Published in final edited form as:
Gastroenterology. 2012 April ; 142(4): 957–966.e12. doi:10.1053/j.gastro.2011.12.039.

Integrative Genomic Identification of Genes on 8p Associated with Hepatocellular Carcinoma Progression and Patient Survival

Stephanie Roessler[1], **Ezhou Lori Long**[2], **Anuradha Budhu**[1], **Yidong Chen**[2,§], **Xuelian Zhao**[1], **Junfang Ji**[1], **Robert Walker**[2], **Hu-Liang Jia**[3], **Qing-Hai Ye**[3], **Lun-Xiu Qin**[3], **Zhao-You Tang**[3], **Ping He**[4], **Kent W. Hunter**[5], **Snorri S. Thorgeirsson**[6], **Paul S. Meltzer**[2], and **Xin Wei Wang**[1]

[1]Laboratory of Human Carcinogenesis, National Cancer Institute, NIH, Bethesda, MD, USA

[2]Genetics Branch, National Cancer Institute, NIH, Bethesda, MD, USA

[3]Liver Cancer Institute, Fudan University, Shanghai, China

[4]Division of Hematology, FDA/CBER/OBRR, Bethesda, MD, USA

[5]Laboratory of Cancer Biology and Genetics, National Cancer Institute, NIH, Bethesda, MD, USA

[6]Laboratory of Experimental Carcinogenesis, National Cancer Institute, NIH, Bethesda, MD, USA

Abstract

Background & Aims—Hepatocellular carcinoma (HCC) is an aggressive malignancy; its mechanisms of development and progression are poorly understood. We used an integrative approach to identify HCC driver genes, defined as genes whose copy numbers associate with gene expression and cancer progression.

Methods—We combined data from high-resolution, array-based comparative genomic hybridization (CGH) and transcriptome analysis of HCC samples from 76 patients with hepatitis B virus infection with data on patient survival times. Candidate genes were functionally validated using in vitro and in vivo models.

Results—Unsupervised analyses of array CGH data associated loss of chromosome 8p with poor outcome (reduced survival time); somatic copy number alterations correlated with expression of 27.3% of genes analyzed. We associated expression levels of 10 of these genes with patient survival times in 2 independent cohorts (comprising 319 cases of HCC with mixed etiology) and 3 breast cancer cohorts (637 cases). Among the 10-gene signature, a cluster of 6 genes on 8p,

Correspondence: Correspondence should be addressed to Dr. Wang, Laboratory of Human Carcinogenesis, NCI, NIH, Bethesda, MD 20892, USA; xw3u@nih.gov.
[§]Current address: Greehey Children's Cancer Research Institute, Department of Epidemiology and Biostatistics, University of Texas Health Science Center at San Antonio, 7703 Floyd Curl Drive, San Antonio, Texas 78229-3900

Disclosures: No conflicts of interest exist.

Microarray Profiling: Raw arrayCGH data is accessible through GEO Series accession number GSE14322.

Writing Assistance: No writing assistance was provided.

Author Contributions: S.R. and X.W.W. initiated and designed the study, analyzed and interpreted data, and wrote the manuscript. S.R., A.B., R.W. and H.L.J. performed microarray analyses. S.R. E.L.L., Y.C. K.W.H., S.S.T. and P.S.M. performed data analyses and interpretation. S.R., X.Z. and J.J. performed TSG functional assays. H.L.J., Q.H.Y., L.X.Q. and Z.Y.T. recruited patients, collected clinical specimens, and advised on clinical issues. P.H. performed pathological evaluation. All authors discussed the results and commented on the manuscript.

(*DLC1*, *CCDC25*, *ELP3*, *PROSC*, *SH2D4A*, and *SORBS3*) were deleted in HCCs from patients with poor outcomes. In vitro and in vivo analyses indicated that the products of *PROSC*, *SH2D4A*, and *SORBS3* have tumor-suppressive activities, along with the known tumor suppressor gene, *DLC1*.

Conclusions—We used an unbiased approach to identify 10 genes associated with HCC progression. These might be used in assisting diagnosis and to stage tumors based on gene expression patterns.

Keywords

Liver Cancer; Tumor Profiling; Cancer Driver Genes

INTRODUCTION

Somatic mutations in oncogenes or tumor suppressor genes (TSGs) of one single cell are believed to cause the development of solid tumors. This initial cell proliferates and due to genomic instability, the resulting daughter cells accumulate genomic changes which lead to clonal expansion and tumor development.[1] A recent large scale study analyzing somatic copy number alterations (SCNAs) in 26 histological cancer types found that SCNAs widely overlap in different cancer types.[2] These genomic changes are irreversible and specific to tumor cells. Therefore, they provide ideal targets for the development of new therapies.

Hepatocellular carcinoma (HCC) is the most frequent malignant tumor in the liver and the third leading cause of cancer death worldwide.[3] The high mortality rate of HCC is mainly caused by metastasis or by de novo tumor formation in the diseased liver, the so called “field effect”.[4, 5] Major etiologies associated with HCC are hepatitis B virus (HBV) and hepatitis C virus infection, chronic alcohol consumption and obesity.[6] Although significant progress has been made in the HCC field, the molecular mechanisms and signaling pathways underlying HCC development and progression are still poorly understood. This is probably the case because HCC, like most other solid tumors, is very heterogeneous in terms of clinical presentation, genomic alterations and gene expression patterns.[7–9]

Previous studies utilized low resolution comparative genomic hybridization (CGH) to examine SCNAs in HCC cancer cell genomes and identified frequent DNA copy number gains at 1q, 8q and 20q, or losses at 1p, 4q, 8p, 13q, 16q and 17p in HCC specimens of different etiologies and cell lines [6]. In addition, candidate approaches have identified several oncogenes such as *MDM4* (1q32), *MYC* (8q24), *Jab1* (8q), *cIAP1* and *Yap* (11q22) and TSGs such as *DLC1* (8p22), *RB1* (13q14), *TP53* (17p13) to be associated with HCC.[10–15] Recent RNAi-based studies targeting minimum deletion loci in HCC coupled with a mouse mosaic model have uncovered several additional TSGs.[12, 15] However, studies employing an unbiased genome-wide search for HCC ‘driver’ genes are limited, particularly for those related to cancer prognosis.

In this study, we used an integrative approach of high-resolution array-based CGH (arrayCGH) and gene expression profiling coupled with patient prognosis to identify SCNAs in HCC clinical specimens which may functionally contribute to tumor progression. In order to identify potential tumor ‘driver’ genes that are functionally important and differ from ‘passenger’ genes which do not provide a selective advantage to the tumor cells[16, 17], we first restricted our search to genes which showed (1) recurrent SCNAs, (2) correlation of the SCNA and the transcriptome and (3) a selective retention in HCC with poor prognosis. We found that SCNAs of HCC with good prognosis differed greatly from HCC with poor prognosis. Global correlation analysis of arrayCGH and gene expression data that link to HCC with poor prognosis revealed a 10-gene signature that was validated as a molecular

predictor of patient survival in five independent cohorts. In addition, we functionally validated three new TSGs *in vitro* and *in vivo*. Thus, our integrative genomics approach was effective in identifying new cancer 'driver' genes which may allow for the development of a diagnostic tool for tumor molecular re-staging as well as the discovery of new therapies specific to these cancer 'driver' genes.

