Transcriptome sequencing of malignant pleural mesothelioma tumors

David J. Sugarbaker*[†‡], William G. Richards*[†], Gavin J. Gordon*[†], Lingsheng Dong*[†], Assunta De Rienzo*[†], Gautam Maulik*[†], Jonathan N. Glickman[§], Lucian R. Chirieac[§], Mor-Li Hartman*[†], Bruce E. Taillon[¶], Lei Du[¶], Pascal Bouffard[¶], Stephen F. Kingsmore[‖], Neil A. Miller[‖], Andrew D. Farmer[‖], Roderick V. Jensen**, Steven R. Gullans[††], and Raphael Bueno*[†]

*International Mesothelioma Program, [†]Division of Thoracic Surgery, [§]Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, 75 Francis Street, Boston, MA 02115; [¶]454 Life Sciences, Inc., 20 Commercial Street, Branford, CT 06405; [‖]National Center for Genome Resources (NCGR), 2935 Rodeo Park Drive East, Santa Fe, NM 87505; **Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, Blacksburg, VA 24060; and [††]RxGen, Inc., 100 Deepwood Drive, Hamden, CT 06517

Communicated by Elliott D. Kieff, Harvard University, Boston, MA, December 31, 2007 (received for review July 6, 2007)

Cancers arise by the gradual accumulation of mutations in multiple genes. We now use shotgun pyrosequencing to characterize RNA mutations and expression levels unique to malignant pleural mesotheliomas (MPMs) and not present in control tissues. On average, 266 Mb of cDNA were sequenced from each of four MPMs, from a control pulmonary adenocarcinoma (ADCA), and from normal lung tissue. Previously observed differences in MPM RNA expression levels were confirmed. Point mutations were identified by using criteria that require the presence of the mutation in at least four reads and in both cDNA strands and the absence of the mutation from sequence databases, normal adjacent tissues, and other controls. In the four MPMs, 15 nonsynonymous mutations were discovered: 7 were point mutations, 3 were deletions, 4 were exclusively expressed as a consequence of imputed epigenetic silencing, and 1 was putatively expressed as a consequence of RNA editing. Notably, each MPM had a different mutation profile, and no mutated gene was previously implicated in MPM. Of the seven point mutations, three were observed in at least one tumor from 49 other MPM patients. The mutations were in genes that could be causally related to cancer and included *XRCC6*, *PDZK1IP1*, *ACTR1A*, and *AVEN*.

DNA sequencing | tumor mutations | lung cancer | bioinformatics | loss of heterozygosity

Because cancer arises as a consequence of multiple mutations, human cancer genomes are being sequenced to identify the mechanisms of tumorigenesis. Pilot sequencing studies include recent exon resequencing of tumors and cell lines that revealed somatic mutations in hundreds of genes not previously implicated in oncogenesis. These studies generally focused on a single class of mutations such as point mutations in coding regions of preselected candidate genes, and the results so far indicate that even within similar histological classes, tumors possess unique mutational profiles (1–3). However, there has rarely been an analysis of whether a mutated gene is actually expressed in the tumor cell nor has there been an attempt to use sequencing to identify other types of mutations such as chromosomal deletions or translocation (4, 5) or loss of heterozygosity related to epigenetic silencing (6, 7). Moreover, no unbiased deep sequencing analysis of all expressed genes in cancer tissues has been reported to date.

Malignant pleural mesothelioma (MPM) is an asbestos-related, rapidly fatal cancer. Its genetic basis is unknown but appears to involve multiple types of chromosomal abnormalities (5, 8–14). Central mechanisms underlying MPM are unclear, although MPM tumors evoke a strong inflammatory response thought to contribute to tumorigenesis (15). In addition, tumor cell survival promoted by TNF-α responsive antiapoptotic proteins such as Inhibitor of Apoptosis-1 (IAP-1) facilitates the resistance of MPM to most cytotoxic chemotherapeutic drugs (16). Expression profiling with microarrays has supported the general role of inflammation in MPM etiology and has provided molecular markers for diagnosis and prognosis (17). More extensive genomic analysis, as with deep sequencing of tumors, could identify potential targets for new biological drugs for this devastating cancer.

Newly developed DNA sequencing technologies (18, 19) allow for rapid, less expensive sequencing of large and complex genomes, but their utility in cancer mutation discovery remains unproved. Accordingly, we used whole-transcriptome shotgun 454 pyrosequencing (18) to characterize the full complement of individual tumor mutations and characterize mRNA expression levels in four MPM tumors and two controls. We developed methodology and informatic rules to reliably identify multiple types of mutations in expressed genes and applied this approach to discover 15 hitherto unknown mutations including somatic mutations, deletions, and epigenetic silencing.

