAL1, FUS2, SEMA3F, and Gene16/RBM6/DEF-3/Ny-Lu-12. Our detailed gene analysis, coupled with an ever-enlarging EST database including information from many different tissues, will facilitate further mRNA splicing form detection. However, our analysis was not able to rule out the possibility that this 3p21.3 region encodes one or more mRNA-like noncoding RNA type genes that may function as TSG(s). We do not know of any currently available informatics approach that can identify such genes. Alignment of the human genomic sequence with the syntenic mouse sequence may provide clues to the location of additional genes such as these by detecting conserved sequences in the intergenic/intronic regions where these genes could reside (59).

Nevertheless, continued searches of the ever-growing EST database as well as comparing the sequence of the homologous region in the mouse (when the whole sequence is available) would provide additional support to the assumption that we have identified all of the genes in this region.

Another possibility is that we have identified the gene and that it is inactivated by a combination of mutation (uncommon) and loss of expression (frequent) mechanisms. Recently, methylation of the promoter region and allele loss leading to loss of expression of p16INK4A have been identified as a combination of mechanisms leading to the inactivation of this accepted TSG in lung cancer (108). In addition, several other TSGs, such as RB, VHL, and BRCA1, as well as a series of genes related to tumorigenesis such as E-cadherin (E-CAD), death-associated protein kinase (DAKP), glutathione S-transferase Pi 1 (GSTP1), O-6-methylguanine-DNA-methyltransferase (MGMT), and CACNA1G, a T-type calcium channel gene, also have been found to have their expression turned off by tumor-acquired promoter hypermethylation, some of which occur in lung cancers (6, 23, 108, 110–112). The definition of the precise structure for each of the 19 genes and their associated CpG islands in putative promoter regions identified by the current work will greatly aid in our current search for promoter hypermethylation as a mechanism of inactivation of expression of genes in this region. Recent evidence indicates this to be the mechanism for loss of expression of RASSF1A. In this regard, 5 of the 19 genes showed loss of expression in a large fraction of at least one histologically typical type of lung cancer (CACNA2D2, RASSF1A, SEMA3F, BLU, and HYAL1). A more complex version of this possibility is that more than one gene in this 3p213 region can serve as a TSG. Because >15 different mutations occurred independently of one another (in 6 genes, BLU, NPRL2/Gene21, FUS1, HYAL1, FUS2, and SEMA3B) and independently of the loss of expression (in 5 genes) in the 40–80 lung cancers tested, it is possible that nearly all of the lung tumor lines had either a mutation or loss of expression for one of these 8 genes. Thus, these 8 genes in particular need to undergo detailed functional tests to rule out their potential tumor suppressor function.  

Internet address: http://bio.cse.psu.edu/pipmaker/
A final possibility is that the TSG in this 3p21.3 region is 1 (or more) of the 19 cloned genes we have identified but does not conform to the classical two-mutation model for TSGs requiring homozygous inactivation in the tumor (113, 114). Instead, the gene may represent an example of the emerging class of haploinsufficient TSGs inferred from mouse models of human cancer (115–118). Recently it has been discovered that mice hemizygous (+/−) for p27/Kip1 or TGFβ show a higher spontaneous tumor rate compared with mice ++ at these loci and rapidly develop multiple tumors when challenged with carcinogens. Analysis of these tumors demonstrated that the remaining wild-type alleles for these genes remained intact and expressed in the tumors (115, 116). Mice hemizygous for the Pch gene frequently develop the rare tumor medulloblastoma with continued expression of one wild-type Pch allele. This provides evidence that haploinsufficiency of Pch leads to medulloblastoma in mice, and a similar circumstance appears to occur in humans as well (117). Similarly, haploid loss of the tumor suppressor Smad4/Dpc4 initiated gastric polyposis and cancer in mice, suggesting that haploinsufficiency of Smad4 is sufficient for tumor initiation (118). In addition, the autosomal dominant syndrome of familial platelet disorder with predisposition to acute myelogenous leukemia (FPD/AML and MIM 601399) in humans was found to be related to haploinsufficiency of the CBF2A gene (119). Thus, we need to consider the possibility that the TSG in the 3p21.3 region may be an example of an acquired abnormality in such a haploinsufficient gene.

This brings us to the discussion of lung cancer gene model(s). Inherited (120), initiating (114), gate-keeper (121), and cancer-causing (our term) genes follow the two-hit proposition for cancer causation (113, 114, 120–123). These genes (TSG or oncogenes) are usually discovered in an inherited family cancer syndrome through genetic mapping and positional cloning (113, 114). Consistent with the two-hit model, the corresponding tumors show loss of one allele and inactivating mutations of the remaining wild-type allele (in the case of a TSG). In the case of an inherited activated oncogene (such as MET), the tumors may show loss of the wild-type allele with a simultaneous duplication of the mutant allele or a second activating mutation in the remaining wild-type allele (124). The net result of these rate-limiting genetic changes (122, 123) is a homozygous status for these genes in tumors. In contrast, the haploinsufficient TSGs lose one copy in the germ-line or somatic tissues, predispose as a result of this loss to cancer but remain hemizygous (and express the wild-type allele) in tumors. Because the acquisition of other genetic alterations is still rate-limiting in developing cancer (122, 123), another mutation in a different gene would appear to be required in the haploinsufficient cancer model. It is tempting to assume that it could be a mutation in a second known TSG such as p53, RB, or p16, all of which are frequently mutated in many common cancers with 3p21.3 allele loss including lung cancers (6). There is no reason to exclude the possibility that classical TSGs and the haploinsufficient cancer-causing genes both could be involved in the development of sporadic malignancies arising from a common stem cell. In addition, it is likely that 3p21.3 allele loss is involved in the development of all histological types of lung cancer. Premalignant lesions in the lung have 3p21.3 LOH that is also found in a large fraction of SCLCs and NSCLCs (12, 15, 16, 100, 125). These include clones with 3p21.3 allele loss in histologically normal-appearing epithelium found in the lungs of current and former smokers (16, 100). Recently, we have been able to estimate the size of these clonal patches with 3p21.3 and other sites of LOH to be ~90,000 cells, as well as determine that the smoking-damaged lung contains thousands of such patches (126). We also know from studies of multiple markers that allele loss at classic TSG loci, such as at RB (13q14) and p53 (17p13), occurs after 3p allele loss at a later stage of carcinogenesis when histologically dysplastic or carcinoma in situ lesions are evident (12, 15, 16, 100). Lung embryology identifies lung stem cells as the primitive columnar epithelia derived from the primordial upper gut at days 40–45 of development (127). These columnar cells first differentiate into multipotent neuroendocrine cells (probably, the cells of origin of SCLC) and a variety of multipotent bronchial epithelia cells (probable the cells of origin of NSCLC). In fact, some lung cancers, within the same tumor, exhibit histological features of several lung cancer types, indicating a possible common stem cell of origin (128, 129). The embryology of the lung would therefore suggest that the same genes could be involved in both SCLC and NSCLC carcinogenesis.

