2012

Growth gone awry: exploring the role of embryonic liver development genes in HCV induced cirrhosis and hepatocellular carcinoma

Martha K. Behnke
Virginia Commonwealth University

Growth gone awry: exploring the role of embryonic liver development genes in HCV induced cirrhosis and hepatocellular carcinoma

Martha K. Behnke

Dissertation submitted to the faculty of Virginia Commonwealth University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Integrative Life Sciences

November 2012

Robert Fisher

Carleton Garrett

Mark Reimers, Chair

Arun Sanyal

Richard Sterling

Timothy York

Acknowledgements

I would like to sincerely thank my supervisor Donna George for allowing me the opportunity to combine my doctoral studies with my work in the Hume-Lee Transplant Center, and my director Dr. Robert A. Fisher for graciously allowing me to work with the data from the A2ALL/GR2 study, as well as his unflagging enthusiasm and belief in me. I also appreciate the support encouragement that all of my colleagues in transplant have given me over the years!

I would particularly like to thank my advisor, Mark Reimers, for taking me under his wing during a difficult period of starting over, and for allowing me the freedom to chase those interesting rabbits without letting me go too far down the rabbit hole!

Finally, I need to thank my wonderful husband and best friend Bill, and my beautiful children Faith, Braedon, Nate, and Ryan. From special meals to taking the kids out on a Saturday so that I could study, and especially these last few months where you've all been "science orphans", I could not have made this journey without your love, encouragement, and belief. Thank you.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Abbreviation	Meaning
ACVR1	Activin receptor type 1
ACVR1B	Activin receptor type 1B
ACVR2A/B	Activin Receptor type 2, A/B
AFP	alpha-fetoprotein
aHSC	activated hepatic stellate cell
AIBP	Aurora-A binding protein
AKT	RAC-alpha serine/threonine-protein kinase
APC	Adenomatous polyposis coli
ARF6	ADP-ribosylation factor 6
ARFP	alternate reading frame protein
ARID5B	AT rich interactive domain 5B
ATF2	Activating transcription factor 2
ATF7	Activating transcription factor 7
ATG7	ATG7 autophagy related 7 homolog
ATP10B	ATPase, class V, type 10B
BDNF	brain-derived neurotrophic factor
BEC	biliary epithelial cells
BIRC5	baculoviral IAP repeat containing 5
BMP2	Bone morphogenic protein 2
BMP4	Bone morphogenic protein 4
BMPR1A	Bone morphogenetic protein receptor, type IA
BMPR1B	Bone morphogenetic protein receptor, type IB
BMPR2	Bone morphogenic protein receptor 2
BP	Biological Processes
BSG	Basigen
C3A	complement component 3
C5A	complement component 5
CADM1	Cell adhesion molecule 1
CADM1	CCAAT/enhancer binding protein, alpha
CBP	Creb binding coactivator

CCL	Chemokine (C-C motif) ligand
CCND1	cyclin D1
CCND2	cyclin D2
CCNE2	Cyclin E2
CDC25C	cell division cycle 25 homolog C
CDH1	E-cadherin
CDK1	cyclin-dependent kinase 1
CDKN2A	cyclin-dependent kinase inhibitor 2A
cDNA	complimentary DNA
CEBPA	CCAAT/enhancer binding protein (C/EBP), alpha
CELA3A	chymotrypsin-like elastase family, member 3A
CER1	Cerebrus
CHGA	chromogranin-A
CHUK	conserved helix-loop-helix ubiquitous kinase
CIR	Cirrhosis
CITED2	CBP/p300-interacting transactivator
COL1A	Collagen type 1A
COL3A	Collagen type 3A
COL4A	Collagen IV alpha
CP	Ceruloplasmin
CRP	c-reactive protein
CSNK1D	Casein kinase I isoform delta
CTNNB1	b-catenin
CXCL10	Chemokine (C-X-C motif) ligand 10
CXCL11	Chemokine (C-X-C motif) ligand 11
CXCL9	Chemokine (C-X-C motif) ligand 9
DC	differential co-expression
DEG	differentially expressed genes
DHH	Desert hedgehog
DKK1	Dickkopf-related protein 1
DLK1	delta-like 1 homolog (Drosophila)
DM	Diabetes Mellitis
DNA	deoxyribonucleic acid
DNMT1	DNA (cytosine-5)-methyltransferase 1
DUSP6	dual specificity phosphatase 6
DVL2	dishevelled homolog
ECM	extracellular matrix
EGF	Epidermal Growth Factor
EGR1	Early Growth Response Factor

EHCC	Early HCC
ELF5	E74-like factor 5
EMT	epithelial-mesechymal transition
EPCAM	Epithelial cell adhesion molecule
ER	endoplasmic reticulum
ERBB2	Epidermal Growth Factor Receptor 2
ERBB4	v-erb-a erythroblastic leukemia viral oncogene homolog 4
ERK	extracellular signal-regulated kinase
ES	enrichment score
EtOh	Alcohol
FC	fold change
FDR	False discovery rate
FGF1	Fibroblast Growth Factor
FGFR1	Fibroblast Growth Factor Receptor 1
FGFR2	Fibroblast Growth Factor Receptor 2
FN1	Fibronectin
FOXA1	Forkhead homeobox A1
FOXA2	Forkhead homeobox A2
FOXM1	Forkhead box M1
FST	Follistatin
FSTL3	Follistatin-like protein 3
FZD	Frizzled
G6PC	glucose-6-phophatase
GAPDH	glyceraldehyde-3-phosphate dehydrogenase
GATA4	GATA binding protein 4
GATA6	GATA binding protein 6
GC	guanine-cytosine
GD	gestational day
GDNF	glial cell line derived neurotrophic factor
GEO	Gene Expression Omnibus
GFER	Human augmenter of liver regeneration
GFRA2	GDNF family receptor alpha 2
GFRα2	GDNF family receptor alpha 2
GO	Gene ontology
GPC3	Glypican 3
GRB2	Growth factor receptor-bound protein 2
GREM1	Gremlin
GSCA	Gene Set Co-expression Analysis
GSEA	Gene Set Enrichment Analysis

GSK3A	GSK3A glycogen synthase kinase 3 alpha
H19	H19, imprinted maternally expressed transcript (non-protein coding)
HAND2	Heart- and neural crest derivatives-expressed protein 2
HBV	Hepatitis B Virus
HCC	Hepatocellular carcinoma
HCV	Hepatitis C Virus
HDGF	Hepatoma-derived growth factor
HEYL	Hairy/enhancer-of-split related
HGF	Hepatocyte Growth Factor
HH	Hedgehog signaling pathway
HHEX	Hematopoietically expressed homeobox
HIF1A	Hypoxia-inducible factor 1a
HLX	H2.0-like homeobox
HMGA2	High-mobility group protein A2
HMGB2	High-mobility group protein B2
HNF1A	Hepatocyte nuclear factor 1 homeobox A
HNF1B	Hepatic Nuclear Factor 1 beta
HNF4A	Hepatic Nuclear Factor 4 alpha
HOXA7	Homeobox A7
HSC	hepatic stellate cell
HSPG2	Heparin sulfate proteoglycan
HTR2B	5-hydroxytryptamine (serotonin) receptor 2B, G protein-coupled
ICMT	isoprenylcysteine carboxyl methyltransferase
ID3	Inhibitor of differentiation 3
IF	Intermediate filament
IFN	Interferon
IGF2	Insulin-like Growth Factor 2
IHH	Indian hedgehog
IKBKB	inhibitor of kappa light polypeptide gene enhancer in B-cells, kinase beta
IKBKG	inhibitor of kappa light polypeptide gene enhancer in B-cells, kinase gamma
IKK	IkB kinase
IL	Interleukin
IL6	interleukin 6
IL6ST	Interleukin 6 signal transducer (gp130, oncostatin M receptor)
IL-8	Interleukin 8
INHA	Inhibin, alpha
INHBA	Activin
INHBB	Inhibin, beta B
INHBC	Inhibin, beta C

INHBE	Inhibin, beta E
INPP1	Inositol polyphosphate 1-phosphatase
IR	insulin resistance
IRS2	insulin receptor substrate 2
ITGA3	Integrin alpha 3
ITGA5	Integrin alpha 5
ITGA6	Integrin alpha 6
ITGB1	Integrin beta 1
ITGB4	Integrin beta 4
IVT	in-vitro transcription
JAG1	Jagged 1
JARID2	Jumonji
JNK	c-Jun NH2-terminal kinase
JUN	Jun protoncogene (c-JUN)
KDR	kinase insert domain receptor (vegfr2)
KEGG	Kyoto Encyclopedia of Genes and Genomes
KIT	c-kit
KLF6	Kruppel-like factor 6
KRAS	k-RAS
KRT19	cytokeratin-19
LAMA2	Laminin alpha 2
LAMA3	Laminin alpha 3
LAMA4	Laminin alpha 4
LAMB1	Laminin beta 1
LAMB2	Laminin beta 2
LAMB3	Laminin beta 3
LAMB4	Laminin beta 4
LAMC1	Laminin gamma 1
LAMC2	Laminin gamma 2
LAMC3	Laminin gamma 3
LEF1	Lymphoid enhancer-binding factor
LHCC	Late HCC
LHX2	LIM/homeobox protein
LPS	lipopolysaccharide
MAP2K4	Mitogen-activated protein kinase kinase 4 (SEK1)
MAP4K4	Mitogen-activated protein kinase kinase kinase kinase 4
MAPK	mitogen activated protein kinase
MAPK1	mitogen-activated protein kinase 1
MAPK14	p38

MAPK8	Mitogen-activfated protein kinase 8 (JNK)
MAS5	Affymetrix Microarray Suite 5.0)
MDK	Midkine
MET	Met proto-oncogene (c-MET)
MF	molecular function
MKi67	antigen identified by monoclonal antibody Ki-67
MM	mismatch
MMP	matrix metalloproteinase
mRNA	messenger RNA
MST1	Macriogage stimulating 1 (hepatocyte growth factor- like)
MTF1	Metal-regulatory transcription factor
MYC	c-Myc
MYCN	N-myc
NAMPT	nicotinamide phosphoribosyltransferase
NASH	Non-alcoholic steatohepatitis
NBL1	Neuroblastoma, suppression of tumorigenicity 1
NCAM	neural-cell-adhesion molecule
NDN	Necdin
NF1	Nuclear factor 1
NF-KB	nuclear factor kappa B
NFKB1	Nuclear factor of kappa light polypeptide gene enhancer in B-cells 1
NFKB2	Nuclear factor of kappa light polypeptide gene enhancer in B-cells 2
NGF	nerve growth factor
NGFR	nerve growth factor receptor
NID1	Nidogen
NKX2-8	NK2 homeobox 8
NODAL	Nodal
NOG	Noggin
NOR	Normal control liver samples
NOTCH2	Neurogenic locus notch homolog protein 2
NOTCH3	Neurogenic locus notch homolog protein 3
NR5A2	liver receptor homolog 1
NRF1	Nuclear respiratory factor 1
NRP1	neurophilin-1
NRTN	Neurturin
NT3	neurotrophin 3
NTF	neurotrophin
NTRK	neurotrophic tyrosine kinase, receptor, type
OC	onecut

ODF1	outer dense fiber of sperm tails 1
ONECUT1	Onecut 1
ONECUT2	Onecut 2
OSM	Oncostatin M
p53	protein 53
P53BP2	p53 binding protein 2
p73	tumor protein p73
PA2G4	Proliferation-assocaited 2G4
PCA	Principal Components Analysis
PDGF	platelet-derived growth factor
PGC	progastricsin
PHF2	PHD finger protein 2
Pi3K	Phosphoinositide-3-kinase
PIK3CA	Phosphoinositide-3-kinase, catalytic, alpha polypeptide (Pi3K)
PIK3R1	Phophoinositide-3-kinase, regulatory subunit 1 (alpha)
PLAU	uroplasminogen activator
PM	perfect match
pRb	retinoblastoma protein
PROX1	prospero homeobox 1
PSG1	pregnancy specific beta-1-glycoprotein 1
PTHLH	parathyroid-hormone-related peptide
PTN	pleiotrophin
PV	portal vein
QA	quality assessment
RA	retinoic acid
RAF1	v-raf-1 murine leukemia viral oncogene homolog 1
RASD1	RAS, dexamethasone-induced 1
REL	v-rel reticuloendotheliosis viral oncogene homolog
RMA	Robust Multi-array Average
RNA	ribonucleic acid
RNS	Reactive nitrogen species
ROS	reactive oxygen species
RXRA	Retinoic Acid Receptor alpha
S100	S100 calcium binding protein
SAA	serum amyloid A
SD	standard deviation
SERPINE	Plasminogen activator inhibitor
SETDB1	SET domain bifurcated 1
SFRP5	secreted frizzled-related protein 5

SHH	Sonic hedgehog
SMA	smooth muscle actin
SMAD2	Smad2
SMAD3	Smad3
SMAD4	Smad4
SMAD5	SMAD family member 5
SMAD6	Smad6
SMAD7	Smad7
SOCS3	Suppressor of cytokine signaling -3
SOD2	superoxide dismutase 2, mitochondrial
SOX	SRY-box
SPP1	Osteopontin
SRPK1	Serine/threonine-protein kinase
STAT3	Signal transducer and activator of transcription 3
STEAP3	STEAP family member 3, metalloreductase
STK3	serine/threonine kinase 3
STM	septum transversum mesenchyme
SYPL1	synaptophysin
T3	triiodothyronine
TBX3	T-box transcription factor 3
TCF3	Transcription factor 3
TEAD	TEA domain family member
TEF	Thyrotroph embryonic factor
TF	transcription factor
TGFB1	Transforming growth factor, beta 1
TGFB2	Transforming Growth Factor beta 2
TGFB3	Transforming Growth Factor beta 3
TGFBR1-3	Transforming growth factor receptor
TGFBRIII	Transforming Growth Factor beta receptor 3
TIMP	TIMP metallopeptidase inhibitor
TLR	Toll-like receptor
Tm	melting temperature
TNF	Tumor Necrosis Factor
TNM	tumor-node-metastasis staging system
TTR	transthyrein
uPAR	plasminogen activator receptor
UPF2	regulator of nonsense transcripts homolog (yeast)
VEGF	Vascular endothelial growth factor
VIM	Vimentin

VSN	Variance Stabilization and Normalization
WNT3A	Wnt 3A
WNT5A	Wnt 5A
WNT9A	Wnt 9A
WT1	Wilms tumor protein
XBP1	X-box binding protein 1
YAP1	Yes-associated protein 1
ZBTB20	Zinc finger factor
ZHX2	Zinc finger and homeoboxes factor 2
ZNHIT3	thyroid hormone receptor interacting protein 3

GROWTH GONE AWRY: EXPLORING THE ROLE OF EMBRYONIC LIVER DEVELOPMENT GENES IN HCV INDUCED CIRRHOSIS AND HEPATOCELLULAR CARCINOMA

Martha Behnke, PhD

Dissertation submitted to the faculty of Virginia Commonwealth University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Integrative Life Sciences

Virginia Commonwealth University

2012

Mark Reimers, PhD, Advisor

Virginia Institute for Psychiatric and Behavioral Genetics

ABSTRACT

Introduction and methods: Hepatocellular carcinoma (HCC) remains a difficult disease to study even after a decade of genomic analysis. Metabolic and cell-cycle perturbations are known, large changes in tumors that add little to our understanding of the development of tumors, but generate "noise" that obscures potentially important smaller scale expression changes in "driver genes". Recently, some researchers have suggested that HCC shares pathways involving the master regulators of embryonic development. Here, we investigated the involvement and specificity of developmental genes in HCV-cirrhosis and HCV-HCC. We obtained microarray studies from 30 patients with HCV-cirrhosis and 49 patients with HCV-HCC and compared to 12 normal livers.

Differential gene expression is specific to liver development genes: 86 of 202 (43%) genes specific to liver development had differential expression between normal and cirrhotic or HCC samples. Of 60 genes with paralogous function, which are specific to development of other organs and have known associations with other cancer types, none were expressed in either adult normal liver or tumor tissue.

Developmental genes are widely differentially expressed in both cirrhosis and early HCC, but not late HCC: 69 liver development genes were differentially expressed in cirrhosis, and 58 of these (84%) were also dysregulated in early HCC. 19/58 (33%) had larger-magnitude changes in cirrhosis and 5 (9%) had larger-magnitude changes in early HCC. 16 (9%) genes were uniquely altered in early tumors, while only 2 genes were

uniquely changed in late-stage (T3 and T4) HCC. Together, these results suggest that the involvement of the master regulators of liver development are active in the pre-cancerous cirrhotic liver and in cirrhotic livers with emerging tumors but play a limited role in the transition from early to late stage HCC.

Common patterns of coordinated developmental gene expression include: (1) Dysregulation of BMP2 signaling in cirrhosis followed by overexpression of BMP inhibitors in HCC. BMP inhibitor GPC3 was overexpressed in nearly all tumors, while GREM1 was associated specifically with recurrence-free survival after ablation and transplant. (2) Cirrhosis tissues acquire a progenitor-like signature including high expression of Vimentin, EPCAM, and KRT19, and these markers remain over-expressed to a lesser extent in HCC. (3) Hepatocyte proliferation inhibitors (HPI) E-cadherin (CDH1), BMP2, and MST1 were highly expressed in cirrhosis and remained over-expressed in 16 HCC patients who were transplanted with excellent recurrence-free survival (94% survival after 2 years; mean recurrence-free survival = 5.6 yrs), while loss in early HCC was associated with early recurrence and (2 year). Loss of HPI overexpression was also correlated with overexpression of c-MET and loss of STAT3, LAMA2, FGFR2, CITED2, KIT, SMAD7, GATA6, ERBB2, and NOTCH2.

Chapter 1: Introduction and overview

1.1 Motivation

Hepatocellular carcinoma (HCC) is the third most common cancer in the world [1] and 600,000 new cases are diagnosed each year [2]. One-year survival rates remain less than 50% in the United States, despite advances in therapy (McGivern 2011). Because of its poor prognosis, HCC is the third leading cause of cancer death worldwide [2]. Chronic Hepatitis B (HBV) is the dominant risk factor in China, while chronic Hepatitis C virus (HCV) is predominant in Japan and North America. HCC develops over decades of chronic infection and is generally thought to be a multistep process resulting from hepatocyte turnover, chronic inflammation, regeneration, oxidative stress, DNA damage, and cirrhosis, as well as direct viral injuries. Unfortunately, the specific molecular mechanisms underlying carcinogenesis remain unclear.

In the last ten years, microarray technology has been a powerful tool to study the molecular basis of disease. By measuring whole-genome transcript levels, expression patterns associated with liver dysfunction have been examined. However, HCC remains a difficult disease to study, with widely variable findings between studies and several proposed non-overlapping gene signatures [3-13]. This is likely due not only to the biological heterogeneity of HCC pathogenesis, but also reflects the varied clinical background of patients and variation in

statistical technique. There are significant statistical challenges which plague the analysis and interpretation of microarray experiments. Differences in technique in every stage of data pre-processing have been demonstrated to dramatically affect the end results, including background correction [14], normalization [15, 16], and probe set summarization [17].

Another difficulty stems from the heterogeneity of cancer processes, in which changes in the expression of important genes occur only in subsets of tumors. This results in skewed density curves (sometimes even bi-modal) that may not be easily detected by means-based tests. Most statistical tests in common use are based on comparing the magnitude of mean change relative to the variation. These tests also place focus on the largest magnitude changes which are often products of tumor behavior, such as increased metabolism and cell proliferation/turnover, rather than drivers that often have smaller fold-changes [18]. We suspect that there are modest changes in the expression of critical genes that may be difficult to distinguish from 'noise' in the data, but may have a significant impact on tumor development [19, 20].

The main Aims of this study were:

Aim 1: Show that liver cells under the stress of chronic infection, inflammation, and other injuries preferentially activate genes and pathways identified a priori as specifically involved in embryonic liver development, including those later involved in healing and regeneration, over genes with similar function that were not involved in liver development.

Aim 2: Identify recurrent patterns of activation that are either common to most tumors, or particular to a subset of tumors, and identify any clinical or prognostic characteristics of those subsets.

Aim 3: Compare our knowledge-driven methodology with standard approaches such as GSEA (Gene Set Enrichment Analysis and determine whether, in fact, the new method is more successful at identifying important patterns in the development and progression of liver tumors.

In the remaining sections of Chapter 1, we review what is known about the molecular mechanisms that are important in HCV infection, liver cirrhosis, and HCC, as well as those that drive liver development, wound healing, and regeneration. The genes identified from this review were used to create the gene sets used to explore Aims 1 and 2. In Chapter 2, Microarray and data processing methods are reviewed. Chapters 3, 4, and 5 address the questions in Aims 1, 2, and 3, and the results are discussed in Chapter 6.

1.2 Hepatitis C infection, mechanisms of liver damage and hepatocarcinogenesis

In this section we review the molecular mechanisms that underlie the development and progression of chronic HCV infection to the development of HCC. The critical genes that drive these processes are identified in order to define gene sets that might be expected to

characterize the genetic signature of cirrhosis samples and provide a basis for comparison to the genetic trends identified in our data.

Acute HCV infection is often clinically unapparent, with only about 1% of acute cases causing life-threatening hepatitis. However, the majority (75-85%) of acute adult HCV infections result in chronic disease (defined as persistent HCV RNA in the bloodstream for at least 6 months after onset of acute infection) [21]. Symptoms associated with chronic infection may not be apparent for years, but eventually present as fatigue, malaise, and the symptoms of hepatitis [22]. The Hepatitis C Virus is a positive-stranded RNA virus of the Flaviviridae family [2] that does not integrate into the host genome [23]. As such, it is the only known RNA virus whose lifecycle takes place in the cytoplasm [24]. Replication occurs in the cytoplasm, using the endoplasmic reticulum (ER) as primary site of genomic replication and virion assembly. Newly synthesized HCV RNA binds to HCV core protein and buds into the ER to form the viral envelope that then leaves the cell through the host cell's secretory pathway.

The HCV virus consists of a structural region and a non-structural region. The structural region contains the core protein and envelope glycoproteins E1, E2, and p7 protease. The non-structural region consists of six proteins that form the viral replicase complex: NS2, NS3, NS4a, NS4b, NS5a, and NS5b [25]. An F (for frameshift protein) or ARFP (for alternate reading frame protein), generated by an overlapping reading frame in the core protein coding sequence, has been proposed [2].

Because the virus does not integrate into the host DNA, mechanisms of liver damage and carcinogenesis are indirect. The HCV proteins have known direct interaction with over 30

host proteins, which over many years results in progressive damage from chronic inflammation, intrahepatic lipid accumulation (steatosis), fibrosis, oxidative stress, and direct oncogenic effects of the HCV proteins [24]. HCV core and non-structural proteins also localize in the outer mitochondrial membrane of the hepatocytes, which induces systemic oxidative stress and related mitogen-activated protein kinase (MAPK) signaling (p38, JNK, ERK, and NF-κB pathways). This leads to enhanced hepatocyte proliferation [26]. Oxidative stress induces production of Reactive Oxygen Species (ROS), leading to mitochondrial DNA damage [1], further increasing oxidative stress and insulin resistance [27]. Insulin resistance (IR) is also mediated directly by HCV core protein interaction with Tumor Necrosis Factor (TNF) receptors. Elevated insulin levels directly stimulate hepatic stellate cell proliferation and secretion of extra-cellular matrix (ECM) and connective tissue growth factors, contributing to fibrosis and cirrhosis development. The interdependence between steatosis, IR, and oxidative stress is important for disease progression and induces tissue damage and inflammation with activation of hepatic stellate cells (HSCs) and increased production of TNF and inteleukin-6 (IL6). Activated HSCs become responsive to both proliferative and fibrogenic cytokines and undergo an epithelial-mesenchymal transition (EMT) into myo-fibroblast-like cells that synthesize ECM components. These accumulate over time to form fibrosis. Eventually, regenerating hepatocytes become enclosed by scar tissue and form the nodules that define cirrhosis [24].

Cirrhosis is the end result of a long period of chronic liver disease, and eventually there is a decrease in hepatocyte proliferation that may be indicative of an exhaustion of the regenerative capacity of the liver [27]. Cirrhosis is also characterized by the continuous activation of hepatic stellate cells (HSC) and sustained production of cytokines, growth factors,

and products of oxidative stress [28]. This may in part be mediated by Toll-like receptors (TLRs). TLR2 and TLR4 are upregulated in the hepatocytes and Kupffer cells of patients with chronic HCV [1]. Molecular processes associated with cirrhosis include down-regulation of ECM production and cell proliferation regulators, down-regulation of genes involved in the regulation of differentiation [29], and up-regulation of JAG1 (pro-angiogenic factor in the Notch pathway), STAT1 and CXCL9-11 (involved in interferon immune response), and insulin growth factor [30]. Other important signaling pathways in the development of cirrhosis include PDGF and TGFβ dependent HSC recruitment mediated by neurophilin-1 (NRP1) [31].

Hepatocytes in a chronically injured liver have altered growth responses compared to hepatocytes in the healthy liver. Although TFGβ is up-regulated, cirrhotic hepatocytes have reduced sensitivity to it and are resistant to TFGβ – induced apoptosis [32], and this may be partly due to increased oxidative stress [33]. Nitta et al (2008) showed that cirrhotic hepatocytes also resist apoptosis via a MAPK-dependent survival pathway [32]. Cirrhotic hepatocytes express high levels of vimentin (VIM; a mesenchymal marker) and decreased expression of E-cadherin and occludin compared to healthy hepatocytes and have a fibroblast-like phenotype consistent with EMT (epithelial mesenchyme transition) [32]. Hepatocyte damage in the context of chronic liver inflammation and necrosis, such as occurs in HCV-cirrhosis, may invoke repair mechanisms involving hepatic progenitor (stem) cells [34]. Pathways involved in stem cell renewal include Notch, Hedgehog, and Wnt, which may also be seminal events in hepatocarcinogenesis [34].

It is estimated that 70-90% of all HCCs develop in a cirrhotic liver [2]. The molecular mechanisms outlined above suggest multiple factors contributing to carcinogenesis: direct action of HCV proteins leading to unregulated behavior in hepatocytes or progenitor cells; genetic alterations or DNA repair defects that in turn may inactivate tumor suppressors, activate oncogenes, and lead to epigenetic alterations; and a gradually decreasing ability to properly balance growth and cytokine signaling. Although the "classic" model of carcinogenesis requires a set of accumulated mutations, the direct effect of HCV proteins allows the pre-neoplastic cell to skip some of these steps. In addition, the HCV genome has hyper-variable regions that generate multiple quasi-species. Different viral variants have been isolated in tumor and non-tumor regions of the liver, suggesting that certain quasi-species may confer competitive advantage for some hepatocyte populations and contribute to carcinogenesis [2].

The main pathways associated with HCC development include Wnt/β-catenin, TGF-β, pRb, and p53 [35], Pi3K/AKT, MYC, MET, and Hedgehog [27]. pRb, p53, TGF-β, and β-catenin regulate cell proliferation or death, and these have been shown to have loss of heterozygosity due to aberrant methylation in some HCC [36]. As noted above, hepatocytes become less responsive to TGF-β induced apoptosis, even though TGF-β levels are increased. This may also play a role in hepatocarcinogenesis [35]. p53 alterations are rarely seen in HCV-HCC, but when present are associated with poor prognosis [35].

Wnt plays multiple roles in cell differentiation, proliferation, and apoptosis as well as embryogenesis, along with stem cell renewal, EMT, and cell adhesion [37]. The roles of Wnt and β-catenin are similarly complex in HCC, with contrasting and contradictory roles reported

[38]. For instance, Wnt5A has been shown to repress canonical Wnt signaling in HCC cell lines [39]. Geng et al (2012) report that loss of Wnt5A has been associated with elevated serum AFP and poor prognosis in patients with HCC [40], while increased Wnt5A expression was associated with poor differentiation (along with increased AFP) in HCC cell lines [41]. A July 2012 report indicates that during embryonic development, Wnt5A can both activate and repress Wnt/ β-catenin depending on which receptors are expressed at various stages of development in a mouse model [42]; this provides a potential explanation for the complexity of Wnt signaling in HCC as well. HCV core protein promotes WNT3A-induced tumor growth [43], and over-expression of GPC3 stabilizes β-catenin-frizzled complexes to activate signaling pathways and promote tumor formation [44]. Activating mutations of the CTNNB1 gene that codes for β-catenin have been found in 20-40% of HCC in studies with mixed etiology [37]. β-catenin also links E-cadherin to the actin cytoskeleton, and loss of either of these molecules results in tumor progression and cell invasion. As important as β-catenin is in the initiation and progression of HCC, determining its effects can be difficult with microarray studies because aberrant behavior is often a result of constitutive activation, stabilization, and/or nuclear localization as opposed to increased transcription [35].

Hedgehog and Notch pathways are developmental pathways that persist into adulthood maintaining the self-renewal capacity of stem cell populations. These similarities between embryonic and oncogenic pathways suggest that some HCCs may develop from liver stem cells [27, 45], or that there is a mechanism for de-differentiation of hepatocytes [46]. Although the existence of cancer stem cells has been controversial, current thinking is that both processes may occur and define prognostic sub-groups in HCC [47, 48].

1.3 Focus on liver development, healing, and regeneration

HCV-induced HCC has been shown in multiple genetic studies to continuously and highly express genes associated with antigen presentation and response to infection, including interferon-inducible genes, immunoglobulin genes, IL-8, and inflammatory response, as well as dysregulation of metabolic processes [29, 30, 49]. However, these expression changes might be viewed as consequences of tumor activity rather than causes of tumor formation and their large-magnitude expression changes make it difficult in unsupervised GeneChip studies to identify the potentially more subtle effects of master regulatory genes. In addition, recent studies have shown that genes related to apoptosis, metabolic processes, and DNA damage repair are altered in deceased donor livers (which typically come from patients on life support) compared to either living donor biopsies or samples from patients with sudden death [50], thus care should be taken when interpreting differential expression results using deceased donor control groups (as many studies do, including this one). To elucidate the roles that important regulators and effectors (such as those regulating liver development, maintenance, and healing) might play, a different approach is needed. Previous approaches have examined the biological function of those genes that are "most significantly changed" (i.e., have the largest magnitude expression changes), or looked for enrichment of significantly changed genes in canonical pathways. Instead, we proposed to a priori identify genes that have important roles in the development, maintenance, and health of the liver. We hoped that an intensive examination of the changes in expression, and in particular the patterns of co-expression of multiple genes through progressive disease states, would shed light on important seminal processes whose

roles were obscured in the shadow of the background of response to infection, metabolic disarray, cell cycle dysregulation and proliferation that is typical of tumors.