Results

Tumor Transcriptome Sequencing and Analyses. To discover mutations in expressed genes of MPM specimens, we sequenced cDNA from tumors of four MPM patients (Patients 1–4). For comparison, we sequenced cDNA from an adenocarcinoma (ADCA) tumor of the lung (Patient 5) and from normal lung of a MPM patient (Patient 6). The process we used for mutation discovery and validation is schematically shown in supporting information (SI) Fig. 2. Briefly, polyadenylated RNA was prepared from microaliquoted (20) tumor specimens to ensure >85% tumor cell content. For each of the six samples, >260 Mb of transcriptome sequence were obtained by shotgun, clonal pyrosequencing using 454 technology (Table 1). Approximately 15 million cDNA sequence reads with lengths of ≈105 bp each were informatically mapped by using BLAST to human mRNA and DNA databases, and overall 98% of the reads matched known human RNA, DNA, and mitochondrial DNA sequences (see SI Table 4). Of the ≈39 Mb that did not map to human databases, ≈720 Kb mapped to chimpanzee, suggesting the existence of additional, previously uncharacterized expressed

Author contributions: D.J.S., W.G.R., G.J.G., S.R.G., and R.B. designed research; G.J.G., L. Dong, A.D.R., G.M., J.N.G., L.R.C., M.-L.H., B.E.T., L. Du, and P.B. performed research; L. Dong, B.E.T., L. Du, S.F.K., N.A.M., A.D.F., and R.V.J. contributed new reagents/analytic tools; D.J.S., W.G.R., G.J.G., L. Dong, A.D.R., G.M., J.N.G., L.R.C., L. Du, S.F.K., N.A.M., A.D.F., R.V.J., S.R.G., and R.B. analyzed data; and D.J.S., W.G.R., G.J.G., S.F.K., R.V.J., S.R.G., and R.B. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

[‡]To whom correspondence should be addressed. E-mail: dsugarbaker@partners.org.

This article contains supporting information online at www.pnas.org/cgi/content/full/0712399105/DC1.

Table 1. Global sequence, gene, and variant analysis

Parameter	Patient 1	Patient 2	Patient 3	Patient 4	Patient 5	Patient 6	Average
Total no. of DNA sequence reads*	2,516,790	2,494,009	2,507,617	2,847,854	2,515,787	2,907,917	—
Average read length	104.72	105.69	108.44	106.99	101.40	105.53	—
Total bases sequenced†	263,551,310	263,588,845	271,921,933	304,684,153	255,096,283	306,867,278	—
Number of Known RefSeq Genes observed							
Total reads mapped to 19,306 Known RefSeq Genes	1,757,631	1,588,315	1,627,354	2,365,665	1,719,634	1,751,853	1,801,742
No. of genes observed with ≥1 read	15,354	15,455	15,512	15,503	15,577	16,082	15,581
No. of genes observed with ≥10 reads	11,530	11,962	11,910	11,942	11,914	12,718	11,996
No. of genes observed with ≥ 20 reads	9,771	10,216	10,114	10,208	10,180	10,922	10,235
No. of genes with ≥ 100 reads (≈4–5× coverage)	3,728	3,853	3,650	3,765	4,094	4,180	3,878
Number of Known RefSeq Genes with known or previously uncharacterized SNVs compared with RefSeq RNA and dbSNP databases							
Coding Region SNVs							
With ≥ 1 SNV	800	734	659	831	1,155	882	844
With ≥ 1 sSNV	514	476	436	539	751	574	548
With ≥ 1 nsSNV	392	363	314	417	569	437	415
With a stop codon	2	6	5	3	6	4	4
Noncoding region SNVs							
With ≥ 1 SNV	2,855	3,124	2,995	2,866	2,564	2,923	2,888

Patients 1–4, MPM; Patient 5, ADCA; Patient 6, normal lung from MPM patient. The term "Known RefSeq Genes" refers to 19,306 well annotated known mRNAs available within the RefSeq mRNA database at NCBI.

*Grand total of DNA sequence reads, 15,789,974.

†Grand total of bases sequenced, 1,665,709,802.

sequences, and preliminary analysis (data not shown) suggested that they were largely noncoding sequences. No reads aligned to SV40 sequences or to any other viral or bacterial genomes.

For variant and mutation discovery, we analyzed only transcript sequences that mapped to the 19,306 well curated human reference mRNAs present in the "RefSeq mRNAs" database (www.ncbi.nlm.nih.gov/RefSeq/). We excluded the 9,456 LOC genes that have been identified informatically from the human genome sequence. The LOC genes have uneven coverage (likely misannotation of splice variants in RefSeq), an overabundance of putative single-nucleotide polymorphisms (SNPs), and premature truncation of alignments. Alignment of sequence reads to all 29,761 RefSeq mRNAs can be visually inspected (www.impmeso.org).

Gene Coverage and Expression. In each patient sample, ≈15,000 Known RefSeq Genes were detected by one or more reads (Table 1 and Fig. 1*A*). When all four MPM samples were pooled informatically, ≈17,000 or ≈90% of the Known RefSeq Genes could be observed (data not shown). Many of the 15,000 mRNAs observed in each sample were represented by only a few reads, likely indicative of either low-abundance cell types in the sample or transcriptional "leakage." A prior study using massively parallel signature sequencing (MPSS) of mRNAs in human tissues concluded that ≈50% of the genes in the human genome are expressed in a tissue (21). Consistent with this, we observed ≈10,000 genes at a threshold of 20 reads per gene (Table 1 and Fig. 1*A*), corresponding to a minimum of ≈1× coverage, assuming an average transcript length of 2 Kb and a read length of 105 bases. Gene expression profiling of these samples with GE CodeLink microarrays (data not shown) supported this number of expressed genes in each tissue.