MATERIALS AND METHODS

Liver Samples and Clinical Data

Hepatic tissues were obtained with informed consent from patients who underwent radical resection between 2002 and 2003 at the Liver Cancer Institute (LCI) (Fudan University, Shanghai, China) and from the Liver Tissue Cell Distribution System (LTCDS) at the University of Minnesota (Minneapolis, MN). The study was approved by the Institutional Review Board of the participating institutes. A total of 256 HCC patients were recruited. The majority of patients (96.31%; Table S1) had a history of HBV infection or HBV-related liver cirrhosis; all were HCC diagnosed by two independent pathologists, with detailed information on clinical presentation, pathological characteristics and survival status. Gene expression profiles were conducted on primary HCC and the corresponding paired non-tumor hepatic fresh frozen tissues from cohort2 and 64 cases of cohort1. The normal liver pool consisted of total mRNA from seven disease-free liver donors which were obtained through LTCDS funded by NIH Contract #N01-DK-7-0004/HHSN267200700004C.

Gene Expression Microarrays

Our analyses involved 256 patients with hepatocellular carcinoma for whom Affymetrix U133A2.0 gene expression data was available and is accessible though GEO accession number GSE14520 (http://www.ncbi.nlm.nih.gov/geo), as described previously.[18] An additional HCC gene expression data set (Table S1) and breast cancer datasets are described in the supplemental text.

ArrayCGH

Agilent Human-Genome-CGH-105A Oligo Microarrays G4412A were carried out according to manufacturer's instructions (Agilent, Santa Clara, CA). A detailed protocol is available in the Supplemental Experimental Procedures. Raw arrayCGH data is accessible through GEO Series - GSE14322.

Statistical Analysis

Detailed information describing microarray data processing is available in the Supplemental Procedures. For unsupervised hierarchical clustering of the arrayCGH data, segmented data was converted to 1, −1, 0 according to their respective status of gain, loss and no change, and then weighted by the squares of the frequencies of copy number gain or loss at a particular genomic location. Adjacent probes with identical DNA copy number profiles across all samples were combined to form unique segments. Average-linkage clustering was performed based on the Euclidean distance metric. Pearson correlation was used for Multidimensional Analysis (MDS). The significance of the difference in gain/loss status between HCC subgroups in unique segments was determined by the Fisher's exact test and the p-values were adjusted using Benjamini-Hochberg correction. Adjacent significant regions with gaps less than 50kb were combined and considered as one large region with differential genomic aberrations between subgroups.

Class prediction of the gene expression data was performed in BRB-Array Tools (Version 3.7.0). Four class prediction algorithms, i.e., Support Vector Machines (SVM), Compound Covariate Predictor (CCP), Linear Discriminant Analysis (LDA), or Nearest Centroid (NC),

were used to determine whether mRNA expression patterns could accurately discriminate good and poor survival HCC groups in an independent data set. In these analyses, arrayCGH cases for which gene expression data was available (cohort1; N=64) were chosen to build a classifier which was then used to predict the cases of two independent HCC cohorts (N=319). The models incorporated genes that were differentially expressed among genes at the 0.001 significance level as assessed by the random variance t-test. In these analyses, 90% of the samples were randomly chosen to build a classifier which was then used to predict the remaining 10% of the cases in the training/test set. The accuracy of the prediction was calculated after 1000 repetitions of this random partitioning process to control the number and proportion of false discoveries. For prediction of the LEC cohort, we converted the gene expression data of both cohorts into z-scores and then conducted class prediction in BRB-Array Tools. We similarly used cohort1 as training/testing set to predict the survival groups G1 and G2 in the LEC cohort. We also used the same method to predict 5 additional breast cancer cohorts. Additional statistical methods as well as *in vitro* and *in vivo* studies are included in the supplementary text.

RESULTS

Copy Number Aberrations and Gene Expression in HCC Exhibit High Correlation

We applied a genome-wide search for functional ‘driver’ genes whose disruption is linked to patient outcome among 256 HCC cases obtained from the Liver Cancer Institute (LCI) at Fudan University. We randomly partitioned these cases to a training/test set (cohort1, N=76, 30%) and an independent validation set (cohort2, N=180, 70%) whose clinical parameters did not differ (Table S2). We performed arrayCGH on cohort1 using the high resolution Agilent 105A array platform (Figure 1A). Consistent with previous publications,[10, 11] we found recurrent gains and losses on chromosomes 1q, 6p, 8q and 4q, 8p, 13q, 16, 17p, respectively (Figure 1D). A total of 2666 genes (1130 gained; 1536 lost) mapped to these regions and were found in more than 20% of specimens assayed.

We next restricted the gene list to potential ‘driver’ genes using two criteria: (1) their expression in tumor, but not adjacent non-tumor specimens, should correlate with SCNA (adjacent non-tumor tissues were included to account for a possible expression contribution by infiltrating non-cancerous cells); (2) their expression should be associated with patient prognosis. These criteria were based on the following reasoning: if a significant correlation is present between a somatic aberration and gene expression, then the gene is likely to be functionally related to HCC development. Furthermore, if this gene is also selectively retained in a subset of HCC with poor survival, this suggests that it is biologically selected during HCC progression. Gene expression profiles of the tumor and non-tumor tissues were available for 64 samples in cohort1. We plotted the density distribution of Pearson correlation coefficients (see Experimental Procedures) for 10841 genes present on both the arrayCGH and mRNA microarrays (Figure 1B). The mean of all Pearson’s coefficients was 0.18 (95% CI: 0.178 to 0.185). A correlation coefficient of 0.3, corresponding to the 99th percentile of the 1000-fold random permutation, was used as the cutoff threshold for positive correlation. A total of 2959 genes (27.3% of all genes) met these criteria and were considered positively correlated. To ensure that the observed correlations were tumor-specific, we calculated the Pearson correlation of the tumors’ SCNA and the paired non-tumor tissue expression values. The distribution of the resulting Pearson coefficients from the non-tumor overlapped with the random distribution and only 95 genes (0.9%) had correlation coefficients exceeding 0.3, suggesting that the positive correlation was tumor-specific (Figure 1C). Overall, among the 2959 correlating genes, 743 were up-regulated and 287 genes were down-regulated based on a 2-fold cutoff when compared to normal liver pools (Figure 1D).