1.4 Developmental biology of the liver

In this section, we review the processes that direct development of the liver. In particular, we identify those genes that are critical drivers of specific stages of development or that are markers for important cell types.

The mechanisms that control the initiation of liver development are well conserved among vertebrates and hepatogenesis occurs through a progressive series of interactions between the embryonic endoderm and nearby mesoderm. Fate mapping studies indicate that the liver originates from the ventral foregut endoderm. The endoderm delineates the primitive gut and gives rise to the epithelial compartment of the gastrointestinal tract and the thyroid, liver, and pancreas. The anterior portion develops into the liver while the posterior portion gives rise to the gall bladder and bile ducts. There are five main stages of development as illustrated in Figure 1.1. Each stage has a unique combination of master regulators that orchestrate the proper timing and location of growth and differentiation:

1. Specification of hepatic fate from the endoderm (hepatic specification)

2. Liver bud formation (liver diverticulum)

3. Rapid liver bud growth

4. Differentiation into either hepatocytes (hepatic fate) or cholangiocytes (biliary fate).

5. Hepatocyte maturation phase continues past birth, culminating with metabolic 'zoning' of the liver lobes.

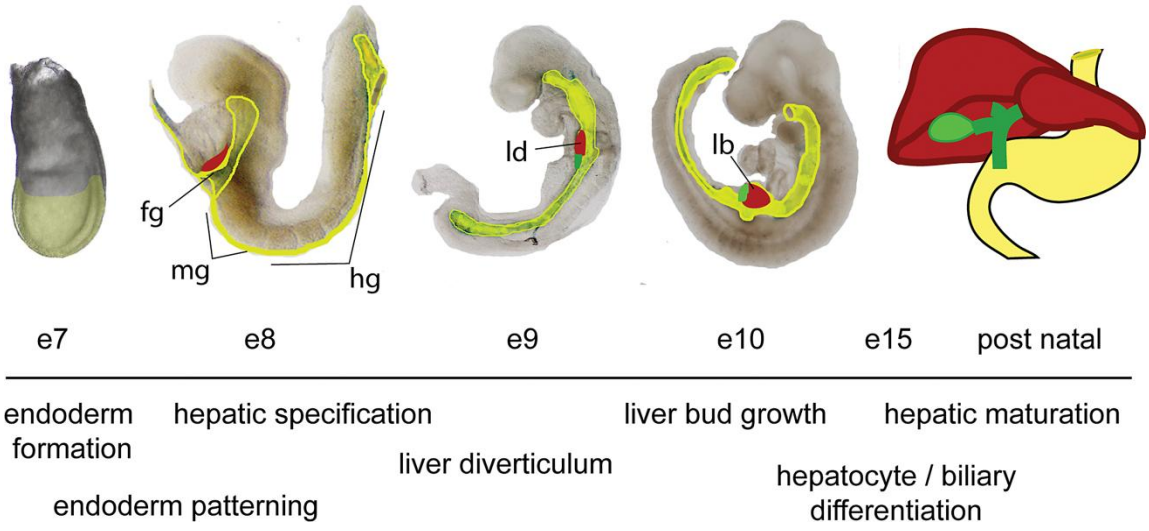


Figure 1.1 Overview of the stages of liver development (mouse model). [51]

1.4.1 Specification of hepatic fate

Wnt/ β-catenin signaling promotes Nodal and Activin initiation of both endoderm and mesoderm formation [52]. Following Activin/Nodal signaling, a Smad2/3/4 complex translocates to the nucleus and stimulates expression of a core group of endoderm transcription factors including Sry-related HMG box SOX17 and forkhead transcription factors

FOXA1-3 (previously known as Hepatic Nuclear Factors α, β, γ), which regulate the signaling cascade driving endoderm differentiation. Matrix metalloproteinases MMP2, 4, and 24 are highly expressed in the mesodermal tissues and induce expression of FOX and GATA transcription factors. SOX17 partners with β-catenin to transcribe Hepatic Nuclear Factor 1 homeobox B (HNF1B), FOXA1, and FOXA2 [52]. At the same time, the gut tube is differentiating into the foregut, which contains the precursors of the liver, gall bladder, pancreas and lungs.

At this point, the pre-hepatic endoderm is developmentally 'competent' – lineage is not yet specified but cells have acquired the capacity to respond to specification-inducing signals. In the chick model this has been shown to be via expression of Fibroblast Growth Factor (FGF) receptors FGFR1 and FGFR2 by the hepatic endoderm [53]. The regional identity of the endoderm is regulated by overlapping temporal and spatial gradients of FGF2/4 from the nearby heart; Wnt and Bone Morphogenic Proteins (BMP) 2 and 4 from the developing Septum Transversum Mesoderm (STM); and retinoic acid from the mesoderm. Only the foregut endoderm is able to develop into the liver. Recent evidence suggests that FGF4 and Wnts secreted from the posterior mesoderm repress foregut fate and promote hindgut development, and FGF4 and Wnt inhibition in the anterior endoderm are required to establish foregut identity. This appears to be mediated by expression of Wnt inhibitors SFRP5 and DKK1 by the foregut endoderm [54]. BMP signaling is required, but not sufficient, for hepatic induction and may act by inducing and maintaining GATA4/6 expression. HNF1B also stimulates expression of FOXA1 and FOXA2 in the pre-hepatic endoderm. FOXA1-3 and GATA4 open the compact chromatin and bind the promoter region of Albumin. This binding provides access for other

transcription factors such as nuclear factor 1 (NF1) and C/EBP-β, initiating albumin transcription. In addition to the FGFs and BMPs, Wnt in the lateral plate mesoderm is required for hepatic specification [55]. The specific Wnt family members required for hepatic specification are still unknown, but WNT3A, WNT5A, and WNT9A are candidates [42, 56, 57].

These factors induce hepatoblast specification, inducing expression of the earliest markers identifying hepatoblasts from the surrounding endoderm cells: Prox1, HHEX, Albumin, transthyrein (TTR), and AFP [58]. Hepatoblasts are bi-potential cells that are morphologically similar to adult oval cells, and are capable of differentiating into either hepatocytes or biliary epithelial cells (BEC) [59]. GATA6 maintains hepatoblast differentiation [60].

1.4.2 Formation of the hepatic bud

Shortly after hepatic specification, the epithelium begins to express liver-specific genes Albumin (ALB), Alpha feto-protein (AFP), and Hepatic Nuclear Factor 4α (HNF4A), and thickens to form the liver diverticulum (around day 9 in mice and day 22 in humans). The liver diverticulum is lined by endodermal cells. Proliferating hepatoblasts strongly express EPCAM and DLK1 [61] and form a tissue bud delineated by a basement membrane containing laminins, collagen IV, nidogen, fibronectin, and heparin sulfate proteoglycan (HSPG). HAND2 regulates the gut-looping process that defines the beginning of the liver bud by remodeling the extra-cellular matrix (ECM) through MMP-mediated reduction of laminin deposition [62] .

The basal layer surrounding the hepatic endoderm begins to break down under the regulation of ONECUT1 and ONECUT2 expressed in the foregut endoderm and hepatoblasts. MMP2 is secreted from the STM and activated by MMP14, a membrane-bound protein expressed exclusively by the hepatoblasts at the onset of basement membrane degradation [63]. At the same time, the hepatoblasts undergo a transition to a pseudostratified epithelium as a result of nuclear migration promoted by HHEX. It has been proposed that T-box transcription factor 3 (TBX3) stimulates PROX1 expression at this point, which then functions as a co-receptor of liver receptor homolog 1 (NR5A2) to induce delamination and migration of the hepatoblasts through the weakened basement membrane into the STM (hepatic mesenchyme) to form the beginnings of the liver bud. This process is similar to an epithelial-mesenchymal transition (EMT) in that the hepatoblasts temporarily lose their epithelial morphology and reduce expression of E-cadherin as they move away from the endoderm. Isoprenylcysteine carboxyl methyltransferase (ICMT), basigen (BSG), and several MMPs (1, 7, 11,12, 15, 16, 17, 19, 23, and 25, as well as TIMP2 and TIMP4) also participate in remodeling the basement membrane. During this phase, GATA4 maintains the integrity of the septum transversum, while GATA6 is required to maintain differentiation of the hepatoblasts and FGF1, FGF4, and FGF8 prevent further differentiation of the hepatoblasts into hepatocytes [64]. VEGFR-2 (KDR) is required for blood vessel formation as hepatoblasts migrate into the stroma. A Glial-derived neurotrophic factor called Neurturin may also be required for hepatoblast migration and/or proliferation. Neurturin is secreted from blood vessels and acts as a hepatoblast chemoattractant via GFRα2 receptors on the hepatoblast surface membrane [65].

1.4.3 Liver bud growth

Once the liver bud has formed, it begins to grow rapidly via hepatoblast proliferation under the control of multiple signaling pathways including Hepatocyte Growth Factor (HGF), Transforming Growth Factor β (TGFβ), Hepatoma Derived Growth Factor (HDGF), and Wnt (Figure 1.2). Growth factor ligands are secreted by the hepatic mesenchyme (STM) and bind to receptors on the surface of the hepatoblasts, triggering expression of transcription factors such as ELF5, ARF6, ATF2/7, RAF1, c-jun, TBX3 , NFκβ, FOXM1B, XBP1, and MTF-1 that control proliferation, migration, and survival.

HGF is expressed by the STM, endothelial cells, and hepatoblasts. Its receptor c-met (MET) is found on the hepatoblast surface, and initiates a cascade to activate ATF2 and ATF7, resulting in transcription of genes that initiate cell cycle progression [66]. ATF2 and ATF7 also dimerize with JUN and other proteins to form the AP1 transcription factor, which is essential in providing a negative feedback loop to protect the hepatoblasts from apoptosis [66]. HGF/MET also promotes hepatoblast migration in part by activating the small GTPase ADP-ribosylation factor 6 (ARF6) [67]. Because HGF also promotes hepatocyte differentiation, TGFβ/TGFβ-RIII and Hedgehog (Hh) signaling is necessary to inhibit differentiation of the hepatoblasts during this stage of rapid growth [59]. TGFβ also stimulates proliferation via Smad2/Smad3 signaling. Hepatoma-derived growth factor (HDGF) is produced by hepatoblasts and stimulates their proliferation. Once hepatoblasts mature into hepatocytes, expression of HDGF ceases. However, HDGF is not required for normal liver development so it may be an as yet unidentified compensatory pathway.

Wnt signaling is necessary for hepatoblast proliferation, but the complexity of the Wnt network in the liver makes Wnt signaling difficult to study. For instance, 11 Wnt ligands and 8 Frizzled receptors are expressed in the mouse liver. Thus, the exact Wnt ligands involved in humans are still unknown, but WNT5A and WNT9A are candidates and are expressed by mesenchymal, sinusoidal, and stellate cells. β-catenin plays a definite role in stimulating hepatoblast proliferation and differentiation and Wnt/β-catenin control the global liver morphology. β-catenin also seems to be a key node at the intersection of multiple signaling cascades and interacts with the HGF receptor c-MET, SMAD2/3, and ELF5. FGF-10 (secreted by myofibroblastic cells) controls β-catenin activation and also stimulates proliferation of hepatoblasts.

Liver bud growth also requires retinoic acid (RA) signaling, which is controlled in part by the zinc finger transcription factor WT1 expressed in STM and stellate cells. The retinoic acid receptor RXRα is expressed in mesodermal cells scattered between the hepatoblasts and are often in contact with sinusoids. This suggests that RA stimulates hepatoblast proliferation by inducing production of trophic factors by the mesodermal cells rather than a direct effect on the hepatoblasts.

Several transcription factors are also involved in regulating hepatoblast proliferation: PROX1 promotes proliferation via suppression of p16 (CDKN2A, a cyclin dependent kinase inhibitor) [68]. PROX1 activity is regulated by liver receptor homolog 1 (LRH1, also called NR5A2). FoxM1B activates expression of CDK1 and Cyclin B, regulators of the G2/M phase of the cell cycle. X-box binding protein 1 (XBP1) controls the expansion of the ER surface in

growing hepatoblasts. Inhibitor of differentiation 3 (Id3), which may act downstream of FGF and/or BMP signals, is transiently expressed and enhances hepatoblast proliferation by inhibiting the protein TCF3. TCF3 is a Wnt-effector TF that limits levels of several proliferation promoters [69]. AT-hook 2 (HMGA2) is also involved in transcriptional activation of proliferation genes and maintaining cells in an undifferentiated state [70].

The STM expresses homeobox transcription factors HLX, LHX2 and N-MYC that promote hepatoblast proliferation and suppress apoptosis, perhaps by regulating production of paracrine signals from the mesenchyme. The exact mechanism of regulation is still unknown, but there is extensive cross talk. For instance, HGF and TFGβ signaling act in parallel and converge on β1-integrin regulation. FGF and HGF signaling stimulate many of the same intracellular kinase cascades, and both stimulate the activity of β-catenin in the liver bud.

Lee et al (2012) [70] examined differential expression of genes at different time points in mouse development. Genes expressed from GD11.5-12.5 ("early expression") included several that are expressed in embryonic stem cells, including Midkine (MDK), pleiotrophin/heparin-binding growth-associated molecule (PTN), Necdin (NDN), and Proliferation-associated 2G4 (PA2G4). MDK and PTN are essential for development of the catecholamine and rennin-angiotensin pathways. MDK regulates PTN expression. PTN may be secreted from mesenchymal cells as a mitogen of parenchymal cells in the embryonic liver. NDN is expressed in primitive stem cells and is involved in hematopoietic stem cell regulation. Other genes found highly expressed in the proliferative phase of liver development (in the mouse model) include MAP4K4 (which activates JNK/MAPK8), WNT9B, SRPK1 (regulates alternative splicing), CSNK1D

(activates several important developmental genes including HIF1A, P53, DVL2/3, DNMT1, and YAP1), H19 (a long non-coding RNA that regulates expression of IGF2), and SET domain bifurcated 1 (SETDB1).

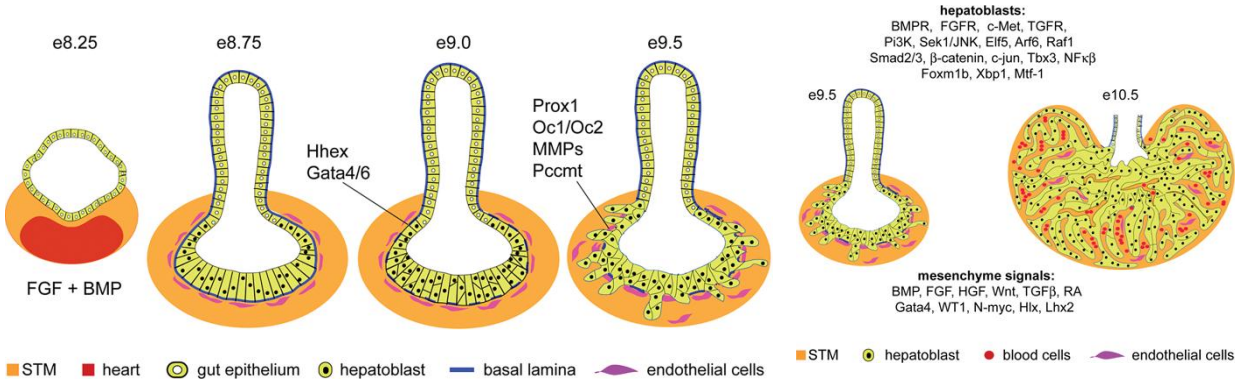


Figure 1.2 Growth of the liver bud [51].

1.4.4 Hepatocyte/ cholangiocyte cell fate determination

Mechanisms of cell fate determination have not been completely characterized. The 'start point' has not been accurately determined, and it appears to occur over more than one developmental stage. During liver bud growth, hepatoblasts begin to express metabolic genes that are active in mature hepatocytes. By the time the liver reaches full size (for the fetus), a

subpopulation of hepatoblasts has generated abundant rough ER and lipid vesicles, indicating that differentiation has begun. To complete differentiation, inhibitors of differentiation must be turned off, including Sonic Hedgehog (SHH) [71].

The differentiation of hepatoblasts into either hepatocytes or Biliary Epithelial Cells (BECs) begins with the expression of Hepatic Nuclear Factor 4a(HNF4A), Albumin (ALB), CEBPA, and AFP in hepatocyte precursors, and cytokeratin-19 (KRT19) and SOX9 in biliary precursors. Hepatoblasts in contact with the portal vein form a layer of biliary precursors that increase KRT19 expression. SOX9 also is re-expressed in the cells near the portal vein branches, and in later developmental stages expression is limited to biliary cells. Vimentin (VIM) is an intermediate filament protein of mesenchymal cells, expressed in the ductal plate and BECs but not hepatoblasts or hepatocytes. Hepatoblasts that are not in contact with portal veins gradually differentiate into mature hepatocytes expressing HNF4A, ALB, and AFP (Figure 1.3).

Regulators of differentiation include TBX3, TGFβ, and Onecut 1/2 (OC1, OC2). TBX3 and OC1 appear to determine the timing of hepatoblast lineage decision, but the exact mechanism is still unknown. TGF-β promotes differentiation of hepatoblasts to biliary cells and represses hepatocyte differentiation. TGF-β signaling is highest near the portal vein, most likely as a result of the high expression of TGF-β2 and TGF-β3 in the periportal mesenchyme and TGFB Receptor 3 (TGFBR3) on the hepatoblasts. OC1 and OC2 modulate the gradient of TGF-β signaling activity, inhibiting TGFβ signaling in the parenchyma to allow differentiation into hepatocytes [72]. NOTCH (JAG1 and its receptor NOTCH2) signaling is also important in biliary

differentiation, and induces expression of HNF1B and SOX9, which modulates TGFβ signaling [73].

These factors act in part by regulating several liver-enriched transcription factors including C/EBPα, HNF1α, FOXA1-3, HNF4α and nuclear hormone receptors. HGF stimulates expression of C/EBPα to promote differentiation toward hepatocyte lineage, while FGF2 and FGF7 induce differentiation towards biliary lineage in cooperation with BMP4 and ECM components. FGFR1 and FGFR2 expression, which disappeared after hepatic specification of the endoderm, reappear in the ductal plates and the developing intrahepatic bile ducts (IHBD), but not in hepatocytes. BMP4 may also be involved in bile duct formation by controling FGF2/7-induced epithelial branching [53]. Wnt signaling (possibly Wnt3a) represses hepatocyte differentiation and promotes biliary differentiation. Specific mechanisms are still unknown but Smad5 is expressed at high levels in early differentiating cholangiocytes. Type IV collagen and laminin are expressed in and accumulate in the basement membrane of the mesothelium, portal vein, and BECs. Fibronectin (FN1) and type I collagen are expressed in connective tissues surrounding the bile ducts and veins.

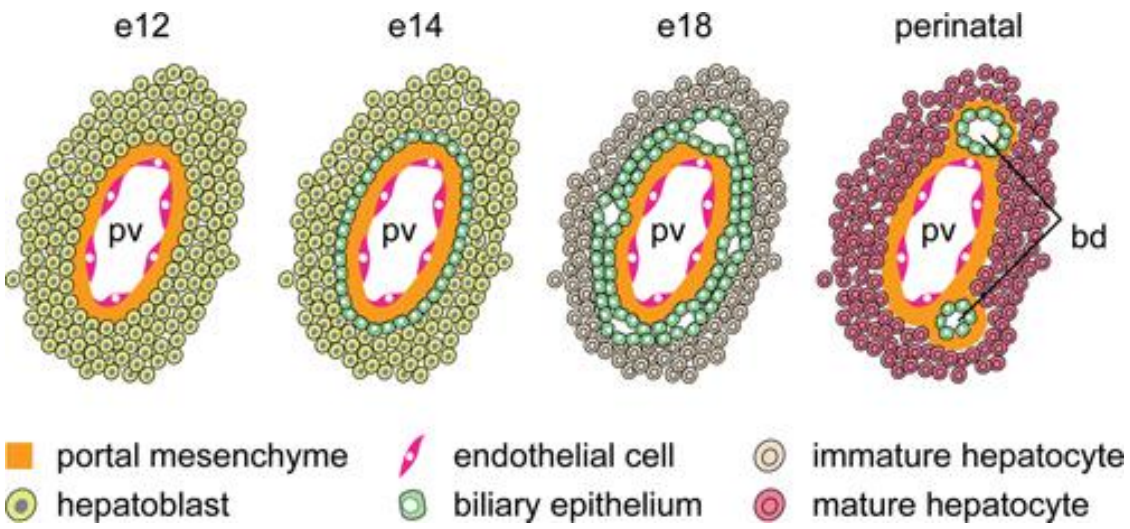


Figure 1.3 Spatial dynamics in hepatocyte/cholangiocyte differentiation [51].

1.4.5 Hepatocyte maturation

Hepatocyte maturation is a process that extends throughout development and after birth. A set of six transcription factors (HNF-1α, HNF-1β, HNF-4α1, OC1,NR5A2, and FoxA2) form a network of auto- and cross-regulatory loops whose interactions increase in number and complexity as maturation proceeds [51]. These six factors occupy the gene regulatory regions of each other and of other factors to form an inter-dependent network that becomes more stable over time. They also cooperate with other cofactors. For instance, OC1 and HNF-4α are both required to stimulate the expression of glucose-6-phophatase (G6PC), a key metabolic protein. Increased levels of C/EBPα and OC1 proteins are required to stimulate association of these factors with the Creb Binding coactivator protein (CBP) then bind the FoxA2 promoter.

Co-activators which initiate expression at particular time points allow for an increasing variety of interactions with the transcription factors.

These six transcription factors play important roles in metabolism in the maturing liver, as well as determining hepatocyte morphology. In particular, HNF-4α is required for normal expression of at least 25 genes whose products are involved in cell junction assembly and adhesion [74], and in regulating the hepatocytes response to the accumulation of unfolded proteins in the endoplasmic reticulum (the endoplasmic stress response). Battle et al (2006) demonstrated that over 550 genes are down-regulated more than 3.5-fold and about 25 genes were up-regulated >2.5 fold in HNF4α-null embryonic mouse livers [74]. These genes are involved in transport, signal transduction, protein folding, nucleic acid metabolism, metabolism, immune response, electron transport, cell adhesion, and cell death. It acts on so many targets via a multitude of functional domains and several cofactors, including CITED2 and thyroid hormone receptor interacting protein 3 (ZNHIT3) (which also interacts with retinoid X receptors).

Hepatocyte maturation also requires repression of a number of genes during the prenatal and postnatal periods. Zinc finger factors ZHX2 and ZBTB20 repress AFP and GPC3 post-natally [75]. ZBTB20 binds to the AFP promoter to inhibit transcription. Organization of hepatocytes into cord-like structures is driven by the small guanosine trphosphatase adenosine diphosphate-ribosylation factor 6 (ARF6). ARF6 is activated in response to HGF. Wnt signals are also involved in hepatocyte maturation.

A third component of hepatocyte maturation involves specialization of zones in response to a variety of extracellular signals. Differentiating hepatocytes are closely associated with hematopoietic precursor cells which colonize the embryonic liver. Near the end of gestation and into the post-natal period, the hematopoietic cells leave the liver and migrate to the bone marrow. These cells are essential to hepatocyte maturation because they secrete oncostain M (OSM), a cytokine related to IL-6 which binds the IL6ST (gp130) receptor at the hepatocyte membrane, inducing a STAT-3 mediated signaling cascade [51]. This stimulates expression of the terminal hepatocyte differentiation markers glucose-6-phosphatase (G6PC) and phosphoenolpyruvate carboxykinase. Jumonji is a transcription factor expressed in several cell types that is necessary for activation of OSM. It also promotes morphological maturation into polarized epithelium via K-ras and E-cadherin. HGF and OSM activity is balanced by TNFα, which inhibits maturation and maintains the proliferative capacity of the fetal hepatocytes. This is necessary to allow the liver to grow to the appropriate size before differentiating. TNFα production decreases after birth.

The HIPPO signaling pathway is emerging as a critical regulator of proper organ size. Evidence suggests that it plays roles in cell contact inhibition and organ size control via cell proliferation inhibition and promoting apoptosis throughout the body. YAP1 overexression leads to reversible liver enlargement (up to as much as 25% of body size) in embryonic mice (Zeng and Hong, 2008). YAP has been shown to bind to several transcription factors including p73, p53BP2, RUNX2, SMAD7, ERBB4, and TEAD/TEFs, resulting in (at least) upregulation of MKi67, c-Myc, SOX4, H19, AFP, BIRC5 (survivin), and BIRC2. YAP1, in turn, is down regulated by MST1 and STK3. Inactivation of MST1 or STK3 at any stage of development leads to multiple

large liver tumors via oval cell induction, and they are also required to maintain hepatocyte quiescence in adult livers.

1.4.6 Post-natal development

Hepatocyte maturation is not complete at birth. Cytochrome p450 genes and HGF appear to be involved but this has not been well-studied. In the weeks after birth, metabolic zonation of the hepatic lobes begins. Within each lobe a periportal and pericentral zone are established on the basis of their expression of different metabolism -regulating genes. HNF-4α contributes to this zonation by repressing periportal expression of glutamine synthase via deacetylase type I. Wnt signaling also contributes to zonation. β-catenin is found only in the pericentral area and its negative regulator APC is found only in the periportal hepatoctyes. Because APC is the key regulator of β-catenin levels along the lobular axis it has been proposed as the master regulator of zonation, but the identity and source of the Wnt ligand(s) are still unknown. It has recently been shown that HNF4α also contributes to liver zonation, acting through cross-talk with the Wnt pathway [76].

1.5 Molecular mechanisms in liver wound healing

Wound healing is the normal response of tissue to an injury, and liver fibrosis occurs as a result of repeated cycles of injury and repair. Normal hepatic wound healing involves 7

distinct phases: inflammation, production of cytokines and growth factors, myofibroblast activation, ECM production, angiogenesis,maturation, and remodelling. Inappropriate repair and scarring occurs if any element is interrupted or overactivated [77]. In this section, we identify the genes involved in liver wound healing that will be used to define the gene sets that might identify signatures of liver wound healing in our cirrhosis or tumor samples.

Inflammation begins with the local production of MMPs at the site of injury that result in disruption of the basement membrane, allowing inflammatory cell infiltration, mostly neutrophils then macrophages and lymphocytes. Leukocytes eliminate invading organisms and remove dead cells. Inflammation produces profibrogenic cytokines and chemokines which activate hepatic stellate cells (HSCs), causing transdifferentiation to a myofibroblast phenotype. Myofibroblasts are the key effectors of wound contraction and repair, and inappropriate activation is the central mechanism of fibrosis. Activated myofibroblasts migrate to the site of injury and proliferate, producing type I and type III collagen (COL1A1,COL1A2,COL3A1), several MMPs and TIMPs [77].

Fibroblasts are mesenchymal in origin, but in the liver there are four distinct sources of myofibroblasts: HSCs, bone-marrow-derived mesenchymal cells, portal fibroblasts (near the biliary tree), and epithelial cells (hepatocytes and cholangiocytes) via EMT [77]. HSC activation results in their transdifferentiation into myofibrobasts in two phases. During initiation phase the cells become responsive to cytokines and growth factors, followed by the perpetuation phase. Initiation is paracrine while perpetuation is both autocrine and paracrine. Injured hepatic cells produce necrotic cell debris and reactive oxygen and nitrogen species (ROS and

RNS). Myofibroblasts express toll-like receptors TLR4 and TLR9. DNA from apoptotic hepatocytes can activate TLR9 in HSC and contribute to activation. HSCs are also antigen-presenting cells to activate immune processes [77].

Activated HSCs develop new autocrine pathways to maintain the activated state, including TGF-β, Angiotensin II, PDGF, monocyte chemoattractant I (CCL2), and VEGF. HSCs also express new membrane receptors that prime them to respond to inflammatory mediators and growth factors, including IL-6, TGF-β, and PDGF receptors. TIMP1, integrins and other adhesion molecules contribute to HSC survival and perpetuation of the myofibroblast phenotype. The activated HSC migrates to the site of injury, secreting large amounts of ECM and regulating ECM degradation. In the early phases of liver injury, they transiently express MMP3, MMP13, and uroplasminogen activator (PLAU). In later stages of injury and activation, the cells express a combination of MMPs including pro-MMP2 and membrane type I MMP (MMP14), which drives generation of active MMP2 and local degradation of the matrix to facilitate replacement with a high density interstitial matrix. TIMP1 expression is also increased [77].

Chronic injury alters the normal healing process and prevents return of the tissue to the preinjury state. Constant inflammation/infection leads to permanent myofibroblast activation, either directly by acting on HSCs or indirectly through paracrine-dependent factors. During chronic hepatic injury, different types of liver cells also may acquire a neuroendocrine phenotype, which may contribute to cell growth, migration, and angiogenesis during wound healing. The hepatic neuroendocrine system is upregulated in the liver following injury and can regulate the pattern of wound healing and regeneration in several ways. Atypically

proliferating cholangiocytes, also known as reactive bile ductules, aquire a neuroendocrine phenotype and are major contributors to the production of a number of neuroendocrine factors in areas of maximal cell death and inflammation [77]. Hepatic progenitor cells (oval cells) lie in or adjacent to the canal of Herring and express neuroendocrine proteins including chromogranin-A (CHGA), neural-cell-adhesion molecule (NCAM1, NCAM2), parathyroid-hormone-related peptide (PTHLH), S-100 protein (family of 21 proteins S100A1-16, S100B, S100P, S100Z), neurotrophins (NGF, BDNF, NT3, NTF5) and their receptors (NTRK1-3, NGFR). These cells are activated to proliferate in chronic liver damage situations where proliferation of hepatocytes is inhibited (NASH, cholestatic liver disease, alcoholic hepatitis and viral hepatitis). Progenitor cells differentiate into hepatocytes, and newly formed intermediate hepatocytes continue to express CHGA [77].

HSCs also express a number of neuroendocrine markers, including synaptophysin (SYPL1) which is correlated with neuroendorine differentiation, neutrophins and neural cell adhesion molecules, along with their receptors, which makes them responsive to neuroendocrine regulation in wound healing [77]. This neuroendocrine differentiation in the liver is associated with cellular stress and inflammation and is regulated by IL-6 and TNF. Differentiation can also be induced by interaction with type IV collagens and HSPG [77]. Once differentiated, the cells produce several neuropeptides including serotonin, endocannabinoids, opioids, and neutrophins that contribute to contraction, migration, proliferation, and ECM production in activated HSCs. Activation of the serotonin receptor HTR2B on fibrogenic HSCs suppresses hepatocyte proliferation through augmented production of TGFβ1 [78].