To assess whether read counts (i.e., the number of reads that mapped to a given mRNA) can quantify relative transcript levels, we focused on six specific transcripts known to be differentially expressed in MPM and ADCA, which in combination can be used as a genomic test to distinguish the two tumor types (22, 23). The calculated geometric combination of the three expression ratios of these six genes obtained through real-time PCR or microarrays can be used to distinguish MPM from ADCA. The use of ratios of read counts of these six genes from the transcriptome sequencing data correctly distinguished MPM from ADCA specimens (Fig. 1*B*), supporting the use of 454 sequencing to quantitatively characterize the tumor transcriptome. These ratio results were independently confirmed in the same specimens by expression analysis with GE CodeLink microarrays and real-time RT-PCR (data not shown). These observations suggest that, at a minimum, read counts can provide a semiquantitative analysis of mRNA transcript levels in a tissue sample, but further validation work will be needed.

Rules for Single Nucleotide Substitution Variant (SNV)[‡‡] Discovery and Determination of Zygosity. Software systems for DNA sequence variant discovery based on Sanger chemistry and base-calling algorithms are inadequate for new DNA sequencing technologies that feature short read lengths, novel base-calling, quality score determination methods, and relatively poorly characterized error profiles (18). To facilitate visualization and automated analysis of 454 sequencing data, Alpheus, an internet-accessible software system that maps individual reads to the National Center for Biotechnology Information (NCBI) RefSeq mRNA database and identifies sequence level variants was created and is accessible at www.impmeso.org. Filter parameters include patient sample, gene name, read coverage, variant frequency, variant type, variant location and hyperlinks to NCBI sequence and gene function databases.

Assessment of putative sequence variants identified by analysis of unfiltered 454 sequencing data revealed an unacceptably high number of false-positive SNVs. To minimize this problem, an empiric rule set was developed as a tool for true mutation discovery in human tumors. These rules require that the variant must have at least four reads covering the base position; be present in at least 30% of the total number of reads covering the variant; be of GS20 quality score ≥20 (18) for the

[‡‡] SNVs are defined as single base substitutions that differ from the human mRNA reference sequence obtained from the NCBI RefSeq mRNA database. SNPs refer to those SNVs either present in the NCBI dbSNP database or observed in the patient's nontumor DNA or cDNA from blood.

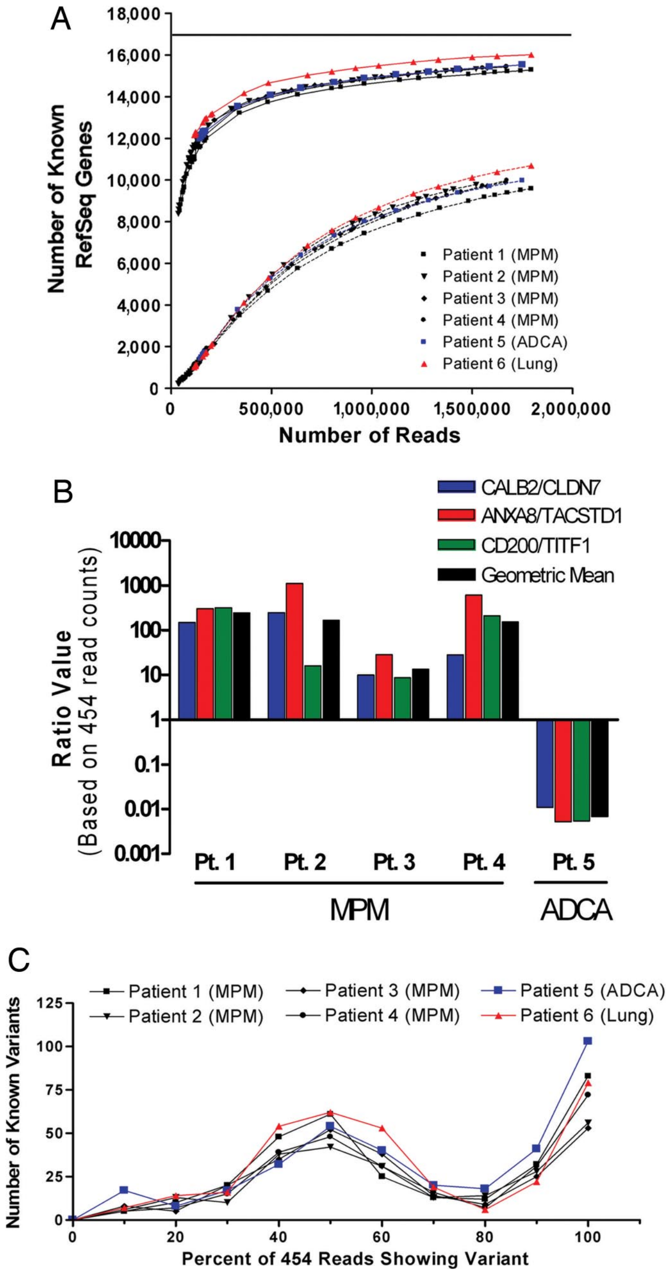

Fig. 1. Transcriptome characteristics. (*A*) Number of Known RefSeq Genes detected by at least 1 read (solid lines) and 20 reads (dashed lines) as a function of increasing depth of transcriptome sequencing (i.e., Number of Reads) for the six patient specimens. The horizontal asymptote represents ≈17,000 of the Known RefSeq Genes detected by at least one read in any of the four MPM samples, which encompassed a total of 7 million reads. (*B*) Classification of tumor specimens using read counts to calculate gene expression ratios for six known diagnostic genes and their geometric mean (23). Ratios correctly identified each tumor type (i.e., >1, MPM; <1, ADCA). (*C*) Analysis of percentage of reads containing known coding region SNVs in the six tissue samples. Known variants were selected based on >16 reads of coverage in the region of interest (see SI Table 6 for data). The distribution of reads ≈50% showing heterozygous expression of the variant is consistent with a binomial distribution.

relevant nucleotide; be observed in reads obtained from both orientations; and be located within a read that is >90% identical along its entire length to the target RefSeq mRNA sequence. These rules exhibited 96% sensitivity in identification of 2,465 well annotated SNPs found in dbSNP (www.ncbi.nlm.nih.gov/projects/SNP/) among 1,415 genes with ≥4× coverage in the normal lung control sample (SI Table 5), and 100% specificity by confirmation of 94 SNVs that were independently confirmed by using conventional Sanger methods. Less stringent rules identified additional putative variants but with diminished likelihood of being true positives.

