nature genetics

Application of comparative functional genomics to identify best-fit mouse models to study human cancer

Ju-Seog Lee[1], In-Sun Chu[1], Arsen Mikaelyan[1], Diego F Calvisi[1], Jeonghoon Heo[1], Janardan K Reddy[2] & Snorri S Thorgeirsson[1]

Genetically modified mice have been extensively used for analyzing the molecular events that occur during tumor development. In many, if not all, cases, however, it is uncertain to what extent the mouse models reproduce features observed in the corresponding human conditions[1–3]. This is due largely to lack of precise methods for direct and comprehensive comparison at the molecular level of the mouse and human tumors. Here we use global gene expression patterns of 68 hepatocellular carcinomas (HCCs) from seven different mouse models and 91 human HCCs from predefined subclasses[4] to obtain direct comparison of the molecular features of mouse and human HCCs. Gene expression patterns in HCCs from *Myc, E2f1* and *Myc E2f1* transgenic mice were most similar to those of the better survival group of human HCCs, whereas the expression patterns in HCCs from *Myc Tgfa* transgenic mice and in diethylnitrosamine-induced mouse HCCs were most similar to those of the poorer survival group of human HCCs. Gene expression patterns in HCCs from *Acox1*$^{-/-}$ mice and in ciprofibrate-induced HCCs were least similar to those observed in human HCCs. We conclude that our approach can effectively identify appropriate mouse models to study human cancers.

The success of comparative sequence analysis in identifying and characterizing genomic regulatory regions with important functional roles is due to the fact that these regions evolve at a slower rate than less important regions[5–8]. Although many of the functional genomic elements are protein-coding sequences, a large number of conserved sequences are probably regulatory elements with roles in modulating gene expression[9,10]. We therefore hypothesize that if regulatory elements of evolutionarily related species are conserved, gene expression signatures reflecting similar phenotypes in the species would also be conserved. To test this hypothesis, we investigated whether comparison of global expression patterns of orthologous genes in human and mouse HCCs would identify similar and dissimilar tumor phenotypes, and thus allow the identification of the best-fit mouse models for human HCC.

We characterized gene expression patterns of 68 HCCs from seven different mouse models: two chemically induced (ciprofibrate and diethylnitrosamine, DENA)[11–13], four transgenic (targeted overexpression of *Myc*, *E2f1*, *Myc* and *E2f1*, and *Myc* and *Tgfa* in the liver)[14–16] and one knockout (*Acox1*$^{-/-}$)[17]. We first applied hierarchical clustering analysis of gene expression patterns to assess the relative similarities among different mouse HCC models. We identified three distinctive HCC clusters, indicating that gene expression patterns of mouse HCC are clearly heterogeneous (**Fig. 1**). As expected, ciprofibrate-induced HCCs and HCCs from *Acox1*$^{-/-}$ mice were closely clustered (cluster 3) and well-separated from the other mouse models. Ciprofibrate is a synthetic peroxisome proliferator that is a nongenotoxic hepatocarcinogen[12]. *Acox1*$^{-/-}$ mice develop HCCs due to accumulation of unmetabolized very long-chain fatty acids that serve as endogenous ligands of Ppara receptor[17]. Cluster 2 largely consisted of HCCs from *Myc*, *E2f1* and *Myc E2f1* transgenic mice, indicating that overexpression of *Myc* and *E2f1* may support similar signaling networks during hepatocarcinogenesis. HCCs induced by DENA, a genotoxic hepatocarcinogen, closely clustered with those from *Myc Tgfa* transgenic mice (cluster 1). This may indicate that gene expression patterns in cluster 1 reflect extensive chromosomal damage during tumor development, which known to occur in both DENA-induced liver tumors and liver tumors in *Myc Tgfa* transgenic mice[18].

Given the three distinctive subgroups of mouse HCC models, we sought to examine how well these models recapitulate human HCC phenotypes as defined by gene expression patterns. In our previous study using similar microarray technology, we identified two distinctive subclasses of human HCCs that are highly associated with the survival of individuals with HCC[4]. Because we used two different microarray platforms to study mouse and human HCC, we selected orthologous genes that were present in both microarrays by using curated mammalian orthology from The Jackson Laboratory. A total of 4,036 orthologous genes were present in both microarrays. We selected orthologous genes whose expression changed nontrivially for further analysis (1,650 genes). We then standardized gene expression ratios separately to a mean $\pm$ s.d. of 0 $\pm$ 1 in each data set. In hierarchical clustering analysis of the integrated data, the three

[1]Laboratory of Experimental Carcinogenesis, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892-4262, USA. [2]Department of Pathology, Northwestern University, the Feinberg School of Medicine, Chicago, Illinois 60611-3008, USA. Correspondence should be addressed to S.S.T. (snorri_thorgeirsson@nih.gov).