The Hedgehog pathway, once thought to be exclusively embryonic, is now known to be activated in response to some injuries, including the growth of hepatic progenitor populations, hepatic accumulation of myofibroblasts, repair-related inflammatory responses, vascular remodeling, liver fibrosis, and hepatocarcinogenesis [79]. BMP2 and BMP4 are transiently expressed in the oval cells, but not Kupffer or macrophage cells in the early stage of liver injury. It plays an as-yet-unknown role in the proliferation and differentiation of progenitor cells in response to liver injury [80].

NFKB is recognized as a regulator of hepatic inflammation and wound healing. The classic pathway is induced in response to inflammatory mediators and microbial or host ligands of the Toll-like receptor system [81]. These stimuli activate the inhibitor of NF-KB kinase complex (CHUK, IKBKB, IKBKG), leading to phosphorylation of the inhibitor IκBα and nuclear transport of active NF-κB. NF-κB is actually several different homo- or hetero-dimers of five different subunits (REL, RELA, RELB, NFKB1, and NFKB2) that have non-overlapping functions. c-Rel (REL) is expressed in adult mouse liver and knockout mice display defects in liver wound healing and regeneration [81]. REL induces expression of CCL5 (RANTES), which remains elevated until healing is complete. CCL5 recruits neutrophils to sites of injury, and can also target HSC to promote their proliferation and migration. REL may also regulate expression of collagen I and α-SMA in HSCs [81].

1.6 Molecular mechanisms in liver regeneration

Regeneration of hepatocytes (liver mass) has been hypothesized to play a role in liver carcinogenesis. In this section we review the molecular events that occur during regeneration after partial hepatectomy in order to identify genetic signatures that might indicate whether these processes are also occurring in our cirrhosis or tumor samples.

Liver regeneration after loss of functional mass has three main phases: initiation or priming with progression of the quiescent hepatocytes to repeated division; proliferative phase, restoring liver volume; termination of growth and balancing functional regions of the liver [82]. In normal tissue, hepatocytes are long lived and rarely divide, with a replication rate of 1 in 20,000. Under normal conditions hepatocytes are unresponsive to growth stimuli. Triggering events include not only partial hepatectomy, but also blunt injury, metabolic stress due to toxins, disruption to intercellular contacts, or digestion of the ECM. Priming signals include lipopolysaccharide (LPS), produced by gut flora and released through a deteriorated intestinal barrier (ie surgical stress), which activates Kupffer and stellate cells to increase production of TNFα and IL-6. Complement factors C3a and C5a from circulating blood act as quickly as LPS. TNFα and IL-6 cause transcription factors NF-Kb, STAT3, c-JUN and CEBPβ to bind DNA rapidly by means of posttranslational modifications. Within 30 minutes, expression of "immediate early release" genes are up-regulated, including c-FOS, c-JUN, c-MYC, and c-MET. HGF, TGFα, and Epidermal Growth Factor (EGF) allow cells to overcome the G1 restriction point and enter mitosis. Priming signals also come from the pancreas (insulin), duodenum or salivary glands (EGF), thyroid gland (T3), and adrenal glands (norepinephrine). Additional priming phase

upregulation of urokinas-type plasminogen activator (PLAU) and its receptor (uPAR) leads to activation of HGF and ligands for EGFR.

The Proliferative phase is characterized by mitotic waves of hepatic cells. Hepatocytes reach the S phase first, with DNA synthesis rising at about 12 hrs after injury and peaking at about 24 hrs. S phase occurs later in nonparenchymal cells - 48 hours for Kupffer and biliary cells, and 96 hours for endothelial cells. Injury via necrosis or apoptosis of hepatoctyes involves similar cell priming, but replicative waves are less coordinated. During proliferative phase, almost all of the hepatocytes undergo mitosis (95% in young rats, 70% in old animals; unknown in humans). The proportion of binucleate cells increases, and some hepatocytes become polyploid but undivided. Early Growth Response Factor (EGR1) is elevated 6-fold by 12 hours after partial hepatectomy and may act by promoting TNF expression [83]. REL (a subunit of NF-κB) is also required for hepatocyte DNA synthesis during hepatocyte proliferation, and may control the timing of FOXM1 expression, which is required for normal mitosis in both development and regeneration [81]. FOXM1 is a direct target of REL, but only in response to injury/regeneration. Subsequent targets for transcriptional stimulation of DNA replication by FOXM1 are Cyclin B1 and CDC25C. c-JUN up-regulates a hepatotrophic factor stimulating hepatocyte proliferation, Human augmenter of liver regeneration (GFER), which protects hepatocytes from apoptosis [84]. Hepatocyte proliferation inhibitors must also be repressed during the proliferative phase, including CDH1, MST1, TGFB, and BMP2.

The Termination phase is still not well-understood. It is not yet known if onset of inhibitory genes or withdrawal of stimulatory genes stops regeneration. Reappearing ECM may

play an important role by renewed binding of pro-HGF. TGF-β1 has a proposed role. Disappearance of TGF-β1 from the periportal to pericentral region of lobules enables progression of hepatocyte mitotic wave in the same direction at the onset of regeneration. TGF-β1 released in the plasma shortly after injury is probably inactivated by binding to α2 macroglobulin, and hepatocytes are transiently resistant to the mito-inhibitory effects of TGF-β1 during the proliferative phase. After the refractory period, TGF-β1 could play a role in ending the regeneration. Plasminogen activator inhibitor (SERPINE1/2, SERPINB2) is induced by IL-6 and blocks HGF action by inhibiting cleavage of pro-HGF into active HGF. Suppressor of cytokine signaling -3 (SOCS3), also upregulated by IL-6, causes down-regulation of STAT3, ultimately terminating the original IL-6 signal. Apoptosis may also play a role in correcting the final size of the liver.

Ho, et al (2007) studied gene expression profiles following human partial hepatectomy and identified a set of differentially regulated genes including immune response genes SAA1-2, CRP, and SOD2, cell growth genes SOCS3, RASD1 and NAMPT, along with genes involved in signal transduction, biosynthesis, and metabolism [85].

Many of the early response genes in liver regeneration are also critical regulators of embryonic development (HGF, NF-κB, STAT3, and c-JUN). NF-κB and c-JUN protect hepatocytes from the apoptotic effects of TNFα during liver bud proliferation, and likely serve the same purpose in the regenerating liver. STAT3 suppresses Cyclin D1 expression to control the rate of hepatocyte proliferation, and HGF has multiple functions throughout liver development.

1.7 Conclusion

In this chapter we have reviewed the important genetic changes that drive the processes that we suspect are involved in the development of HCC. Using this knowledge of developmental and healing processes, we can define gene sets that characterize each specific stage of development. We will also define gene signatures from this review to identify whether particular processes such as hepatoblast proliferation or hepatocyte proliferation were occurring in our cirrhosis and tumor samples. Our hypothesis was that these important genes may be working in a coordinated fashion in liver disease but that the signal strength from these genes in microarray experiments might be difficult to discern against the background of metabolic disturbances that have much larger fold-changes. We suspected that some of the gene expression changes were occurring in activated hepatic stellate cells, stem cell niches, or sub-populations of tumor cells. When gene expression changes occur in cell populations that make up a small proportion of the total sample (as the above scenarios do), then overall signal strength will be fairly low compared to even modest gene expression changes that occur in the majority of cells in the sample. Since standard analysis of microarray experiments focus on the largest magnitude mean expression changes, these comparatively suble signals may not be recognized. A targeted approach of specifically examining changes in important driver genes may allow a deeper understanding of tumor biology.

Chapter 2: Microarrays and processing methods

2.1 Introduction to microarray technology

First introduced in 1995, microarray chips using hybridization of fluorescently labeled targets to cDNA probes have revolutionized the study of genomics. There are several different microarray systems, but the two main chip types are the one-channel arrays (ie, Affymetrix) and two-channel arrays. Two channel arrays are made by attaching pre-made oligonucleotides of fixed length onto slides and simultaneously hybridizing experimental and control samples which have been labeled with different color fluoresence. One channel arrays, which have become more popular in recent years, are made in situ and hybridize a single sample per slide, or GeneChip. The Affymetrix HGU133A2 chips used in this study use probes that are 25 bases long, with about 11 probes per probe set attached to random locations on the chip. Many genes have multiple probe sets that map to different locations on the gene. Each Perfect Match (PM) probe has an accompanying MisMatch (MM) probe formed by switching the middle base of the PM sequence. This was intended to measure non-specific hybridization but is not generally used anymore. The general experimental process is to extract messenger RNA (mRNA) from a sample, reverse-transcribe and convert to double-stranded complimentary DNA (cDNA), then amplify to complimentary RNA (cRNA) that is tagged with a fluorescent label that

can be detected with a scanning device. The labeled cRNA is chemically fragmented then hybridized to the GeneChip.

Although microarray technology is a powerful tool for studying molecular biology, there are inherent limitations that limit the effectiveness of microarray experiments. All microarray experiments have both biological and technical sources of variation. Biological variation results from differences in tissue samples, cell type mix between samples, genetic polymorphisms, differences in mRNA levels among individuals and their cells due to gender, age, disease state, and genotype-environment interactions, among others. This biological variation is the component that is of interest to researchers. Technical variation, or "noise" that obscures detection of biological signals, results from differences in sample preparation, labeling, hybridization, and other steps of sample processing. Even inconsistencies in the environmental conditions (room temperature, humidity, and ozone levels for example) can introduce technical variation.

Further, microarray experiments have some biological limitations. First, it only measures relative expression values in the form of intensity of fluorescence. There is background fluoresence in every experiment, and the amount varies between chips. This can be fairly well corrected for using statistical "background correction" models, described below, however, without an absolute measure of expression it is difficult to determine which RNA products have "no" expression vs. "low" expression. A more serious problem is that the technology only measures the relative abundance of mRNA, which is not a direct indicator of corresponding protein abundance or activity. mRNA expression is often assumed to be low

because there is no demand for the corresponding protein, but differences in protein stability and turnover rates may affect the correlation between mRNA and protein abundance. However, several studies have demonstrated that most mRNA levels generally correspond to protein abundance [86, 87]. For this reason, microarray experiments are still quite useful but should be interpreted carefully.

2.2 Microarray data preparation and analysis

In order to maximize the measurement of actual biological variation between experimental and control groups, it is necessary to remove as much of the technical variability as possible. This involves extensive pre-processing of the data. Four or five separate steps are generally required:

- Quality assessment of each chip must be done to remove chips (samples) with significant systematic bias caused by technical errors in one or more sample processing steps.
- Background correction is intended to remove nonspecific background intensities of scanner images.
- PM correction is to correct for the effect of nonspecific hybridization. However, several algorithms ignore this step and only use the perfect match (PM) signals.
- Normalization attempts to reduce most of the non-biological differences between chips so that the signal intensities are comparable across chips.

- Summarization is the final stage in pre-processing, where the expression values from all probes in a probe set are summarized into a single expression value.

It has been shown elsewhere [14-17] that, for each step, using different methods can have a profound impact on the resulting list of differentially expressed genes (DEG) for an experiment. Moreover, no single technique has been shown to be superior to all others in all situations [88, 89]. Differences in performance depend on several factors, including how well the data conforms to the assumptions of the statistical models employed, the degree of correlation between important and unimportant genes, and whether the kind of technical variation that a method is designed to correct is the dominant source of variation in the dataset of interest. Therefore care should be taken to choose the most appropriate techniques for each experiment.

2.2.1 Quality Assessment metrics

The measurement of gene expression by microarray technology, as with any laboratory procedure, possesses an associated error due to both random error and technical differences in sample processing. There are many sources of technological error in microarray processing that can introduce significant bias into the statistical analysis if not recognized and corrected. If a chip has systematic sources of technical variation it may dilute the ability to extract meaningful information from the sample, and could also distort the results of subsequent pre-

processing steps, so such chips must be identified and excluded before proceeding with further processing. The quality of the chips strongly influences the diagnostic and predictive power obtained from the data, and a poor quality "training set" may lead to misclassification of future samples. Investigators desire large sample sizes for maximum statistical power and biological information, but conservatively limiting one's dataset to only include the chips with the least technical error will minimizes potential bias or false results. The decision of how to balance quality and quantity varies widely among studies, and there is still no consensus regarding the best methods or objective criteria for assessing chip quality.

In the case where technical replicates can be obtained (performing the microarray experiment 2-5 times on a single divided biological sample), technical variation can be fairly well modeled by a variety of techniques, assuming that the variation in conditions within the replicate groups is similar to the variation in conditions between samples. However, it is not always possible to generate technical replicates. Microarray technology is expensive, and the researcher faces the decision of whether to process more samples (higher N) or replicate samples simply to assess technical variance. In addition, certain types of biological samples are difficult to obtain in quantity (i.e. biopsy tissue) and patient safety must be balanced with the desire to obtain large quantities of tissue for analysis. Finally, early microarray experiments were run before the full extent of the impact that technical variability can have on results was realized, so older datasets may not have technical replicates available.

Commonly used methods for Quality Assessment (QA) include metrics from the SimpleAffy package in the R programming system. The "qc" function is applied to generate

scale factors, percent present calls, and min/max/avg background calculations, 3' to 5' ratios for GAPDH and β-actin, and values for spike-in controls. The scale factor is used to scale all probe sets to a target value (usually arbitrarily set to 100). Large-scale differences between chips may indicate cases where the normalization assumptions are likely to fail due to issues with sample quality or amount of starting material, or issues with RNA labeling, scanning or chip manufacture. Affymetrix recommends that scale factors be within 3-fold of each other. Percent Present calls measure the difference between Perfect Match (PM) and MisMatch (MM) values for each probe pair in a probeset. Probesets are only called present when the PM value is significantly above the MM probes. Significant variation in the % Present call in an array compared to other arrays of the same type of tissue may indicate a problem in hybridization on that chip. 3':5' ratios of housekeeping genes such as GAPDH and β-Actin near one indicate successful cDNA and cRNA synthesis [90].

Reimers and Weinstein (2005) recommend further methods to examine quality of chips, including the correlation between a probe and its neighbor, correlation between rows on a chip, and the log(PM/MM) ratio, which are all generated using the bias.display R package [91]. Because probes are placed randomly on a chip, the correlation in expression values between neighboring probes should be zero when no technical variation in present. Less than 30% correlation is considered acceptable, 30-40% is of questionable value, and any chip with an average 40% correlation or more has considerable systematic bias which may be too severe to correct for in the normalization step. Similarly, the correlation between rows should be close to 1. Like the % Present calls, the log(PM/MM) should be comparable across chips.

2.2.2 Background correction and normalization

Background fluorescence can arise from many sources, such as deposits left after the wash stage and optical noise from the scanner [92]. There is also a good deal of non-specific hybridization of labeled mRNA, both to the chip surface and to probes with similar sequences. Removal of this ambient non-specific signal from the total intensity readings is called background correction. Affymetrix and other modern high density chips have probes placed so densely that a "local" background measurement is not possible. Instead, the background must be estimated from the probe signals themselves.

The two most commonly used background correction methods are MAS5 and RMA. MAS5 (Affymetrix Microarray Suite 5.0) is a commonly used regional adjustment method. The entire array area is divided into 16 rectangular zones and the lowest 2nd percentile of the probe values are chosen to represent the background value in given zones [93]. The background value is computed as the weighted sum of the background values of the neighboring zones. Robust Multi-array Average (RMA) is actually a three step process of background correction, normalization, and probe set summarization. The background correction step uses a signal/noise convolution model in which PM intensity distribution is modeled as an exponentially distributed signal and a normally distributed background component [94].

Normalization

The intent of normalization is to remove, as much as possible, the differences in signal intensity due to technical variation in the physical processing of the GeneChips so that values for particular probe sets can be effectively compared across chips. Some of the sources of technical (systematic) variation include differences in the amounts of sample exposed to a particular chip, the fluorescent label used, differences in the settings of the equipment used, and a host of other environmental sources which may be difficult or impossible to completely control such as the local humidity and ozone levels on the day a chip is run.

It has been demonstrated that different normalization procedures can change modeled expression values enough to result in significant differences in what genes are called as significantly differentially expressed [95]. Because one of the hallmarks of good science is the reproducibility of results, this is a significant concern. Several recent papers have compared a variety of pre-processing combinations in an effort to find the "best" procedure, or at least define guidelines as to appropriate usage of the different methods [94-97]. Normalization strategies of high density oligonucleotide array chips such as the Affymetrix GeneChip are different from that of spotted oligonucleotide or cDNA arrays. The Affymetrix GeneChip uses multiple probes for a gene and a single-color detection system with one sample per chip. Therefore, GeneChip normalization is done between arrays and is the focus of this discussion. Hundreds of strategies have been proposed so only a few of the most commonly used techniques are briefly described here.

RMA (Robust MultiArray Average) proposed by Irizarry et al performs three of the pre-processing steps in one algorithm: first it does global background correction using Perfect Match probes only, then performs quantile normalization on the log-transformed values before summarizing the probes in the probe set [94]. Quantile normalization forces the probe intensities to have equal density distribution across sample by setting probes with the same ranked intensity in each sample to the same value. It is by far the most common normalization technique in use today, primarily because it is conceptually simple and easily implemented via Bioconductor R packages. Variance Stabilization and Normalization (vsn) is based on the idea that the variance of microarray data is dependent on the signal intensity and that a transformation based on "shift" and "scale" can be found to generate approximately constant variances [98]. The authors recommend using RMA background correction and normalization first, then applying the vsn transformation before summarizing the probesets. GCRMA combines RMA with physical modeling of sequence information of the probes [99].

Most normalization methods, including the ones described above, are constructed on the assumption that the majority of genes are not differentially regulated and/or the number of up-regulated genes is roughly equal to the number of down-regulated genes. The bias introduced when these assumptions are not met, which often occurs when studying cancer, can be serious [100]. Several authors have compared the results from different algorithms and found very poor overlap in DEG lists [101-103] .

A promising approach to the normalization problem is local regression on technical covariates, which derives from the observation that probes with similar physical characteristics

appear to be distorted by similar amounts. Technical regression does not depend on assumptions regarding the number or direction of changed probes; instead, it uses physical information about the probes themselves. Although all of the factors causing distortion cannot be known, some likely candidates are guanine-cytosine (GC) content, location of the probe on the chip, and the probe melting temperature. Guanine-cytosine content relates to melting temperatures and thus its binding affinity [104], and is correlated with measured expression values [105]. Probe location is informative because there are often visible 'patterns' on the scanned chip image that likely relate to non-uniform washing, small smudges and scratches. Melting temperature is defined as the temperature at which 50% of all the molecules of a given DNA sequence are hybridized into a double strand. Higher Tm correlates with the ease of forming a double strand, and is affected not only by the DNA sequence (G-C forms stronger bonds than A-T and thus has higher Tm), but probe length, DNA concentration (which is fixed for probes and variable for samples), and ion concentration in the solution. Because DNA concentration of the sample cannot be completely controlled for, Tm must be estimated and assumed to be equal across samples. There are a number of methods for estimating Tm (http://www.entelechon.com/2008/08/dna-melting-temperature/).

Even today after more than a decade of debate, there is no universal agreement as to which methods are best. New microarray normalization methods continue to be published in September 2012 [106, 107], along with head to head method comparisons [108], while bloggers plaintively ask "When can we expect the last damn microarray paper?" (http://jermdemo.blogspot.com/2012/01/when-can-we-expect-last-damn-microarray.html)

Summary expression methods

Affymetrix GeneChips have several probe pairs for each probe set. Summarization condenses the intensity measures from each of the probe pairs in a probe set into a single intensity for each gene. It was originally thought that the signal intensity from each probe should be very similar since the same gene hybridizes to each of them. In reality, there are large differences between individual probes in a probe set. Differences in the proportion of nucleotides in probe sequences alter the thermodynamic binding affinity to each probe, meaning that the sample fragments will bind more efficiently to different probes at different temperatures. Probe differences may also result from alternative splicing events that were unknown at the time the probes were designed.

MAS5 and Median Polish are the common probe summary techniques. MAS5 attempts to reduce the impact of outlier probes by replacing the value of those mismatch probes (MM) with higher intensity than the PM probe with a modeled MM value derived from the other MM intensities in the probeset. MM values are then subtracted from PM values and a robust average of the probe intensities is calculated using Tukey's biweight algorithm. Most other models disregard MM values, as they have not proven to accurately measure cross-hybridization [109].

Median Polish is a popular summarization technique and has been incorporated into the RMA algorithm. It improves over MAS5 by incorporating information from multiple arrays in an additive model. Probe behavior is compared over many chips, and outliers are excluded from

the expression summary [101]. Median polish is an iterative process of removing outliers from the row and column medians until values stabilize. When considering probes as rows and samples as columns in a matrix, the row effect is defined as the median of each row subtracted from the row values, and the column effect is the column median subtracted from the column values. This is repeated until the row and column medians are 0, then the row effects and column effects are added to give the "all effect" variable that is subtracted from the row effect and column effect variables. Because iterations are based on median values in each step, probes with extreme values (compared to other probes in the probe set or other samples) will not distort the summarized value.

2.3 Conclusion

In this chapter we reviewed some technical considerations inherent in the performance and analysis of microarray experiments. Our focus throughout this study was to maximize data quality, and our opinion is that one must understand the limitations and technical artifacts that can be introduced not only in the experiments themselves but also through the choice of pre-processing techniques. In particular, normalization technique has a considerable influence on which genes are ultimately identified as significantly changed between study groups. Our intent here was to illustrate some of the important choices that must be made . Chapter 3 describes the pre-processing choices that we made as a result of careful consideration of the issues discussed here, including the application of strict Quality exclusion criteria and novel normalization techniques.

Chapter 3. Analysis of developmental and regeneration gene expression in HCV-induced cirrhosis and HCC

3.1 Introduction

In Chapter 3, we present details of the study population, data pre- processing, statistical methods, and initial results. This addresses the first part of Aim 2, to identify those liver development, healing, and regeneration genes that were differentially expressed in cirrhosis and HCC. Section 3.3 examines Aim 2 (identification of common patterns of activation of genes) using the approach of defining gene sets *a priori* based on knowledge of the genes involved in specific developmental stages and particular biological processes.

3.2 Patients and data collection methods

Since 1997, HCV patients diagnosed with cirrhosis and HCC have been evaluated and treated at the Hume-Lee Transplant Center at VCUHS according to an Institutional Review Board approved study protocol [110]. Informed consent was obtained from all patients. Patients were clinically staged according to the American Tumor Study Group modified tumor-

node-metastasis (TNM) classification, and histopathological classification was performed according to the Edmondson grading system where possible. After staging, HCC patients had their tumors ablated and were evaluated for liver transplant according to the United Network for Organ Sharing criteria. Tissue samples were collected from biopsies and explanted livers according to protocols established by the Liver Tissue Cell Distribution System (Richmond, Virginia, funded by NIH Contract #N01-DK-7-0004 / HHSN267200700004C). Control liver samples were obtained from explanted donor livers. Donor livers were shown to have normal function and were negative for hepatitis C virus antibodies. Microarray studies were performed on 180 normal, cirrhosis, and HCC samples.

An independently published dataset of 75 samples with 10 normal controls, 13 HCV-cirrhosis, 17 dysplastic nodules, 18 early stage HCC and 17 advanced HCC was also obtained for verification of results [30] (Wurmbach, et al, 2007; NCBI GEO database accession GEO6764). This dataset was pre-processed using the same methods used for our own data (described below). In addition, the absolute expression levels of target genes in a normal adult human liver were obtained from the BodyMap gene expression database [111], a tissue-specific database of gene expression generated on the Illumina HiSeq 2000 RNA sequencing platform. We aligned raw sequence reads to human reference sequence HG19 using the Burrows-Wheeler Alignment tool (BWA) with default parameters [105]. In cases where BodyMap results were inconclusive (counts of 0-40), a literature search was performed to confirm adult expression of target genes.

3.2.1 Sample preparation

Pre-transplant biopsies and explanted livers were sectioned and grossly examined. Samples from tumors and cirrhotic liver tissue (according to diagnosis and pathological examination) were freshly snap-frozen and processed in the Hume-Lee Transplant Center Molecular Diagnostic Laboratory. Liver tissue samples were collected in RNAlater solution (Ambion, Austin, TX, USA) and stored at -80° C until use. Explanted livers were sliced at intervals of 4-5mm, and all nodules suspicious for HCC processed for light microscopy. Only tumor samples with more than 85% tumor cell content were used for the microarray studies. Normal and necrotic tissues were macro-dissected from the sample.

With minor modifications, the sample preparation protocol follows the Affymetrix GeneChip Expression Analysis Manual (Affymetrix, Santa Clara, CA). After hybridization and scanning, the microarray images were checked for major chip defects or abnormalities in the hybridization signal. Total RNA quality and integrity of each sample were analyzed using the Agilent 2100 Bioanalyzer (Agilent Technologies), and products of cDNA synthesis and in vitro transcription (IVT) were tested before being considered for microarray analysis using the Agilent 2100 Bioanalyzer (cDNA synthesis 1.5 kb < cDNA < 5.0kb; IVT 1.0kb < cRNA < 4.5 kb).

3.2.2 Data pre-processing methods

Data files were read into the R (version 2.13) programming environment and first examined with several quality control tests [91, 112]. Any chip that fell well outside the recommendations for any of the quality assessment tests was excluded from further analysis. Affymetrix- recommended QA tests of hybridization quality which were examined included scale factor (should be near 1.00), percentage of probe sets called as "present" (>40%), and min/max/avg background [112]. The bias.display program [91] was then used to identify chips with unacceptable regional artifacts. Correlation in expression values between neighboring probes should be zero when no technical variation is present. Any chip with an average 40% correlation or more has considerable systematic bias which may be too severe to correct for in the normalization step. Similarly, the correlation between rows should be close to 1. Chips with more than 40% average correlation between neighboring probes or less than 70% correlation between rows on the chip were excluded. Like the % Present calls, the log(PM/MM) should be comparable across chips, and chips with log(PM/MM) >50% higher or lower than the average of all samples were excluded (see Appendix B for list of excluded chips).

As noted in Chapter 2, Robust Multichip Average (RMA) pre-processing is broadly accepted as robust, easy to implement, and widely applicable. However, due to the heterogeneity of the cancer samples we were concerned that the assumptions of RMA may not be met [100]. Specifically, RMA assumes that a relatively small percentage of genes are differentially expressed (5-10%), and that roughly equal number of genes are over-expressed and under-expressed. We processed a test dataset of 58 samples using RMA normalization and

assessed differential expression with a moderated t-test with FDR<0.05 using limma in the R environment. Comparison of group contrasts identified 25-45% genes that were called as differentially expressed, and two-thirds of those genes were over-expressed in tumors. This suggested to us that RMA may not be an appropriate method for this dataset. Instead, we first performed background correction, then normalized the data using non-parametric, distribution-free regression on technical covariates of probes (GC content, melting temperature, and the X-Y coordinates on the GeneChip for each probe) to estimate and correct for systematic bias [105]. Normalized expression probe values were saved as corrected .CEL files, then read back into the R environment using updated probe annotations from the BrainArray project (version 14.1.0, HGU133A2_Hs_REFSEQ), which have been shown to improve accuracy of probe – gene mapping over the standard Affymetrix annotation [113]. Probe sets were summarized using Median Polish.

3.2.3 Description of patient population

Microarray studies were obtained over a 10 year period from 180 samples of cirrhotic tissue and tumors collected from 140 patients with chronic HCV infection. As described in section 3.2.2, stringent quality control criteria were applied to minimize technical artifacts. 73 chips were excluded based on the following criteria: probe-neighbor correlation >40%, row-neighbor correlation <70%, or log(PM/MM) >50% different from average log(PM/MM). Appendix B contains the list of the 73 excluded chips and their exclusion criteria. The final

dataset included 30 HCV-cirrhosis (CIR) and 49 HCV-HCC (HCC) tumors (31 stage T1 and T2, and 17 stage T3 and T4). These were compared to a control group of 12 non-diseased, deceased donor livers (NOR). Twenty-nine HCC patients were transplanted, 6 died on the transplant waiting list and 14 were never listed for transplant due to age, stage of cancer, or other co-morbidities.

3.2.4 Statistical methods

Dysregulated genes in cancer samples are often not normally distributed. Poorly regulated genes may have a broad, flat distribution of expression values or long thick tails and high variability, while genes that are differentially expressed in a subset of tumor samples will have a skewed or bi-modal distribution. In these cases the mean expression over all cancer samples may not be significantly different from controls. However, their variation (standard deviation) will be comparatively higher. In order to identify either of these situations, we did group comparisons using both a t-test to identify shifts in mean expression, and Bartlett's F-test of unequal variance to identify high-variance genes. Combined significance was calculated for each gene using Fisher's combined p-value and FDR < 0.01 [114].

We used scaled Principal Components Analysis (PCA) on specific gene sets to explore the relationship between sets of differentially expressed genes and disease behavior [115]. PCA constructs components as linear combinations of the variables and identifies the genes that account for the most variance in the samples. The PCA plot displays the clustering of samples

based on the principal components. Because it is easily distorted by a few outliers in the data, we Winsorized the data by capping extreme values at 3 median absolute deviations above or below the median value for each gene. Additionally, because variability between groups for each gene may not be the same, we scaled the data to have unit variance. We used the function "prcomp" with scale.=TRUE in the R environment [115].

3.2.5 Identification of test genesets

In this study, we investigated the hypothesis that genes critical to the development and maintenance of the liver are poorly regulated or preferentially activated in HCV-induced cirrhosis and HCV-induced HCC. Genes were identified from an extensive literature review (summarized in Chapter 1), and biologically meaningful gene sets were created based on stage of development or participation in particular biological processes. While many studies of coordinated gene activity are based on KEGG canonical pathways or GO Biological classifications, there is no such resource for the specific developmental processes that we wished to examine, so we manually curated our developmental gene sets from the genes contained in our review of liver development as described in Chapter 1. See Appendix 1 for a complete list of liver development genes that had probe sets on the Affymetrix HG-U133Av2 GeneChip and their inclusion in stage-specific genesets.

3.3 Results

3.3.1 Expression of liver regeneration genes in cirrhosis and HCC

We first examined the expression of the liver regeneration genes reviewed in section 1.6 in 30 cirrhosis (CIR), 31 early HCC (EHCC), and 17 late HCC (LHCC) compared to normal control samples (NOR). Group contrasts were performed between CIR-NOR, EHCC-CIR, and LHCC-EHCC in order to assess the differential expression of each gene between progressively worse disease status. Early HCC was also compared to normal controls.