By focusing on SNVs, clonal sequencing technology has the potential to observe and quantify allelic differences in genes. Analysis of transcript-based allele frequencies of 350 well annotated coding region SNPs in abundant transcripts (>16 reads covering the SNP in each specimen, with the SNP present in dbSNP (www.ncbi.nlm.nih.gov/projects/SNP/) detected homozygosity and heterozygosity (Fig. 1*C* and SI Table 6). In most cases of heterozygosity, the wild-type and variant alleles were expressed at similar levels—note peak at 50% (Fig. 1*C*). For a subset of alleles, one variant was expressed over the other (i.e., >80%). This may variously reflect the homozygous presence of a SNP, preferential transcription/stability of one allele (24), or copy number variation (25). Although the small number of tumors evaluated precludes generalization, it appears that transcriptome analysis with 454 sequencing technology is suitable for gross analysis of allele copy number, particularly loss of heterozygosity (LOH) or duplication for heterozygous alleles.

SNVs and Mutations in MPM Tumors. Each of the six tissue specimens contained, 659–1,155 Known RefSeq Genes with at least one coding domain SNV (cSNV), many of which are known human polymorphisms based on their presence in dbSNP (www.ncbi.nlm.nih.gov/projects/SNP/). Interestingly, each of the four MPM tumors had 153–220 genes that contained at least one previously uncharacterized cSNV (Table 2). Previously undiscovered SNVs are of greatest interest, because they represent candidate mutations. *In toto*, the four MPM tumors contained 619 nonredundant, previously uncharacterized cSNVs (SI Table 7) and 2,369 known SNPs. Interestingly, the ratio of nonsynonymous (ns) SNVs to synonymous (s) SNVs in the four MPM tumor samples was 1.5 for previously unidentified variants and only 0.75 for known SNVs, suggesting that the pool of previously unidentified variants contained somatic mutations not subjected to site-specific purifying selection pressure (Table 2). Both known and previously uncharacterized SNVs were detected in UTRs at a substantially higher prevalence than in coding regions.

nsSNVs not present in dbSNP were explored further because they are considered more likely to be functionally relevant, tumor-related mutations (3). Twelve nsSNVs were common to all five tumor samples, but absent from the normal lung, and four were common to the four MPM tumors only. Upon sequencing the normal genomic DNA (gDNA) from these patients, all proved to be germ-line variants and not mutations, supporting observations that most somatic mutations are specific to individual tumors.

Next, we focused on nsSNVs unique to each MPM specimen. Of ≈100–150 genes per MPM specimen with at least one previously uncharacterized nsSNV, 67 genes (12–20 per sample) had a total of 69 patient-specific previously uncharacterized nsSNVs after exclusion of highly polymorphic HLA and ABO loci (Table 2 and SI Table 8). Only one gene (*GOLGA8B*) had a potential mutation in two MPM patients, but the nsSNVs were present in different regions of the transcript, underscoring the unique mutational profile of each tumor. Sanger sequencing after PCR amplification was used to determine whether the 69 previously uncharacterized nsSNVs represented somatic tumor mutations in MPM. In other words, by sequencing the variant in the tumor cDNA, gDNA, normal adjacent cDNA, and/or gDNA from host peripheral blood lymphocytes (PBL), we differentiated between somatic mutations and germ-line variants, which enabled us to distinguish different types of mutations.

Table 2. Number of genes with candidate mutations in MPM tumors

	Known SNPs				Candidate mutations–previously uncharacterized SNVs			
	Patient 1	Patient 2	Patient 3	Patient 4	Patient 1	Patient 2	Patient 3	Patient 4
All variants								
	No. of genes with ≥1 SNP				No. of genes with ≥1 SNV			
Total	643	580	544	675	209	208	153	220
sSNP or sSNVs	432	391	384	463	101	108	69	102
nsSNPs or nsSNVs	286	264	229	300	130	122	105	142
	No. of coding region SNPs				No. of coding region SNVs			
Total SNVs	1,040	920	906	1,073	296	303	242	322
sSNP or sSNVs	584	523	522	623	120	133	95	126
nsSNPs or nsSNVs	456	397	384	450	176	170	147	196
Ratio ns/s	0.78	0.76	0.74	0.72	1.47	1.28	1.55	1.56
Patient-specific variants (excluding HLA and ABO genes)								
	No. of genes with ≥1 SNP				No. of genes with ≥1 SNV			
Total	58	49	47	84	36	32	15	36
sSNP or sSNVs	35	27	31	59	21	13	4	17
nsSNPs or nsSNVs	23	24	16	27	16	20	12	19
	No. of coding region SNPs				No. of coding region SNVs			
Total	63	53	50	91	38	34	16	36
sSNP or sSNVs	39	28	33	64	21	13	4	17
nsSNPs or nsSNVs	24	25	17	27	17	21	12	19