Published online 21 November 2004; doi:10.1038/ng1481

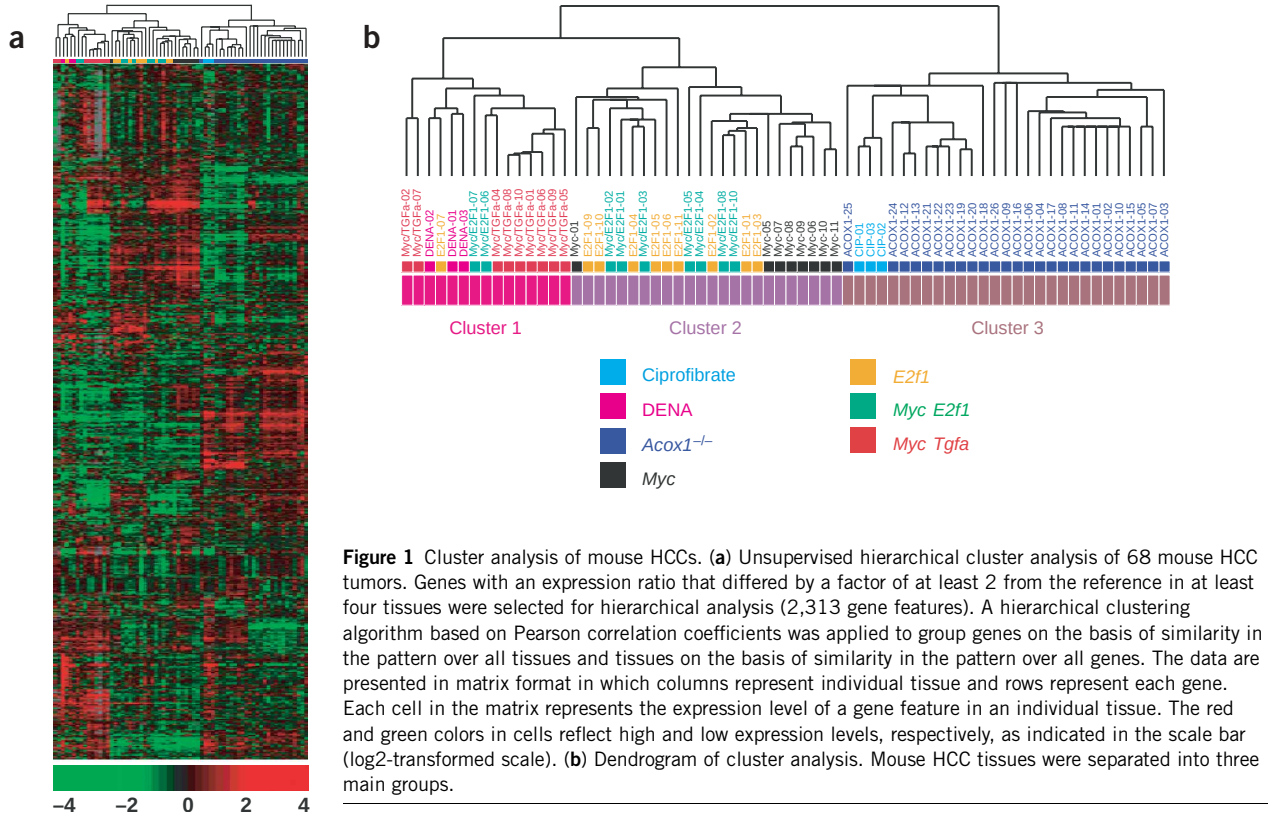

Figure 1 Cluster analysis of mouse HCCs. (**a**) Unsupervised hierarchical cluster analysis of 68 mouse HCC tumors. Genes with an expression ratio that differed by a factor of at least 2 from the reference in at least four tissues were selected for hierarchical analysis (2,313 gene features). A hierarchical clustering algorithm based on Pearson correlation coefficients was applied to group genes on the basis of similarity in the pattern over all tissues and tissues on the basis of similarity in the pattern over all genes. The data are presented in matrix format in which columns represent individual tissue and rows represent each gene. Each cell in the matrix represents the expression level of a gene feature in an individual tissue. The red and green colors in cells reflect high and low expression levels, respectively, as indicated in the scale bar (log2-transformed scale). (**b**) Dendrogram of cluster analysis. Mouse HCC tissues were separated into three main groups.

previously identified subgroups of mouse HCC and two subclasses of human HCC were still well separated from each other (**Fig. 2**). Gene expression patterns of HCCs from *Myc*, *E2f1* and *Myc E2f1* transgenic mice were most similar to those of the better survival group of human HCCs (subclass B), whereas the expression patterns of HCCs from *Myc Tgfa* transgenic mice and DENA-induced mouse HCCs were most similar to those of the poorer survival group of human HCCs (subclass A). Gene expression patterns of HCCs from *Acox1*$^{-/-}$ mice and ciprofibrate-induced HCCs were least similar to most human HCCs, and clustered with only a small fraction of them (**Fig. 2b** and **Supplementary Fig. 1** online). We observed similar results when we used the 'survival genes'[4] found among the orthologous genes for cluster analysis (**Supplementary Note** and **Supplementary Fig. 2** online). These data strongly suggest that hepatocarcinogenesis driven by peroxisome proliferation in mice proceeds through a carcinogenic pathway not frequently observed in humans, and they support previous studies suggesting that the human liver is insensitive to peroxisome proliferators[19,20].

We next applied supervised learning methods to validate the unsupervised cluster analysis. We applied five independent prediction methods to determine which of the mouse models might best mimic the human phenotypes. We used the gene expression data sets from the two subclasses of human HCCs to train prediction methods. All methods predicted that most HCCs from *Myc Tgfa* transgenic mice are relatively similar to subclass A, whereas HCCs from the other models are relatively similar to subclass B (**Table 1**). By χ^2 test of each predicted pattern, we determined that the predicted outcome of HCCs from *Myc Tgfa* transgenic mice significantly differs from that of the rest of models ($P < 0.005$), whereas the predicted outcome of HCCs from *Myc E2f1* transgenic mice does not differ significantly from those of HCCs from *Myc* or *E2f1* transgenic mice ($P > 0.05$). In addition, when we examined the subclass memberships of the tumors as determined by various prediction methods, we observed only a few discrepancies (**Supplementary Fig. 3** online). Taken together, these results support the notion that better- or best-fit mouse models for human studies can be identified by applying genome-scale comparison of gene expression patterns.