Good and Poor Outcome HCCs Differ in Genomic Regions of Copy Number Loss

We postulated that 'driver' genes should be functionally selected and retained in tumors with an aggressive outcome. To test this hypothesis, we performed unsupervised hierarchical clustering analysis by SCNA (Figure 2A). We found that cases separated by the first dendrogram branch, which separated the cases into two major clusters, differed mainly by their chromosome 1q status (Figure 2A and data not shown), but showed no overall survival difference (Figure S1A; p=0.92). However, further subdivision yielded four subgroups, C1 to C4, which differed in their survival outcome but not in their clinical characteristics (Tables S3 and S4). The estimated median survival and Kaplan-Meier analysis showed that overall and disease-free survival of clusters C1 and C3 was longer than that of C2 and C4 (Figure S1A and B; Table S4). To test the robustness of the four clusters, we introduced 100 perturbations with standard deviation 0.003 and measured the proportion of pairs of specimens within a cluster for which the members of the pair remain together in the re-clustered perturbed data.[19] We found that the cluster was highly stable with an obtained Robustness index R of 0.986. Moreover, multidimensional scaling analysis based on genomic profiles revealed a close proximity of clusters C1 and C3 as well as C2 and C4 (Figure 2B). Thus we divided the samples into two major survival subtypes, i.e., G1 (good survival; C1 and C3) and G2 (poor survival; C2 and C4). Consistently, Kaplan-Meier survival analysis revealed that G2 had significantly worse survival than G1 (Figure 2C). To assess the prognostic significance of these subgroups, we performed empirical survival analysis by randomly permuting (10,000 times) the label of the survival data. The empirical p-value was calculated by the total number of p-values smaller than 0.014 (the p-value of the survival test between G1 and G2). This test confirmed that it is unlikely to obtain significant survival difference by random sub-setting the data into two groups (p=0.015). Thus, the two subgroups identified by SCNAs are significantly associated with survival.

Comparison of the SCNA frequency in the two subtypes showed that the gained regions are similar in G1 and G2 but the lost regions differ by up to 60% (Figure 3). Chromosome 4q loss was mainly associated with G1 while 8p loss was mainly associated with G2. We searched for regions with significant difference between G1 and G2 by applying two criteria: (1) The frequency of SCNAs had to differ by at least 20% between G1 and G2 and (2) the adjusted p-value had to be less than 0.05 (Benjamini-Hochberg correction). The regions we identified were located on 1p, 4q, 8p and 9p (Table S5). Interestingly, all of the regions that were significantly different between G1 and G2 were genomic loss regions. This suggested that survival-related regions were associated with genomic loss and were therefore, potential locations containing TSGs, whereas, amplified regions which might contain oncogenes did not differ between the survival groups. Among 578 genes that mapped to these regions of loss, 419 had expression data and among these, 134 (31.98%) showed significant correlation with genomic changes.

The 'Driver' Gene-based Signature Predicts HCC Survival

The arrayCGH results led us to hypothesize that prognosis-related HCC subtypes may be biologically distinct. We sought to build a survival prediction signature based on the 134 potential cancer 'driver' genes using Affymetrix gene expression data. Class comparison analysis resulted in ten significantly differentially expressed genes between G1 and G2 in cohort1 ($p<0.001$; FDR<0.05; Table S6). Among these ten genes, six genes mapped to 8p and were associated with the poor outcome group G2 while four genes mapped to 4q and were associated with the good outcome group G1 (Figure 3E). Quantitative RT-PCR of the 8p genes *SH2D4A*, *CCDC25*, *DLC1*, *PROSC* and *SORBS3* showed high correlation with the microarray gene expression data ($p<0.0001$; Figure S2).

A multivariate class prediction analysis using 10-fold cross validation was performed on cohort1 cases (N=64) and then applied to predict 180 independent HCC cases in cohort2. The 10-gene signature could significantly discriminate G1 from G2 cases in cohort1 by random selection of 90% of cases and testing on the remaining 10% (multivariate cross-validated $p<0.05$). Similar to cohort1, applying the SVM algorithm to cohort2 from the LCI followed by Kaplan-Meier survival analysis revealed that the predicted G1 and G2 subgroups had significant survival differences (log-rank p=0.008; Figure 4A). However, the 10-gene signature could not differentiate patient survival groups when tested on the paired adjacent non-tumor gene expression data (log-rank p=0.421; Figure S3). Similar results were observed with three additional class prediction algorithms (data not shown). Furthermore, the 10-gene signature significantly predicted outcome in a second independent validation cohort with mixed etiology (Table S1; Figure 4B). In contrast, a survival-related gene signature could not be found when this search was restricted to non-correlated but somatically altered loci in HCC (data not shown). Thus, the 10-gene signature was tumor-specific and could predict survival independent of etiology in two separate HCC cohorts.

Next, we performed Cox proportional hazards regression analysis to determine whether the genomic predictor was confounded by underlying clinical parameters. Univariate analysis showed that the signature was a significant predictor of survival (p=0.004; Table S7). Multivariate analysis controlling for potential confounding covariates (serum AFP levels, cirrhosis; microvascular invasion and BCLC staging) demonstrated that the genomic predictor was significantly associated with a 2.1-fold increased risk of death for patients with a G2 gene aberration/expression profile (Table S7). Similar results were obtained for final models including CLIP or TNM staging (data not shown). Thus, the 10-gene signature is an independent and significant predictor of survival. Strikingly, the 10-gene signature was able to predict outcome in patients with early stage disease, i.e. TNM stage I or BCLC stage 0-A (Figure S4).

The 10-Gene Signature Can Predict Survival in Breast Cancer

Interestingly, chromosome 8p deletion has been detected in many solid tumor types such as breast, lung, colon, liver and pancreas.[14] Therefore, we further tested whether the 10-gene signature was associated with survival of other solid tumors of an epithelial origin. We restricted our analyses to large cohorts with publicly available gene expression profiles using the same Affymetrix microarray platform and with available follow-up data. We identified six breast cancer datasets (see supplemental text). Unsupervised hierarchical clustering followed by Kaplan-Meier and Cox proportional hazard regression analyses based on the resulting subgroups showed that the 10-gene signature was significantly associated with overall survival in the Uppsala-1 and Karolinska cohorts and disease-free survival of the Uppsala-2 cohort, which were all composed of mixed node-positive and negative cases. However, the signature failed to predict disease-free survival in the Rotterdam, TRANSBIG or Mainz cohorts, which only contained node-negative cases, suggesting that the survival predictive capacity of the 10-gene signature is associated with tumor cell dissemination (Figure 4C). Taken together, the 10-gene set was validated in multiple independent cohorts as a signature to predict outcome of HCC patients and breast cancer patients with mixed node status.

SORBS3, SH2D4A and PROSC Suppress HCC Tumor Growth In Vitro and In Vivo

Since *DLC1*, a known TSG, was included in our poor outcome-associated ‘driver’ gene set, we tested whether the remaining five genes on 8p, *SH2D4A*, *SORBS3*, *CCDC25*, *ELP3* or *PROSC* could act as HCC TSGs. We hypothesized that if our screening approach led to TSG enrichment, a majority of the genes in the poor outcome signature would have the classic TSG impact of negative tumor cell growth. In this vein, expression vectors encoding each of

the six genes on 8p were introduced via transfection into Hep3B cells which harbor 8p deletions[20] and HuH1 cells which have reduced gene expression levels of all six genes (Figure S5). The re-expression of PROSC, SORBS3 and SH2D4A inhibited HCC cell colony formation and cell migration in both cell lines (Figure 5). DLC1 was included as a positive control and the results were largely consistent with published data.[14, 21, 22] Thus, these results suggest a functional link between tumor cell growth and four of the six 'driver' genes. In addition, we applied four different prediction algorithms and found that similarly to the 10-gene signature these four genes alone are sufficient to predict patient outcome (Figure S6).