All four MPM tumors contained mutations, but each had a unique mutational profile (Table 3). Of 69 nsSNVs, 54 were present in the paired normal gDNA, indicating they were polymorphisms and not mutations. Although these 54 germline nsSNPs may predispose patients to MPM, they were not further explored. The remaining 15 nsSNVs (22%) were found to be tumor-specific variants representing multiple types of mutations including somatic mutations ($n = 7$), RNA editing ($n = 1$), and LOH due to chromosomal deletions ($n = 3$) and epigenetic silencing ($n = 3$), including X chromosome inactivation ($n = 1$, likely due to clonality). The seven nsSNVs that represented LOH mutations were heterozygous in the normal gDNA. LOH variant allele frequencies in the tumor gDNA differentiated silencing (heterozygous) and deletion (homozygous).

The frequency of the seven nsSNV somatic mutations was evaluated in 49 additional MPM tumors by genotyping cDNA and gDNA in the specific exons affected by the individual mutations. The *COL5A2* mutation (c2773t, NM_000393.3) in Patient 3 was found in two additional patients, both of whom had MPM tumors with nonepithelial histology (i.e., total frequency 3 of 53, or ≈6%). The *UQCRC1* mutation (g851a, NM_003365.2) was also found in two additional patients (≈6%), and the *MXRA5* mutation (c7862a, NM_015419.1) was found in one additional patient (≈4%). Thus, despite being relatively uncommon in MPM tumors, at least three of these

Table 3. Cancer-associated genetic lesions in MPM patients

MPM patient	Gene symbol	Accession no.	Chromosome	Variant	Amino acid change	BLOSUM score	Entrez gene name
Somatic mutation							
1	*ACTR1A*	NM_005736.2	10q24.32	a413 g	K → R	2	ARP1 actin-related protein 1 homolog
1	*MXRA5*	NM_015419.1	Xp22.33	C7862a	A → V	0	Matrix-remodelling associated 5
1	*PDZK1IP1*	NM_005764.3	1p33	c403t	T → I	−1	PDZK1-interacting protein 1
1	*PSMD13*	NM_175932.1	11p15.5	C1254a	L → M	2	Proteasome 26S subunit 13
1	*UQCRC1*	NM_003365.2	3p21.3	g851a	R → H	0	Ubiquinol-cyto *c* reductase core protein I
3	*COL5A2*	NM_000393.3	2q14-q32	c2773t	P → L	−3	Collagen, type V, α2
3	*XRCC6*	NM_001469.3	22q13.2–13.31	g956a	V → M	1	X-ray repair (Ku autoantigen, 70 kDa)
LOH due to deletion							
1	*LRP10*	NM_014045.3	14q11.2	G1998a	R → Q	1	LDL receptor-related protein 10
2	*C14orf159*	NM_024952.4	14q32.12	T1727 g	V → G	−3	Chromosome 14 ORF 159
2	*TM9SF1*	NM_006405.5	14q11.2	c2014t	R → W	−3	Transmembrane 9 superfamily member 1
LOH due to epigenetic silencing							
4	*C9orf86*	NM_024718.2	9q34.3	C2110 g	P → R	−2	Chromosome 9 ORF 86
4	*AVEN*	NM_020371.2	15q13.1	a784c	E → A	−1	Apoptosis, caspase activation inhibitor
4	*PSMD8BP1/NOB1*	NM_014062.1	16q22.3	A1074 g	Q → R	1	NIN1/RPN12-binding protein 1 homolog
LOH due to X inactivation							
2	*CXorf34*	NM_024917.4	Xq22.1	G1780a	G → R	−2	Chromosome X ORF 34
RNA editing							
4	*FLJ00312/CTGLF6*	XM_374801.3*	10q11.22	T1721a	D → E	2	Centaurin, γ-like family, member 6

*Replaced by accession no. XR_015233.

genetic mutations were present in 4–6% of a larger cohort of MPM tumors.

Discussion

This study demonstrates that transcriptome sequencing of patient tumors can result in discovery of previously uncharacterized human cancer mutations. By using an integrated approach that includes specimen enrichment for tumor cells, pyrosequencing, and rule-driven informatics, rare mutations were discovered among thousands of expressed genes. In addition to the advantages of speed and cost, this approach enriches for mutations in expressed genes and identifies multiple classes of mutations. The use of tumor tissue avoids artifactual mutations generated in cell culture. In addition, transcriptome sequencing provides information about mRNA expression levels not available with exon resequencing (1–3).

The identification of multiple types of genetic variants contributes to an expanded understanding of MPM and demonstrates that such an approach is essential to discovering the full complement of genomic changes associated with tumorigenesis. The four MPM patients had unique mutational profiles. Patient 1 had five somatic mutations and one LOH mutation, whereas Patient 2 had LOH mutations due to deletions in chromosome 14 and an X inactivation mutation that may have been a clonal event. Patient 3 had two somatic mutations only, one of which was present in two other patients' tumors. Patient 4 had three LOH mutations due to silencing and one due to RNA editing. This diversity of mutations emphasizes that defining correlations between tumor genotypes, histology, and various risk factors such as asbestos exposure will require sequencing a much larger cohort of MPM patients.