By directly comparing the relative expression ratio of orthologous genes between human and mouse, we assessed how closely the mouse models mimic the gene expression activity of two human HCC subclasses. From the integrated gene expression data set, excluding HCCs from *Acox1*$^{-/-}$ and ciprofibrate-treated mice, human and mouse HCCs were divided into two groups based on outcomes from the prediction methods. We selected the top 500 genes that are differentially expressed between subclass A and subclass B. We calculated and compared relative average gene expression ratios between subclass A and subclass B in each species. With few exceptions, the relative difference of the expression of the 500 orthologous genes between two subgroups of mouse models is highly similar to those in humans (**Supplementary Fig. 4** online). We used independent *t*-tests to select orthologous genes that had significant differences in expression between subclass A and subclass B in both species ($P < 0.05$ in both *t*-tests) and yielded 329 genes. We used knowledge-based annotation of 329 genes based on a public database search. The genes fell into several functional groups (**Table 2** and **Supplementary Table 1** online). As observed in previous studies[4,21], genes involved in the regulation of cell growth and proliferation were the best predictors of an unfavorable outcome of human cancers. All orthologous mouse genes in this category were more highly expressed in *Myc Tgfa* transgenic and DENA-treated mice (subclass A–like) than the rest of the mice (subclass B–like). Expression of positive regulators of cell cycle, such as *CDK4* (*Cdk4*), *CDC25A* (*Cdc25a*), *CDC7* (*Cdc7*) and

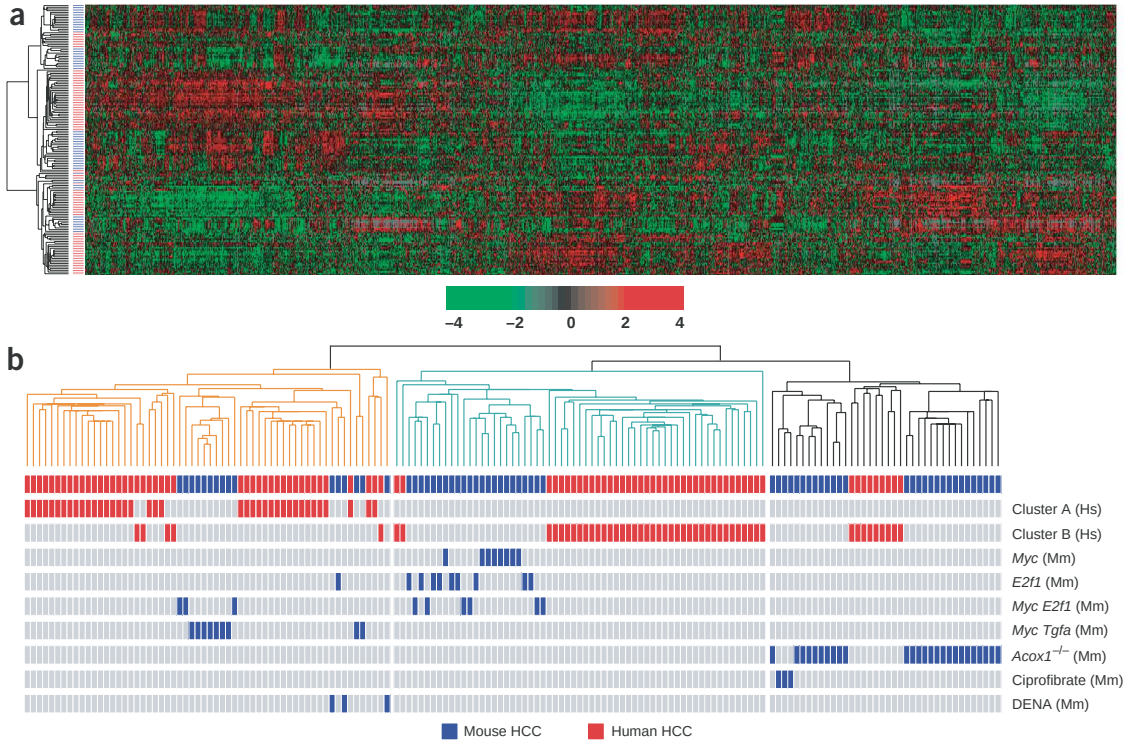

Figure 2 Cluster analysis of integrated human and mouse HCC. (**a**) Unsupervised hierarchical cluster analysis of integrated 68 mouse and 91 human HCC tumors. Orthologous genes with an expression ratio that differed by a factor of at least 2 from the reference in at least 10% of tissues in one of the data sets were selected for hierarchical analysis (1,650 genes). The data are presented in matrix format in which columns represent individual gene and rows represent each tissue. (**b**) Dendrogram of cluster analysis. Red and blue bars represent human and mouse HCC tissues, respectively. The identity of each HCC tissue is shown at the end of each row.