The molecular events in liver regeneration after acute hepatic injury have been well studied in animal models in the context of partial hepatectomy. The "first responder" genes are IL-6 and TNFα. IL-6 has increased expression in cirrhosis but not HCC, while TNFα is not differentially expressed in cirrhosis or HCC. Early response genes stimulated by IL-6 and TNFα include c-FOS, c-JUN, c-MYC, and c-MET (FOS is involved but not required for normal liver development while JUN, MYC, and MET are crucial liver development genes). Table 1 lists mean fold-change (compared to normal controls) of the differentially expressed regeneration genes in our data. FOS, JUN, and MYC were up-regulated in cirrhosis and returned to normal levels in most HCC samples, while MET was slightly down-regulated in cirrhosis then up-regulated in a subset of HCC samples. Early release growth signals that stimulate hepatocytes to enter mitosis include HGF, TGFα, and EGF, which were not differentially expressed in either cirrhosis or HCC, nor were their activators PLAU and PLAU receptor. Furthermore, inhibitors of

hepatocyte proliferation that are down-regulated during liver regeneration (CDH1, MST1, TGFB, and BMP2) are over-expressed in cirrhosis. These data suggest that cirrhotic livers, but not tumors, are signaling for regenerative repair but the downstream targets are not responding appropriately and that hepatocyte proliferation is being actively suppressed.

The proliferative phase might be hypothesized to be recapitulated in tumors, since tumors are characterized by uncontrolled proliferation. Epithelial Growth Factor (EGR1) and TNFα, which initiate the proliferative phase, showed no change in expression or were down-regulated in cirrhosis (q= 0.6, 0.02, respectively) and HCC samples (q=0.00001, 0.08). This indicates that tumors are not "continuously initiating" hepatocyte proliferation, at least not using regenerative mechanisms. However, proliferation inhibitors BMP2, CDH1, and TGFB1 are not expressed as highly in HCC and MST1 is down-regulated, while MET is no longer under-expressed in tumors, which suggests that tumors may be capable of overcoming the suppression of hepatocyte proliferation present in cirrhosis.

Table 1. Differentially expressed liver regeneration genes in cirrhosis, early HCC and late HCC compared to normal controls (Fold-change relative to normal controls). . * denote genes that are differentially expressed compared to normal (for CIR) or cirrhosis (for HCC) (FDR<0.01); ** denotes genes with q<0.00001; V denotes genes that are significantly more variable by F-test of variance.

Gene	CIR	Early HCC	Late HCC
IL6	1.5*	1.2	1.0
FOS	4.3*	1.7**	0.9
JUN	2.3*	1.4**	1.1
MET	0.8*	1.1**	1.2
PLAU	1.2*	1.3	1.4
EGR1	1.1	0.7*	0.5

CDH1	1.7*	1.4[V]	1.3
MST1	0.9	0.6*	.05
BMP2	1.5*	1.2*	1.2
TGFB1	2.7*	1.6	1.6
REL	1.4*	1.4	1.2
FOXM1	0.9*	1.0**	1.2**
SAA1	0.1*	0.1	0.0
SAA2	0.1*	0.1	0.0
CRP	0.3*	0.2	0.1
SOD2	0.4*	0.5	0.6
SOCS3	1.2	1.0	0.9[V]
NAMPT	0.6*	0.5	0.4

3.3.2 Differential expression of liver healing genes

Liver wound healing is distinct from large-scale regeneration and is characterized by activation and proliferation of progenitor cells instead of hepatocytes. We examined CIR, EHH, and LHH samples compared to normal controls for differential expression of the important wound healing genes identified in section 1.5. This process is directed by activated HSCs, and several markers of activated HCSs are elevated in cirrhosis (Table 2). PDGFA/C, CCL2, VEGFC maintain the activated state of HSCs, while IL6, TGFB1, PDGFR, and TIMP1 respond to inflammatory mediators and contribute to HSC survival. Early phase genes MMP3, MMP13, and PLAU are not differentially expressed in either cirrhosis or HCC, although injury-related collagens COL1A1, COL1A2, and COL3A1 are elevated in both cirrhosis and HCC. Markers of progenitor cells (EPCAM, VIM, and KRT19).are highly over-expressed in both cirrhosis and HCC (Table 2).

During chronic hepatic injury, several types of liver cells can acquire a neuroendocrine phenotype, including proliferating cholangiocytes, oval cells, and activated HSCs. None of the neuroendocrine proteins CHGA, NCAM, PTHLH, the neurotrophins, or their receptors are differentially expressed in cirrhosis or HCC. However, several members of the S-100 protein family are dysregulated: S100A4, 6 , 10, 11, 14, and 16 are over-expressed, while S100A8, 9 and 12 are under-expressed in cirrhosis and HCC (Table 2). S100A4 and A6 promote apoptosis, S100A8, A9, A11, and A12 promote inflammation, and S100A8 and A9 promote chemotaxis. S100 proteins have been implicated in tumorogenesis in several cancer types, but have not previously been associated with liver cirrhosis.

Serotonin receptor HTR2B, associated with hepatocyte proliferation suppression, was increased in both cirrhosis and HCC. Similarly, REL and CCL5 (which promote HSC proliferation and migration to sites of injury) are elevated in both cirrhosis and HCC (Table 2), perhaps indicating that sustained wound healing is active throughout cirrhosis and carcinogenesis. However, other hallmarks of hepatic wound healing are not differentially expressed, including hedgehog pathway genes, and the other NF-κB subunits.

Table 2. Differentially expressed liver regeneration genes (fold-change relative to normal controls). Genes that are differentially expressed compared to the earlier stage of disease ($q<0.01$) are indicated with an *.

Gene	CIR	Early HCC	Late HCC
PDGFC	1.5*	0.9*	0.8
CCL2	4.4*	1.8*	1.7
PDGFRA	5.3*	1.6*	0.9
VEGFC	1.8*	1.5	1.4
IL6	1.5*	1.2*	1.0

TGFB1	2.7*	1.6	1.6
TIMP1	1.7*	1.2*	1.2
COL1A1	2.0*	2.2	2.4
COL1A2	4.3*	4.7	5.4
COL3A1	2.6*	2.8	2.7
S100A4	1.6*	1.6	1.7
S100A6	6.0*	3.1	4.0
S100A8	0.4*	0.3	0.3
S100A10	2.3*	2.1	2.5
S100A11	2.5*	2.3	2.8
S100A14	1.8*	1.4	1.3
S100A9	0.5*	0.7	0.6
S100A12	0.6*	0.5	0.4
HTR2B	2.2*	2.0	1.8
REL	1.4*	1.4	1.2
CCL5	5.5*	4.2	3.1
EPCAM	14.0*	7.1	2.9
VIM	5.3*	4.1	4.6
KRT19	5.0	2.1	1.6

3.3.3 Potential covariate effects on developmental gene expression in HCV-cirrhosis and HCC

Smoking, alcohol, advanced age and diabetes are independently associated with a higher risk of developing HCC, with synergistic acceleration of cirrhosis and HCC development in HBV and HCV patients [116-119]. Our main question is whether these risk factors result in differential expression of our genes of interest within the cirrhosis and HCC cohorts. Most patients had at least one of these risk factors and many had several (Table 3).

Although a role for Diabetes Mellitis (DM) in development and outcomes in viral hepatitis-induced HCC has been demonstrated [120], and genetic alterations in DM have been extensively studied [121], the molecular impact of DM in HCC or in interactions with HCV have not. In general, DM has been associated with alterations in mitochondrial phosphorylation and

oxidative metabolism (oxidative stress), up-regulation of pro-inflammatory genes, dysregulation of lipid metabolism, and sustained release of acute phase proteins [122].

Similarly, several studies have shown that smoking increases the risk of developing HCC, particularly in patients infected with HBV or HCV. However, the effect of smoking on gene expression in liver disease has not been studied.

Mechanisms for synergistic activity of alcohol and HCV are best known and are thought to include four main mechanisms: impaired adaptive immunity; reduced antigen presentation on viral infected cells; reactive oxygen species (ROS) induction by both HCV core protein and alcohol injury; and inflammation associated with both chronic HCV infection and alcohol. Mas and others have noted that alcohol abuse is associated with reduced HCV clearance and an accelerated disease course [118, 123].

At the molecular level, chronic alcohol consumption impairs the secretion of TNF, IFN-γ, and IL12 [118]. Alcohol consumption also increases IL10 production, which also shifts immune response to TH2-type. Mas et al (2010) compared the mean fold change compared to normal control samples for HCV-cirrhosis , EtOh-cirrhosis, and HCV-EtOh cirrhosis samples and the significantly changed genes identified included some liver development genes (JUN, 2.43x; IGF2, 1.62x; FZD5, 1.56x; MMP25, 1.52x; LAMC2, 1.33x; TCF2, 1.43x; JAG1, 1.17x, SMAD6, 1.67x; LAMA4, 2.02; TBX1, 1.4x). However, this compared single-etiology samples rather than the additional effect of alcohol consumption on HCV-HCC gene expression.

Table 3. Demographic characteristics of cirrhosis and early HCC patients. There were no significant differences between cirrhosis and early HCC. Age is presented as mean ± sd. Minimum q-value is the smallest q-value of all results for the 202 developmental genes when analyzed separately for gender, age, alcohol abuse, history of smoking, and diabetes co-morbidity.

Covariate	Cirrhosis (n=31)	Early HCC (n=30)	Minimum q-value
Gender - male	23 (74%)	23 (77%)	0.79
Mean age	52.4 ± 5.2 (range 42-62)	56.3 ± 5.5 (range 48-68)	0.26
History of alcohol abuse	20 (64.5%)	19 (63%)	0.85
History of cigarette smoking	18 (58%)	14 (47%)	0.99
Diabetes	7 (23%)	5 (17%)	0.74

Univariate effects of each covariate on each of the 202 developmental genes in Appendix A were tested using a moderated t-test with limma. Age was categorized based on the mean age in the HCC group (55 and above vs. below 55) and also tested as a continuous variable. KIT was the only developmental gene identified in the literature as potentially increased with advanced age [124] and appeared to have a slight effect in our data, but this turned out to be driven by 2 outliers (Sample T3_400 had FC 3.7 compared to normal and Sample T2_388 had FC 6.4, when these two samples were excluded, mean FC was 1.3 for HCC, compared to FC 1.7 when included). Alcohol abuse was defined as a history of heavy drinking or a diagnosis of alcohol-induced liver disease based on the patient's transplant evaluation. History of social drinking was not included. Smoking was defined as being a current smoker, or having a history of smoking, at the time of diagnosis of tumor or time of transplant evaluation

for cirrhosis. Gender, mean age, history of alcohol abuse, history of smoking, and diabetes were not significantly different between cirrhosis and early HCC groups ($p>0.05$). None of the developmental genes had differential expression in cirrhosis or HCC based on age, EtOh, tobacco use, or diabetes at FDR<0.25.

3.3.4 Differential expression of liver development genes in HCV-cirrhosis and HCV-HCC

We then examined expression of the 202 liver development genes (listed in Appendix A) in HCV-CIR and HCV-HCC compared to normal control samples. Fifty-seven genes with low variance (standard deviation <0.3) were filtered out. To capture changes in either mean or variation of the remaining genes, we assessed significance with a combined p-value from both t and F tests. Of the remaining 118 genes, 37 were not significantly changed in any group compared to normal controls (94 total filtered or non-significant, see Appendix C1).

In cirrhotic tissue compared to normal controls, 68 genes had a significant shift in mean expression by t-test, and 1 was highly variable by F-test. Of those 69 differentially expressed genes (DEG) in cirrhosis, 42 (61%) had significant mean shift in early HCC compared to cirrhosis and 16 (23%) had similar mean expression to the cirrhosis samples but were significantly more variable. The remaining 11 (16%) genes had the same expression pattern in early HCC as in cirrhosis by Fisher's combined test (Table 4). COL4A4, CSNK1D, and HNF1B were only up-regulated in cirrhosis and returned to normal expression levels in HCC, while 18 genes were still over-expressed in HCC, but not as highly as in cirrhosis. These included EPCAM and several

extra-cellular matrix (ECM) genes (COL4A2, MMP7, Laminin-α2, Laminin-γ3) and members of the Wnt/BMP axis (SFRP5, FSTL3, FGFR2, and SMAD7). Transcription factors following this expression pattern include SOX9, GATA6, ARID5B, ID3, and CITED2. Additionally, tumor suppressor KLF6, growth suppressor Necdin, mesenchymal marker KRT19, and the heparin-binding growth factor Pleiotrophin (PTN) were more highly expressed in cirrhosis than in tumors.

Table 4. Liver development genes with significantly higher expression in cirrhosis than tumor samples. FC = Fold-change relative to normal samples. * denote genes that are differentially expressed compared to normal (for CIR) or cirrhosis (for HCC) (FDR<0.01); ** denotes genes with q<0.00001; V denotes genes that are significantly more variable by F-test of variance. Abbreviations: TF= transcription factor; ECM= Extra-cellular matrix; IF= intermediate filament; GF= growth factor; FC = Fold-Change.

GENE	GENE NAME	GENE FUNCTION	Mean FC CIR	Mean FC Early HCC	Mean FC Late HCC
EPCAM	Epithelial cell adhesion	ECM	14.0 **	7.1^{V**}	2.9*
MMP7	Matrix metalloproteinase 7	ECM	6.3 **	3.0*	3.6*
KRT19	Cytokeratin-19	Epidermal IF	5.0**	2.1*	1.6*
MMP2	Matrix metalloproteinase 2	ECM	4.9**	4.6^{V**}	3.0*
VIM	Vimentin	Mesenchymal	5.5**	3.3*	2.7
SOX9	SRY-box 9	TF	4.7**	2.5**	2.8*
LAMA2	Laminin alpha 2	ECM	4.2**	1.9*	1.4*
FGFR2	Fibroblast Growth Factor	GF receptor	4.1**	2.0**	1.3*
KLF6	Kruppel-like factor 6	TF	3.9**	2.4**	1.6*
COL4A2	Collagen IV alpha 2	ECM	3.6**	2.3*	2.2*
LAMB1	Laminin beta 2	ECM	3.2**	1.8^{V**}	1.5*
ARID5B	AT rich interactive domain 5B	TF	3.4**	1.8**	1.6*
FSTL3	Follistatin-like protein 3	GF antagonist	2.9**	1.5**	1.5*
SMAD7	SMAD family member 7	Signal transduction	2.8**	1.5**	1.1*
GATA6	GATA binding protein 6	TF	2.7**	1.2**	0.7*
CITED2	CBP/p300-interacting	TF	2.5*	1.6*	1.5*
SFRP5	secreted frizzled-related	Wnt inhibitor	2.4**	1.6**	1.3*
ID3	Inhibitor of DNA binding 3	TF antagonist	2.3*	1.5*	1.2*
LAMC3	Laminin gamma 3	ECM	2.2**	1.3**	1.2*

NDN	Necdin	TF	2.1**	1.2**	1.1*
PTN	pleiotrophin	GF	2.0**	1.5**	1.3*
ZBTB20	zinc finger and BTB domain	TF	2.0**	1.3**	1.1*
CDH1	Cadherin 1	ECM	1.7**	1.4^{V**}	1.3*
FGF7	Fibroblast growth factor 7	GF	1.5**	1.2^{V*}	1.1*
COL4A4	Collagen IV alpha 4	ECM	1.5**	1.1*	1.2*
CSNK1D	Casein kinase I isoform delta	kinase	1.4**	1.0*	1.1*
HNF1B	Hepatic Nuclear Factor 1β	TF	1.3*	1.0*	1.1

Only five genes were more higly expressed in HCC than cirrhosis, including Osteoponton (SPP1), GPC3, Midkine, MMP9, and Integrin α-6 (Table 5).

Table 5. Genes over-expressed in cirrhosis and more highly over-expressed in HCC. FC= Fold-Change relative to normal controls.

GENE	GENE NAME	GENE FUNCTION	FC cirrhosis	FC early HCC	FC late HCC
SPP1	Osteopontin	Mediates integrins and CD44 signaling	9.8	9.5	16.4
GPC3	Glypican 3	BMP inhibitor	2.1	7.8	10.1
MDK	Midkine	Regulates PTN	1.7	2.7	2.4
MMP9	matrix metalloproteinase 9	Type IV collagenase	1.2	2.5	4.4
ITGA6	Integrin alpha 6	Cell-cell adhesion	1.4	1.8	1.8

Seven genes were down-regulated in cirrhosis and remained low in tumors: transcription factors FOXA1, FOXA2, XBP1, and GATA4; Activin receptor ACVR2B, Retinoic Acid Receptor RXRA; and signaling molecule neurturin (NRTN). Originally identified as a neuron outgrowth factor, in 2007 neurturin was identified as a critical factor in directing embryonic liver bud migration [65]. Its function in adult liver tissue has not been studied. BMP4, FOXM1,

NR5A2, and SRPK1 were down-regulated exclusively in cirrhosis. c-MET, the hepatocyte growth factor receptor, was the only developmental gene that was down-regulated in cirrhosis and up-regulated in tumors.

Table 6. Genes down-regulated in cirrhosis and HCC. * denote genes that are differentially expressed compared to normal (for CIR) or cirrhosis (for HCC) (FDR<0.01). ; V denotes genes that are significantly more variable by F-test of variance.

GENE	GENE NAME	Gene function in adult liver	FC CIR	FC early HCC	FC late HCC
BMP4	Bone morphogenic protein 4	Maintains biliary differentiation	0.76*	0.88*	1.02
FOXM1	Forkhead box M1	Activates cell cycle regulators	0.87*	1.02*	1.2^{V}
NR5A2	Liver receptor homolog 1	Antagonizes progenitor cell proliferation; promotes hepatocyte maturation	0.56*	0.85*	0.7
SRPK1	Serine/threonine protein kinase	Regulates alternative splicing	0.75*	0.85^{V}	0.97
MET	Met proto-oncogene	Hepatocyte growth factor receptor	0.79*	1.08*	1.22
FOXA1	Forkhead box A1	Regulates neoglucogenesis	0.4*	0.46^{V}	0.37
FOXA2	Forkhead box A2	Lipid metabolism, bile homeostasis	0.58*	0.63	0.63
XBP1	X-box binding protein 1	Regulates lipogenesis	0.65*	0.82*	0.69
GATA4	GATA binding protein 4	Inhibit proliferation; tumor suppressor	0.66*	0.74	0.79
ACVR2B	Activing receptor 2B	Activin and BMP receptor	0.49*	0.57*	0.59
RXRA	Retinoic acid receptor α	promotes hepatocyte survival	0.50*	0.51	0.52
NRTN	Neurturin	Not yet determined	0.33*	0.38	0.35

Fifteen genes were differentially expressed uniquely in early HCC (Table 6). Only 2 genes were uniquely changed in late-stage tumors (IGF2 and YAP1 were down-regulated), while FOXM1, ITGA3, and CP were significantly more variable.

Table 7. Genes uniquely changed in HCC (FDR<0.01). Abbreviations: TF= transcription factor; GF= growth factor. Genes with significant mean change compared to normal and cirrhosis samples are marked with an *; genes that are significantly more variable in tumors than cirrhosis or normal samples are marked with a $^{V.}$

GENE	GENE NAME	PATHWAY	GENE FUNCTION	FC early HCC	FC late HCC
AFP	Alpha FetoProtein		uncertain	1.8	1.6
ATF2	Activating transcription factor 2	JUN,JNK	TF	1.2	1.4
CCNE2	Cyclin E2	Cell cycle		1.4^{V}	1.7
DKK1	Dickkopf 1	Wnt	Wnt inhibitor	1.3	1.3
DKK4	Dickhopf 4	Wnt	Wnt inhibitor	1.2^{V}	1.4
GREM	Gremlin	BMP	BMP inhibitor	1.9v	1.5
GSK3B	Glycogen synthase kinase-3	Wnt	kinase	0.6	0.6
IGF2	Insulin-like growth factor 2	IGF2	GF	0.8	0.4*
KRAS	v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog	multiple	GTPase	0.7	0.7
MMP12	matrix metalloproteinase 12			2.0	3.1
MST1	Macrophage stimulating 1 (hepatocyte growth factor- like)	Hippo, JNK	GF	0.6	0.5
NID1	Nidogen		ECM	0.8^{V}	0.8
NOTCH2	Neurogenic locus notch homolog protein 2	NOTCH	JAG1 receptor	0.8	0.8
STAT3	Signal transducer and activator of transcription 3	multiple	TF	0.6	0.5
TGFBR3	Transforming growth factor beta receptor 3	TGFB	receptor	0.6	0.5
WNT5A	Wnt 5a	Wnt		1.3^{V}	1.6
YAP1	Yes-associated protein 1	Hippo	TF	0.9	0.7

3.4 Functional gene sets that discriminate between normal, cirrhosis, and tumor samples

Since genes specific to a particular stage of development are working coordinately in the liver during development, we wished to investigate whether the genes dysregulated in cancer were specific to stage of development. We defined five gene sets corresponding to the main phases of development: hepatic fate specification, liver bud formation (hepatoblast migration), liver bud growth, hepatocyte/cholangiocyte differentiation, and maturation (See Appendix A). We used PCA to evaluate the amount of variation between disease types that was explained by the stage-specific gene sets and looked at the loadings of the first few principal components to identify which genes were the most important contributors (Figure 3.1). We found that genes from each stage of development appeared to separate normal, cirrhosis, and HCC samples, and that the PCs for each gene set were not dominated by only a few genes, but instead driven by several genes with modest fold-changes. This fits our original hypothesis that small changes in multiple regulatory genes may be important drivers of tumor progression that might look unimportant in a genome-wide “smallest q-value” analysis.

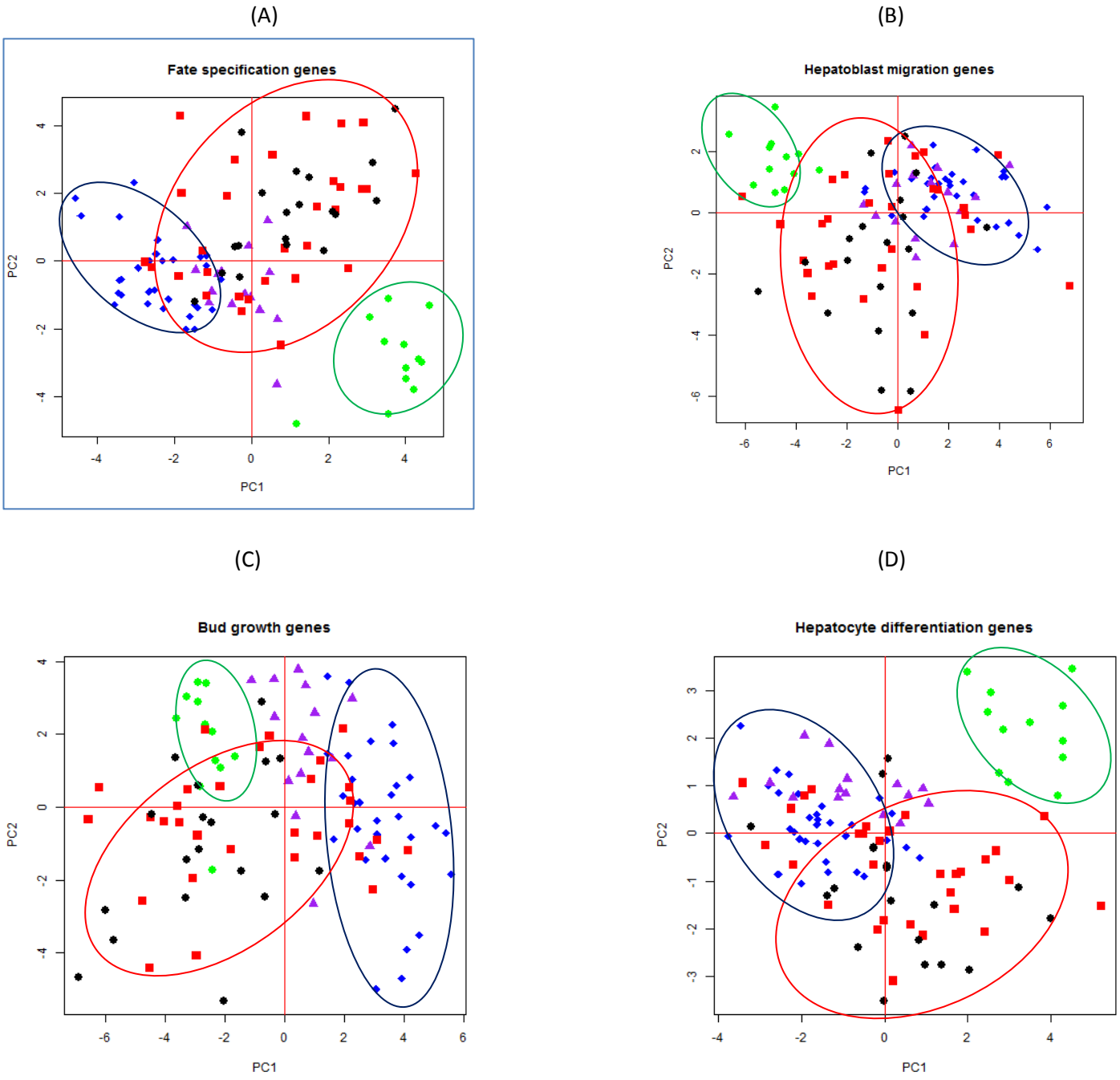


Figure 3.1. PCA plots of first two principal components of stage-specific liver development genes. Green = normal; Blue = cirrhosis; Purple = cirrhotic tissue surrounding tumor; Red = early stage HCC; Black = late stage HCC. (A) Hepatic specification genes (B) Hepatoblast migration genes (C) Bud growth genes (D) Hepatocyte/cholangiocyte differentiation genes.

Since the important genes were not specific to particular developmental stages, we turned our attention to specific functional groups. Genes related to extra-cellular matrix (ECM) maintenance or remodeling demonstrated major changes in both cirrhosis and tumors. PCA of the significantly changed genes demonstrate that these genes also independently discriminate between normal, cirrhosis, and tumor samples (Figure 3.2 (A)). PC1 explained 30% of the total variance and the largest contributors were COL4A1, COL4A2, LAMA2, LAMB1, LAMBC3, MMP2, MMP7, and EPCAM. PC2 was dominated by MMP12, which is uniquely expressed in HCC and explained 13% of the total variance.

The BMP signaling pathway is also highly dysregulated in HCV-cirrhosis and HCC. BMP2, its receptors, and BMP inhibitors are all differentially expressed in cirrhosis and HCC. BMP2 was over-expressed in cirrhosis, while the BMP inhibitors (GPC3, GREM1, FST) were more highly expressed in HCC. A PCA plot demonstrates the gene set's ability to discriminate most tumors from normal and cirrhosis samples (Figure 3.2 (B)).

A. Extra-cellular matrix genes

B. BMP2, BMP receptors and inhibitors (GREM1, FST, FSTL3,GPC3)

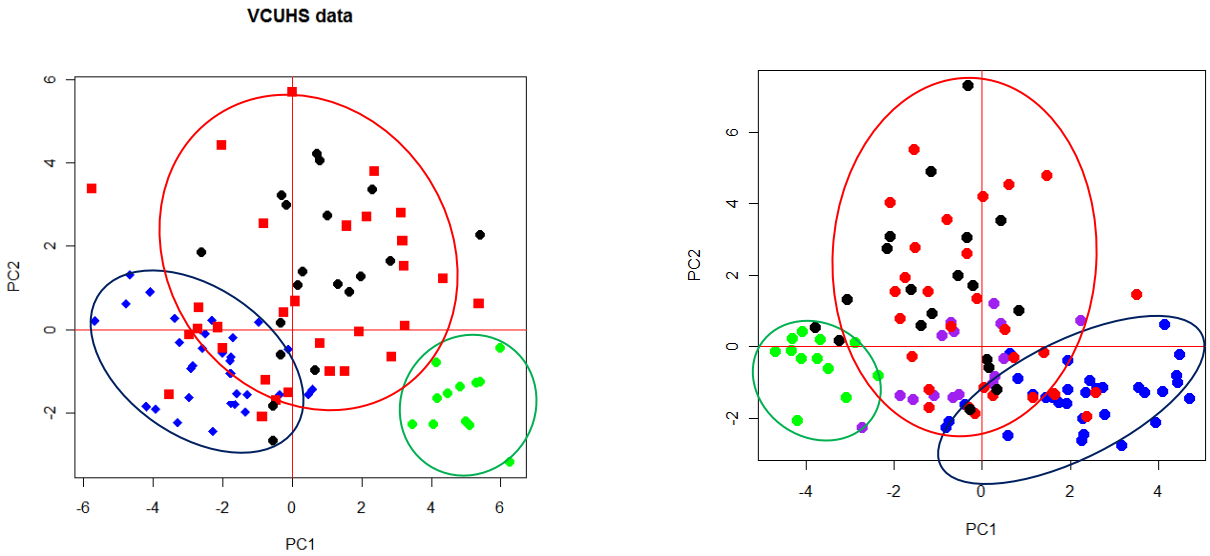


Figure 3.2. PCA plots of (A) ECM genes and (B) BMP2 and its receptors and inhibitors. Green = normal control livers; Blue = cirrhosis; Red = early stage HCC; Black = late stage HCC. Normal tissues cluster well away from either cirrhosis or tumors. Both the ECM genes (A) and BMP inhibitors (B) were able to discriminate between cirrhosis and many of the tumor tissues. ECM geneset includes: CADM1, CDH1, COL4A1, COL4A2, COL4A5, EPCAM, ICMT, ITGA3, ITGA6, LAMA2, LAMA3, LAMA4, LAMB1, LAMC1, LAMC3, MMP1, MMP2, MMP7, MMP9 MMP12, MMP15, MMP17, MMP19, and NID1. The BMP geneset includes BMP2, BMPR2, GREM1, GPC3, FST, FSTL3.

Because the specific Wnt family members involved in liver development and disease have not been completely determined [39], Wnt genes were not included in the initial testing. However, because several of the differentially expressed genes related to regulation of the Wnt pathway, we tested differential expression of all Wnt pathway genes (Figure 3.3). Most of the canonical Wnt effectors were low variance (not DEG), including APC, AXIN, GSK3A, DVL1/2/3, and most of the Frizzled receptors. WNT5A, a ligand in the non-canonical pathway which has been suggested as a candidate liver development Wnt, was the only significantly changed Wnt

ligand. Most of the Frizzled genes were dysregulated. FZD4 and FZD5 were down-regulated in both cirrhosis and HCC, while FZD6 and FZD7 were over-expressed. The biological implication of these changes is not clear. FZD5 and FZD7 are specific to canonical Wnt signaling, while FZD4 and FZD6 are specific to non-canonical signaling. FZD6 and FZD7 have previously been reported to be over-expressed in primary HCC tumors [39]. The down-regulation of co-receptors LRP5 and LRP6 may make the cells less responsive to Wnt signaling. The significantly changed genes are shown in the table below, and PCA of these genes demonstrate that they can discriminate most HCC from cirrhosis and normal tissue.