Of 15 mutations, seven were somatic point mutations, representing ≈1.75 ns somatic mutations per ≈3,800 Known RefSeq Genes sequenced with 4–5× coverage. By extrapolation to the ≈10,000 expressed Known RefSeq Genes detected in MPM transcriptomes (Table 1), it is estimated that individual MPM tumors harbor, on average, 6 transcribed genes with somatic mutations, or ≈10–14 genes with a nsSNV in the entire genome, in accord with a recent exon-resequencing survey of cancers (3). At this depth of sequencing, ≈38% of the expressed genes could be exhaustively analyzed for mutations. Additional sequencing and better characterization of the LOC transcripts will be necessary to characterize the full mutational spectrum of each tumor.

For the 15 mutated genes observed in MPM, this study provides the evidence that they can be mutated in cancer, in keeping with recent mutational surveys that also uncovered many previously uncharacterized mutated genes in other tumor types (1–3, 28). However, is there evidence that these mutations could be functionally meaningful? Mutated genes often exhibit abnormal levels of expression, and a retrospective analysis of published MPM profiling data revealed that most of these 15 mutated genes are over-expressed in a majority of MPM tumors [(23) data not shown]. A literature survey of gene function and expression in tumor cells reveals that the seven genes affected by somatic mutations (Table 3) are plausibly related to oncogenesis, although additional functional studies are needed to examine the physiological relevance of the observed mutations.

Of particular note, the protein product of *XRCC6* (*Ku70*) forms a heterodimer with that of *XRCC5* (*Ku80*) and mediates the repair of DNA double-strand breaks via nonhomologous end-joining. In non-small-cell lung cancer *XRCC5* is often hypermethylated and underexpressed at the mRNA and protein levels and is generally correlated with p53 changes (5). The g956a mutation that was observed in the *XRCC6* gene results in a V296M amino acid substitution in a protein region that directly contacts Ku80 (27). This specific amino acid substitution could be a cancer-driver mutation based on computational analysis of protein domain structure (28).

ACTR1A encodes the most abundant subunit of dynactin, which is associated with transport of p53 to the nucleus (29). Disruption of this complex via mutations in *ACTR1A* could potentially result in p53 inactivation, which is intriguing, given the absence of known inactivating mutations in p53 in MPM tumors. *PDZK1IP1* is overexpressed in human carcinomas of diverse origin and exhibits a tumor-suppressor phenotype in cultured colon cancer cells by negatively affecting proliferation and tumor growth (30–32).

COL5A2 encodes the α-chain for a low-abundance fibrillar collagen that is up-regulated in colon cancer (33) and normally has antitumor effects in breast cancer, including the induction of apoptosis (34). *UQCRC1* is a component of the mitochondrial ubiquinol–cytochrome-*c* reductase complex that was mutated in three MPM patients. It is known to be overexpressed in breast and ovarian cancer and has been suggested to play a role in tumorigenesis (35). The *PSMD13* gene encodes subunit 11 of the 26S proteasome (36), which is the target of a new class of anticancer drugs (37). Although functionally uncharacterized, *MXRA5* is overexpressed in colon cancer (38).

In addition to somatic mutations, gene deletions, gene silencing, and RNA editing were identified as common lesions in the MPM tumors. Among these, LOH mutations were observed in three genes situated on chromosome 14 in Patient 2. This genomic region was previously implicated in MPM tumors (11), and a notable gene within this region is *AVEN*. The genetic lesion identified in *AVEN* is of potential interest because this gene impairs Apaf-1-mediated activation of caspases, and thus apoptosis (39). We and others (9, 15–17, 40) have previously identified other (non-Bcl-2) antiapoptotic survival pathways as being particularly important in MPM tumorigenesis and drug resistance, and *AVEN* was recently implicated in acute leukemias (41). Elucidating the functional relevance of the previously uncharacterized variant alleles rendered homozygous by LOH in MPM (Table 3) is a promising avenue for further exploration.

Although less well studied than somatic mutations and LOH, both X chromosome inactivation (42, 43) and RNA editing (44) (i.e., a posttranscriptional process that alters the information encoded in gene transcripts, in this case a nsSNV present in the mRNA but not the gDNA) have been previously linked to cancer but not MPM (45). The actual editing is site-specific and occurs through specific mechanisms that are thought to be altered in cancers (44). Interestingly, RNA editing has also been postulated to be responsible for at least some proportion of the SNPs deposited into dbSNP at NCBI (45).

Transcriptome pyrosequencing permits comprehensive, unbiased, mutational analysis of expressed genes. This technique can also provide additional genetic information, such as insertion and deletion (indel) variant identification, read-count-based gene expression profiling, SNP allele frequencies, haplotype frequencies, novel isoform identification, and relative isoform abundance. In addition, transcriptome sequencing yields a more comprehensive set of gene-tagging SNPs that will be of considerable utility in disease-association studies. Nonetheless, cancer mutations can arise in ways that are not evident from transcriptome sequencing, and uncovering these will require supplementary approaches, such as karyotyping or whole-genome sequencing, to provide a more comprehensive understanding of the mechanisms that underlie tumorigenesis.