MAPK3 (*Mapk3*), was greater in subclass A than subclass B in both species. As expected from our previous study[4], many orthologous genes that are more highly expressed in subclass A in both species are antiapoptotic. Many poor prognostic markers in human cancers were also more highly expressed in subclass A in both species.

We next examined whether the predicted biological similarities between human HCC and mouse models were faithfully reflected in measurable phenotypes of each subclass. Proliferation rates were significantly higher in subclass A than subclass B in both species ($P < 1.0 \times 10^{-4}$ in human, $P < 1.0 \times 10^{-9}$ in mouse), and apoptosis rates were significantly lower in subclass A than subclass B in both species ($P < 1.0 \times 10^{-6}$ in human, $P < 1.0 \times 10^{-4}$ in mouse; **Fig. 3a**,**b**). Because previous studies indicated that the degree of ubiquitination in HCCs is highly associated with prognosis of affected

Table 1 Outcomes of the gene expression–based prediction methods

	CCP		1NN		3NN		NC		SVM		LDA	
Predicted subclass	A	B	A	B	A	B	A	B	A	B	A	B
Human HCC												
Subclass A ($n = 41$)	41	0	40	1	39	2	41	0	41	0	41	0
Subclass B ($n = 50$)	2	48	2	48	2	48	2	48	2	48	2	48
Percentage correctly classified[a]	98		97		96		98		98		98	
Mouse HCC												
DENA ($n = 3$)	1	2	1	2	1	2	1	2	1	2	1	2
Myc ($n = 8$)	0	8	0	8	0	8	0	8	0	8	0	8
E2f1 ($n = 10$)	1	9	1	9	1	9	1	9	1	9	1	9
Myc E2f1 ($n = 9$)	3	6	2	7	3	6	3	6	2	7	3	6
Myc Tgfa ($n = 9$)	7	2	7	2	7	2	7	2	7	2	7	2
P (χ^2)[b]	0.0048		0.0031		0.0048		0.0048		0.0031		0.0048	
P (χ^2) without *Myc Tgfa*[c]	0.053		0.11		0.053		0.053		0.11		0.053	

[a]Percentage for correct prediction during leave-one-out cross-validation. [b]P values of χ^2 test were computed for contingency tables of all mouse HCCs: for example, (1,2), (0,8), (1,9), (3,6) and (7,2) were used to compute P value for CCP. [c]P values of χ^2 test were computed for contingency tables of mouse HCCs without *Myc Tgfa*: for example, (1,2), (0,8), (1,9) and (3,6), were used to compute P value for CCP.
CCP, compound covariate predictor; 1NN, one nearest neighbor; 3NN, three nearest neighbor; NC, nearest centroid; SVM, support vector machines; LDA, linear discriminator analysis.