GENE	CIR	Early HCC	Late HCC
WNT5A	1.04	1.25	1.57
FZD1	1.55	1.15	1.16
FZD4	0.74	0.85	0.90
FZD5	0.53	0.61	0.52
FZD6	2.42	2.00	2.32
FZD7	3.16	1.79	1.47
TCF4	2.57	1.88	1.56
LRP5	0.67	0.61	0.54
LRP6	0.74	0.82	0.81
DKK1	1.05	1.34	1.34
DKK3	5.19	3.19	2.39
DKK4	1.00	1.16	1.44
GSK3B	0.78	0.58	0.62

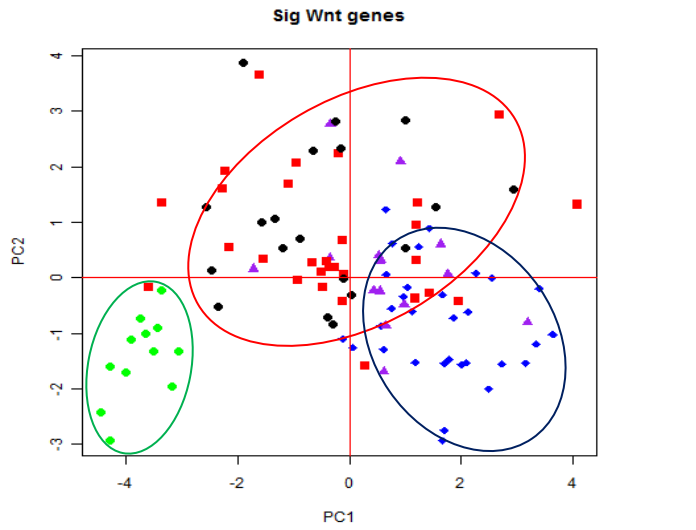


Figure 3.3. Fold-change of differentially expressed Wnt pathway genes compared to normal controls and PCA plot comparing normal, cirrhosis, and tumor samples.

3.4 Discussion

In this chapter we presented results of differential expression of the genes critical for liver regeneration, liver wound healing, and liver development. Many of the "early response" genes that initiate liver regeneration are also important liver development genes, and these genes were differentially expressed in cirrhosis. However, the down-stream targets of these genes during regeneration were not differentially expressed. In a small study of five living donors undergoing partial hepatectomy, Ho et al (2007) identified several genes up-regulated five hours after resection, including SAA1, SAA2, CRP, SOD2, SOC3, and NAMPT (involved in immune response and cell growth) [125]. All of these genes were down-regulated in our data, providing further evidence that the processes active in cirrhosis and HCV- induced tumors do not share important characteristics with the regenerative processes in otherwise healthy livers.

Overall, 90 of the 202 (45%) genes critical to liver development had altered expression patterns in cirrhosis and HCC. A complete list can be found in Appendix C. EPCAM, an intermediate filament specific to epithelial cell types, is the most highly over-expressed development gene in cirrhosis (15x) and early HCC (14.4x), and less highly over-expressed in late HCC (5.7x). Yoon et al (2011) suggest that EPCAM markers seen in cirrhosis represent immature hepatocytes that have recently derived from progenitor cells, which retain many characteristics of embryonic hepatoblasts [126]. Progenitor cell markers include KRT19, Vimentin (VIM), and c-Kit (KIT). Interestingly, KRT19 is an epithelial IF, while Vimentin is an IF marker of both mesenchymal cells and those undergoing Epithelial-Mesenchymal transition. The progenitor cell markers are also very highly expressed in cirrhosis: KRT19 (5.0x), VIM (5.5x),

and KIT (1.9), but markers of hepatocyte proliferation were not up-regulated (HGF) or down-regulated (c-Met 0.8x). This supports the idea that hepatocyte replacement in advanced cirrhosis is accomplished by proliferation and differentiation of progenitor cells [127] as a form of wound healing. This process closely resembles the initial formation of the embryonic liver bud by hepatoblast migration through the STM and subsequent proliferation and differentiation into hepatocytes (as described in Section 1.4.2 and 1.4.3).

In healthy livers, hepatocyte turnover is very slow, and the liver expresses several inhibitors of hepatocyte proliferation including BMP2, MST1 and CDH1. During liver regeneration after partial hepatectomy, all of these genes are down-regulated, and the promoters of hepatocyte proliferation are (MET and HGF) up-regulated. In our cirrhosis samples, hepatocyte proliferation inhibitors were up-regulated (BMP2 1.5x, CDH1 1.7x) or expressed at normal levels (MST1), providing further evidence that HCV-induced cirrhosis is characterized by "wound healing" as opposed to regenerative processes. BMP2 and CDH1 were less higly over-expressed in tumors (BMP2 1.2x, CDH1 1.4x) and MST1 was down-regulated (0.6x), suggesting that tumors may acquire hepatocyte proliferative capabilities in addition to progenitor cell proliferation.

Aberrant wound healing has also been suggested as a mechanism in several types of cancer [128] including renal cell carcinoma [129] and lung cancer [130]. In our data, both early and late tumors also over-expressed markers of proliferating progenitor cells including VIM, KRT19, and EPCAM. EPCAM has also been noted previously as highly expressed in pre-malignant hepatic tissue and a subset of HCC with poor prognosis [131]. Budhu et al

distinguished a subtype of HCC that displayed a molecular signature with features of progenitor cell markers (EPCAM, c-KIT, KRT19, VIM, PROM1, and AFP) and the activation of Wnt signaling [132]. However, in our data EPCAM, KRT19, VIM, and PROM1 were over-expressed in every cirrhosis sample and less highly expressed in HCC. Early HCC with poor prognosis (death or recurrence within 2 years) and late-stage HCC had EPCAM (<3x) loss in 6/9 and 12/19 (67%, 63%) of samples, compared to 2/21 (9%) of early HCC with good prognosis, which seems to contradict the previous association of EPCAM+ subtype with poor prognosis. However, Kumar et al (2011) distinguished an HCC EPCAM+/AFP- subtype that has good prognosis compared to EPCAM+/AFP+ [133], and our data is largely AFP- (the six AFP+ patients in our data had recurrence or died, regardless of EPCAM levels). It has also recently been shown by Wang et al (2012) that Hepatitis B virus X induces EPCAM expression and aggressive clinicopathologic features [134], so high EPCAM levels may be more prognostic for HBV than for HCV induced HCC.

Other developmental genes that have recently been demonstrated to be either expressed by or activate progenitor cells include Midkine (MDK), pleotrophin (PTN), SOX9, FGF7, FGFR2b [135], which were all over-expressed in cirrhosis and tumor samples. In the developing embryo, MDK regulates PTN expression, which is a hepatoblast growth factor. SOX9 has been previously associated with activated HSC, fibrosis, and cirrhosis [136] and recently associated with poor prognosis and tumor progression in HCC [137]. Embryonically, SOX9 guides hepatoblast differentiation towards a biliary fate. It was in 2011 also suggested as a marker of liver progenitor cells or their recent progeny cells along with EPCAM and PROM1 after liver injury [138-140].

FGFR2 is a receptor for that is highly expressed on hepatocytes and progenitor cells and plays a role in liver homeostasis [141]. FGFR2 is capable of dimerizing when over-expressed, which could lead to ligand-independent signaling. In cell lines, inhibition of FGFR2 signaling led to enhanced apoptosis, suggesting that FGFR2 over-expression may protect against apoptosis [142]. Interestingly, isoform IIIb is expressed exclusively on epithelial cells, whereas the IIIc isoform is expressed on mesenchymal cells, so isoform switching can indicate EMT and contribute to unbalanced autocrine signaling [142].

Although microarray data cannot be used to evaluate alternative splicing events, several FGFR2 probe sets show differential expression from each other in cirrhosis and HCC, suggesting the possibility of isoform switching that we may be able to investigate in future RNA-seq studies. The epithelial isoform FGFR2IIIb binds specifically to FGF7, while the mesenchymal isoform FGFR2IIIc has binding specificity to liver FGFs 2, 4, and 8. FGF2 proteins are long-lived and present in significant levels in adult livers and maintain hepatocyte differentiation [143]. FGF2 was not differentially expressed from normal samples in our data.

MMP2 and MMP7, which have been shown to correlate with degree of fibrosis and liver function tests [144] were also highly over-expressed in cirrhosis compared to HCC. However, MMP2 and MMP7 over-expression have also been associated with invasiveness and poor prognosis in HCC and other cancers [145-149], and no studies have addressed the difference in expression or role between cirrhosis and HCC. Both MMP2 and MMP7 specifically degrade the type IV collagens that make up the basement membrane in healthy livers both embryonically and in adulthood [150]. Several collagens and laminins that make up the ECM are up-regulated

in cirrhosis and HCC, along with their receptors (integrin α3 and α6) and the MMPs that degrade and remodel the matrix. MMP12 and MMP9 were exclusively up-regulated in tumors. This is in agreement with Han et al (2004), who showed that MMP9 was not produced by fully activated HSC in cirrhosis, but present during early stages of fibrosis then again in early HCC [151]. MMP9 expression from damaged hepatocytes promotes the release of progenitor/stem cells from the bone marrow, which migrate to the liver and participate in ECM remodeling, and are capable of differentiating into hepatocytes [152-154]. It is thought to have multiple roles in the development of tumors, including activation of TGFB1, degradation of collagen IV, promote tumor invasion into blood and lymph vessels, and resistance to natural killer cells [155]. MMP12 has been recently associated with venous infiltration and poor prognosis in HCC [156]. In our data, MMP12 was either not over-expressed at all, or over-expressed at least 10-fold over normal controls. Five of 30 (17%) early stage and 5/17 (30%) of late stage tumors over-expressed MMP12.

Embryonically, BMP signaling is antagonistic to FGF signaling and this balance is controlled by the DAN family of BMP antagonists from mesenchymal cells and GPC3 expressed by hepatocytes. BMP2, BMPR1A, FGF7, FGFR2, and ID3 were more highly expressed in cirrhosis than HCC, while the BMP inhibitors were more highly expressed in tumors. At least one of the inhibitors GPC3, GREM1, FSTL3, and/or FST were expressed nominally higher (FC>1.5) in 100% of tumor samples. MST1 is a negative regulator of YAP1 in the Hippo pathway responsible for maintenance of organ size. In the liver, it maintains hepatocyte quiescence and is considered a tumor suppressor, and loss of MST1 and MST2 (STK3) is sufficient to initiate hepatocyte proliferation and development of HCC [157]; however in our data, MST1 is down-regulated

while STK3 is up-regulated [158] . YAP1 expresion is unchanged in cirrhosis and early HCC and down-regulated in late tumors.

STAT3 is an oncogene, and overexpression or constitutive activation is generally associated with HCC and tumor progression [159]. However, in our data STAT3 was highly expressed in normal and cirrhosis samples and down-regulated in HCC. High levels of normal expression were confirmed in the BodyMap dataset (counts > 15,000). STAT3 is an essential effector of acute phase response to IL-6 signaling in the liver, and STAT3 inactivation leads to serious impairment of the acute-phase response. It is a complex protein with six functional regions that appears to play diverse and contradictory roles in different cell types within the liver, including inflammatory responses, promoting hepatocyte survival, regulating hematopoietic stem and progenitor cell proliferation and survival [160]. The apparent loss of STAT3 in tumor samples may reflect the difference in proportion of mature hepatocytes, or the presence of persistently activated STAT3 may lead to reduced transcript levels [161].

RXRA down-regulation has been reported in both HCV- and HBV- induced HCC [9, 29, 30, 162, 163]. In healthy livers, RXRA is expressed by hepatoctyes and promotes their survival, as well as being involved in several metabolic functions. It also mediates inhibition of cell cycle progression and induction of apoptosis by vitamin A derived retinoic acid [164]. Therefore, loss of RXRA expression in cirrhotic tissues may contribute to carcinogenesis. The mechanism of RXRA down-regulation is not yet known, but a recent ChIP-Seq study by Chorley et al (2012) [165] implicates Nuclear Factor NRF2, which is activated by oxidative stress and inhibited by the Hepatitis C virus [166].

Chapter 4. CORRELATED EXPRESSION MODULES

A main goal of microarrays is to identify genes or gene sets that are differentially regulated across biological conditions. Even more interesting is to identify which genes are working together in health and disease, how patterns change as disease progresses or to identify diagnostic or prognostic sub-groups. One approach to this question is to examine patterns of correlated expression within and between disease states. Cluster analysis is frequently used, and can identify genes that have highly correlated patterns of expression, but if those genes are not highly correlated in another biological state then the genes would not be associated with another by a clustering algorithm [167]. Variation may exist in the expression of a gene in different sub-populations, leading to incorrect grouping. In addition, cluster analysis identifies groups of genes that are correlated above a certain cutoff, but with no indication of which particular pairs are interacting. Another potentially useful idea is that of differential co-expression (DC). Two genes are DC if their correlation in one condition differs from their correlation in another condition [167]. Similarly, a set of genes is differentially co-expressed if the correlation structure among the group's genes in one condition differs from the correlation structure in another condition [168]. We examined both correlated expression

and differential co-expression to address Aim 2, identifying recurrent patterns and clinical/prognostic sub-types in cirrhosis and HCC.

4.1 Methods

Correlation patterns were examined using Spearman's correlation coefficient on Winsorized data (so that few extreme values did not drive results) for each pair of liver development genes, within cirrhosis and early HCC. Functional gene sets and genes within the same pathway were examined for mutual correlation. Any pair of genes with correlation |>0.5| was considered correlated. To investigate what unanticipated changes may occur between cirrhosis and early tumors, we also built 'naïve' gene sets independent of known functional relationships (see Appendix D). Any gene with at least five other genes correlated at least |>0.5| was defined as a gene set. By doing this for both cirrhosis and HCC separately, we could identify those genes sets that "changed pattern" between cirrhosis and HCC, and determine what those changes were.

We used Gene Set Co-expression Analysis (GSCA) to examine differential co-expression of these gene sets between cirrhosis and early tumors. This method calculates a dispersion index from all of the pair-wise correlations in each gene set for each group (cirrhosis vs. early HCC). Specifically, pairwise correlations were calculated for all gene pairs, then a dispersion index was defined as the Euclidean distance, adjusted for the size of the gene set under consideration:

$$D\ \rho_c^{T_1}, \rho_c^{T_2} \ = \ \overline{\frac{1}{P_c} \sum_{p=1}^{P_c} \rho_p^{T_1,T_2}}\ ,$$

where $\rho_p^{T_1,T_2} = \rho_p^{T_1} - \rho_c^{T_2}$, p=1,…,Pc = $\frac{n_c}{2}$ indexes gene pairs with the gene set c of size nc, and $\rho_p^{T_k}$ denotes the co-expression calculated for the gene pair p within condition Tk, k=1,2.

Samples were permuted across conditions to simulate the null of equivalent correlation between conditions. A score was calculated from the permuted dataset and repeated on 10,000 permuted datasets to yield gene set specific p-values. The correlation method is chosen by the user and we used Spearman's correlation for consistency.

4.2 Correlated gene pattern results

4.2.1 Correlated genes in cirrhosis

Expression analysis of developmental genes in Chapter 3 identified markers of hepatic progenitor cells (EPCAM, KRT19, and VIM) as a consistent pattern in both cirrhosis and tumors. We were first interested in identifying what other developmental genes were correlated with expression of progenitor cell markers for two reasons: first we wished to establish what processes were associated with the progenitor cell signature, and second to establish the "baseline" for comparison, so that we could identify what distinguishes tumor tissue from the cirrhosis samples that did not develop tumors. It was recently suggested that EPCAM+ hepatocytes are the recent progeny of hepatic progenitor cells [126], while KRT and VIM are

markers of the progenitor cells themselves. EPCAM was not correlated with any other developmental gene in cirrhosis, perhaps because it was highly over-expressed in all samples (FC range 5-24). KRT19 and VIM had mutually correlated expression with SMAD7, FSTL3, and ECM genes COL4A2 and LAMB1. VIM expression was also correlated with TGBF1. TGFB1 and SMAD7 are expressed at very low levels in normal tissue and up-regulated in most cirrhosis tissues, and the degree of over-expression is correlated with VIM levels. TGBF1/SMAD7 signaling induces EMT and the expression of VIM in hepatocytes. Vimentin is also correlated with MMP2 (which degrades Type IV collagens), and with a number of laminins, which are produced by activated HSCs and high levels are associated with more severe fibrosis and inflammation. Thus, there appears to be a highly correlated network that show markers of proliferating progenitor cells, activated hepatic stellate cells, and possibly EMT of hepatoctyes.

Necdin is a growth suppressor that interacts with p53, and recent evidence suggests that it activates canonical Wnt signaling in activated HSCs to promote a myogenic phenotype [169]. In our data, Necdin is over-expressed only in cirrhosis, and expression is correlated with SMAD2, SMAD7, TGFB1, and TIMP2, supporting a Wnt connection.

Extra-cellular matrix genes are also highly expressed and highly correlated in cirrhosis, including LAMA2, LAMB1, LAMC1, LAMC3, COL4A1, COL4A2, ITGA3, and ITGA6. These ECM genes are co-expressed with a major component of ECM remodeling in cirrhosis, MMP2, and TGF-β signaling (TGFB3 and TGFBR3.

The other major characteristic of cirrhosis identified from differential expression analysis was the expression of hepatocyte proliferation inhibitors BMP2, CDH1, and MST1.

Surprisingly, these genes were not co-expressed with each other or their regulatory partners. MST1, which is highly expressed in normal tissue and cirrhosis but down-regulated in HCC, was correlated in cirrhosis with Fibronectin, RXRA, Ceruplasmin, and ERBB2, but negatively correlated with KIT and ECM genes LAMA2, LAMB1, LAMC1, LAMC3, ITGA, and MMP9. The biological interpretation of this pattern isn't completely clear, but every cirrhosis sample either over-expressed BMP2 or CDH1 or maintained high expression of MST1, which may imply that any of the proliferation inhibitors is sufficient to maintain hepatocyte quiescence.

4.2.2 Correlated genes in early HCC

In tumors, the same markers of progenitor cells and ECM remodeling described above remain highly correlated (EPCAM, VIM, KRT19, MMP2, MMP7, COL4A1, LAMA2, LAMB1, LAMC3), but this network is no longer correlated with TGFB1 and SMAD7. TGFB1 and SMAD7 have been implicated in EMT, along with NOTCH signaling and loss of E-Cadherin (CDH1). NOTCH and TWIST are not differentially expressed in cirrhosis or HCC, while CDH1 is highly expressed in cirrhosis and most tumors. Loss of CDH1 is associated with both EMT and acquisition of hepatocyte proliferation, and in a subset of tumors was correlated with expression of NOTCH2, SMAD7 and EPCAM, but not with VIM or KRT19. CDH1 and MST1 loss were also correlated in HCC

Of the genes that were uniquely over-expressed in HCC, only TBox3 (TBX3) had correlated expression with other developmental genes, including IRS1, LAMA3, NR5A2, HDGF, YAP1, CADM1, and ITGA6. TBX3 is a downstream mediator of β-catenin signaling that is closely

associated with β-catenin mutational status in HCC [170]. Insulin receptor substrate 1 (IRS1), is up-regulated by constitutively activated β-catenin [171], and NR5A2 is also a β-catenin target gene that may play a role in acquiring pluripotency [172]. IRS2 also interacts with YAP1 and is associated with YAP1 nuclear retention. YAP1 is the end-product of the Hippo signaling pathway and inhibits β-catenin signaling [173]. YAP1 expression is, in turn, correlated with Midkine expression, another Wnt inhibitor. This correlated gene set may suggest the presence of mutationally activated β-catenin is a small subset of our HCC patients.

4.3 Co-expression analysis

In the previous section, correlations among the most highly expressed genes were examined. We found that several genes without overall differential expression significance had correlated expression to DEG, and that some of these pairings had biological relevance. In order to potentially discover other correlated gene sets that occur in HCC sub-populations, a naïve approach was also taken. Rather than limiting our analysis to the expression patterns of the genes with overall significant changes in cirrhosis and tumors, all highly correlated genes were included. Using gene co-expression analysis, gene sets that were highly correlated within one disease group were compared to their correlation structure in the other group, in order to identify what changes in gene expression patterns may accompany the transition from cirrhosis to HCC.

4.3.1 Co-expression patterns in cirrhosis

Within cirrhosis, 30 genes were correlated | >0.5| with at least 5 other genes (some gene correlation groups were a subset of a larger group and these were disregarded). 19 of these gene sets were differentially co-expressed compared to early HCC, that is, the pattern of which genes were co-expressed together in cirrhosis was different than those genes that were co-expressed in HCC. Several of the gene sets overlap to form a co-expression network that was differentially co-expressed compared to HCC (Table 8). This network includes progenitor cell markers VIM and KRT19, and Wnt-inhibitory genes SRPK1, FSTL3, and SMAD7. Casein kinase Iδ (CSNK1D) what highly correlated to each of them. It interacts with both Wnt and YAP1 signaling as well as DNA-repair proteins. TGF-β1 is commonly up-regulated in cirrhosis but hepatocytes are resistant to TGFB1-mediated apoptosis, and this may possibly be partly due to the co-expression of SMAD7, which blocks TGFB1 receptor binding. NFKB1 is a positive regulator of Wnt via direct binding to β-catenin. COL4A2 is induced by TGFB1. It is less clear how the other genes are biologically related in this grouping.

Table 8. Network of co-expressed genes in cirrhosis. Empty cells are those not correlated at least 0.5.

	CSNK1D	SRPK1	FSTL3	TGFB1	PA2G4	SMAD7	COL4A2	NFKB1	BSG	STAT3	YAP1	MAP2K4	ARF6
CSNK1D	1.00	0.79	0.72	0.71	0.67	0.55	0.60	0.67	0.66	0.65	0.63	0.61	0.61
SRPK1	0.79	1.00	0.52	0.54	0.58			0.52	0.52	0.54	0.59		
FSTL3	0.72	0.52	1.00	0.66	0.64	0.59	0.72	0.74		0.58			0.51
TGFB1	0.71	0.54	0.66	1.00	0.79	0.77	0.72	0.58	0.75		0.58	0.61	0.60
PA2G4	0.67	0.58	0.64	0.79	1.00	0.53	0.52	0.60	0.59	0.57	0.59	0.59	0.50
SMAD7	0.55		0.59	0.77	0.53	1.00	0.65	0.53	0.50		0.50	0.52	0.58
COL4A2	0.60		0.72	0.72	0.52	0.65	1.00	0.62	0.52				
NFKB1	0.67	0.52	0.74	0.58	0.60	0.53	0.62	1.00					
BSG	0.66	0.52		0.75	0.59	0.50	0.52		1.00				
STAT3	0.65	0.54	0.58		0.57					1.00	0.55		
YAP1	0.63	0.59		0.58	0.59	0.50				0.55	1.00	0.79	0.60
MAP2K4	0.61			0.61	0.59	0.52					0.79	1.00	0.64
ARF6	0.61		0.51	0.60	0.50	0.58					0.60	0.64	1.00

4.3.2 Co-expression patterns in Early HCC

In general, there were many more groups of co-expressed genes in early HCC compared to cirrhosis: 42 genes had at least 5 genes correlated >0.5. 25 of these gene sets were differentially co-expressed in HCC compared to cirrhosis, and several sets are quite large: LAMA2 (40), MMP2 (38), NDN (37), MET (35), CDH1 (32), FSTL3 (32), STAT3 (30), LAMB2 (27), KIT (26), ZNHIT3 (21), KLF6 (17). Several of these gene sets overlap to form a large co-expression network. As in cirrhosis, the Wnt-related genes CSNK1D, TGFB1, FSTL3, SMAD7, PA2G4, and COL4A2 remain correlated in HCC, but also become part of a much larger network including genes that were not co-expressed in cirrhosis: Necdin, several ECM genes including MMP2/7/19, LAMA2, LAMB1/2, LAMC3, and ITGA3/5, mesenchymal markers KLF6, EPCAM, and

Vimentin, HGF receptor c-MET, FGF receptor 2, EGF receptor ERBB2, and several transcription factors (ARID5B, CITED, LHX2, GATA6) (Table 4.2). Necdin is an imprinted gene associated with increased expression in hepatic progenitor cells (Chang 2009). KIT, a stem cell factor receptor and marker for several types of stem cells, is highly correlated with NDN in early HCC but not cirrhosis.

KLF6 is a nuclear protein thought to be a tumor suppressor. In our data it is highly over-expressed in cirrhosis and less highly over-expressed in HCC. Although consistently over-expressed in our cirrhosis tissues (FC range 2.1-5.9), its expression is not correlated with other developmental genes. However, in early HCC KLF6 expression becomes much more variable and loss of expression is correlated to a subset of the NDN/KIT/MMP network, including tumor suppressors GATA6 and ARID5B, STAT3, LAMB1/2, SFRP5, FSTL3, and MET (Table 10). In addition, expression is positively correlated with ARF6, PIK3R1, and ID2. Early HCC patients who had recurrence/death had similar expression levels of KLF6, STAT3, and ID2 to late-stage HCC samples (lower than that of early HCC patients who did well), while GATA6, ARID5B, LAMB1, and LAMB2 were under-expressed compared to both Early HCC/good outcomes and Late HCC.

Table 9. Genes co-expressed with KLF6 in Early vs. Late HCC (compared to normal controls)

Gene	Early HCC/ poor outcome	Late HCC/ poor outcome	Early HCC/ good outcome
KLF6	1.9	1.9	3.0
STAT3	0.4	0.5	0.6
ID2	0.6	0.7	0.8
GATA6	0.7	1.1	2.1
ARID5B	1.4	2.0	2.4
LAMB1	1.5	2.2	2.9
LAMB2	0.8	0.9	1.0
MET	1.4	1.2	1.0

c-MET, the Hepatocyte Growth Factor Receptor, promotes hepatocyte proliferation and is down-regulated in cirrhosis but not in HCC. E-Cadherin (CDH1), ID2, and EGF receptor ERBB3 are positively correlated to MET in cirrhosis but negatively correlated in tumors (Table 10). E-Cadherin is an inhibitor of hepatocyte proliferation up-regulated in cirrhosis and highly variable in HCC. Co-expression analysis suggests that it "changes allegiance" in early HCC. TGFBR3 and MET are positively correlated to CDH1 in cirrhosis and negatively correlated in tumors. In tumors, CDH1 and MST1 are correlated, and these two genes are mutually correlated with several other developmental genes including receptors ERBB2, FGFR2, and NOTCH2, transcription factors STAT3, LHX2, GATA6, CITED2, and Wnt antagonists SFRP5 and SMAD7. It is also negatively correlated with GREM1, CTNNB1, FOXM1, and ZNHIT3 expression (Table 9). Cirrhosis samples consistently up-regulated CDH1 while MET had normal or below normal expression. The E-cadherin network appears to be protective in HCC patients: of the 16 early HCC with CDH1 up-regulated (and higher expression of associated genes), 15 were transplanted

and still alive 2-10 years post-tx (94%). The other thirteen early HCC patients had normal or decreased expression of CDH1 and MST1 and increased expression of MET. Of these, 8 (62%) died (4 due to recurrence, 4 from other causes). Embryonically, E-cadherin expression is lost by hepatoblasts as they acquire a migratory phenotype. It has a long-established role in malignant cell transformation and is regarded as a "suppressor of invasion" as loss of function correlates with increased invasiveness and metastasis in many cancer types [174]. Similarly, MET amplification has a well-established association with invasion and recurrence in HCC [175, 176]. Lee et al (2009) identified a gene signature for lymph node invasion in mixed HBV/HCV-HCC that included MET overexpression and CDH1 under-expression [177]. This data appears to provide evidence that there is a sub-population of HCC samples that acquires the capacity for hepatocyte proliferation, and that this subgroup has poor outcomes compared to those HCC that maintain hepatocyte quiescence.

YAP1 signaling stimulates growth of both hepatoblasts and maturing hepatocytes in the developing liver, and over-expression leads to liver overgrowth and tumors. In our data, YAP1 over-expression is associated with poor outcomes in early HCC and most of the long-term survivors have down-regulated YAP1. However, samples taken from late-stage tumors also have low levels of YAP1. MST1 is a key regulator of YAP1, and they might be expected to be correlated. Surprisingly, this is not the case. Not only are they un-correlated (ρ=0.06), they have non-overlapping sets of correlated genes. MST1, which is down-regulated in HCC, is dicussed above. YAP1 co-expresses with MDK, NR5A2, TBX3, ITGA6, MET, HDGF, SRPK1, and MAP4K4. Several of these genes are associated with β-catenin mutational status as discussed in Section 4.2.2. Thus it appears that YAP1 is part of a co-expression module associated with β-

catenin mutations, while YAP1 inhibitor MST1 is associated with the CDH1 signature of samples with poor prognosis.

Table 10. Correlated gene network in early HCC. Only correlations >0.5 are displayed.