This study confirmed the accuracy of pyrosequencing for 94 of 94 previously uncharacterized variants when the empirically derived filtering rules for SNV discovery were used and suggested that this overall approach could become a standard for discovery and validation of genetic variants and tumor mutations. Solid tumors represent a major cause of morbidity and mortality in developed nations. Therapies for advanced cancer are limited because of the genetic complexity and variability among tumors. Large-scale pyrosequencing of the tumor transcriptome may be useful for determination of patient-specific mutational profiles enabling discoveries that have the potential to impact individual patient care. We envision that this approach may ultimately form the basis of molecular subtyping of patients with cancer, allowing combinational multitherapy designed individually for each patient based on mutational profile.

Materials and Methods

Detailed methods are presented in *SI Text*. Tumors were harvested in the operating room from consenting patients, four representing the clinical spectrum of MPM and two controls, and immediately dissected to generate high-quality fresh-frozen specimens. To examine the prevalence of specific mutations discovered in these six tumors, 49 additional MPM specimens were selectively analyzed.

The selected tumor specimens were processed by using a microaliquoting technique (20) to identify and subselect, using cryosections, samples with high tumor cell content (>85%) and little necrosis. High-quality mRNA (Agilent Bioanalyzer RNA integrity number >7.8 for total RNA) was used (2 μg of polyadenylated RNA from each tumor were used to make cDNA; see *SI Text*) for 12 runs of shotgun 454 sequencing with 454 Life Sciences GS20 technology (18). An internet-based information resource was developed to perform MegaBLAST alignments against RefSeq mRNA and to permit the selection and display of all sequence variants and relevant metadata (www.impmeso.org). We also conducted stringent MegaBLAST searches of reads that did not map to the 19,306 Known RefSeq Genes against the 52,935 ''Main Genes'' in AceView (www.ncbi.nlm.nih.gov/IEB/Research/Acembly/index.html), the human genome, and the *Pan troglodytes* (chimpanzee) genome. We developed and applied rules to identify candidate mutations among the Known RefSeq Genes that were unique to each of the four MPM patients. All candidate mutations derived from the four MPM samples (Patients 1–4) were selected for confirmation and characterization by using conventional Sanger sequencing, both in the discovery samples and additional tumors (the validation set) and matched normal tissue from patients with MPM.

ACKNOWLEDGMENTS. We thank R. Barker, M. K. Boateng, H. Edenfield, J. Esposito, K. Gill, J. Godleski, A. Imrich, L. Kobzik, H. Mercer, M. Mani, L. Mukhopadhyay, and S. B. Van Oss (Brigham and Womens Hospital); J. Xu and D. Gurgul of the Enterprise Research IS Group at Partners Healthcare for their support and use of the HPC facilities; Ann Adams for editorial assistance; B. A. Desany, A. J. Goncalves, J. Tsai, and M. Benitez (454 Life Sciences); S. M. Baxter, C. F. Black, L. C. Mader, M. K. Myers, and S. M. Virk (National Center for Genome Resources); and M. Lawrence (RxGen) for technical help. This work was supported by the International Mesothelioma Program (www.impmeso.org). D.J.S. was supported by National Cancer Institute (NCI) Grants P20CA9057801-A1, U01 CA65170-07, and U24CA114725-01; W.G.R. was supported by NCI Grants P20CA9057801-A1 and U24CA114725-01; and R.B. was supported by NCI Grants R01 CA 120528 and R21/R33 CA 100315.