To examine the tumor suppressive function of PROSC, SORBS3 and SH2D4A *in vivo* we subcutaneously injected Hep3B cells transduced with PROSC, SORBS3, SH2D4A or vector control into nude mice. Expression of SH2D4A and SORBS3 significantly reduced the tumor incidence rate (Figure 6A) and tumor volume (Figure 6B). PROSC expression also decreased the tumor incidence rate and tumor volume, but the results were not statistically significant. Representative animals and tumors are shown in Figure 6C and D (Table S8). Three months after tumor cell inoculation, the control animals developed ten tumors out of ten injections, whereas SORBS3 expressing cells led to the development of eight out of ten tumors, PROSC expression to seven out of ten tumors and SH2D4A to four out of ten tumors. To analyze the cause for the failure of tumor suppression by these genes, we performed quantitative RT-PCR for the expression of the three TSGs in the tumor tissues. Strikingly, the tumors which escaped from suppression exhibited a drastic reduction of SH2D4A, SORBS3 or PROSC expression with at least two orders of magnitude as compared to the levels prior to subcutaneous injection (Figure 6E). Thus, it appeared that the tumor cells that escaped from an inhibitory effect of these TSGs were able to form tumors. Taken together, our results indicate that loss of SH2D4A, SORBS3 and PROSC contributes to HCC tumor growth and that their re-expression can inhibit HCC cell growth, migration and tumorigenesis.

DISCUSSION

The HCC patient population has very poor outcome and is generally underserved due to ineffective therapies, making this tumor type one of the most aggressive worldwide. Surgical resection or liver transplantation are the only curative treatments for HCC, but eligibility is sparse due to advanced disease presentation[23] and the post-surgical tumor relapse rate is high due to recurrence or metastasis.[24] Recently, sorafenib, an oral multi-kinase inhibitor, has been described to improve survival in advanced HCC, but the survival benefit is still modest.[25] Thus, there is an urgent need to develop better molecular tools to assist in patient stratification and to identify new drugs to prevent relapse and prolong patient survival. Of promise, we recently found that microRNA-26 can serve as a biomarker to predict HCC survival and response to adjuvant IFN therapy.[26]

In this study, we applied genome-wide arrayCGH and mRNA profiling to identify survival-related cancer 'driver' genes in HCC. This is the largest, unbiased study to date identifying SCNAs that correlate with expression and that are linked to HCC survival. In addition, our study was independently conducted in test and validation cohorts. We found that SCNAs differ between HCCs with good or poor prognosis. There were significant differences in recurrent deleted regions between prognostic groups whereas recurrent amplified regions appeared similar.

High-resolution profiling of SCNAs has recently been employed to identify new oncogenes and TSGs. However, due to increased genomic instability, tumor cells can accumulate aberrations of large genomic segments and even whole chromosomes, thus increasing the

difficulty of identifying genes within these regions with downstream carcinogenic effects. Several recent studies have successfully applied a genome-wide screen followed by a candidate approach whereby selected regions of interest are chosen for further validation.[15, 27] Here, we successfully used an unbiased approach by employing an integrative genomic and gene expression profiling strategy to identify cancer 'driver' genes in HCC with poor prognosis. We found that a large percentage (27.3%) of the genes analyzed displayed SCNAs and gene expression correlation. By selecting correlating genes located in genomic regions of significant difference between good and poor prognostic tumors, we identified a 10-gene signature that could predict prognosis in two independent HCC cohorts. We found that within this signature, four genes on 4q and six genes on 8p were linked to good or poor prognosis, respectively. Consistently, deletion of 8p has been suggested to be associated with HCC metastasis in earlier studies.[28] Therefore, we focused on the function of the six 8p genes and the analysis of the 4 genes on 4q will be the aim of future studies.

Among the six poor prognosis genes, only *DLC1*, a confirmed and frequently deleted TSG, has been previously linked to breast, lung, colon, pancreas and liver cancer.[14, 29] *DLC1* encodes a Rho-GTPase activating protein and when deleted, leads to increased GTP-bound RhoA and carcinogenesis.[14, 30] Interestingly, a recent integrative study by Woo et al. using arrayCGH and gene expression data identified 50 potential driver genes including the chr 8p genes ELP3 and HMBOX1.[31] In addition, the chr 8p gene EPHX2 has also been linked to HCC survival.[32] Therefore, we tested whether these three chr 8p genes can individually predict patient outcome in our cohort. Kaplan-Meier survival analysis showed that ELP3 and EPHX2 are significantly associated with patient survival, whereas, HMBOX1 showed a trend for poor outcome of the low expressing cases but it was not significant (p=0.087; Figure S7). ELP3 was also identified in our study, however, it did not show any tumor suppressive function but rather seemed to increase colony formation and cell migration. Among the six genes on 8p, *SORBS3*, also known as Vinexin, encodes two isoforms of vinculin-binding cytoskeletal proteins involved in focal adhesion and cell-cell adhesion.[33] The function of SH2D4A, CCDC25 and PROSC is poorly understood. SH2D4A is an adapter protein with high homology to T cell specific adaptor protein (TSAd) which has been shown to regulate T cell receptor (TCR) signal transduction in T cells.[34] Colony formation and cell migration analyses revealed that, analogous to *DLC1*, re-expression of *SH2D4A*, *SORBS3* and *PROSC* inhibited HCC cell growth and migration, suggesting that their loss contributes to aggressive HCC. In addition, either *SH2D4A* or *SORBS3* expression was able to reduce tumor formation in a xenograft mouse model. We did not include CCDC25 and ELP3 in the xenograft mouse model because they did not alter colony formation or cell migration. Therefore, additional studies are needed to test their function in HCC, i.e. metastasis promotion. Sequencing analysis of SORBS3 in 16 matched tumor and non-tumor tissues revealed the absence of somatic mutations implying that SORBS3 acts as a haploinsufficient tumor suppressor (data not shown). Interestingly, DLC1 has been shown to be a haploinsuffient tumor suppressor since no somatic mutations in DLC1 have been observed in HCC.[35, 36] To date it is unclear why multiple TSGs are clustered together on chromosome 8p. Future functional studies of the chromosome 8p TSGs are required to test whether these genes function independently or collaboratively in HCC development and progression.

In conclusion, our strategy of combining and correlating genomic, transcriptomic and survival data successfully identified new HCC driver genes, especially those related to HCC with poor outcome. We demonstrated the feasibility by using an integrative genomic and transcriptomic approach and developing a 'driver' gene signature which might be useful in predicting patient survival in HCC and breast cancer. Further characterization of the function and downstream signaling pathways of the three newly identified 8p genes,

especially *SH2D4A* and *SORBS3*, may provide insight into the mechanisms of HCC progression and lead to the development of new therapeutic agents for genotype specific treatment.

Acknowledgments

Grant Support: This work was supported in part by the Intramural Research Program of the Center for Cancer Research, the US National Cancer Institute (Z01 BC 010313 and Z01 BC 010876).

We thank Nicholas Popescu for providing the DLC1 construct; Xiaolin Wu and his team members at the Laboratory of Molecular Technology of NCI-SAIC for high-through-put microarray analyses; Sean Davis for statistical expertise; Karen MacPherson for bibliographic assistance; Luhe Mian for technical assistance.