	NDN	MMP2	KIT	LAMA2	FSTL3	LAMB2	VIM	EPCAM	MMP7	JAG1
NDN	1.00	0.84	0.80	0.80	0.65	0.67	0.65	0.60	0.56	0.65
MMP2	0.84	1.00	0.74	0.74	0.71	0.66	0.71	0.57	0.62	0.72
KIT	0.80	0.74	1.00	0.74	0.70	0.61	0.61		0.50	0.62
LAMA2	0.80	0.74	0.74	1.00	0.69	0.56	0.63	0.58	0.51	0.70
COL4A2	0.77	0.76	0.73	0.74	0.65	0.54	0.62	0.64	0.51	0.75
ARID5B	0.76	0.78	0.73	0.80	0.66		0.86	0.63	0.66	0.81
FGFR2	0.71	0.56		0.70				0.71		0.58
ITGA5	0.70	0.67	0.65	0.58	0.64	0.67				0.59
PTN	0.70	0.52	0.56	0.73			0.59		0.54	
SMAD7	0.70	0.62		0.60	0.56	0.75		0.52		
LAMC3	0.68	0.66	0.72	0.82	0.72	0.58				0.50
LAMB2	0.67	0.66	0.61	0.56	0.69	1.00				
NOTCH2	0.65	0.75	0.58	0.64	0.73	0.70				0.57
FSTL3	0.65	0.71	0.70	0.69	1.00	0.69				0.60
LAMB1	0.65	0.74	0.65	0.73	0.54		0.75	0.67	0.71	0.74
JAG1	0.65	0.72	0.62	0.70	0.60		0.79	0.72	0.68	1.00
VIM	0.65	0.71	0.61	0.63			1.00	0.59	0.73	0.79
CITED2	0.64	0.55	0.55	0.56	0.57	0.63				
TGFB1	0.61	0.54	0.56		0.60	0.57				
EPCAM	0.60	0.57		0.58			0.59	1.00		0.72
GATA6	0.59	0.68	0.63	0.66	0.81	0.74				
CDH1	0.58			0.60				0.53		
MMP19	0.56	0.68	0.75	0.72	0.74		0.65		0.50	0.62
MMP7	0.56	0.62	0.50	0.51			0.73		1.00	0.68
ERBB2	0.56					0.54				
ITGA3	0.54	0.59		0.57	0.52	0.74				
STAT3	0.54			0.50		0.64				
NFKB1	0.53									
KLF6	0.52	0.53	0.54		0.56	0.56				
COL4A1	0.52	0.57	0.57	0.56			0.70	0.52	0.51	0.71
SFRP5	0.50	0.59		0.73	0.63	0.55		0.57		
KRT19	0.50	0.67	0.54	0.67	0.59		0.70	0.69	0.63	0.76
FOXM1	-0.51			-0.51	-0.50	-0.61				
MET	-0.61	-0.61	-0.51	-0.54		-0.74		-0.51		

Chapter 5. VALIDATION

5.1 Comparison with moderated t-test using limma

we confirmed 1,311 genes with no documented liver expression. Of these, four were DEG in the limma analysis, which is well within the false discovery rate. These four were CELA3A and PGC (digestive enzymes), ODF1 (a sperm protein), and PSG1 (a pregnancy-specific glycoprotein).

More specifically, we wished to determine whether the general developmental pathways altered in HCC were using genes specific to liver development, or whether any member of the gene family might be engaged. To examine this, we identified 26 paralog genes that have highly related developmental functions in other tissues, that are not expressed in normal healthy livers (based on the RNA sequencing data and verified with a literature search). In our data, no paralogs were expressed in disease compared to normal samples (FDR <0.01) (Table 10, column A). We validated this to an independently collected HCV-HCC dataset from Wurmbach et al (see methods), which also had no expression of these paralog genes (Table 10, column B). Density plots illustrating the common patterns seen in the liver development compared to non-liver paralog genes are shown in Figure 5.1. Normal and HCC samples have no expression of the non-liver gene (RXRG and SOX1), which have narrow expression

distribution. RXRA is expressed in normal tissue and down-regulated in HCC, while SOX9 is over-expressed compared to normal samples.

Unfortunately there is currently no authoritative, comprehensive annotation of all tissue-specific expression. The liver has the ability to activate many metabolic and detoxifying mechanisms only when needed, and these would not be expressed in most "normal liver" samples used for comparison. However, we identified over a thousand genes that were not expressed in the reference healthy liver sample by RNA sequencing, and only 4 of them showed differential expression in our tumor samples. These results suggest that the changes occurring in cirrhosis and HCC are driven by aberrant expression of genes normally expressed in the liver, or expressed at some time during the normal life history of the liver. Activation of genes not normally expressed by the liver might be expected to occur via such processes as copy number variation or DNA replication damage to promoter regions of random genes, but we found no evidence of such activation.

5.1.1 Genome-wide testing results

Low variance genes (st. dev. < 0.3) were filtered, leaving 11,731 probesets. The 7,270 probe sets with unique gene names and expression were retained. 3,170 probe sets (43.6%) were differentially expressed between cirrhosis and normal samples, and 1,543 were differentially expressed between early HCC and cirrhosis. This is far too many to evaluate

individually, and these probe sets were evaluated using GSEA (results below). All 90 differentially expressed liver development genes were also identified in this analysis as well.

Only 8 probe sets were differentially expressed between early and late HCC including IGF2, a developmental gene that was identified in our targeted analysis. The other late-stage genes were INS-IGF2 read-through, an open reading frame that contains alternative splicing regions for Insulin and IGF2; Aurora-A binding protein (AIBP), which is involved in chromosome alignment during cell division; DUSP6, which inactivates MAPK1; Inositol polyphosphate 1-phosphatase (INPP1), a general signal transduction membrane protein; ATP10B, an ATPase; and Tetranectin, which stimulates muscle differentiation during embryonic development. Tetranectin is intriguing because it is involved with plasminogen activation and ECM remodeling, and is associated with poor prognosis in several cancer types, including oral [178], ovarian [179], bladder [180], and colorectal [181]. Although it has not been previously associated with HCC, it has recently been proposed as an important factor in the survival of pancreatic islet cells after transplantation into the liver [182, 183]. Since plasminogen and tetranectin are both produced by hepatocytes, this raises the tantalizing idea that tetranectin might have an un-recognized role in liver development.

5.1.2 Differentially expressed genes are specific to liver development

We hypothesized that genes not normally expressed in adult livers are less likely to be transcriptionally activated in HCV-HCC. To test this, we identified a set of genes with zero

counts in an RNA sequencing study on a normal liver sample from the BodyMap project. 1,399 of these were represented on the Affymetrix U133Av2 genechip. Genes that are not expressed in normal liver tissue have varied measured values caused by variations in background non-specific hybridization and by technical noise that was not corrected by normalization. Because these distributions are highly non-normal (a high peak around "zero" expression and narrow variation), we assessed expression of these genes in disease samples with a one-sided, non-parametric two-sample Kolmogorov-Smirnov Test to test differences in both location and shape of the distributions.

There were 36 genes with significant expression in cirrhosis samples and 31 DEG in HCC samples. A search of online databases of tissue-specific expression (immunobase.org, nextprot.org, BioGps.org, and bGee.unil.ch) confirmed that 32 of these genes can be expressed in the liver, along with a further 22 genes that had no expression in either the BodyMap sample or our samples. This is reasonable because not all possible genes will be expressed at all times in a given sample.

Table 11. Liver development genes compared to their non-liver paralogs. (A) one-sided K_S test of identical distribution comparing HCC to normal samples; (B) one-sided K-S test in the Wurmbach dataset; (C) K-S test comparing the liver development gene to its non-liver paralog.

Liver development gene	Expression in normal adult liver	Non-liver paralog	Non-liver gene, tumor vs. normal VCU data A	Non-liver gene, tumor vs. normal Wurmbach data B	Liver vs. non-liver gene in tumors VCU data C
ACVR2A	+	AMHR2	0.18	0.009	0.14
BMP2	+	BMP3	0.014	0.57	9.9 x10-10
BMP4	+	BMP3	0.014	0.57	0.0091
CDH1	++	CDH3	0.20	0.007	1.7x10-6
ELF5	-	SPDEF	0.003	0.29	5.2x10-08
FGF1	-	FGF3	0.73	0.37	0.0023
FGF2	+	FGF3	0.73	0.37	1.7x10-6
FGF7	+	FGF12	0.81	0.20	1.7x10-6
FGF8	-	FGF17	0.90	0.69	0.14
FOXA1	++	FOXB1	0.02	0.97	2.2x10-16
FOXA2	++	FOXD2	0.06	0.72	1.2x10-12
GATA4	++	GATA1	0.05	0.18	2.4x10-10
GATA6	+	GATA1	0.05	0.18	1.7x10-6
GPC3	-	GPC4	0.29	0.22	1.1x10-11
HHEX	++	VENTX	0.002	0.32	3.6x10-5
HLX	+	BARX1	0.02	0.02	0.00049
IL6ST	+++	IL12RB2	0.25	0.59	7.3x10-6
KIT	+	FLT3	0.87	0.09	0.0011
KRT19	+	KRT17	0.12	0.59	3.7x10-8
LHX2	+	LHX1	0.31	0.55	2.7x10-6
MET	++	MST1R	0.19	0.006	0.00026
MMP7	+	MMP10	0.59	0.07	2.2x10-16
MMP12	-	MMP10	0.59	0.07	0.0012
MMP14	+	MMP10	0.59	0.07	1.2x10-7
MMP19	+	MMP10	0.59	0.07	0.0025
MMP2	+	MMP10	0.59	0.07	2.2x10-16
NR5A2	++	NR5A1	0.32	0.37	2.2x10-16
NRTN	+	PSPN	064	0.36	2.2x10-16
RXRA	+++	RXRG	0..05	0.15	2.2x10-16
SOX9	*	SOX1	0.27	0.05	4.4x10-16
SOX17	+	SOX11	0.01	0.26	0.0005
TBX3	++	TBX2	0.34	0.02	5.6 x10-7
WT1	-	EGR4	0.14	0.009	3.6x10-5

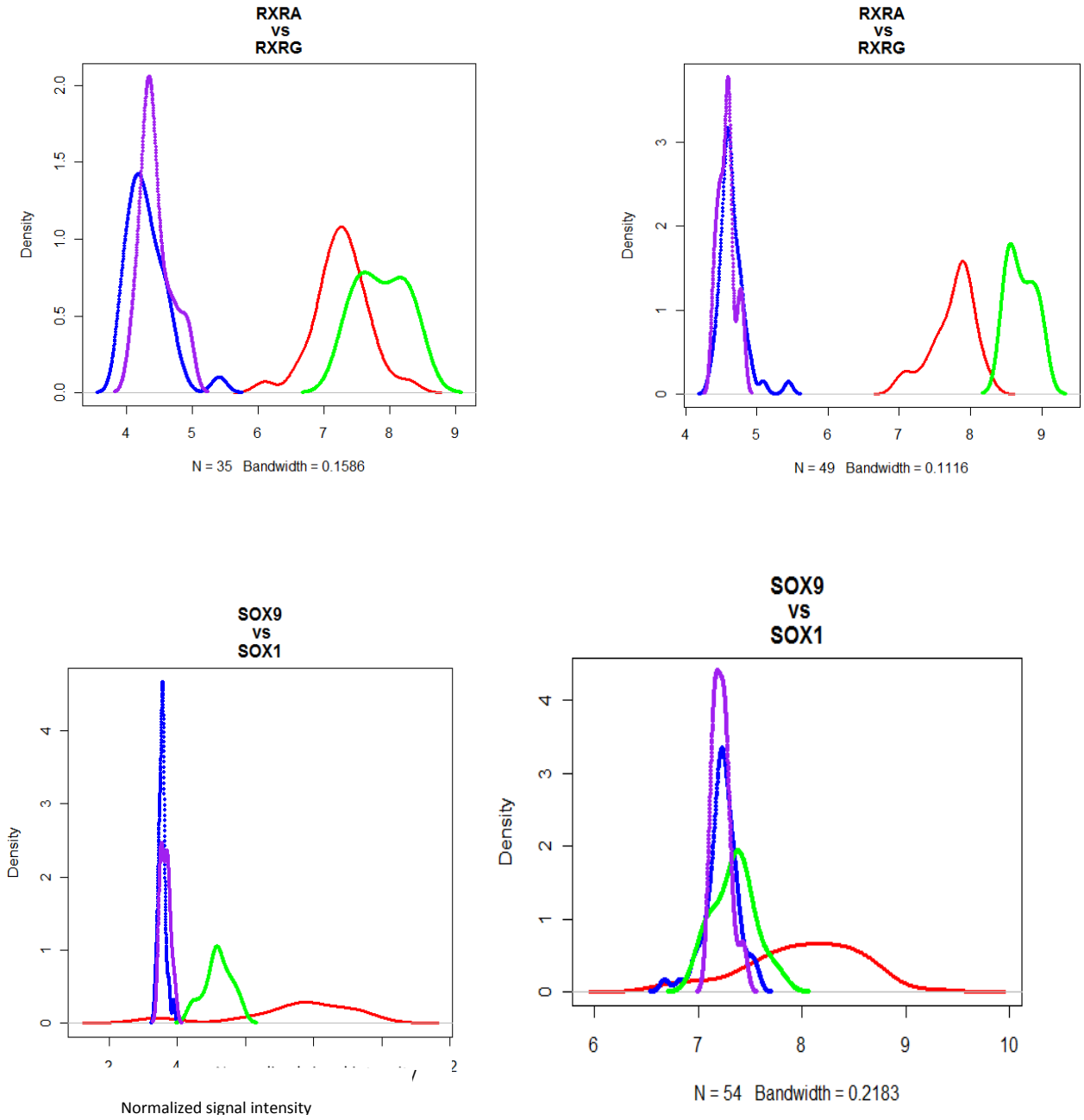


Figure 5.1. Selected density plots of liver development vs. paralog non-liver development genes. Expression densities are shown for gene pairs in normal and HCC samples from our data and in the Wurmbach dataset. Red = expression of the liver gene in HCC; Green = expression of liver gene in normal controls; Blue- expression of non-liver gene in HCC; Purple= expression of non-liver gene in normal controls. Paralog genes (RXRG and SOX1) were not expressed in HCC or normal samples, while liver development genes RXRA and SOX9 were differentially expressed in HCC. These patterns were also observed in the Wurmbach dataset.

5.2 Comparison with Gene Set Enrichment Analysis (GSEA)

GSEA generates scores based on possibly small-amplitude but coherent changes in the expression of sets of genes. These gene sets can be based on pathways, common biological function, chromosomal location or common regulations. GSEA works by determining whether members of a gene set S tend to occur near the top or bottom of the list L of ranked gene expressions from samples with two phenotypes, and comparing this to the null hypothesis that the genes are randomly distributed throughout the ranked gene list.

Specifically, the method proceeds in three steps. First, an enrichment score (ES) is calculated by taking a running sum across L of the scores of all the genes in S (increasing the score) and those genes not in S (decreasing the score). If S is randomly distributed throughout L, it's enrichment score will be small, but if many members of S are clustered at the top or bottom of L, then the ES will be high. ES corresponds to a weighted Kolmogorov-Smirnov- like statistic, where the weights correspond to the expression value for each gene. The significance of ES is estimated by comparing to a null distribution generated by a permutation test, where the samples are permuted. Finally, multiple hypothesis testing is accounted for by normalizing the ES against the size of the gene set and calculating the FDR by comparing the tails of the observed and permuted null distributions for the normalized ES.

We first filtered probe sets to only include those with standard deviation of greater than 3 and excluded probe sets from the same gene that had identical expression values in all samples. Using the Gene Set Enrichment Analysis Java interface (version 2.0), we submitted the

remaining 7,270 probe sets for GSEA analysis using several publically available and commonly used gene sets: GO terms for molecular function, GO terms for biological processes, Biocarta canonical pathways, and KEGG pathways. Gene sets with at least 8 genes were included. We also tested the liver development genes as a gene set for comparison.

5.2.1 Analysis of public gene sets and pathways

GO molecular function (MF) terms describe the molecular tasks that gene products perform. MF gene sets include all the genes that are annotated with these functions. Some examples of MF gene sets that might be predicted to be enriched include Antigen_binding, Integrin_binding, Interleukin_receptor_activity, Protein_kinase_binding, Cytokine_activity, Cytokine_receptor_binding, or Smad_binding, for example. Analysis of 325 gene sets with an FDR<0.25 resulted in no significant genesets in any pairwise comparison of Normal, Cirrhotic, Early HCC, or Late HCC samples.

GO Biological Processes (BP) terms describe what type of process a gene product is part of , for instance, Cell Cycle, Development, Metabolic processes, Response to stimulus, or Signaling. Some BP gene sets that might be predicted to be enriched are Activation_of_MAPK_activity, Intracellular_signaling_cascade, JAK_STAT_cascade, Regulation_of_angiogenesis, or Cell_cell_adhesion, to name a few. 598 BP sets with >8 genes were analyzed with an FDR<0.25. Seven gene sets were significant when comparing Cirrhosis to Normal controls, 6 involving regulation of immune response, especially T-cell activation and

proliferation (Table 12). The 7th gene set was “nuclear localization signal-bearing substrate into the nucleus”. Comparison of Early HCC to Normal tissue identified two significant gene sets, Endoplasmic Reticulum Nuclear Signaling, and Biogenic amine metabolism. No gene sets were significantly different between Early or Late HCC and Cirrhosis samples.

Table 12. GO Biological Processes with significant enrichment at FDR<0.25.

Comparison	Significant Gene Set	q-value
CIR-NOR	T-cell proliferation	0.209
CIR-NOR	Positive regulation of immune response	0.019
CIR-NOR	Positive regulation of lymphocyte activation	0.09
CIR-NOR	Positive regulation of immune processes	0.239
CIR-NOR	T-cell activation	0.24
CIR-NOR	Regulation of lymphocyte activation	0.24
CIR-NOR	Response to virus	0.243
CIR-NOR	NLS substrate nuclear import	0.24

Comparison of the canonical pathways from Biocarta identified 18 (of 215) significant pathways between Cirrhosis and Normal samples. As with the Biological Process GO terms, most of these pathways involved immune response and T-cell regulation. Two unexpected pathways were “Dream” (involved in repression of pain sensation), and “Vitamin C in the Brain”. Comparison of Early HCC to Normal identified only the T-Cell Receptor activation pathway, and there were no differences between Early HCC and Cirrhosis, or Late HCC to Early HCC (Table 13).

Table 13. Biocarta pathways significantly enriched at FDR < 0.25

Comparison	Pathway	q-value
CIR-NOR	CCR5 pathway	0.058
CIR-NOR, Early-NOR	T-Cell Receptor Activation pathway	0.071, 0.24
CIR-NOR	T-cytotoxic pathway	0.091
CIR-NOR	CTL pathway	0.099
CIR-NOR	IL17 pathway	0.138
CIR-NOR	T-helper pathway	0.138
CIR-NOR	Lym pathway	0.182
CIR-NOR	ArenRF2 pathway	0.166
CIR-NOR	B-lymphocyte pathway	0.154
CIR-NOR	Fibrinolysis pathway	0.196
CIR-NOR	Platelet app pathway	0.212
CIR-NOR	TC apoptosis pathway	0.247
CIR-NOR	CTLA4 pathway	0.24
CIR-NOR	Granuloctyes pathway	0.228
CIR-NOR	Monocyte pathway	0.234
CIR-NOR	VitC in the Brain pathway	0.235
CIR-NOR	AS B-cell pathway	0.24
CIR-NOR	DREAM pathway	0.228

The Kyoto Encyclopedia of Genes and Genomes (KEGG) contains not only canonical pathways, but also processes and gene sets associated with particular diseases. 186 KEGG genesets were analyzed. 8 pathways were identified in the Cirrhosis-Normal comparison: Asthma, Allograft rejection, Type I Diabetes Mellitus, Intestinal Immunity, Graft Vs. Host Disease, Viral Myocarditis, the Leithmania Infection. Some relevant genesets that had nominal, but not FDR-corrected significance, included Cell Adhesion, Extra-Cellular Matrix Receptors, Antigen Processing, and Leukocyte Migration. The Early HCC vs. Normal comparison identified the same pathways as found in CIR-NOR, with the addition of Autoimmune Thyroid Disease and Folate Biosynthesis. None of the other comparisons were significant at FDR<0.25, although the "Prostate Cancer" gene set was significant at p=0.009 between Early HCC and Cirrhosis.

5.2.2 GSEA of liver development genes

A gene set was constructed of all the liver development genes that passed the filtering of SD>0.3 (116 genes). The Cirrhosis-Normal, Early HCC- Normal, and Early HCC-Cirrhosis comparisons were significant (q=0.03, q=0.08, and q=0.078, respectively). The 10 highest ranking genes discriminating Cirrhosis from Normal were EPCAM, KRT19, MMP7, SOX9, MMP2, SPP1, LAMA2, COL4A1, and FGFR2, TGFB1. The highest ranking genes discriminating Early HCC from Cirrhosis were KRT19, MMP12, LAMA2, GREM1, GPC3, GATA6, SMAD7, FGFR2, FSTL3, and MMP7.

5.3 Validation to the Wurmbach dataset

One of the frustrating outcomes after a decade of investigation into the molecular basis of HCC has been the lack of reproducibility. We used a publicly available dataset of HCV-induced cirrhosis and HCC from Wurmbach et al (2007) [30] to evaluate whether the patterns that we identified in our data were also evident in this independently collected dataset.

5.3.1 Developmental gene sets

As noted in Section 5.1.2 above, we have shown that the developmental genes dysregulated in HCC are specific to liver development in both our data and in the Wurmbach data. Gene sets that we identified from our data were validated against the Wurmbach dataset

by applying the PCA loadings from our data to their dataset, and we observed similar patterns of separation between normal, cirrhosis, and HCC tissues (Figure 5.1, 5.2). Normal vs. cirrhosis tissues were not as well distinguished but late stage tumors show better separation from early tumors. As in our data, there were no major genes driving the principal components for each developmental stage, rather several genes contributed.

.

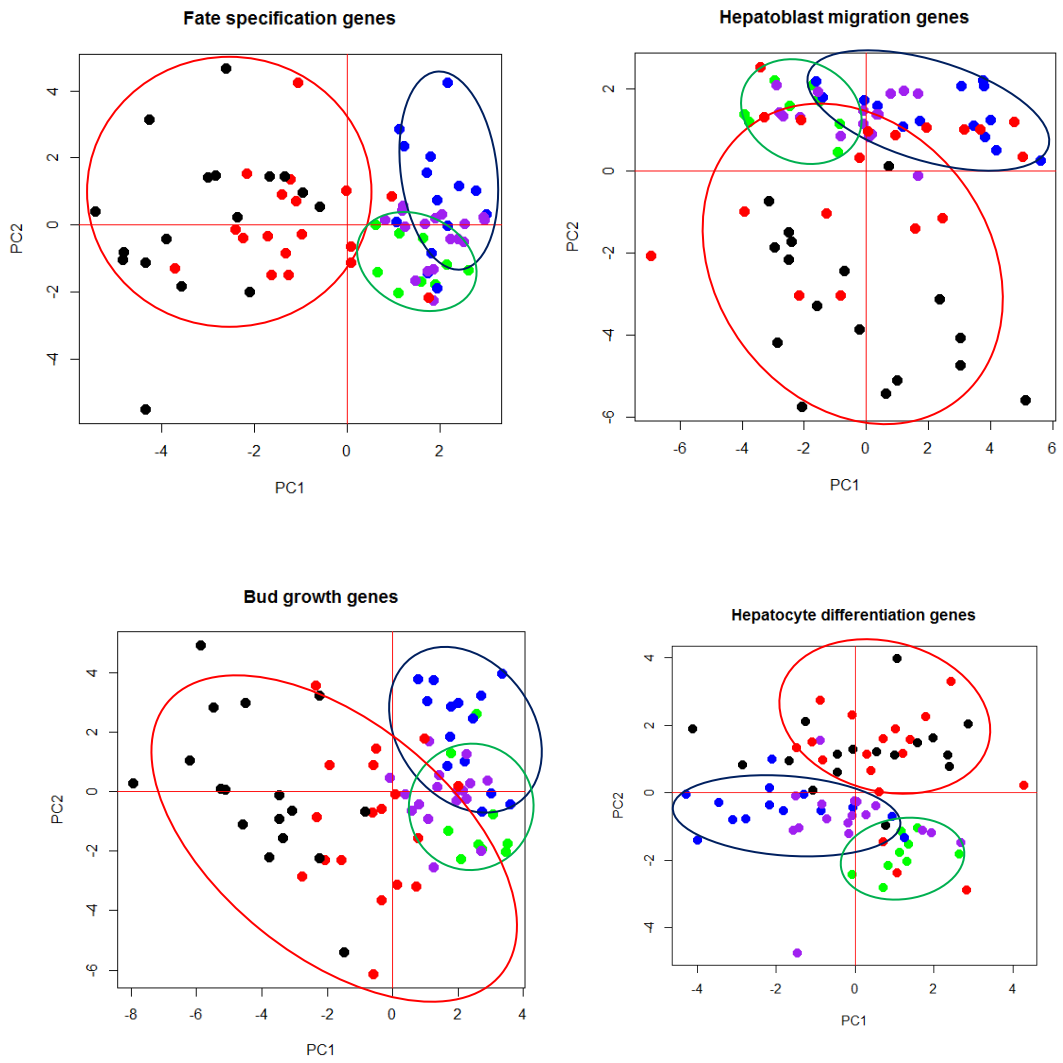
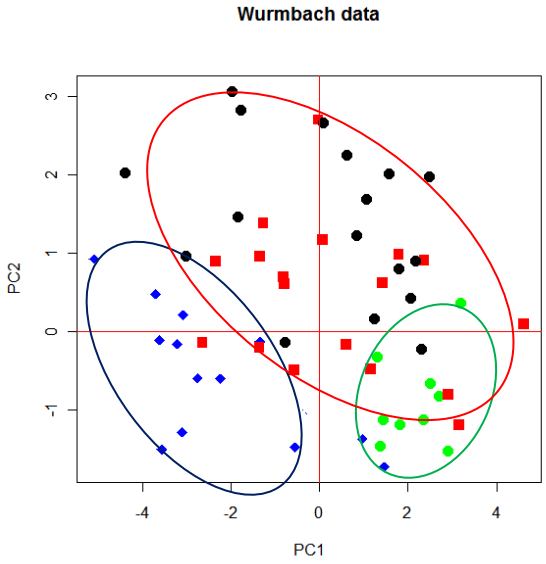
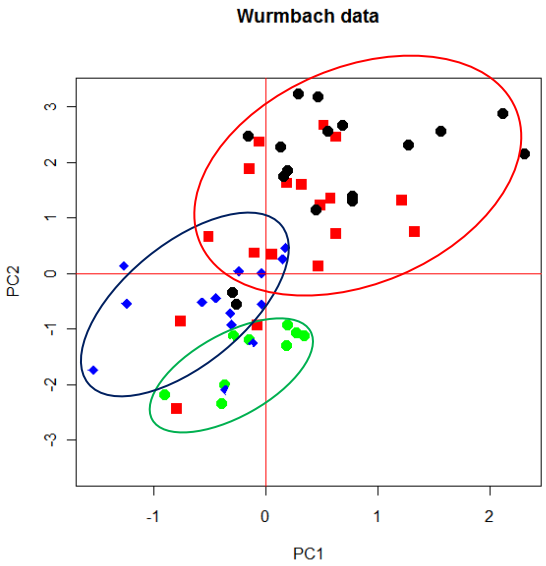


Figure 5.2 PCA plots of developmental gene sets by stage of development in the Wurmbach dataset. Green = normal; Blue = cirrhosis; Purple = cirrhotic tissue surrounding tumor; Red = early stage HCC; Black = late stage HCC.

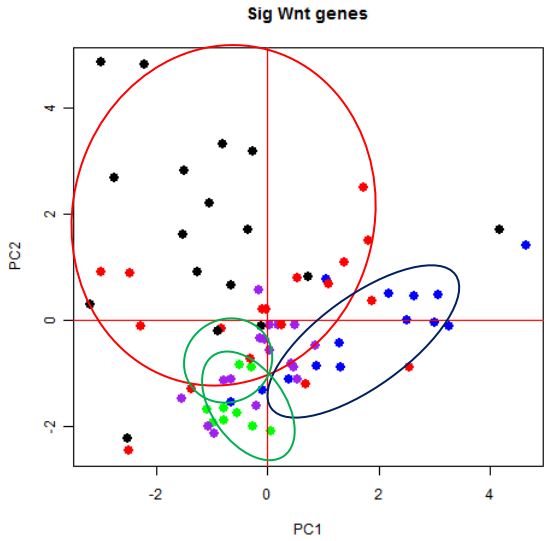
(A) ECM genes



(B) BMP genes, receptors, and inhibitors



(C) Wnt genes



(D) Progenitor-Hepatocyte Signature

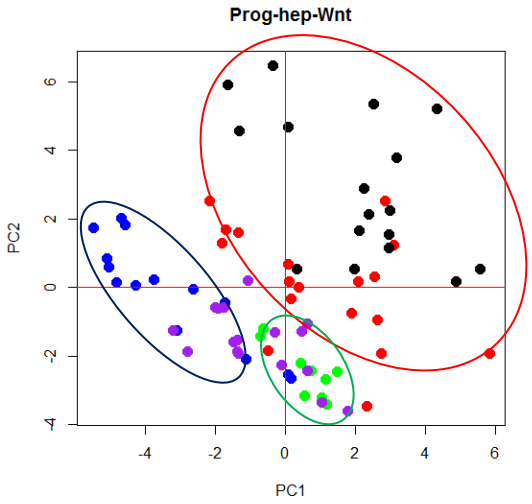


Figure 5.3. PCA of gene sets identified in our data in the Wurmbach dataset. Green = normal; Blue = cirrhosis; Purple = cirrhotic tissue surrounding tumor; Red = early stage HCC; Black = late stage HCC.

The progenitor-cell signature (VIM, EPCAM, and KRT19) was also up-regulated in cirrhosis samples and in the dysplastic nodules in the Wurmbach data. Inhibitors of hepatocyte proliferation CDH1 and BMP2 were up-regulated and MST1 expression remained an normal (high) levels in cirrhosis, as in our data. However, expression of EPCAM and KRT19 was lost in most of the tumor samples, although secondary markers of progenitor cells were over-expressed (SOX9, Midkine, and Nidogen). CDH1 and MST1 expression was much more down-regulated in Wurmbach's data compared to our data, while MET was over-expressed in a subset of tumors, suggesting that in the Wurmbach dataset hepatocyte proliferation is a more common HCC sub-type. Since Wurmbach's data contained more advanced tumors, and included more metastatic tumors than our data, this supports the observation that hepatocyte proliferation seems to signify poor prognosis.

5.3.2 Regeneration genes

Genes involved with liver regeneration were also examined in the Wurmbach data. As in our data, early response gene IL-6 was up-regulated in cirrhosis but not HCC, while TNF did not change expression compared to their normal controls. Several of the intermediated response genes are also liver development genes (STAT3, HGF, JUN, MYC, and MET) and are up-regulated during regeneration. In the Wurmbach dataset, HGF, JUN, MYC, and MET are down-regulated in most of the HCC samples. However, proliferation genes FOXM1, CCNB1, and CDC25C were up-regulated in late-stage HCC while signals thought to terminate the proliferative phase were down-regulated (SERPINE1, SERPINE2). In our data, proliferative

signals were not present, however, Wurmbach's data includes more metastatic late-stage samples.