1. Sjoblom T, *et al.* (2006) The consensus coding sequences of human breast and colorectal cancers. *Science* 314:268–274.
2. Futreal PA, Wooster R, Stratton MR (2006) Somatic mutations in human cancer: Insights from resequencing the protein kinase gene family. *Cold Spring Harbor Symp Quant Biol* 70:43–49.
3. Greenman C, *et al.* (2007) Patterns of somatic mutation in human cancer genomes. *Nature* 446:153–158.
4. Balsara BR, *et al.* (1999) Comparative genomic hybridization and loss of heterozygosity analyses identify a common region of deletion at 15q11.1–15 in human malignant mesothelioma. *Cancer Res* 59:450–454.
5. Lee M, *et al.* (2007) Epigenetic inactivation of the chromosomal stability control genes BRCA1, BRCA2, and XRCC5 in non-small cell lung cancer. *Clin Cancer Res* 3:832–838.
6. Jones PA, Baylin SB (2007) The epigenomics of cancer. *Cell* 128:683–692.
7. Haber DA, Settleman J (2007) Cancer: Drivers and passengers. *Nature* 446:145–456.
8. Britton M (2002) The epidemiology of mesothelioma. *Semin Surg Oncol* 29:18–25.
9. Whitson BA, Kratzke RA (2006) Molecular pathways in malignant pleural mesothelioma. *Cancer Lett* 239:183–189.
10. Baldassarre M, *et al.* (2003) Dynamin participates in focal extracellular matrix degradation by invasive cells. *Mol Biol Cell* 14:1074–1084.
11. De Rienzo A, Jhanwar SC, Testa JR (2000) Loss of heterozygosity analysis of 13q and 14q in human malignant mesothelioma. *Genes Chromosomes Cancer* 28:337–341.
12. Huncharek M (1995) Genetic factors in the aetiology of malignant mesothelioma. *Eur J Cancer* 31A:1741–1747.
13. Musti M, *et al.* (2006) Cytogenetic and molecular genetic changes in malignant mesothelioma. *Cancer Genet Cytogenet* 170:9–15.
14. Pylkkanen L, *et al.* (2002) Concurrent LOH at multiple loci in human malignant mesothelioma with preferential loss of NF2 gene region. *Oncol Rep* 9:955–959.
15. Yang H, *et al.* (2006) TNF-α inhibits asbestos-induced cytotoxicity via a NF-κB-dependent pathway, a possible mechanism for asbestos-induced oncogenesis. *Proc Natl Acad Sci USA* 103:10397–10402.
16. Gordon GJ, *et al.* (2002) Inhibitor of apoptosis protein-1 promotes tumor cell survival in mesothelioma. *Carcinogenesis* 23:1017–1024.
17. Gordon GJ, *et al.* (2007) Inhibitor of apoptosis proteins are regulated by tumor necrosis factor-alpha in malignant pleural mesothelioma. *J Pathol* 211:439–446.
18. Margulies M, *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376–380.
19. Green RE, *et al.* (2006) Analysis of one million base pairs of Neanderthal DNA. *Nature* 444:330–336.
20. Richards WG, *et al.* (2007) Microaliquoting: A technique for precision histologic annotation and cell content optimization of frozen tissue specimens. *Biotech Histochem* 1–9.
21. Jongeneel CV, *et al.* (2005) An atlas of human gene expression from massively parallel signature sequencing (MPSS). *Genome Res* 15:1007–1014.
22. Gordon GJ, *et al.* (2002) Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Res* 62:4963–4967.
23. Gordon GJ, *et al.* (2005) Identification of novel candidate oncogenes and tumor suppressors in malignant pleural mesothelioma using large-scale transcriptional profiling. *Am J Pathol* 166:1827–1840.
24. Benz CC, *et al.* (2006) Altered promoter usage characterizes monoallelic transcription arising with ERBB2 amplification in human breast cancers. *Genes Chromosomes Cancer* 45:983–994.
25. Redon R, *et al.* (2006) Global variation in copy number in the human genome. *Nature* 444:444–454.
26. Weir BA, *et al.* (2007) Characterizing the cancer genome in lung adenocarcinoma. *Nature* 450:893–898.
27. Walker JR, Corpina RA, Goldberg J (2001) Structure of the Ku heterodimer bound to DNA, its implications for double-strand break repair. *Nature* 412:607–614.
28. Kaminker JS, *et al.* (2007) Distinguishing cancer-associated missense mutations from common polymorphisms. *Cancer Res* 67:465–473.
29. Galigniana MD, Harrell JM, O'Hagen HM, Ljungman M, Pratt WB (2004) Hsp90-binding immunophilins link p53 to dynein during p53 transport to the nucleus. *J Biol Chem* 279:22483–22489.
30. Kocher O, Cheresh P, Brown LF, Lee SW (1995) Identification of a novel gene, selectively up-regulated in human carcinomas, using the differential display technique. *Clin Cancer Res* 1:1209–1215.
31. Kocher O, Cheresh P, Lee SW (1996) Identification and partial characterization of a novel membrane-associated protein (MAP17) up-regulated in human carcinomas and modulating cell replication and tumor growth. *Am J Pathol* 149:493–500.
32. Kocher O, *et al.* (1999) PDZK1, a novel PDZ domain-containing protein up-regulated in carcinomas and mapped to chromosome 1q21, interacts with cMOAT (MRP2), the multidrug resistance-associated protein. *Lab Invest* 79:1161–1170.
33. Fischer H, Stenling R, Rubio C, Lindblom A (2001) Colorectal carcinogenesis is associated with stromal expression of COL11A1 and COL5A2. *Carcinogenesis* 22:875–878.
34. Luparello C, Sirchia R (2005) Type V collagen regulates the expression of apoptotic and stress response genes by breast cancer cells. *J Cell Physiol* 202:411–421.
35. Kulawiec M, *et al.* (2006) Proteomic analysis of mitochondria-to-nucleus retrograde response in human cancer. *Cancer Biol Ther* 5:967–975.
36. Hoffman L, Gorbea C, Rechsteiner M (1999) Identification, molecular cloning, and characterization of subunit 11 of the human 26S proteasome. *FEBS Lett* 449:88–92.
37. Adams J (2004) The proteosome: A suitable antineoplastic target. *Nat Rev Cancer* 4:349–360.
38. Zou TT, *et al.* (2002) Application of cDNA microarrays to generate a molecular taxonomy capable of distinguishing between colon cancer and normal colon. *Oncogene* 21:4855–4862.
39. Chau BN, Cheng EH, Kerr DA, Hardwick JM (2000) Aven, a novel inhibitor of caspase activation, binds Bcl-xL and Apaf-1. *Mol Cell* 6:31–40.
40. Gordon GJ (2007) Expression patterns of inhibitor of apoptosis proteins in malignant pleural mesothelioma. *J Pathol* 211:447–454.
41. Paydas S, *et al.* (2003) Survivin and aven: Two distinct antiapoptotic signals in acute leukemias. *Ann Oncol* 14:2045–2050.
42. Kristiansen M, *et al.* (2005) High incidence of skewed X chromosome inactivation in young patients with familial non-BRCA1/BRCA2 breast cancer. *J Med Genet* 42:877–880.
43. Li G, *et al.* (2006) Skewed X chromosome inactivation of blood cells is associated with early development of lung cancer in females. *Oncol Rep* 16:859–864.
44. Scholzova E, Malik R, Sevcik J, Kleibl Z (2007) RNA regulation and cancer development. *Cancer Lett* 246:12–23.
45. Eisenberg E, *et al.* (2005) Identification of RNA editing sites in the SNP database. *Nucleic Acids Res* 33:4612–4617.