Abbreviations

AFP	alpha-fetoprotein
ALT	alanine transferase
arrayCGH	array-based comparative genomic hybridization
CCP	Compound Covariate Predictor
HBV	hepatitis B virus
HCC	Hepatocellular carcinoma
LCI	Liver Cancer Institute
LDA	Linear Discriminant Analysis
LEC	Laboratory of Experimental Carcinogenesis
MDS	Multidimensional Analysis
NC	Nearest Centroid
SCNA	somatic copy number alteration
SVM	Support Vector Machines
TSG	tumor suppressor gene

References

1. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell. 2000; 100:57–70. [PubMed: 10647931]
2. Beroukhim R, Mermel CH, Porter D, et al. The landscape of somatic copy-number alteration across human cancers. Nature. 2010; 463:899–905. [PubMed: 20164920]
3. Parkin DM, Bray F, Ferlay J, et al. Global cancer statistics, 2002. CA Cancer J Clin. 2005; 55:74–108. [PubMed: 15761078]
4. Libbrecht L, Craninx M, Nevens F, et al. Predictive value of liver cell dysplasia for development of hepatocellular carcinoma in patients with non-cirrhotic and cirrhotic chronic viral hepatitis. Histopathology. 2001; 39:66–73. [PubMed: 11454046]
5. Sherman M. Recurrence of hepatocellular carcinoma. N Engl J Med. 2008; 359:2045–2047. [PubMed: 18923166]
6. Farazi PA, DePinho RA. Hepatocellular carcinoma pathogenesis: from genes to environment. Nat Rev Cancer. 2006; 6:674–687. [PubMed: 16929323]
7. Thorgeirsson SS, Grisham JW. Molecular pathogenesis of human hepatocellular carcinoma. Nat Genet. 2002; 31:339–346. [PubMed: 12149612]

8. Yamashita T, Ji J, Budhu A, et al. EpCAM-positive hepatocellular carcinoma cells are tumor-initiating cells with stem/progenitor cell features. Gastroenterology. 2009; 136:1012–1024. [PubMed: 19150350]
9. Yamashita T, Forgues M, Wang W, et al. EpCAM and alpha-fetoprotein expression defines novel prognostic subtypes of hepatocellular carcinoma. Cancer Res. 2008; 68:1451–1461. [PubMed: 18316609]
10. Schlaeger C, Longerich T, Schiller C, et al. Etiology-dependent molecular mechanisms in human hepatocarcinogenesis. Hepatology. 2008; 47:511–520. [PubMed: 18161050]
11. Patil MA, Gutgemann I, Zhang J, et al. Array-based comparative genomic hybridization reveals recurrent chromosomal aberrations and Jab1 as a potential target for 8q gain in hepatocellular carcinoma. Carcinogenesis. 2005; 26:2050–2057. [PubMed: 16000397]
12. Zender L, Spector MS, Xue W, et al. Identification and validation of oncogenes in liver cancer using an integrative oncogenomic approach. Cell. 2006; 125:1253–1267. [PubMed: 16814713]
13. Yuan BZ, Miller MJ, Keck CL, et al. Cloning, characterization, and chromosomal localization of a gene frequently deleted in human liver cancer (DLC-1) homologous to rat RhoGAP. Cancer Res. 1998; 58:2196–2199. [PubMed: 9605766]
14. Xue W, Krasnitz A, Lucito R, et al. DLC1 is a chromosome 8p tumor suppressor whose loss promotes hepatocellular carcinoma. Genes Dev. 2008; 22:1439–1444. [PubMed: 18519636]
15. Zender L, Xue W, Zuber J, et al. An oncogenomics-based in vivo RNAi screen identifies tumor suppressors in liver cancer. Cell. 2008; 135:852–864. [PubMed: 19012953]
16. Vogelstein B, Kinzler KW. Cancer genes and the pathways they control. Nat Med. 2004; 10:789–799. [PubMed: 15286780]
17. Mitelman F, Johansson B, Mertens F. The impact of translocations and gene fusions on cancer causation. Nat Rev Cancer. 2007; 7:233–245. [PubMed: 17361217]
18. Roessler S, Jia HL, Budhu A, et al. A unique metastasis gene signature enables prediction of tumor relapse in early-stage hepatocellular carcinoma patients. Cancer Research. 2010; 70:10202–10212. [PubMed: 21159642]
19. McShane LM, Radmacher MD, Freidlin B, et al. Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data. Bioinformatics. 2002; 18:1462–1469. [PubMed: 12424117]
20. Zimonjic DB, Keck CL, Thorgeirsson SS, et al. Novel recurrent genetic imbalances in human hepatocellular carcinoma cell lines identified by comparative genomic hybridization. Hepatology. 1999; 29:1208–1214. [PubMed: 10094966]
21. Goodison S, Yuan J, Sloan D, et al. The RhoGAP protein DLC-1 functions as a metastasis suppressor in breast cancer cells. Cancer Res. 2005; 65:6042–6053. [PubMed: 16024604]
22. Zhou X, Zimonjic DB, Park SW, et al. DLC1 suppresses distant dissemination of human hepatocellular carcinoma cells in nude mice through reduction of RhoA GTPase activity, actin cytoskeletal disruption and down-regulation of genes involved in metastasis. Int J Oncol. 2008; 32:1285–1291. [PubMed: 18497990]
23. McCormack L, Petrowsky H, Clavien PA. Surgical therapy of hepatocellular carcinoma. Eur J Gastroenterol Hepatol. 2005; 17:497–503. [PubMed: 15827439]
24. Llovet JM, Schwartz M, Mazzaferro V. Resection and liver transplantation for hepatocellular carcinoma. Semin Liver Dis. 2005; 25:181–200. [PubMed: 15918147]
25. Llovet JM, Ricci S, Mazzaferro V, et al. Sorafenib in advanced hepatocellular carcinoma. N Engl J Med. 2008; 359:378–390. [PubMed: 18650514]
26. Ji J, Shi J, Budhu A, et al. MicroRNA expression, survival, and response to interferon in liver cancer. N Engl J Med. 2009; 361:1437–1447. [PubMed: 19812400]
27. Chiang DY, Villanueva A, Hoshida Y, et al. Focal gains of VEGFA and molecular classification of hepatocellular carcinoma. Cancer Res. 2008; 68:6779–6788. [PubMed: 18701503]
28. Qin LX, Tang ZY, Sham JS, et al. The association of chromosome 8p deletion and tumor metastasis in human hepatocellular carcinoma. Cancer Res. 1999; 59:5662–5665. [PubMed: 10582679]
29. Durkin ME, Yuan BZ, Zhou X, et al. DLC-1:a Rho GTPase-activating protein and tumour suppressor. J Cell Mol Med. 2007; 11:1185–1207. [PubMed: 17979893]

30. Durkin ME, Avner MR, Huh CG, et al. DLC-1, a Rho GTPase-activating protein with tumor suppressor function, is essential for embryonic development. FEBS Lett. 2005; 579:1191–1196. [PubMed: 15710412]

31. Woo HG, Park ES, Lee JS, et al. Identification of potential driver genes in human liver carcinoma by genomewide screening. Cancer Res. 2009; 69:4059–4066. [PubMed: 19366792]

32. Woo HG, Park ES, Cheon JH, et al. Gene expression-based recurrence prediction of hepatitis B virus-related human hepatocellular carcinoma. Clin Cancer Res. 2008; 14:2056–2064. [PubMed: 18381945]