In the light of these observations, the following model of lung cancer pathogenesis is proposed. Consequent to smoking damage, 3p21.3 allele loss occurs in thousands of different sites throughout the respiratory epithelium, leaving the putative 3p21.3 TSG haploinsufficient. These initiated cell(s) grow to form clonal patches of ~50–100,000 cells (~15–17 doublings) throughout the lung. The next hit could occur in another cancer-causing gene (such as RB, p53, p16, or “gene X”), leading to the next stages of invasive cancer. The RB and p53 genes are mutated in nearly 100% of SCLC samples, whereas mutations of p53 occur in 50% of NSCLC and some form of p16 inactivation (also inactivating the “RB pathway”) occurs in a very large fraction of NSCLCs (6). Alternatively, it is possible that the 3p21.3 TSG undergoes allele loss and then a second inactivating event (either a common mutation or loss of expression through promoter hypermethylation), which is required to allow the clonal outgrowth of these initiated cells. In any event, the nearly universal presence of 3p21.3 allele loss in SCLC and squamous cell carcinomas of the lung and its occurrence in >50% of adenocarcinomas of the lung suggest that this alteration is an obligatory rate-limiting step in the pathogenesis of many lung cancers (6, 8, 14, 16). In this regard, it will be very interesting to study SCLCs arising in the rare families segregating a mutant RB allele in the germ-line in which the majority of affected individuals surviving to adulthood have developed SCLC to see whether tumors arising under these conditions also have to undergo 3p21.3 allele loss (130). It should also be noted that allele loss that includes 3p21.3 and immediately surrounding 3p21 regions would lead to the hemizygotic state for several other possible cancer-predisposing genes residing in the area, including, MLH1, TGFBR1, β-catenin, RON, and Wnt5, which also could contribute to malignant transformation. Functional testing by gene transfer into lung cancer cells and gene disruption strategies in mice are now necessary to test this model and identify the putative 3p21.3 TSG(s).

Studies of familial lung cancer risk, including data from lung cancer occurring in young nonsmoking individuals, are compatible with Mendelian codominant inheritance of a rare major autosomal gene that produces earlier age of onset of lung and probably other cancers (7, 131, 132). A new national consortium (Genetic Epidemiology of Lung Cancer) has been formed that has accumulated enough families with informative cases with an apparently inherited predisposition to lung cancer to begin genetic linkage analysis. In due time, a classical TSG(s) (with homozygous inactivation in tumors) segregating in these lung cancer pedigrees could also be discovered. A causative role for this inherited lung cancer gene in the origin of sporadic lung cancers would also have to be ascertained. Linkage to markers in this 3p21.3 region can be readily tested, and if found, the genes we have reported in this study would need to be screened for germ-line mutations. In any event, the early involvement of chromosome 3p21.3 allele loss in premalignant lesions and sporadic lung cancers argues that one or more of the genes we have found should play a critical role in the development of common, sporadic lung tumors.
ACKNOWLEDGMENTS

We thank Dr. Alfred Knudson for many discussions that shaped this project, Dr. Richard Klausner for arranging the sequencing of the 630-kb 3p21.3 region, and both Genome Sequencing Centers, The Sanger Genome Sequencing Center (Hinxton, United Kingdom), and The Washington University Genome Sequencing Center (Saint Louis, MO) for sequencing the 630-kb contig.

REFERENCES


3p21.3 TUMOR SUPPRESSOR GENE POSITIONAL CLONING


The 630-kb Lung Cancer Homozygous Deletion Region on Human Chromosome 3p21.3: Identification and Evaluation of the Resident Candidate Tumor Suppressor Genes

Michael I. Lerman and John D. Minna

Cancer Res 2000;60:6116-6133.

Updated version
Access the most recent version of this article at:
http://cancerres.aacrjournals.org/content/60/21/6116

Cited articles
This article cites 115 articles, 38 of which you can access for free at:
http://cancerres.aacrjournals.org/content/60/21/6116.full#ref-list-1

Citing articles
This article has been cited by 84 HighWire-hosted articles. Access the articles at:
http://cancerres.aacrjournals.org/content/60/21/6116.full#related-urls

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, use this link http://cancerres.aacrjournals.org/content/60/21/6116.
Click on "Request Permissions" which will take you to the Copyright Clearance Center's (CCC) Rightslink site.