5.4 Discussion

Our approach in this project was to do an in-depth, focused analysis of genes likely to be engaged in liver dysfunction in an effort to address historical problems with high false positive rates, difficulty in interpretation of thousands of results, and reproducibility across datasets. In order to assess our success in this endeavor, we compared our results to a "standard analysis" of moderated t-test of means using limma, and applying Gene Set Enrichment Analysis (GSEA) to the results. We found that, as expected, several thousand genes were differentially expressed using this approach. Although the liver development genes were among these significant results, it would have been impossible to highlight them as more important than the other significant genes except perhaps using GSEA to capture enrichment in particular pathways. However, GSEA also failed to identify the pathways represented by the significant developmental genes, possibly because the KEGG gene set definitions include all members of a pathway without regard to tissue specificity. Thus the contribution of the dysregulated liver genes may have been 'diluted' by the non-contributing genes that are specific to other tissue types.

The only gene sets identified by GSEA involved pathways involved in response to infection, which was expected in this population of patients with active, chronic HCV infection. However, GSEA applied to a gene set of all liver development genes was significant when

comparing cirrhosis to normal, early tumors to normal tissue, and early tumors to cirrhosis. The highest ranking genes included some of the progenitor cell markers, cell adhesion genes, and BMP inhibitors discussed in previous sections.

We also attempted to validate our important results in the Wurmbach dataset, which is one of the few other publically available datasets of HCV-induced cirrhosis and HCC. This dataset contains more late-stage tumors than our data, and we were interested in looking for differences between early and late tumors in their data since we found almost no differences between early and late tumors in our data. We applied the PCA models (loadings) calculated from our dataset to the Wurmbach data, and in every case the patterns were similar. In most cases, the advanced tumors had similar patterns but were the more extreme values (outliers) in the graphs. The exception was found in the hepatocyte regenerative markers, which were up-regulated in Wurmbach's advanced tumors, which included more metatatic samples, in contrast to our dataset which contained large non-metastatic tumors. Some of the surprising negative results, including the lack of p53 and Hedgehog pathway dysregulation, were also reproduced in the Wurmbach dataset.

Chapter 6. Discussion and conclusion

6.1 Data quality

The main theme of this dissertation was to look below the surface of the big changes common to all tumors – metabolic disturbances, cell turnover, inflammatory response, etc - and find the potentially subtle signals of master regulators that drive tumor initiation and progression. Tumor biology is inherently variable, and the technical noise in microarray experiments can obscure actual transcript abundance. Therefore it was critical to remove as much noise as possible from the data with careful attention to data quality and pre-processing in order to maximize the precision of the measured signals. We chose "quality over quantity" and applied stringent quality control criteria, ultimately excluding almost 40% of the GeneChip results with unrecoverable technical artifacts. We were also careful in our normalization in order to minimize adding bias to the results. To test normalization success, we checked for apparent expression of Y-chromosome genes in samples from females, and also compared the density curves of genes known to have zero expression in the liver. In our initial normalization strategy we found that HCC samples had apparent down-regulation of non-expressed genes, indicating that the model was over-correcting the data, and adjusted the model to remove this artifact.

6.2 Dysregulated genes are specific to those normally expressed by the liver

Our main aim was to examine the role that liver development genes play in HCV-induced HCC. To justify this narrow focus on just a few hundred genes out of thousands, we also examined whether the genes activated are specific to the life history of the liver, and whether the patterns of gene expression recapitulate other processes that may occur in the liver such as wound healing and regeneration. We addressed the issue of liver specificity by identifying genes (both developmental and otherwise) that are not normally expressed in the liver. This was a non-trivial exercise because the liver has the ability to express thousands of genes to fulfill a multitude of functions including fat, glucose, iron, and amino acid homeostasis, combating infection, neutralizing drugs and environmental toxins, and manufacturing many hormones and enzymes that are used elsewhere in the body. Ultimately we found a set of about 1,400 genes (present on the Affymetrix HG-U133v2 GeneChip) that are not normally expressed in healthy livers or expected to be induced in response to infection or toxins. Twenty-nine of these genes were developmental genes that play highly paralogous roles in other organs (including BMP3, CDH3, FGF3, FGF12, FOXB1, GPC4, and others). Sixty genes were non-liver members of developmental gene families with less obvious paralog functions, including ATF, FOX, BMP, FGF, TBX, CDH, CDX and MMP genes. None of the 89 non-liver developmental genes and only three of the other non-liver genes were significantly differentially expressed, which is well within the number of false positives predicted by an FDR < 0.05. This provides evidence that the processes involved in the development of liver tumors

in patients with chronic HCV does not involve random activation, but are more likely systematic responses to the stress of chronic infection and cirrhosis that may be poorly controlled or out of balance in tumors [184].

6.3 The role of liver regenerative processes in cirrhosis and HCC

A healthy liver maintains hepatocyte quiescence, enforced by expression of CDH1 and MST1. When injured, the liver regenerates mass by repressing these genes and inducing hepatocyte proliferation in a cascade triggered by IL-6 and TGFα. In our cirrhotic samples, the genes that normally maintain hepatocyte quiescence (CDH1 and MST1) are highly expressed, and inhibitors of hepatocyte proliferation (TGFB and BMP2) were up-regulated. HGF and c-MET, the final link in the chain of events leading to induction of hepatocyte proliferation, were not activated in cirrhosis. However, in nearly half of the early tumors, c-MET was up-regulated and expression of proliferation inhibitors CDH1 and MST1 was lost, suggesting that there may be a sub-type of HCC that gains hepatocyte proliferative abilities. This group has 38% two-year recurrence-free survival compared to 95% survival in patients with elevated CDH1 and low MET expression. This is similar to a prognostic subtype proposed by Yamashita et al, 2009 for HBV-induced HCC [132] .

Li, et al (2010), compared publically available microarray datasets of mouse liver development and mouse liver regeneration expression profiles to Wurmbach's HCV-induced HCC data [185]. Using hierarchical clustering and two-dimensional clustering techniques, they found that liver regeneration samples had very different expression profiles from both HCC and

liver development, while HCC and liver development had similar expression profiles. Their results are interesting because the human HCC profiles were compared to mouse development and regeneration profiles. One might expect that the mouse samples would be more similar to each other than to any human samples, and the cross-species agreement between development and tumors was stronger than the within-species differences between development and regeneration.

In the only study of its kind, Ho et al (2007) obtained gene expression profiles from the livers of five living donors for transplant at two points during the donor operation. Biopsies were taken at the beginning and end of the partial hepatectomy procedure (about five hours later). They identified a signature of cell growth and immune response genes including SAA1/2, CPR, CHST4, S11A8, SOD2, RASD1, PBEF1, RRS, and SOCS3 that are up-regulated, which is in good agreement with known targets of the immediate response genes TGFα and IL-6 [125]. Interestingly, in our data, these genes were either not differentially expressed or were down-regulated. This may suggest that the proliferation resulting from chronic injury is not triggered by the same mechanisms involved in response to acute injury.

6.4 Progenitor cells and liver healing mechanisms recapitulate liver development and play important roles in cirrhosis and HCC

In contrast to the markers of hepatocyte-proliferation regenerative processes which were only present in a potentially prognostic subset of HCC, all of our cirrhosis and HCC samples

displayed markers of progenitor cells, including EPCAM, Vimentin, and KRT19, which are also putative markers of liver cancer stem cells [186]. Sell and Leffert (2008) point out that the activation and proliferation of putative liver stem and progenitor cells occurs in response to injury only when hepatocyte proliferation is inhibited [186]. Although these findings are largely based on animal models of acute liver injury, the patterns of gene expression in our data containing 79 cirrhosis and HCC patient samples fit this model.

There is growing acceptance of the idea that cancer stem cells arise from normal stem cells that have lost control of regulatory mechanisms [187]. The specific regulatory mechanisms are still unknown because liver stem cell research is difficult due to the rarity of hepatic stem cell niches and difficulty establishing cell cultures. However, Hedgehog, Notch, and Wnt signaling are the most important regulators of other types of stem cells. Mutations of stem-cell-related genes have been reported in some liver tumors, suggesting that disruption of the self-renewal process in hepatic stem cells may lead to carcinogenesis [188]. Majumdar et al (2012) suggest Wnt activation and loss of TGF-β signaling in hepatic cancer stem cell formation [189], and this appears to be supported by our data: TGFBR3 is lost (<.8 fold) in 36 of 49 (74%) HCC samples compared to 6/30 (20%) cirrhosis; Wnt inhibitor SMAD7 was over-expressed in 27 (90%) cirrhosis but only 15 (31%) of HCC, and was down-regulated in 13 HCC samples (27%). Similarly, Wnt inhibitor SFRP5 was over-expressed (>1.5 fold) in all cirrhosis samples but only 20 HCC (41%). This is also consistent with the HCC subtype identified by Yamashita et al (2008) with over-expression of EPCAM, KRT19, KIT, and activated Wnt-β-catenin signaling [132].

c-MET is a high affinity receptor for HGF and has a well-established association with tumor growth, invasion, and metastasis in HCC [175, 176, 190]. In our data, MET was over-expressed in a subset of tumors with poor 2-year recurrence free survival. MET over-expression was also correlated with loss of expression several tumor suppressors in HCC: KLF6, STAT3, ARID5B, GATA6, and ID2. The correlated expression pattern of these genes in early stage tumors was similar to their expression pattern in late-stage tumors.

In summary, HCV-cirrhosis samples displayed a consistent pattern of markers of proliferating progenitor cells that is consistent with chronic wound healing processes and reminiscent of hepatoblast proliferation in embryonic develompent. Even though some of the intermediate initiators of liver regeneration were mildly up-regulated, the down-stream effectors that would indicate hepatocyte proliferation were not in evidence, and several inhibitors of hepatocyte proliferation were up-regulated, including Wnt inhibitors, TBGF-β, and BMP2. All of our HCC samples displayed the same markers of progenitor cell proliferation, however the tumors lost expression of Wnt inhibitors and up-regulated at least one BMP inhibitor, suggesting that the proliferative controls present in cirrhotic tissues may be overcome in tumors to allow uncontrolled proliferation of progenitor cells and possibly transformation to cancer stem cells, as proposed by Koike et al, 2012 [187]. This pattern was nearly universal in our HCC samples. In addition, there was a sub-set of 13 (45%) of early stage tumors that also displayed a signature of MET over-expression and the loss of EPCAM, E-cadherin, and several tumor suppressors, and this subset had poor survival, as predicted by Yamashita et al, 2009 [132].

6.5 Not all HCC-associated pathways were engaged in our HCV-induced cirrhosis and HCC data

P53 and pRb are cell-cycle control pathways implicated in HCC that are not associated specifically with liver development. The TP53 gene, is best known as a tumor suppressor but also plays roles in embryonic development, particularly differentiation of neurogenic, osteogenic, and myogenic, meatogoietic, and adipogenic cells [191]. Because it appears to be dispensable in liver development [192], it was not included in our developmental gene set, but since it is highly associated with HCV and HCC we did examine its behavior in our samples. In our data TP53 is not differentially expressed in any sample (std. dev = 0.27). Similarly, the pRb pathway controls cell cycle exit and expression is frequently lost during tumorogenesis but not altered in our HCV-cirrhosis or HCV-HCC samples. These patterns were also seen in the Wurmbach dataset of HCV-induced cirrhosis and HCC.

The Hedgehog pathway is another HCC-associated pathway with important roles in body patterning in nearly every region of the body. Hedgehog signaling is activated in response to epithelial-mesenchymal signaling from BMPs and FGFs, and in general promotes the survival of progenitor cells. In adulthood, Hedgehog signals released from activated HSCs have a well-established role in liver wound healing and regeneration [193], and strong HH signals have been found to persist in some sub-groups of HCC as well [194-196], including HCV-induced HCC (24/60 HCV-HCC samples vs. 1/28 cirrhotic livers; Lemmer et al 2006, AACR Abstract #2676) [197]. However, none of the component genes of the pathway were differentially expressed in

cirrhosis or HCC samples, in our samples or in the Wurmbach dataset. Chen et al (2012) have recently shown that sonic hedgehog signaling induces cell migration and invasion through production and activation of MMP2 and MMP9, which specifically degrade type IV collagens that are the major component of hepatic fibrosis and cirrhosis [198]. MMP2 and MMP9 are over-expressed in both cirrhosis and tumor samples in our data, however, they have other regulators including IL-8, TNFα, NF-κB, and SP1, so it may be that any of these are "sufficient but not necessary" to induce MMP2/9 activity.

Some of the other surprising negative results include the lack of involvement of FGF2, and liver maturation factors HGF, Hepatic Nuclear Factors 1a, 1b, and 4a, Onecut 1 and 2, HHEX, and PROX1. FGF2 is a critical early detmininant of hepatic fate that plays important post-natal roles in wound healing to stimulate angiogenesis and may help maintain hepatocyte differentiation and liver homeostasis. The main FGF receptor FGFR2 is up-regulated in cirrhosis and many tumors. In adulthood the liver maintains substantial levels of inactivated FGF2 as stable proteins and low mRNA transcript levels [199], so it is possible that FGF2 is exerting some effect in cirrhosis and cancer by protein activation rather than increased transcription. However, none of the expected downstream targets of FGF signaling were dysregulated, so it is unclear what effect FGF signaling through up-regulated FGFR2 is having. Similarly, Hepatocyte Growth Factor (HGF) is expressed by activated HSC and promotes hepatocellular regeneration, mediates epithelial-mesenchymal interactions, and is associated with the development and progression of several types of cancer including HCC. However HGF levels remained stable throughout cirrhosis and all stages of HCC in both datasets. k-Ras is an oncogene that can be directly activated by the HCV core protein and has been implicated in hepatic carcinogenesis in

mouse models [200, 201], however in human HCC mutations were more important than over-expression in determining outcome, possibly explaining the lack of differential expression.

As critical regulators of liver maturation, we expected that hepatic nuclear factors HNF1A, HNF1B, HNF4A, ONECUT1, and ONECUT2 would be differentially regulated in HCC. However, this may be explained by the fact that they all require co-activation by FOXA1 and FOXA2 for transcription [202]. In our data both FOXA1 and FOXA2 were down-regulated in cirrhosis and HCC. These transcription factors are inhibited by elevated insulin, so this may be a consequence of the insulin resistance and impaired glucose homeostasis caused by Hepatitis C core protein activity [203].

6.6 Summary and conclusion

Our major hypothesis in this study was that the mechanisms of carcinogenesis in HCV-induced HCC may be unique compared to the development of other cancers. Unlike other types of cancer, HCV-HCC arises against a background of decades of response of chronic infection, inflammation, and increasing fibrosis and ultimately cirrhosis. HCV is also unique in that is it is an RNA virus that does not insert into the genome, so tumors might be expected to not be a result of random mutation to the same extent as other cancers. Therefore we focused on the expression of genes that have been used in the life history of the liver, including embryonic development, healing, and regeneration.

Early embryonic development is characterized by proliferation of bi-potential hepatoblasts, while proliferation of differentiated hepatocytes is the dominant growth mechanism in the final maturation stage of development. In the adult liver, these mechanisms are recapitulated in specific instances. Healing of small scale and chronic injury involves induction of niches of undifferentiated proliferative/stem cell populations which reside in the biliary ducts (Canals of Hering) and migrate to the site of injury, proliferate, and differentiate in a manner reminiscent of hepatoblast migration through the STM and subsequent differentiation. In contrast, regeneration of lost liver mass, as occurs after acute toxicity or partial hepatectomy, involves the proliferation of differentiated hepatocytes that more closely resembles the final steps of liver maturation.

We found that cirrhotic and tumor samples universally expressed markers of proliferating progenitor cells and their offspring (newly differentiated hepatocytes). Cirrhosis samples also ubiquitously over-expressed Wnt inhibitors (which controls the rate of progenitor cell proliferation) and several inhibitors of hepatocyte proliferation. Nearly all of the tumor samples continue to express markers of progenitor cells but lose expression of Wnt inhibitors, indicating that in tumors control of progenitor cell proliferation may be lost. In addition, we identified three sub-populations of early tumors. A group of 13 early stage tumors characterized by loss of E-cadherin and EPCAM expression and over-expression of c-MET had 38% 2 year recurrence-free survival. The 16 early tumors that had levels of of these genes similar to that seen in cirrhosis samples had a 95% 2-year recurrence-free survival rate. Patients with good prognosis tended to express higher levels of BMP inhibitors as well. Interestingly, these signatures were not prognostic for late stage tumors. However, there was a

set of 6 tumor supressors that were down-regulated to similar degrees in late stage tumors and those early stage tumors with poor outcomes. We also identified a group of tumors that over-expressed genes associated with β-catenin mutations, however this group was not associated with either good or poor prognosis.

These patterns were identified using a focused analysis of genes that had either shifts in overall mean expression or high variability. We found several high-variability genes with no change in mean expression, but that had correlated expression patterns that led to the identification of the HCC sub-populations. None of these patterns could have been identified from a global gene expression analysis of "highest magnitude mean shift". We feel that we have proven the utility in using a knowledge-driven approach to identify important disease drivers, and that examining high-variability genes for shifts in co-expression has been a fruitful approach to understanding the heterogeneity inherent in the drivers of hepatocellular carcinoma.

LIST OF REFERENCES

LIST OF REFERENCES

1. Machida, K., et al., *Cancer stem cells generated by alcohol, diabetes, and hepatitis C virus.* J Gastroenterol Hepatol, 2012. **27 Suppl 2**: p. 19-22.
2. Levrero, M., *Viral hepatitis and liver cancer: the case of hepatitis C.* Oncogene, 2006. **25**(27): p. 3834-47.
3. Iizuka, N., et al., *Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection.* Lancet, 2003. **361**(9361): p. 923-9.
4. Iizuka, N., et al., *Differential gene expression in distinct virologic types of hepatocellular carcinoma: association with liver cirrhosis.* Oncogene, 2003. **22**(19): p. 3007-14.
5. Llovet, J.M., et al., *A molecular signature to discriminate dysplastic nodules from early hepatocellular carcinoma in HCV cirrhosis.* Gastroenterology, 2006. **131**(6): p. 1758-67.
6. Hoshida, Y., et al., *Molecular classification and novel targets in hepatocellular carcinoma: recent advancements.* Semin Liver Dis, 2010. **30**(1): p. 35-51.
7. Woo, H.G., et al., *Exploring genomic profiles of hepatocellular carcinoma.* Mol Carcinog, 2011. **50**(4): p. 235-43.
8. Boyault, S., et al., *Transcriptome classification of HCC is related to gene alterations and to new therapeutic targets.* Hepatology, 2007. **45**(1): p. 42-52.
9. Chiang, D.Y., et al., *Focal gains of VEGFA and molecular classification of hepatocellular carcinoma.* Cancer Res, 2008. **68**(16): p. 6779-88.
10. Hoshida, Y., et al., *Integrative transcriptome analysis reveals common molecular subclasses of human hepatocellular carcinoma.* Cancer Res, 2009. **69**(18): p. 7385-92.
11. Lee, D., et al., *Discovery of differentially expressed genes related to histological subtype of hepatocellular carcinoma.* Biotechnol Prog, 2003. **19**(3): p. 1011-5.
12. van Malenstein, H., et al., *A seven-gene set associated with chronic hypoxia of prognostic importance in hepatocellular carcinoma.* Clin Cancer Res, 2010. **16**(16): p. 4278-88.
13. Budhu, A.S., et al., *The molecular signature of metastases of human hepatocellular carcinoma.* Oncology, 2005. **69 Suppl 1**: p. 23-7.
14. Korn, E.L., et al., *Controlling the number of false discoveries: application to high-dimensional genomic data.* Journal of Statistical Planning and Inference, 2004. **124**(2): p. 379-398.
15. Owen, A.B., *Variance of the number of false discoveries.* Journal of the Royal Statistical Society, 2005. **67**(3): p. 411-426.

16. Schwartzman, A. and X. Lin, *The effect of correlation in false discovery rate estimation.* Biometrika, 2011. **98**(1): p. 199-214.
17. Hu, Z. and G.R. Willsky, *Utilization of two sample t-test statistics from redundant probe sets to evaluate different probe set algorithms in GeneChip studies.* BMC Bioinformatics, 2006. **7**: p. 12.
18. Merlo, L.M., et al., *Cancer as an evolutionary and ecological process.* Nat Rev Cancer, 2006. **6**(12): p. 924-35.
19. Sung E Choe*†, M.B., AlanMMichelson*†‡, GeorgeMChurch* and Marc SHalfon, *Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset.* 2005.
20. Jung, K., T. Friede, and T. Beissbarth, *Reporting FDR analogous confidence intervals for the log fold change of differentially expressed genes.* BMC Bioinformatics, 2011. **12**: p. 288.
21. Chen, S.L. and T.R. Morgan, *The natural history of hepatitis C virus (HCV) infection.* Int J Med Sci, 2006. **3**(2): p. 47-52.
22. Block, T.M., et al., *Molecular viral oncology of hepatocellular carcinoma.* Oncogene, 2003. **22**(33): p. 5093-107.
23. Shimotohno, K., *Hepatitis C virus and its pathogenesis.* Semin Cancer Biol, 2000. **10**(3): p. 233-40.
24. Bartosch, B., et al., *Hepatitis C virus-induced hepatocarcinogenesis.* J Hepatol, 2009. **51**(4): p. 810-20.
25. Koike, K., *Pathogenesis of HCV-associated HCC: Dual-pass carcinogenesis through activation of oxidative stress and intracellular signaling.* Hepatol Res, 2007. **37 Suppl 2**: p. S115-20.
26. Aiba, T., et al., *[C-type hepatitis in spontaneous intracerebral hemorrhage].* No To Shinkei, 1996. **48**(12): p. 1116-9.
27. Sanyal, A.J., S.K. Yoon, and R. Lencioni, *The etiology of hepatocellular carcinoma and consequences for treatment.* Oncologist, 2010. **15 Suppl 4**: p. 14-22.
28. Bataller, R. and D.A. Brenner, *Liver fibrosis.* J Clin Invest, 2005. **115**(2): p. 209-18.
29. Mas, V.R., et al., *Hepatocellular carcinoma in HCV-infected patients awaiting liver transplantation: genes involved in tumor progression.* Liver Transpl, 2004. **10**(5): p. 607-20.
30. Wurmbach, E., et al., *Genome-wide molecular profiles of HCV-induced dysplasia and hepatocellular carcinoma.* Hepatology, 2007. **45**(4): p. 938-47.
31. Cao, S., et al., *Neuropilin-1 promotes cirrhosis of the rodent and human liver by enhancing PDGF/TGF-beta signaling in hepatic stellate cells.* J Clin Invest, 2010. **120**(7): p. 2379-94.
32. Nitta, T., et al., *Murine cirrhosis induces hepatocyte epithelial mesenchymal transition and alterations in survival signaling pathways.* Hepatology, 2008. **48**(3): p. 909-19.
33. Black, D., et al., *Primary cirrhotic hepatocytes resist TGFbeta-induced apoptosis through a ROS-dependent mechanism.* J Hepatol, 2004. **40**(6): p. 942-51.
34. Villanueva, A., et al., *Genomics and signaling pathways in hepatocellular carcinoma.* Semin Liver Dis, 2007. **27**(1): p. 55-76.

35. Saffroy, R., et al., *New perspectives and strategy research biomarkers for hepatocellular carcinoma.* Clin Chem Lab Med, 2007. **45**(9): p. 1169-79.
36. Tischoff, I. and A. Tannapfe, *DNA methylation in hepatocellular carcinoma.* World J Gastroenterol, 2008. **14**(11): p. 1741-8.
37. Imbeaud, S., Y. Ladeiro, and J. Zucman-Rossi, *Identification of novel oncogenes and tumor suppressors in hepatocellular carcinoma.* Semin Liver Dis, 2010. **30**(1): p. 75-86.
38. Feng, G.S., *Conflicting roles of molecules in hepatocarcinogenesis: paradigm or paradox.* Cancer Cell, 2012. **21**(2): p. 150-4.
39. Yuzugullu, H., et al., *Canonical Wnt signaling is antagonized by noncanonical Wnt5a in hepatocellular carcinoma cells.* Mol Cancer, 2009. **8**: p. 90.
40. Geng, M., et al., *Loss of Wnt5a and Ror2 protein in hepatocellular carcinoma associated with poor prognosis.* World J Gastroenterol, 2012. **18**(12): p. 1328-38.
41. Behari, J., *The Wnt/beta-catenin signaling pathway in liver biology and disease.* Expert Rev Gastroenterol Hepatol, 2010. **4**(6): p. 745-56.
42. van Amerongen, R., et al., *Wnt5a can both activate and repress Wnt/beta-catenin signaling during mouse embryonic development.* Dev Biol, 2012. **369**(1): p. 101-14.
43. Liu, J., et al., *Enhancement of canonical Wnt/beta-catenin signaling activity by HCV core protein promotes cell growth of hepatocellular carcinoma cells.* PLoS One, 2011. **6**(11): p. e27496.
44. Gao, W., Mitchell Hoa, *The Role of Glypican-3 in Regulating Wnt in Hepatocellular Carcinomas.* 2011.
45. Mishra, L., et al., *Liver stem cells and hepatocellular carcinoma.* Hepatology, 2009. **49**(1): p. 318-29.
46. Bralet, M.P., V. Pichard, and N. Ferry, *Demonstration of direct lineage between hepatocytes and hepatocellular carcinoma in diethylnitrosamine-treated rats.* Hepatology, 2002. **36**(3): p. 623-30.
47. Chang, Q., et al., *Sustained JNK1 activation is associated with altered histone H3 methylations in human liver cancer.* J Hepatol, 2009. **50**(2): p. 323-33.
48. Sun, W., et al., *Gankyrin-mediated dedifferentiation facilitates the tumorigenicity of rat hepatocytes and hepatoma cells.* Hepatology, 2011. **54**(4): p. 1259-72.
49. Mas, V.R., et al., *Genes involved in viral carcinogenesis and tumor initiation in hepatitis C virus-induced hepatocellular carcinoma.* Mol Med, 2009. **15**(3-4): p. 85-94.
50. Conti, A., Simona Scala, Marina Romano, Antonella Izzo, Floriana Fabbrini, Floriana Della Ragione, Maurizio D'Esposito, Lucio Nitsch, Fulvio Calise and Antonio Faiella, *Gene Expression Profile in Liver Transplantation and the Influence of Gene Dysregulation Occurring in Deceased Donor Grafts.* The Open Surgery Journal, 2011. **5**: p. 1-12.
51. Zorn, A.M., *Liver development*, in *StemBook*. 2008: Cambridge (MA).
52. Lade, A.G. and S.P. Monga, *Beta-catenin signaling in hepatic development and progenitors: which way does the WNT blow?* Dev Dyn, 2011. **240**(3): p. 486-500.
53. Yanai, M., et al., *FGF signaling segregates biliary cell-lineage from chick hepatoblasts cooperatively with BMP4 and ECM components in vitro.* Dev Dyn, 2008. **237**(5): p. 1268-83.

54. McLin, V.A., S.A. Rankin, and A.M. Zorn, *Repression of Wnt/beta-catenin signaling in the anterior endoderm is essential for liver and pancreas development.* Development, 2007. **134**(12): p. 2207-17.
55. Ober, E.A., et al., *Mesodermal Wnt2b signalling positively regulates liver specification.* Nature, 2006. **442**(7103): p. 688-91.
56. Sugimura, R. and L. Li, *Noncanonical Wnt signaling in vertebrate development, stem cells, and diseases.* Birth Defects Res C Embryo Today, 2010. **90**(4): p. 243-56.
57. Zeng, G., et al., *Wnt'er in liver: expression of Wnt and frizzled genes in mouse.* Hepatology, 2007. **45**(1): p. 195-204.
58. Niu, X., H. Shi, and J. Peng, *The role of mesodermal signals during liver organogenesis in zebrafish.* Sci China Life Sci, 2010. **53**(4): p. 455-61.
59. Sanchez, A. and I. Fabregat, *Growth factor- and cytokine-driven pathways governing liver stemness and differentiation.* World J Gastroenterol, 2010. **16**(41): p. 5148-61.
60. Lemaigre, F.P., *Mechanisms of liver development: concepts for understanding liver disorders and design of novel therapies.* Gastroenterology, 2009. **137**(1): p. 62-79.
61. Tanaka, M., et al., *Mouse hepatoblasts at distinct developmental stages are characterized by expression of EpCAM and DLK1: drastic change of EpCAM expression during liver development.* Mech Dev, 2009. **126**(8-9): p. 665-76.
62. North, T.E. and W. Goessling, *Endoderm specification, liver development, and regeneration.* Methods Cell Biol, 2011. **101**: p. 205-23.
63. Margagliotti, S., et al., *Role of metalloproteinases at the onset of liver development.* Dev Growth Differ, 2008. **50**(5): p. 331-8.
64. Sandeep S. Sekhon, X.T., Amanda Micsenyi, William C. Bowen, and Satdarshan P.S. Monga, *Fibroblast Growth Factor Enriches the Embryonic Liver Cultures for Hepatic Progenitors.* American Journal of Pathology, 2004. **164**(6).
65. Tatsumi, N., et al., *Neurturin-GFRalpha2 signaling controls liver bud migration along the ductus venosus in the chick embryo.* Dev Biol, 2007. **307**(1): p. 14-28.
66. Breitwieser, W., et al., *Feedback regulation of p38 activity via ATF2 is essential for survival of embryonic liver cells.* Genes Dev, 2007. **21**(16): p. 2069-82.
67. Suzuki, T., et al., *Crucial role of the small GTPase ARF6 in hepatic cord formation during liver development.* Mol Cell Biol, 2006. **26**(16): p. 6149-56.
68. Kamiya, A., et al., *Prospero-related homeobox 1 and liver receptor homolog 1 coordinately regulate long-term proliferation of murine fetal hepatoblasts.* Hepatology, 2008. **48**(1): p. 252-64.
69. Tam, W.L., et al., *T-cell factor 3 regulates embryonic stem cell pluripotency and self-renewal by the transcriptional control of multiple lineage pathways.* Stem Cells, 2008. **26**(8): p. 2019-31.
70. Lee, J.S., et al., *Transcriptional ontogeny of the developing liver.* BMC Genomics, 2012. **13**: p. 33.
71. Hirose, Y., T. Itoh, and A. Miyajima, *Hedgehog signal activation coordinates proliferation and differentiation of fetal liver progenitor cells.* Exp Cell Res, 2009. **315**(15): p. 2648-57.
72. Clotman, F., et al., *Control of liver cell fate decision by a gradient of TGF beta signaling modulated by Onecut transcription factors.* Genes Dev, 2005. **19**(16): p. 1849-54.