33. Kioka N, Sakata S, Kawauchi T, et al. Vinexin: a novel vinculin-binding protein with multiple SH3 domains enhances actin cytoskeletal organization. J Cell Biol. 1999; 144:59–69. [PubMed: 9885244]

34. Lapinski PE, Oliver JA, Kamen LA, et al. Genetic analysis of SH2D4A, a novel adapter protein related to T cell-specific adapter and adapter protein in lymphocytes of unknown function, reveals a redundant function in T cells. J Immunol. 2008; 181:2019–2027. [PubMed: 18641339]

35. Wong CM, Lee JM, Ching YP, et al. Genetic and epigenetic alterations of DLC-1 gene in hepatocellular carcinoma. Cancer Res. 2003; 63:7646–7651. [PubMed: 14633684]

36. Park SW, Durkin ME, Thorgeirsson SS, et al. DNA variants of DLC-1, a candidate tumor suppressor gene in human hepatocellular carcinoma. Int J Oncol. 2003; 23:133–137. [PubMed: 12792785]

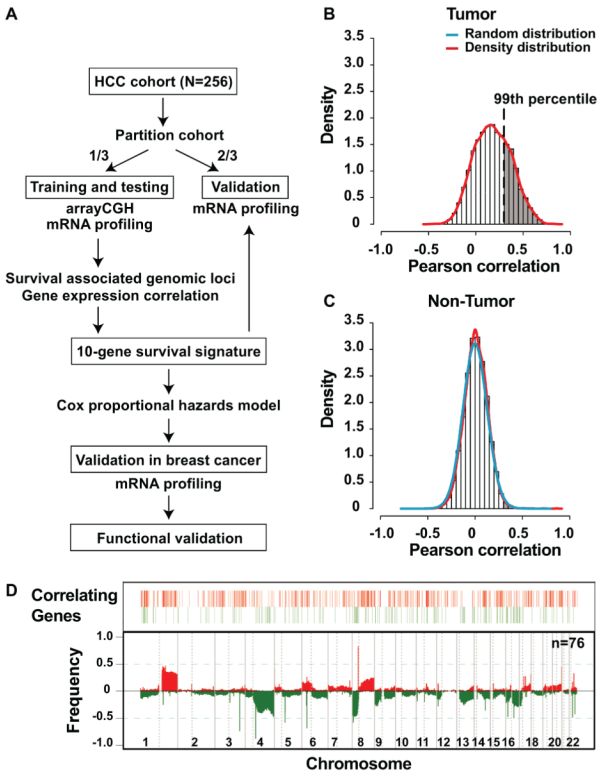

Figure 1.

Integration and Correlation of the Global SCNA and Gene Expression Profiles. **(A)** Schematic overview of the study design. **(B)** The density histogram shows the distribution of the Pearson correlation coefficients of gene expression and arrayCGH data from 60 tumor tissues. The blue line represents the density distribution of the 1000-fold random permutation of the data and the red line represents the density distribution of the Pearson correlation coefficients. **(C)** The density histogram shows the Pearson correlation coefficient of the gene expression of the non-tumor tissue and paired arrayCGH data of the cancerous tissue. **(D)** In the lower panel, frequencies of significant aberration in SCNAs are plotted as a function of genome location for 76 clinical specimens. Positive values indicate frequencies of samples showing copy number increases [log2(copy number)>0.5; shown in red] and negative values indicate frequencies of samples showing copy number decreases [log2(copy number)<−0.5; shown in green]. Chromosome boundaries and centromere position are indicated by vertical solid and dashed lines, respectively. Horizontal dashed blue lines indicate frequency of + and −0.5. The upper panel shows the position of correlating genes which are more than two-fold up (indicated in red) or down (green) regulated compared to normal liver.

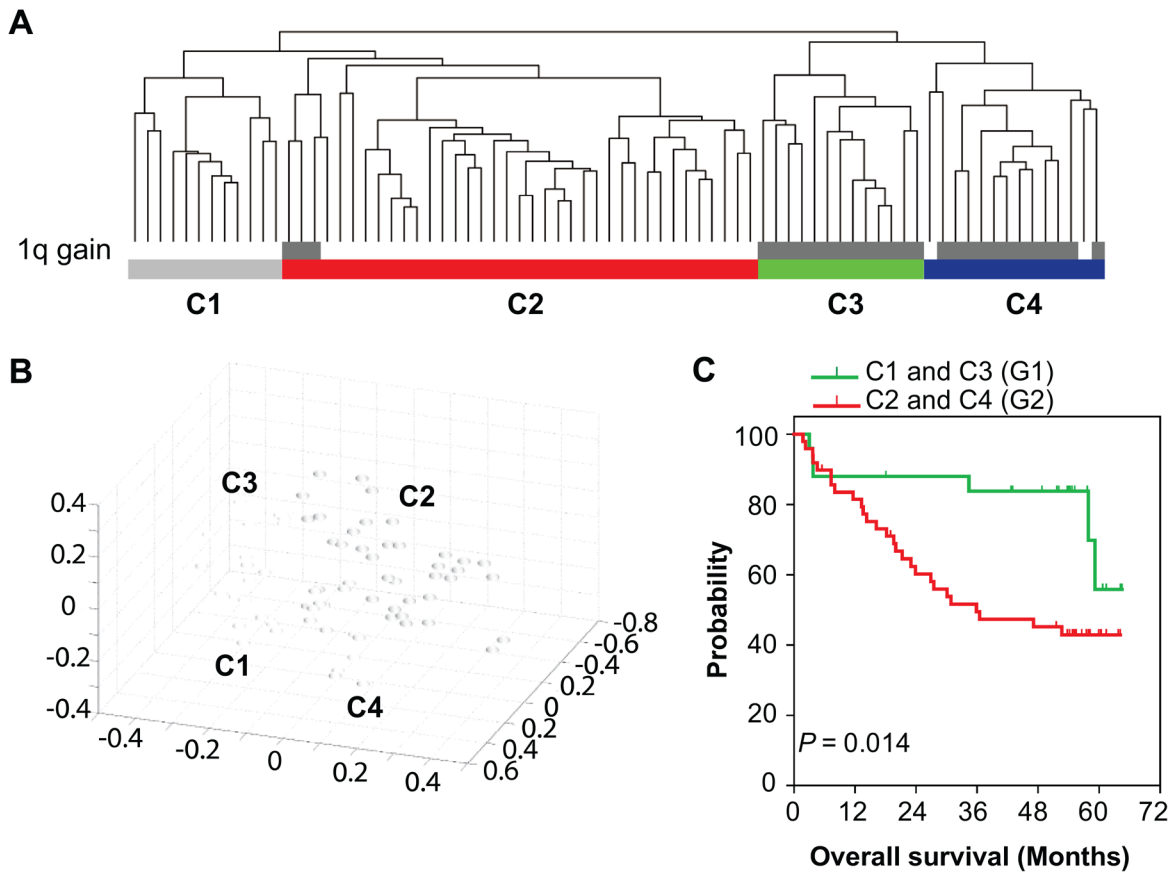

Figure 2.