Table 2 Summary of selected genes

Unigene		Gene symbol			Log ratio (A/B)*	
Human	Mouse	Human	Mouse	Description	Human	Mouse
Nuclear pore transport						
Hs.113503	Mm.151329	*RANBP5*	*Kpnb3*	Karyopherin (importin) beta 3	0.97	0.81
Hs.180446	Mm.16710	*KPNB1*	*Kpnb1*	Karyopherin (importin) beta 1	0.98	0.75
Hs.90073	Mm.22417	*CSE1L*	*Cse1l*	Chromosome segregation 1-like (budding yeast); Exportin	1.13	0.90
Anti-apoptosis						
Hs.75462	Mm.239605	*BTG2*	*Btg2*	BTG family, member 2	0.71	1.46
Hs.171391	Mm.226905	*CTBP2*	*Ctbp2*	C-terminal binding protein 2	0.80	0.75
Hs.75562	Mm.5021	*DDR1*	*Ddr1*	Discoidin domain receptor family, member 1	0.69	1.49
Hs.180414	Mm.197551	*HSPA8*	*Hspa8*	Heat shock 70kDa protein 8	0.77	1.25
Hs.145279	Mm.28805	*SET*	*Set*	SET translocation (myeloid leukemia-associated)	1.03	0.76
Hs.115770	Mm.6426	*TNFSF11*	*Tnfsf11*	Tumor necrosis factor (ligand) superfamily, member 11	0.81	0.85
Hs.373508	Mm.3399	*TRAF2*	*Traf2*	TNF receptor-associated factor 2	0.67	0.78
Hs.349530	Mm.3308	*YWHAH*	*Ywhah*	Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein	1.12	0.88
Cell growth and proliferation						
Hs.152759	Mm.22430	*ASK*	*Ask*	Activator of S phase kinase	1.15	0.96
Hs.1634	Mm.29800	*CDC25A*	*Cdc25a*	Cell division cycle 25A	0.58	1.32
Hs.28853	Mm.20842	*CDC7*	*Cdc7*	CDC7 Cell division cycle 7-like 1 (budding yeast)	0.84	0.95
Hs.95577	Mm.6839	*CDK4*	*Cdk4*	Cyclin-dependent kinase 4	1.31	0.93
Hs.122579	Mm.2995	*ECT2*	*Ect2*	Epithelial cell transforming sequence 2 oncogene	1.39	0.98
Hs.401150	Mm.15918	*MAP3K1*	*Map3k1*	Mitogen-activated protein kinase kinase kinase 1	0.53	1.48
Hs.861	Mm.8385	*MAPK3*	*Mapk3*	Mitogen-activated protein kinase 3	0.85	0.81
Hs.179565	Mm.4502	*MCM3*	*Mcm3*	Minichromosome maintenance deficient 3	1.09	1.02
Hs.155462	Mm.4933	*MCM6*	*Mcm6*	Minichromosome maintenance deficient 6	1.25	1.34
Hs.89901	Mm.36865	*PDE4A*	*Pde4a*	Phosphodiesterase 4A	0.71	1.09
Hs.13501	Mm.28659	*PES1*	*Pes1*	Pescadillo homolog 1, containing BRCT domain (zebrafish)	0.93	0.78
Hs.93837	Mm.1860	*PITPNM1*	*Pitpnm*1	Phosphatidylinositol transfer protein, membrane-associated 1; Nir2	0.64	1.20
Hs.78944	Mm.28262	*RGS2*	*Rgs2*	Regulator of G-protein signaling 2, 24kDa	1.11	0.80
Hs.68061	Mm.20944	*SPHK1*	*Sphk1*	Sphingosine kinase 1	1.30	1.04
Hs.79150	Mm.46781	*CCT4*	*Cct4*	Chaperonin containing TCP1, subunit 4 (delta)	1.07	0.68
Hs.1600	Mm.1813	*CCT5*	*Cct5*	Chaperonin containing TCP1, subunit 5 (epsilon)	0.96	1.16
Hs.108809	Mm.914	*CCT7*	*Cct7*	Chaperonin containing TCP1, subunit 7 (eta)	1.06	0.75
Hs.178551	Mm.30066	*RPL8*	*Rpl8*	Ribosomal protein L8	0.73	1.08
Hs.182825	Mm.16423	*RPL35*	*Rpl35*	Ribosomal protein L35	1.04	0.97
Hs.406682	Mm.3229	*RPL26*	*Rpl26*	Ribosomal protein L26	0.98	0.90
Hs.425293	Mm.2424	*RPL10A*	*Rpl10a*	Ribosomal protein L10a	0.93	0.84
Hs.298262	Mm.103634	*RPS19*	*Rps19*	Ribosomal protein S19	1.22	0.78
Hs.433411	Mm.11376	*RPL36*	*Rpl36*	Ribosomal protein L36	0.93	0.77
Hs.180450	Mm.16775	*RPS24*	*Rps24*	Ribosomal protein S24	1.03	0.77
Hs.301547	Mm.5281	*RPS7*	*Rps7*	Ribosomal protein S7	0.95	0.74
Hs.380843	Mm.1139	*RPS6*	*Rps6*	Ribosomal protein S6	1.09	0.71
Hs.153	Mm.37835	*RPL7*	*Rpl7*	Ribosomal protein L7	0.58	1.17
Proteinases						
Hs.78056	Mm.930	*CTSL*	*Ctsl*	Cathepsin L	0.70	0.73
Hs.18069	Mm.17185	*LGMN*	*Lgmn*	Legumain	0.74	1.04
Hs.2256	Mm.4825	*MMP7*	*Mmp7*	Matrix metalloproteinase 7 (matrilysin, uterine)	0.69	1.71
Hs.151738	Mm.4406	*MMP9*	*Mmp9*	Matrix metalloproteinase 9 (gelatinase B, 92kDa gelatinase, 92kDa	0.83	1.24
Hs.1695	Mm.2055	*MMP12*	*Mmp12*	Matrix metalloproteinase 12 (macrophage elastase)	0.67	0.79
Poor Prognosis Markers						
Hs.375108	Mm.6417	*CD24*	*Cd24*	CD24 antigen (small cell lung carcinoma cluster 4 antigen)	0.99	1.45
Hs.433996	Mm.4426	*CD63*	*Cd63*	CD63 antigen (melanoma 1 antigen)	0.84	1.62
Hs.275243	Mm.100144	*S100A6*	*S100a6*	S100 calcium binding protein A6 (calcyclin)	1.20	1.20
Hs.8036	Mm.260157	*RAB3D*	*Rab3d*	Member RAS oncogene family	0.91	1.53
Hs.155421	Mm.80	*AFP*	*Afp*	Alpha-fetoprotein	0.97	1.45
Hs.82961	Mm.4641	*TFF3*	*Tff3*	Trefoil factor 3 (intestinal)	0.55	1.57

*Average gene expression ratios (log2-transformed) between subclass A and subclass B in human HCCs and between cluster 1 (human subclass A–like) and cluster 2 (human subclass B–like) in mouse HCCs.
We did not include the genes downregulated in subclass A, because most of these genes are involved in the metabolic pathways of the liver, and downregulation of these genes merely reflects the more severe loss of liver function (complete list of genes is available in **Supplementary Table 1** online).

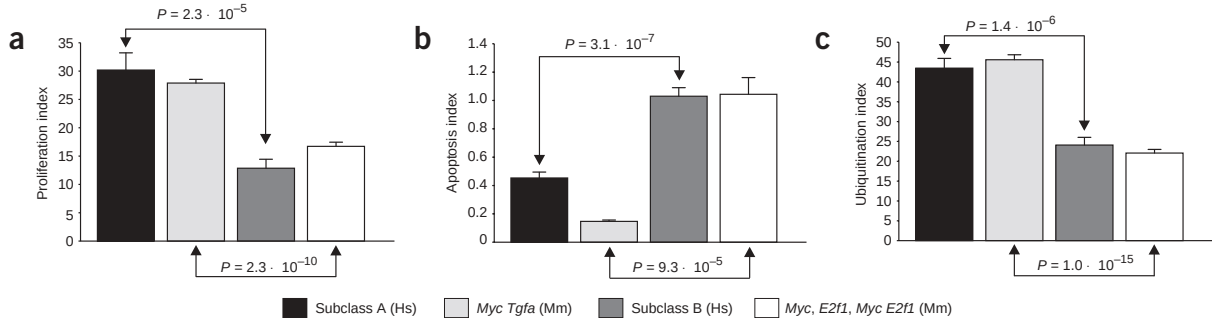

Figure 3 Comparison of measurable phenotypes between two subclasses in human and mouse HCC. (**a**) Cell proliferation index as measured by immunohistochemical staining with antibodies to PCNA (mouse) and to Ki-67 (human). Values shown are mean ± s.e. per 100 cells counted. (**b**) Apoptosis index, measured as the number of apoptotic cells per 100 cells counted. (**c**) Ubiquitination index, as measured by immunohistochemical staining with antibody to ubiquitin. Values shown are mean ± s.e. per 100 cells counted. *P* values were calculated by applying two sample *t*-test from each comparison.

individuals[4,22], we compared the ubiquitination index of HCCs. In both species, the degree of ubiquitination was significantly higher in subclass A than subclass B ($P < 1.0 \times 10^{-5}$ in human, $P < 1.0 \times 10^{-14}$ in mouse; **Fig. 3c**).

Next, we compared gene expression patterns of mouse HCCs with human nonliver cancers to assess whether the gene expression patterns shared in human and mouse HCCs truly reflect biological similarity of tumorigenesis in liver of both species. We used previously published data sets of human diffuse large B-cell lymphoma[23] and ovarian cancer[24]. Both studies showed that the tumors segregated into two subgroups whose gene expression patterns well reflect proliferative properties of the tumor cells. We selected orthologous genes and standardized them as described in **Supplementary Tables 2** and **3** online. In each comparison, we used human data for training prediction methods and assigned the mouse samples to the prediction set. In both analyses, most prediction methods successfully separated subgroups of human cancers clustered together in previous studies during leave-one-out cross validation in training sets, but they failed to segregate *Myc Tgfa* transgenic mice from the rest of mouse models or to produce concordant outcomes among prediction methods (**Supplementary Tables 2** and **3** online). These results using methods trained on nonliver data sets are highly discordant to those using the same prediction methods trained on liver data, indicating that gene expression patterns shared in human and mouse HCCs are liver-specific and do not represent relative similarity of proliferation. Taken together, our data suggest that mouse models that reflect gene expression patterns observed in two subclasses of human HCC may, to a considerable extent, recapitulate the underlying biology of the tumorigenesis in human liver.

In this study, we showed that cross-species comparison of gene expression patterns of HCCs can be used to identify the mouse models that are most similar to human HCCs. Moreover, this approach may be used to identify the most relevant mouse models for subclasses of human HCC. Gene expression–based prediction of mouse models is highly concordant with our earlier observation of phenotypes in mice. *Myc Tgfa* transgenic mice typically have a poor prognosis, including earlier and higher incident rate of HCC development, higher mortality, higher genomic instability and higher expression of poor prognostic markers (*e.g.*, Afp)[18,25]. *Myc* and *Myc E2f1* transgenic mice have a relatively higher frequency of mutations in β-catenin (*Catnb*) and nuclear accumulation of β-catenin that are indicative of lower genomic instability and better prognosis in human HCC[26]. As demonstrated by cross-species similarities in relative expression ratio of orthologous genes between subclasses (**Supplementary Figs. 4** and **5** online) and measurable phenotypes (**Fig. 3**), mimicry of the mouse models of subclasses of human HCC may, to a large extent, be due to the similarity in the underlying biology of the disease. Although the precise molecular mechanism driving hepatocarcinogenesis in both species is yet to be determined, the relative similarity of *Myc Tgfa* transgenic mice to human subclass A HCCs indicates that the signaling pathways driven by the receptor for TGFA or its related receptors have a role in prognosis for human HCC. The clear gain to be realized from this new approach, comparative functional oncogenomics, is to connect molecular pathogenic features of human cancer to mouse models with a greater level of confidence. Establishing this molecular relationship between the mouse models and the human cancers should provide new opportunities to explore research avenues into molecular pathogenesis, treatment and prevention of human cancer.

METHODS

Microarrays. We obtained mouse GEM2 cDNA clones from Incyte Genomics, and arrays were printed on preprepared glass slides at the Advanced Technology Center (National Cancer Institute).

Preparation of RNA and microarray. We isolated total RNAs from frozen liver tissue using CsCl density-gradient centrifugation methods[27].We pooled total RNA from the livers of ten wild-type mice and used them as reference in entire microarray experiments. To obtain gene expression profile data from four transgenic HCC mouse models, we used 20 μg of total RNAs from tissues to drive fluorescently (Cy-5 or Cy-3) labeled cDNA. We carried out at least two hybridizations for each tissue using dye-swap strategy to eliminate dye-labeling bias as described[4]. We used previously published data for HCCs from *Acox1*$^{-/-}$, DENA-treated and ciprofibrate-treated mice[28]. We generated the data from each mouse model using the same microarray platform and reference RNA. Animal housing and care were in accordance with guidelines from the Animal Care and Use Committee of the National Cancer Institute.