HCC with Poor and Good Prognosis Differ in Their Copy Number Pattern. **(A)** Unsupervised hierarchical clustering of the weighted SCNAs profile 76 HCC cases revealed clusters C1, C2, C3 and C4. **(B)** Multidimensional scaling shows close positioning of clusters C1 and C3 as well as of clusters C2 and C4. **(C)** Kaplan-Meier survival analysis of these four clusters reveals that clusters C1 and C3 have good prognosis, whereas, clusters C2 and C4 have poor prognosis. The statistical *p*-value was generated by the Cox-Mantel log-rank test.

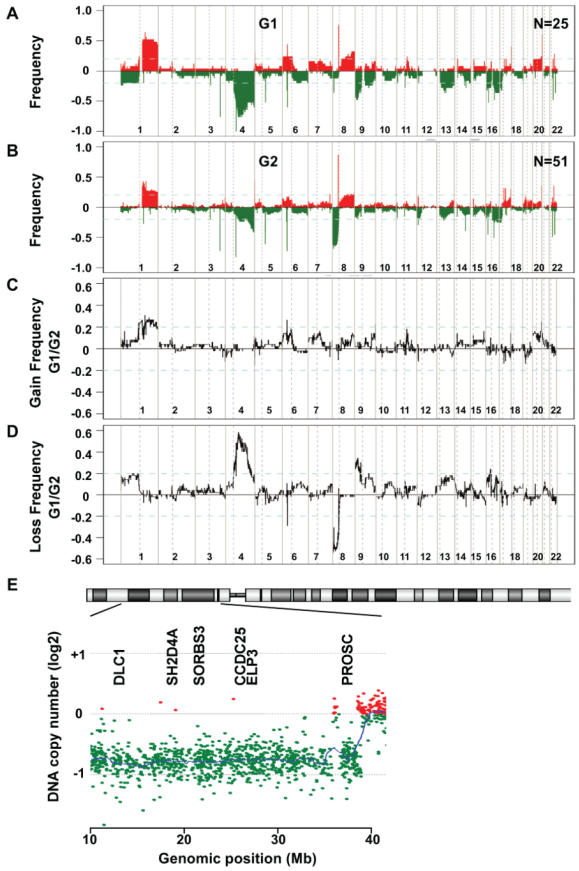

Figure 3.

Good (G1) and Poor (G2) Survival HCC Subgroups Differ in Distinct Genomic Areas **(A and B)** Frequencies of significant increases in SCNAs are plotted as a function of genome location of G1 and G2 HCC subgroups, respectively. Positive values indicate frequencies of samples showing copy number increases [log2 (copy number)>0.5] and negative values indicate frequencies of samples showing copy number decreases [log2 (copy number)< −0.5]. Chromosome boundaries and centromere position are indicated by vertical solid and dashed lines, respectively. Horizontal dashed blue lines indicate a frequency of 0.2. Grey and white boxes indicate tumor samples with or without chromosome 1q gain, respectively. **(C and D)** Differences (y axis) between frequencies of gain and loss across the genome for G1 versus G2 subtypes are shown. SCNA frequencies are plotted as a function of positioned location in the genome with positive values indicating higher frequencies in G1 versus G2. Horizontal dashed blue lines indicate frequency differences of 0.2. **(E)** A representative case with chromosome 8p deletion. Dots represent single probes, red dots represent amplified and green dots represent lost genomic regions.

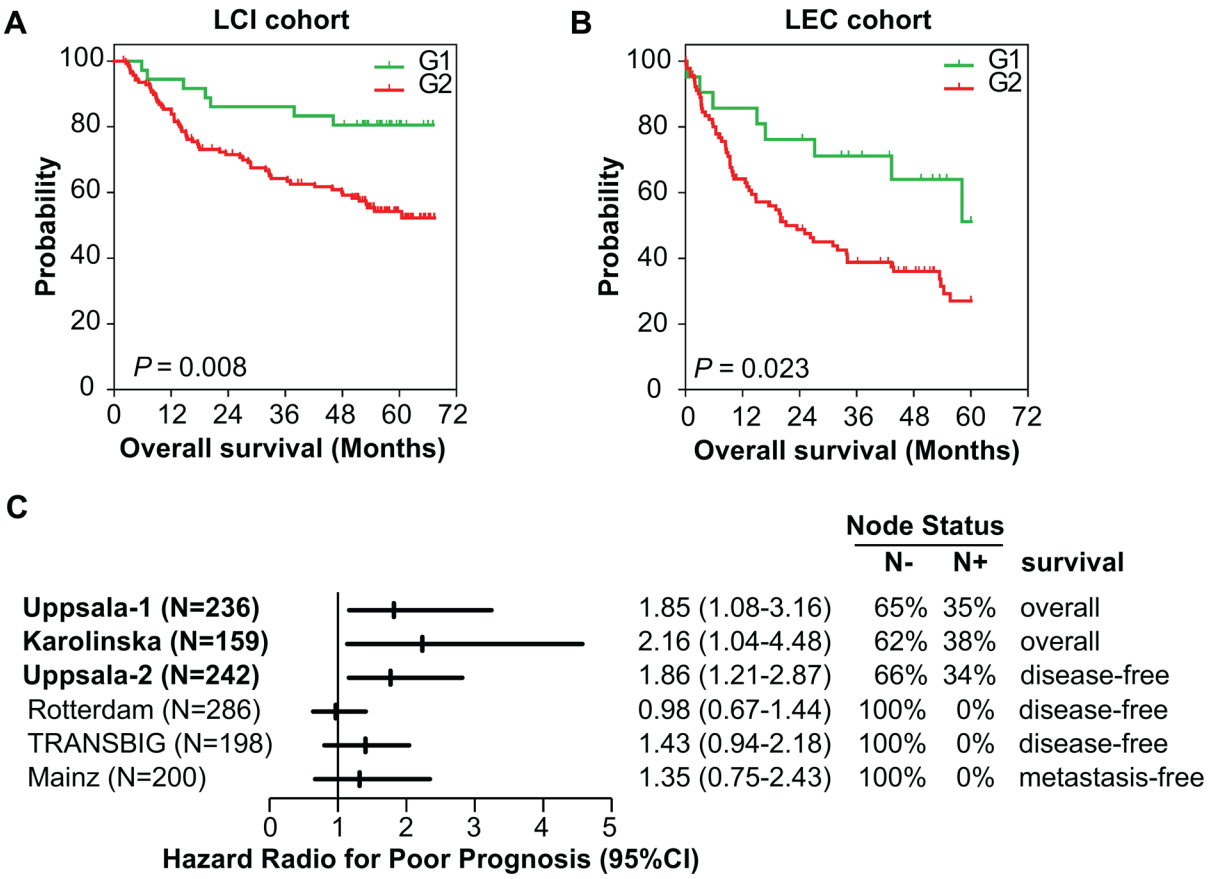

Figure 4.

The 10-'Driver' Gene-Signature Can Predict Survival Outcome in HCC Tumor Specimens and in Breast Cancer. **(A)** Kaplan-Meier overall survival on the independent validation cohort2 from the LCI by predicted classification of G1 and G2 by SVM. **(B)** Kaplan-Meier overall survival based on the predicted classification of G1 and G2 by gene expression of the LEC validation cohort by SVM. **(C)** Forest plot of the hazard ratios for poor survival of six breast cancer studies with varying percentage of node-negative patients. HR (95%CI), hazard ratio (95% confidence interval); N−, node-negative; N+; node-positive.