Data analysis. We transformed and normalized mouse gene expression data as described[4]. We then averaged expression ratios of each gene from replicated experiments and used them in subsequent analysis. When genes were represented more than once in the microarray platform, we used the averaged expression ratios. To identify the genes whose expression changed nontrivially, we selected genes with <30% missing expression data across the tissues in each data set and an expression ratio that differed from reference by a factor of at least 2 in at least 10% of tissues in each data set for further analysis (1,650 genes). Before integrating the two data sets, we standardized the expression of each gene to mean ± s.d. of 0 ± 1 independently in both data sets as

described[29]. We applied hierarchical clustering analysis as described[4]. To select genes that are differentially expressed in two given groups of tissues, we used significance analysis of microarrays[30] as a method for two-sample *t*-test with the estimation of false discovery rate. We chose a cut-off to retain the top 500 genes in the comparison. The predicted number of false discoveries in the first 500 genes was <1. Primary microarray data is available in the National Center for Biotechnology Information's Gene Expression Omnibus public database.

Prediction of mouse models for human cancer study. We applied five different prediction methods: linear discriminator analysis, support vector machines, nearest centroid, nearest neighbor and compound covariate predictor. Before the analysis, we removed mouse HCCs from *Acox1*$^{-/-}$ mice and ciprofibrate-induced HCCs from the mouse data set, because both cluster analyses indicated that they are least similar to human HCCs. We then selected for further analysis orthologous genes with $<30\%$ missing expression data across the tissues in each data set and with an expression ratio that differed by a factor of at least 2 from reference in at least 10% of tissues in each data set (1,950 genes). We used gene expression data from two predefined subclasses of human HCC to develop and train the prediction methods. We started to identify the most differentially expressed genes between subclass A ($n = 41$) and B ($n = 50$) in the human data set. We combined these genes (248 genes, $P < 1.0 \times 10^{-6}$) to form a series of classifiers that estimate the probability that a particular HCC tissue belongs to subclass A or subclass B. The number of genes in the classifiers was optimized to minimize misclassification errors during the leave-one-out cross-validation of the human data set.

Proliferation and ubiquitin indices. We carried out immunohistochemical staining on 10% formalin-fixed, paraffin-embedded tissues. We removed the paraffin from sections and incubated them in 3% H_2O_2 dissolved in 1× phosphate-buffered saline for 30 min and then microwaved them in 10 mM citrate buffer (pH 6.0) for 12 min. We applied mouse monoclonal antibody to PCNA (Santa Cruz Biotechnology; dilution 1:1,000), antibody to Ki-67 (Novocastra Laboratories) and rabbit polyclonal antibody to ubiquitin Ab-1 (Neomarkers, Fremont). We visualized immunoreactivity with the Vectastain Elite ABC kit (Vector Laboratories) and 3,3′ DAB (Dako Corporation) as the chromogen. We counterstained slides with Gill's hematoxylin. We determined PCNA-labeling (for mouse tissues), Ki-67-labeling (for human tissues) and ubiquitin-labeling indices by counting immunostaining-positive cells after counterstaining with hematoxylin. We counted at least 2,000 cells per tissue ($n = 10$ for each mouse model and $n = 15$ for each human subclass). Indices are represented as a percentage (mean ± s.e.) of the total number of cells counted.

Quantification of apoptosis. We calculated apoptotic indices by counting the apoptotic figures per 5,000 hepatocytes on tumor sections from 10 tissues per mouse model and 15 tissues per human subclass. We stained sections with the ApoTag peroxidase *in situ* apoptosis detection kit (Serologicals Corporation) and expressed apoptosis as a percentage (mean ± s.e.) of the total number of counted cells.

Quantitative RT-PCR. We generated first-strand cDNA using SuperScript First-strand synthesis system (Invitrogen) and carried out quantitative PCR using PRISM/7700 Sequence Detector with the SYBR Green PCR Core Reagents Kit (Applied Biosystems) as described in the manufacturer's manual. We designed primers to detect the following human and mouse mRNAs: *ASK* (*Ask*), *GTSE1* (*Gtse1*), *SLC16A2* (*Slc16a2*) and *INHBC* (*Inhbc*). We used *GAPD* (*Gapd*) as the endogenous control. Primer sequences are available on request. We expressed the relative mRNA expression levels in tissues as $-\Delta\Delta$Ct, in which ΔCt is the difference in the threshold PCR cycle (Ct) value of mRNA and the corresponding internal control *GAPD* and $\Delta\Delta$Ct is the difference in the ΔCt value of each tissue and normal liver.

GEO accession numbers. Human microarray platform, GPL1528; human HCC microarray data, GSE1898; mouse microarray platform, GPL1529; mouse HCC microarray data, GSE1897.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank R. Simon for discussions and advice on statistical analysis, J.W. Grisham for critical reading of the manuscript, E. Asaki for managing gene expression database and V.M. Factor and E.A. Conner for help with the mouse colonies.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Received 30 August; accepted 28 October 2004
Published online at http://www.nature.com/naturegenetics/

1. Hann, B. & Balmain, A. Building 'validated' mouse models of human cancer. *Curr. Opin. Cell Biol.* **13**, 778–784 (2001).
2. Klausner, R.D. Studying cancer in the mouse. *Oncogene* **18**, 5249–5252 (1999).
3. Rangarajan, A. & Weinberg, R.A. Opinion: Comparative biology of mouse versus human cells: modelling human cancer in mice. *Nat. Rev. Cancer* **3**, 952–959 (2003).
4. Lee, J.S. *et al.* Classification and prediction of survival in hepatocellular carcinoma by gene expression profiling. *Hepatology* **40**, 667–676 (2004).
5. Kimura, M. Evolutionary rate at the molecular level. *Nature* **217**, 624–626 (1968).
6. King, J.L. & Jukes, T.H. Non-Darwinian evolution. *Science* **164**, 788–798 (1969).
7. Ureta-Vidal, A., Ettwiller, L. & Birney, E. Comparative genomics: genome-wide analysis in metazoan eukaryotes. *Nat. Rev. Genet.* **4**, 251–262 (2003).
8. Cooper, G.M. & Sidow, A. Genomic regulatory regions: insights from comparative sequence analysis. *Curr. Opin. Genet. Dev.* **13**, 604–610 (2003).
9. Eddy, S.R. Computational genomics of noncoding RNA genes. *Cell* **109**, 137–140 (2002).
10. Hardison, R.C. Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet.* **16**, 369–372 (2000).
11. Rao, M.S., Lalwani, N.D., Watanabe, T.K. & Reddy, J.K. Inhibitory effect of antioxidants ethoxyquin and 2(3)-tert-butyl-4-hydroxyanisole on hepatic tumorigenesis in rats fed ciprofibrate, a peroxisome proliferator. *Cancer Res.* **44**, 1072–1076 (1984).
12. Reddy, J.K. & Lalwai, N.D. Carcinogenesis by hepatic peroxisome proliferators: evaluation of the risk of hypolipidemic drugs and industrial plasticizers to humans. *Crit. Rev. Toxicol.* **12**, 1–58 (1983).
13. Poirier, L.A. Hepatocarcinogenesis by diethylnitrosamine in rats fed high dietary levels of lipotropes. *J. Natl. Cancer Inst.* **54**, 137–140 (1975).
14. Conner, E.A. *et al.* Dual functions of E2F-1 in a transgenic mouse model of liver carcinogenesis. *Oncogene* **19**, 5054–5062 (2000).
15. Conner, E.A., Lemmer, E.R., Sanchez, A., Factor, V.M. & Thorgeirsson, S.S. E2F1 blocks and c-Myc accelerates hepatic ploidy in transgenic mouse models. *Biochem. Biophys. Res. Commun.* **302**, 114–120 (2003).
16. Murakami, H. *et al.* Transgenic mouse model for synergistic effects of nuclear oncogenes and growth factors in tumorigenesis: interaction of c-myc and transforming growth factor alpha in hepatic oncogenesis. *Cancer Res.* **53**, 1719–1723 (1993).
17. Fan, C.Y. *et al.* Steatohepatitis, spontaneous peroxisome proliferation and liver tumors in mice lacking peroxisomal fatty acyl-CoA oxidase. Implications for peroxisome proliferator-activated receptor alpha natural ligand metabolism. *J. Biol. Chem.* **273**, 15639–15645 (1998).
18. Calvisi, D.F., Factor, V.M., Ladu, S., Conner, E.A. & Thorgeirsson, S.S. Disruption of beta-catenin pathway or genomic instability define two distinct categories of liver cancer in transgenic mice. *Gastroenterology* **126**, 1374–1386 (2004).
19. Bentley, P. *et al.* Hepatic peroxisome proliferation in rodents and its significance for humans. *Food Chem. Toxicol.* **31**, 857–907 (1993).
20. Gonzalez, F.J., Peters, J.M. & Cattley, R.C. Mechanism of action of the nongenotoxic peroxisome proliferators: role of the peroxisome proliferator-activator receptor alpha. *J. Natl. Cancer Inst.* **90**, 1702–1709 (1998).
21. Rosenwald, A. *et al.* The proliferation gene expression signature is a quantitative integrator of oncogenic events that predicts survival in mantle cell lymphoma. *Cancer Cell* **3**, 185–197 (2003).
22. Shirahashi, H. *et al.* Ubiquitin is a possible new predictive marker for the recurrence of human hepatocellular carcinoma. *Liver* **22**, 413–418 (2002).
23. Alizadeh, A.A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511 (2000).
24. Schaner, M.E. *et al.* Gene expression patterns in ovarian carcinomas. *Mol. Biol. Cell* **14**, 4376–4386 (2003).
25. Sargent, L.M. *et al.* Nonrandom cytogenetic alterations in hepatocellular carcinoma from transgenic mice overexpressing c-Myc and transforming growth factor-alpha in the liver. *Am. J. Pathol.* **154**, 1047–1055 (1999).
26. Laurent-Puig, P. *et al.* Genetic alterations associated with hepatocellular carcinomas define distinct pathways of hepatocarcinogenesis. *Gastroenterology* **120**, 1763–1773 (2001).
27. Sambrook, J., Fritsch, E. & Maniatis, T. *Molecular Cloning* 7.19–7.22(Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 1989).
28. Meyer, K. *et al.* Molecular profiling of hepatocellular carcinomas developing spontaneously in acyl-CoA oxidase deficient mice: comparison with liver tumors induced in wild-type mice by a peroxisome proliferator and a genotoxic carcinogen. *Carcinogenesis* **24**, 975–984 (2003).
29. Ellwood-Yen, K. *et al.* Myc-driven murine prostate cancer shares molecular features with human prostate tumors. *Cancer Cell* **4**, 223–238 (2003).
30. Tusher, V.G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA.* **98**, 5116–5121 (2001).