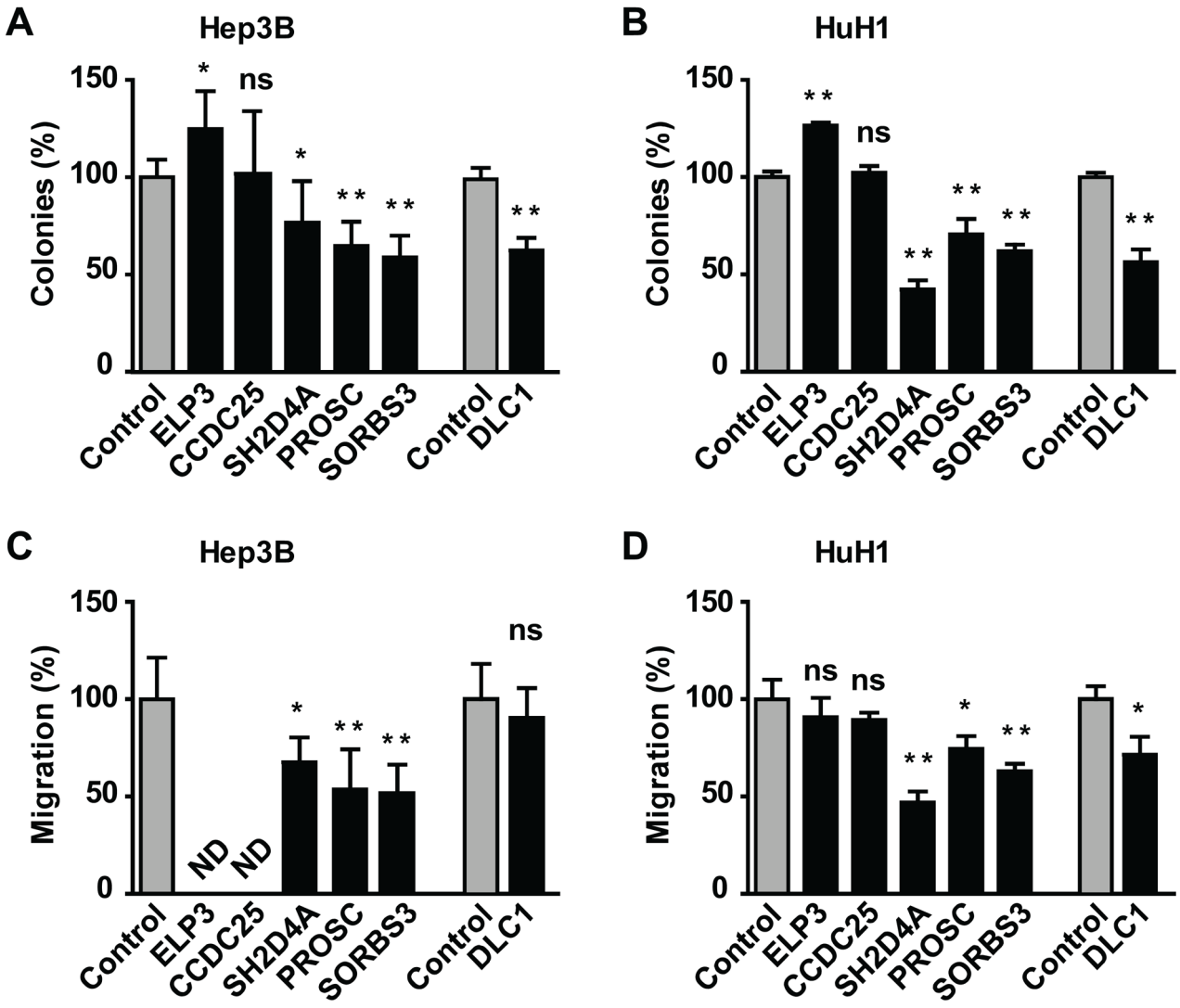

Figure 5.

SH2D4A, SORBS3 and PROSC Inhibit Colony Formation and Cell Migration In Vitro. **(A and B)** Colony formation assay and **(C and D)** cell migration assay of Hep3B and HuH1 transfected with the vector control or chr 8p gene as indicated. Data represent averages ±SD. Colony formation and migration assays were performed in quintuplets for Hep3B and in triplicates for HuH1. ND: not determined.

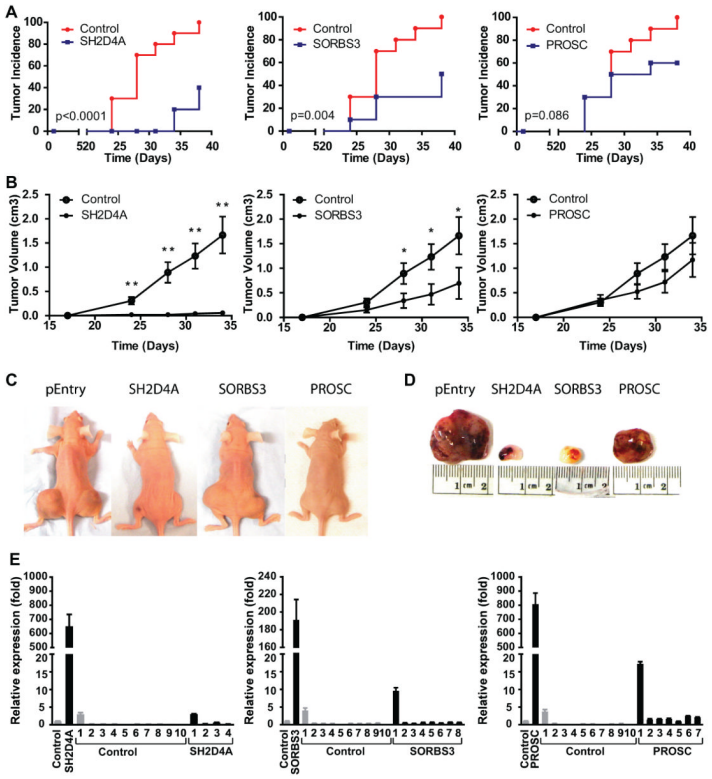

Figure 6.

SH2D4A and SORBS3 Inhibit Tumor Growth *In Vivo*. **(A)** Tumor incidence of Hep3B cells transfected with vector control, SH2D4A, SORBS3 or PROSC cDNA after subcutaneous injection into immune-compromised mice (n=10). Tumor incidence was observed twice per week. The log-rank p-value is indicated. **(B)** Growth curve of tumor xenografts of Hep3B cells transfected with vector control, SH2D4A, SORBS3 or PROSC (n=10). Data represent averages ± SEM. * $p<0.05$, ** $p<0.005$ by two-sided Student's t-test. **(C)** Representative nude mice and **(D)** subcutaneous tumors 34 days after subcutaneous injection. The experiment had to be terminated after 34 days because of tumor burden. **(E)** Relative gene expression of Hep3B cells transfected with vector control, SH2D4A, SORBS3 or PROSC, respectively, and of subcutaneous tumors was determined by real-time qRT-PCR. The first two bars represent Hep3B cell lines prior to subcutaneous injections. Quantitative RT-PCR was performed in triplicates. Data represent averages ±SD.