73. Zong, Y., et al., *Notch signaling controls liver development by regulating biliary differentiation.* Development, 2009. **136**(10): p. 1727-39.
74. Battle, M.A., et al., *Hepatocyte nuclear factor 4alpha orchestrates expression of cell adhesion proteins during the epithelial transformation of the developing liver.* Proc Natl Acad Sci U S A, 2006. **103**(22): p. 8419-24.
75. Peterson, M.L., C. Ma, and B.T. Spear, *Zhx2 and Zbtb20: novel regulators of postnatal alpha-fetoprotein repression and their potential role in gene reactivation during liver cancer.* Semin Cancer Biol, 2011. **21**(1): p. 21-7.
76. Colletti, M., et al., *Convergence of Wnt signaling on the HNF4alpha-driven transcription in controlling liver zonation.* Gastroenterology, 2009. **137**(2): p. 660-72.
77. Ebrahimkhani, M.R., A.M. Elsharkawy, and D.A. Mann, *Wound healing and local neuroendocrine regulation in the injured liver.* Expert Rev Mol Med, 2008. **10**: p. e11.
78. Papadimas, G.K., et al., *The emerging role of serotonin in liver regeneration.* Swiss Med Wkly, 2012. **142**: p. w13548.
79. Choi, S.S., et al., *The role of Hedgehog signaling in fibrogenic liver repair.* Int J Biochem Cell Biol, 2011. **43**(2): p. 238-44.
80. Oumi, N., et al., *A crucial role of bone morphogenetic protein signaling in the wound healing response in acute liver injury induced by carbon tetrachloride.* Int J Hepatol, 2012. **2012**: p. 476820.
81. Gieling, R.G., et al., *The c-Rel subunit of nuclear factor-kappaB regulates murine liver inflammation, wound-healing, and hepatocyte proliferation.* Hepatology, 2010. **51**(3): p. 922-31.
82. David Rychtrmoc1, Antonín Libra2, Martin Bunček2, Tomáš Garnol1, Zuzana Červinková1, *STUDYING LIVER REGENERATION BY MEANS OF MOLECULAR BIOLOGY: HOW FAR WE ARE IN INTERPRETING THE FINDINGS?* 2009.
83. Dussmann, P., et al., *Live in vivo imaging of Egr-1 promoter activity during neonatal development, liver regeneration and wound healing.* BMC Dev Biol, 2011. **11**: p. 28.
84. Polimeno, L., et al., *Decreased expression of the augmenter of liver regeneration results in increased apoptosis and oxidative damage in human-derived glioma cells.* Cell Death Dis, 2012. **3**: p. e289.
85. Cheng-Maw Ho, P.-H.L., Yeun-Tyng Lai, Rey-Heng Hu, Ming-Chih Ho, Yao-Ming Wu, *Gene Expression Profiles in Living Donors Immediately After Partial Hepatectomy— The Initial Response of Liver Regeneration.* 2007.
86. Gry, M., et al., *Correlations between RNA and protein expression profiles in 23 human cell lines.* BMC Genomics, 2009. **10**: p. 365.
87. Waters, K.M., J.G. Pounds, and B.D. Thrall, *Data merging for integrated microarray and proteomic analysis.* Brief Funct Genomic Proteomic, 2006. **5**(4): p. 261-72.
88. Harr, B. and C. Schlotterer, *Comparison of algorithms for the analysis of Affymetrix microarray data as evaluated by co-expression of genes in known operons.* Nucleic Acids Res, 2006. **34**(2): p. e8.
89. Reimers, M., *Statistical analysis of microarray data.* Addict Biol, 2005. **10**(1): p. 23-35.
90. Waxman, S. and E. Wurmbach, *De-regulation of common housekeeping genes in hepatocellular carcinoma.* BMC Genomics, 2007. **8**: p. 243.

91. Reimers, M. and J.N. Weinstein, *Quality assessment of microarrays: visualization of spatial artifacts and quantitation of regional biases.* BMC Bioinformatics, 2005. **6**: p. 166.
92. Ritchie, M.E., et al., *A comparison of background correction methods for two-colour microarrays.* Bioinformatics, 2007. **23**(20): p. 2700-7.
93. Draghici, S., et al., *Noise sampling method: an ANOVA approach allowing robust selection of differentially regulated genes measured by DNA microarrays.* Bioinformatics, 2003. **19**(11): p. 1348-59.
94. Irizarry, R.A., et al., *Exploration, normalization, and summaries of high density oligonucleotide array probe level data.* Biostatistics, 2003. **4**(2): p. 249-64.
95. Do, J.H. and D.K. Choi, *Normalization of microarray data: single-labeled and dual-labeled arrays.* Mol Cells, 2006. **22**(3): p. 254-61.
96. Kadota, K., Y. Nakai, and K. Shimizu, *A weighted average difference method for detecting differentially expressed genes from microarray data.* Algorithms Mol Biol, 2008. **3**: p. 8.
97. Kadota, K. and K. Shimizu, *Evaluating methods for ranking differentially expressed genes applied to microArray quality control data.* BMC Bioinformatics, 2011. **12**: p. 227.
98. Huber, W., et al., *Variance stabilization applied to microarray data calibration and to the quantification of differential expression.* Bioinformatics, 2002. **18 Suppl 1**: p. S96-104.
99. Wu, W., et al., *Evaluation of normalization methods for cDNA microarray data by k-NN classification.* BMC Bioinformatics, 2005. **6**: p. 191.
100. Wang, D., et al., *Extensive increase of microarray signals in cancers calls for novel normalization assumptions.* Comput Biol Chem, 2011. **35**(3): p. 126-30.
101. Millenaar, F.F., et al., *How to decide? Different methods of calculating gene expression from short oligonucleotide array data will give different results.* BMC Bioinformatics, 2006. **7**: p. 137.
102. Park, T., et al., *Evaluation of normalization methods for microarray data.* BMC Bioinformatics, 2003. **4**: p. 33.
103. Fujita, A., et al., *Evaluating different methods of microarray data normalization.* BMC Bioinformatics, 2006. **7**: p. 469.
104. Ning, K., et al., *A post-processing method for optimizing synthesis strategy for oligonucleotide microarrays.* Nucleic Acids Res, 2005. **33**(17): p. e144.
105. Benjamini, Y. and T.P. Speed, *Summarizing and correcting the GC content bias in high-throughput sequencing.* Nucleic Acids Res, 2012. **40**(10): p. e72.
106. Ghandi, M. and M.A. Beer, *Group normalization for genomic data.* PLoS One, 2012. **7**(8): p. e38695.
107. van Iterson, M., et al., *A novel and fast normalization method for high-density arrays.* Stat Appl Genet Mol Biol, 2012. **11**(4).
108. Schmidt, M.T., et al., *Impact of DNA microarray data transformation on gene expression analysis - comparison of two normalization methods.* Acta Biochim Pol, 2011. **58**(4): p. 573-80.
109. Giorgi, F.M., et al., *Algorithm-driven artifacts in median Polish summarization of microarray data.* BMC Bioinformatics, 2010. **11**: p. 553.
110. Robert A Fisher, T.P.M., Ann S Fulcher, Daniel Maluf, John A Clay, Luke G Wolfe, Sherfield Dawson III, Adrian Cotterell, R Todd Stravitz, Velimir A Luketic, Mitchell

Shiffman, Richard K Sterling and Marc P Posnera, *Hepatocellular carcinoma: strategy for optimizing surgical resection, transplantation and palliation.* 2002.
111. Hishiki, T., et al., *BodyMap: a human and mouse gene expression database.* Nucleic Acids Res, 2000. **28**(1): p. 136-8.
112. Miller, C.J., *simpleaffy: Very simple high level analysis of Affymetrix data. R package version 2.28.0.*
113. Sandberg, R. and O. Larsson, *Improved precision and accuracy for microarrays using updated probe set definitions.* BMC Bioinformatics, 2007. **8**: p. 48.
114. Fisher, R.A., *Statistical Methods for Research Workers*, 1925, Oliver and Boyd: Edinburgh.
115. Jolliffe, I.T., *Principal component analysis*. 2nd ed. Springer series in statistics. 2002, New York: Springer. xxix, 487 p.
116. Donato, F., et al., *Hepatitis B and C virus infection, alcohol drinking, and hepatocellular carcinoma: a case-control study in Italy. Brescia HCC Study.* Hepatology, 1997. **26**(3): p. 579-84.
117. Donato, F., et al., *Alcohol and hepatocellular carcinoma: the effect of lifetime intake and hepatitis virus infections in men and women.* Am J Epidemiol, 2002. **155**(4): p. 323-31.
118. Szabo, G., et al., *Alcohol and hepatitis C virus--interactions in immune dysfunctions and liver damage.* Alcohol Clin Exp Res, 2010. **34**(10): p. 1675-86.
119. Marrero, J.A., et al., *Alcohol, tobacco and obesity are synergistic risk factors for hepatocellular carcinoma.* J Hepatol, 2005. **42**(2): p. 218-24.
120. Wang, C.S., et al., *The impact of type 2 diabetes on the development of hepatocellular carcinoma in different viral hepatitis statuses.* Cancer Epidemiol Biomarkers Prev, 2009. **18**(7): p. 2054-60.
121. Keller, M.P. and A.D. Attie, *Physiological insights gained from gene expression analysis in obesity and diabetes.* Annu Rev Nutr, 2010. **30**: p. 341-64.
122. Buechler, C. and A. Schaffler, *Does global gene expression analysis in type 2 diabetes provide an opportunity to identify highly promising drug targets?* Endocr Metab Immune Disord Drug Targets, 2007. **7**(4): p. 250-8.
123. Mas, V.R., et al., *Molecular mechanisms involved in the interaction effects of alcohol and hepatitis C virus in liver cirrhosis.* Mol Med, 2010. **16**(7-8): p. 287-97.
124. Hooten, N.N.A., Kotb, Gorospe, Myriam, Ejiogu, Ngozi, Zonderman, Alan B., Evans, Michele K., *microRNA expression patterns reveal differential expression of target genes with age.* PLOS ONE, 2010. **5**(5): p. e10724.
125. Ho, P.-H.L., Yeun-Tyng Lai, Rey-Heng Hu, Ming-Chih Ho, Yao-Ming Wu, *Gene Expression Profiles in Living Donors Immediately After Partial Hepatectomy— The Initial Response of Liver Regeneration.* 2007.
126. Yoon, S.M., et al., *Epithelial cell adhesion molecule (EpCAM) marks hepatocytes newly derived from stem/progenitor cells in humans.* Hepatology, 2011. **53**(3): p. 964-73.
127. Oishi, N. and X.W. Wang, *Novel therapeutic strategies for targeting liver cancer stem cells.* Int J Biol Sci, 2011. **7**(5): p. 517-35.
128. Chang, H.Y., et al., *Gene expression signature of fibroblast serum response predicts human cancer progression: similarities between tumors and wounds.* PLoS Biol, 2004. **2**(2): p. E7.

129. Riss, J., et al., *Cancers as wounds that do not heal: differences and similarities between renal regeneration/repair and renal cell carcinoma.* Cancer Res, 2006. **66**(14): p. 7216-24.
130. Dauer, D.J., et al., *Stat3 regulates genes common to both wound healing and cancer.* Oncogene, 2005. **24**(21): p. 3397-408.
131. Kim, J.W., et al., *Cancer-associated molecular signature in the tissue samples of patients with cirrhosis.* Hepatology, 2004. **39**(2): p. 518-27.
132. Yamashita, T., et al., *EpCAM-positive hepatocellular carcinoma cells are tumor-initiating cells with stem/progenitor cell features.* Gastroenterology, 2009. **136**(3): p. 1012-24.
133. Kumar, M., X. Zhao, and X.W. Wang, *Molecular carcinogenesis of hepatocellular carcinoma and intrahepatic cholangiocarcinoma: one step closer to personalized medicine?* Cell Biosci, 2011. **1**(1): p. 5.
134. Wang, C., et al., *Hepatitis B virus X (HBx) induces tumorigenicity of hepatic progenitor cells in 3,5-diethoxycarbonyl-1,4-dihydrocollidine-treated HBx transgenic mice.* Hepatology, 2012. **55**(1): p. 108-20.
135. Tanaka, M., et al., *Liver stem/progenitor cells: their characteristics and regulatory mechanisms.* J Biochem, 2011. **149**(3): p. 231-9.
136. Pritchett, J., et al., *Osteopontin is a novel downstream target of SOX9 with diagnostic implications for progression of liver fibrosis in humans.* Hepatology, 2012.
137. Guo, X., et al., *Expression features of SOX9 associate with tumor progression and poor prognosis of hepatocellular carcinoma.* Diagn Pathol, 2012. **7**: p. 44.
138. Dorrell, C., et al., *Prospective isolation of a bipotential clonogenic liver progenitor cell in adult mice.* Genes Dev, 2011. **25**(11): p. 1193-203.
139. Furuyama, K., et al., *Continuous cell supply from a Sox9-expressing progenitor zone in adult liver, exocrine pancreas and intestine.* Nat Genet, 2011. **43**(1): p. 34-41.
140. Friedman, J.R. and K.H. Kaestner, *On the origin of the liver.* J Clin Invest, 2011. **121**(12): p. 4630-3.
141. Amann, T., et al., *Reduced expression of fibroblast growth factor receptor 2IIIb in hepatocellular carcinoma induces a more aggressive growth.* Am J Pathol, 2010. **176**(3): p. 1433-42.
142. Haugsten, E.M., et al., *Roles of fibroblast growth factor receptors in carcinogenesis.* Mol Cancer Res, 2010. **8**(11): p. 1439-52.
143. Haga, H., et al., *Enhanced expression of fibroblast growth factor 2 in bone marrow cells and its potential role in the differentiation of hepatic epithelial stem-like cells into the hepatocyte lineage.* Cell Tissue Res, 2011. **343**(2): p. 371-8.
144. Lichtinghagen, R., Dirk Michels, Christian I Haberkorn, Burkhard Arndt, Matthias Bahr, Peer Flemming, Michael P Manns, Klaus H.W Boeker, *Matrix metalloproteinase (MMP)-2, MMP-7, and tissue inhibitor of metalloproteinase-1 are closely related to the fibroproliferative process in the liver during chronic hepatitis C.* Journal of Hepatolgoy, 2001. **34**(2): p. 239-247.
145. Oka, T., et al., *Overexpression of beta3/gamma2 chains of laminin-5 and MMP7 in biliary cancer.* World J Gastroenterol, 2009. **15**(31): p. 3865-73.
146. Ishii, Y., et al., *A study on angiogenesis-related matrix metalloproteinase networks in primary hepatocellular carcinoma.* J Exp Clin Cancer Res, 2003. **22**(3): p. 461-70.

147. Bu, W., X. Huang, and Z. Tang, *[The role of MMP-2 in the invasion and metastasis of hepatocellular carcinoma (HCC)].* Zhonghua Yi Xue Za Zhi, 1997. **77**(9): p. 661-4.
148. Gianelli, U., et al., *Lymphomas of the bone: a pathological and clinical study of 54 cases.* Int J Surg Pathol, 2002. **10**(4): p. 257-66.
149. Giannelli, G., et al., *Clinical role of MMP-2/TIMP-2 imbalance in hepatocellular carcinoma.* Int J Cancer, 2002. **97**(4): p. 425-31.
150. Wells, R.G., *Textbook of Hepatology. Chapter 2.4.3 Function and metabolism of collagen and other extracellular matrix proteins*, J. Rodes, Behnamou, Jean-Pierre, Blei, Andres, Reichen, Jurg, and Rizzetto, Mario, Editor 2007, GastroHep.com.
151. Han, Y.P., et al., *Essential role of matrix metalloproteinases in interleukin-1-induced myofibroblastic activation of hepatic stellate cell in collagen.* J Biol Chem, 2004. **279**(6): p. 4820-8.
152. Sato, Y., et al., *Macrochimerism of donor type CD56+ CD3+ T cells in donor specific transfusion via portal vein following living related donor liver transplantation.* Hepatogastroenterology, 2003. **50**(54): p. 2161-5.
153. Dalakas, E., et al., *Hematopoietic stem cell trafficking in liver injury.* FASEB J, 2005. **19**(10): p. 1225-31.
154. Tong-Jing, X., et al., *Mechanism and Efficacy of Mobilization of Granulocyte Colony-Stimulating Factor in the Treatment of Chronic Hepatic Failure.* Hepatogastroenterology, 2012. **60**(121).
155. Thieringer, F.R., et al., *Liver-specific overexpression of matrix metalloproteinase 9 (MMP-9) in transgenic mice accelerates development of hepatocellular carcinoma.* Mol Carcinog, 2012. **51**(6): p. 439-48.
156. Ng, K.T., et al., *Overexpression of matrix metalloproteinase-12 (MMP-12) correlates with poor prognosis of hepatocellular carcinoma.* Eur J Cancer, 2011. **47**(15): p. 2299-305.
157. Zhou, D., et al., *Mst1 and Mst2 maintain hepatocyte quiescence and suppress hepatocellular carcinoma development through inactivation of the Yap1 oncogene.* Cancer Cell, 2009. **16**(5): p. 425-38.
158. Avruch, J., et al., *Mst1/2 signalling to Yap: gatekeeper for liver size and tumour development.* Br J Cancer, 2011. **104**(1): p. 24-32.
159. Jarnicki, A., T. Putoczki, and M. Ernst, *Stat3: linking inflammation to epithelial cancer - more than a "gut" feeling?* Cell Div, 2010. **5**: p. 14.
160. Mantel, C., et al., *Mouse hematopoietic cell-targeted STAT3 deletion: stem/progenitor cell defects, mitochondrial dysfunction, ROS overproduction, and a rapid aging-like phenotype.* Blood, 2012. **120**(13): p. 2589-99.
161. Wang, H., et al., *Signal transducer and activator of transcription 3 in liver diseases: a novel therapeutic target.* Int J Biol Sci, 2011. **7**(5): p. 536-50.
162. Lee, J.S., et al., *A novel prognostic subtype of human hepatocellular carcinoma derived from hepatic progenitor cells.* Nat Med, 2006. **12**(4): p. 410-6.
163. Jia, Y.P., et al., *Postoperative complications in patients with portal vein thrombosis after liver transplantation: evaluation with Doppler ultrasonography.* World J Gastroenterol, 2007. **13**(34): p. 4636-40.
164. Tang, X.H. and L.J. Gudas, *Retinoids, retinoic acid receptors, and cancer.* Annu Rev Pathol, 2011. **6**: p. 345-64.

165. Chorley, B.N., et al., *Identification of novel NRF2-regulated genes by ChIP-Seq: influence on retinoid X receptor alpha.* Nucleic Acids Res, 2012. **40**(15): p. 7416-29.
166. Carvajal-Yepes, M., et al., *Hepatitis C virus impairs the induction of cytoprotective Nrf2 target genes by delocalization of small Maf proteins.* J Biol Chem, 2011. **286**(11): p. 8941-51.
167. Watson, M., *CoXpress: differential co-expression in gene expression data.* BMC Bioinformatics, 2006. **7**: p. 509.
168. Choi, Y. and C. Kendziorski, *Statistical methods for gene set co-expression analysis.* Bioinformatics, 2009. **25**(21): p. 2780-6.
169. Zhu, N.L., J. Wang, and H. Tsukamoto, *The Necdin-Wnt pathway causes epigenetic peroxisome proliferator-activated receptor gamma repression in hepatic stellate cells.* J Biol Chem, 2010. **285**(40): p. 30463-71.
170. Renard, C.A., et al., *Tbx3 is a downstream target of the Wnt/beta-catenin pathway and a critical mediator of beta-catenin survival functions in liver cancer.* Cancer Res, 2007. **67**(3): p. 901-10.
171. Bommer, G.T., et al., *IRS1 regulation by Wnt/beta-catenin signaling and varied contribution of IRS1 to the neoplastic phenotype.* J Biol Chem, 2010. **285**(3): p. 1928-38.
172. Wagner, R.T., et al., *Canonical Wnt/beta-catenin regulation of liver receptor homolog-1 mediates pluripotency gene expression.* Stem Cells, 2010. **28**(10): p. 1794-804.
173. Imajo, M., et al., *A molecular mechanism that links Hippo signalling to the inhibition of Wnt/beta-catenin signalling.* EMBO J, 2012. **31**(5): p. 1109-22.
174. Pecina-Slaus, N., *Tumor suppressor gene E-cadherin and its role in normal and malignant cells.* Cancer Cell Int, 2003. **3**(1): p. 17.
175. Kondo, S., et al., *Clinical impact of c-Met expression and its gene amplification in hepatocellular carcinoma.* Int J Clin Oncol, 2012.
176. Irena Ivanovska1, Chunsheng Zhang1,4, Angela M. Liu2,3, Kwong F. Wong3, Nikki P. Lee2, Patrick, U.P. Lewis1, Dimple Bansal4, Carolyn Buser5, Martin Scott4, Mao Mao1¤a, Ronnie T. P. Poon2,, and M.A.C.b. Sheung Tat Fan2, John M. Luk2,3*, Hongyue Dai, *Gene Signatures Derived from a c-MET-Driven Liver Cancer Mouse Model Predict Survival of Patients with Hepatocellular Carcinoma.* 2011.
177. Lee, C.F., et al., *Genomic-wide analysis of lymphatic metastasis-associated genes in human hepatocellular carcinoma.* World J Gastroenterol, 2009. **15**(3): p. 356-65.
178. Arellano-Garcia, M.E., et al., *Identification of tetranectin as a potential biomarker for metastatic oral cancer.* Int J Mol Sci, 2010. **11**(9): p. 3106-21.
179. Begum, F.D., et al., *Serum tetranectin is a significant prognostic marker in ovarian cancer patients.* Acta Obstet Gynecol Scand, 2010. **89**(2): p. 190-8.
180. Brunner, A., et al., *Expression and prognostic significance of Tetranectin in invasive and non-invasive bladder cancer.* Virchows Arch, 2007. **450**(6): p. 659-64.
181. Hogdall, C.K., et al., *Plasma tetranectin and colorectal cancer.* Eur J Cancer, 1995. **31A**(6): p. 888-94.
182. Hermann, M., et al., *In the search of potential human islet stem cells: is tetranectin showing us the way?* Transplant Proc, 2005. **37**(2): p. 1322-5.
183. Hermann, M., R. Margreiter, and P. Hengster, *Molecular and cellular key players in human islet transplantation.* J Cell Mol Med, 2007. **11**(3): p. 398-415.

184. Fabregat, I., C. Roncero, and M. Fernandez, *Survival and apoptosis: a dysregulated balance in liver cancer.* Liver Int, 2007. **27**(2): p. 155-62.
185. Li, T., et al., *Comparison of gene expression in hepatocellular carcinoma, liver development, and liver regeneration.* Mol Genet Genomics, 2010. **283**(5): p. 485-92.
186. Sell, S. and H.L. Leffert, *Liver cancer stem cells.* J Clin Oncol, 2008. **26**(17): p. 2800-5.
187. Koike, H. and H. Taniguchi, *Characteristics of hepatic stem/progenitor cells in the fetal and adult liver.* J Hepatobiliary Pancreat Sci, 2012.
188. Ma, S., et al., *Identification and characterization of tumorigenic liver cancer stem/progenitor cells.* Gastroenterology, 2007. **132**(7): p. 2542-56.
189. Majumdar, A., et al., *Hepatic stem cells and transforming growth factor beta in hepatocellular carcinoma.* Nat Rev Gastroenterol Hepatol, 2012. **9**(9): p. 530-8.
190. Patil, M.A., et al., *Role of cyclin D1 as a mediator of c-Met- and beta-catenin-induced hepatocarcinogenesis.* Cancer Res, 2009. **69**(1): p. 253-61.
191. Molchadsky, A., et al., *p53 is balancing development, differentiation and de-differentiation to assure cancer prevention.* Carcinogenesis, 2010. **31**(9): p. 1501-8.
192. Dumble, M.L., et al., *Hepatoblast-like cells populate the adult p53 knockout mouse liver: evidence for a hyperproliferative maturation-arrested stem cell compartment.* Cell Growth Differ, 2001. **12**(5): p. 223-31.
193. Omenetti, A. and A.M. Diehl, *The adventures of sonic hedgehog in development and repair. II. Sonic hedgehog and liver development, inflammation, and cancer.* Am J Physiol Gastrointest Liver Physiol, 2008. **294**(3): p. G595-8.
194. Omenetti, A., et al., *Hedgehog signaling in the liver.* J Hepatol, 2011. **54**(2): p. 366-73.
195. Che, L., et al., *[Expression of genes related to Sonic Hedgehog signaling in human hepatocellular carcinomas].* Beijing Da Xue Xue Bao, 2008. **40**(6): p. 616-23.
196. Cheng, W.T., et al., *Role of Hedgehog signaling pathway in proliferation and invasiveness of hepatocellular carcinoma cells.* Int J Oncol, 2009. **34**(3): p. 829-36.
197. Pereira Tde, A., et al., *Viral factors induce Hedgehog pathway activation in humans with viral hepatitis, cirrhosis, and hepatocellular carcinoma.* Lab Invest, 2010. **90**(12): p. 1690-703.
198. Chen, J.S., et al., *Sonic hedgehog signaling pathway induces cell migration and invasion through focal adhesion kinase/AKT signaling-mediated activation of matrix metalloproteinase (MMP)-2 and MMP-9 in liver cancer.* Carcinogenesis, 2012.
199. Yu, C., et al., *Role of fibroblast growth factor type 1 and 2 in carbon tetrachloride-induced hepatic injury and fibrogenesis.* Am J Pathol, 2003. **163**(4): p. 1653-62.
200. Nguyen, A.T., et al., *A high level of liver-specific expression of oncogenic Kras(V12) drives robust liver tumorigenesis in transgenic zebrafish.* Dis Model Mech, 2011. **4**(6): p. 801-13.
201. Nguyen, A.T., et al., *An inducible kras(V12) transgenic zebrafish model for liver tumorigenesis and chemical drug screening.* Dis Model Mech, 2012. **5**(1): p. 63-72.
202. Bochkis, I.M., et al., *Genome-wide location analysis reveals distinct transcriptional circuitry by paralogous regulators Foxa1 and Foxa2.* PLoS Genet, 2012. **8**(6): p. e1002770.

203. Wolfrum, C., et al., *Insulin regulates the activity of forkhead transcription factor Hnf-3beta/Foxa-2 by Akt-mediated phosphorylation and nuclear/cytosolic localization.* Proc Natl Acad Sci U S A, 2003. **100**(20): p. 11624-9.

Appendix A. List of all liver development genes present on the Affymetrix HG-U133A v2 GeneChip.

Gene	Symbol	Developmental stage	Adult expression	function
Activin receptor type 1	ACVR1	FATE SPEC	+	Activin receptor
Activin receptor type 1B	ACVR1B	FATE SPEC	++	Activin receptor
Activin Receptor type 2, A/B	ACVR2A/B	FATE SPEC	+/+	Activin/nodal receptor
Adenomatous polyposis coli	APC	LVR MAT	+	regulates zonation post-natally
AT rich interactive domain 5B	ARID5B	LVR MAT	+	coActivator of HNF4A
ADP-ribosylation factor 6	ARF6	BUD GR	++	activated by HGF/c-MET to promote hepatoblast migration; regulates zonation of maturing liver
Activating transcription factor 2	ATF2	BUD GR	+	negative regulator of HGF-initiated SEK1/MKK4 signaling
Activating transcription factor 7	ATF7	BUD GR	+	Required to maintain hepatocyte differentiation
ATG7 autophagy related 7 homolog	ATG7	MULTIPLE	+	required for autophagy of organelles; req for homeostasis of differentiated hepatocytes
Bone morphogenic protein 2	BMP2	FATE SPEC	+	regulates regional identity of the endoderm; maintains GATA4/6 expression
Bone morphogenic protein 4	BMP4	FATE SPEC	-	regulates regional identity of the endoderm; maintains GATA4/6 expression; later promotes differentiation of hepatoblast to biliary lineage
Bone morphogenetic protein receptor, type IA	BMPR1A	FATE SPEC	+	Type 1 BMP receptor
Bone morphogenetic protein receptor, type IB	BMPR1B	FATE SPEC	+	Type 1 BMP receptor
Bone morphogenic protein receptor 2	BMPR2	FATE SPEC	++	Type 2 BMP receptor
Basigen	BSG	HEPB MIGR	++	stimulates MMP2 and MMP14
Cell adhesion molecule 1	CADM1	HEP DIFF	++	biliary epithelial adhesion; bile duct development
cyclin D1	CCND1	LVR MAT	++	regulate hepatocyte proliferation

cyclin D2	CCND2	LVR MAT	+	regulate hepatocyte proliferation
Cyclin E2	CCNE2	LVR MAT	+	regulate hepatocyte proliferation
CCAAT/enhancer binding protein, alpha	CADM1	HEP DIFF	++	promotes differentiation to hepatocyte
E-cadherin	CDH1	HEPB MIGR	++	must be down-regulated to allow hepatoblast migration through the liver bud
CCAAT/enhancer binding protein (C/EBP), alpha	CEBPA	HEP DIFF	++	hepatocyte transcription factor
Cerebrus	CER1	FATE SPEC	-	Nodal antagonist
CBP/p300-interacting transactivator	CITED2	LVR MAT	+	co-factor for HNF4A in maturing hepatocytes
Collagen IV alpha 1	COL4A1	HEPB MIGR	++	forms basement membrane of hepatic endoderm
Collagen IV alpha 2	COL4A2	HEPB MIGR	++	forms basement membrane of hepatic endoderm
Collagen IV alpha 3	COL4A3	HEPB MIGR	++	forms basement membrane of hepatic endoderm
Collagen IV alpha 4	COL4A4	HEPB MIGR	+	forms basement membrane of hepatic endoderm
Collagen IV alpha 5	COL4A5	HEPB MIGR	+	forms basement membrane of hepatic endoderm
Collagen IV alpha 6	COL4A6	HEPB MIGR	-	forms basement membrane of hepatic endoderm
Ceruloplasmin	CP	FATE SPEC	+++	copper transport; iron metastasis; marker of hepatoblast differentiation
b-catenin	CTNNB1	FATE SPEC	++	mediates Sox17 and Smad signaling
Casein kinase I isoform delta	CSNK1D	BUD GR	+	activates HIF1A, P53, DVL2/3, DNMT1, and YAP1
Desert hedgehog	DHH	BUD GR	-	Inhibits hepatoblast differentiation during bud growth
Dickkopf-related protein 1	DKK1	FATE SPEC	-	Represses Wnt signaling to allow foregut specification into liver and pancreas
delta-like 1 homolog (Drosophila)	DLK1	BUD GR	-	Expressed by proliferating hepatoblasts
DNA (cytosine-5)-methyltransferase 1	DNMT1	BUD GR	+	required for chromatin alterations
dishevelled homolog 2	DVL2	BUD GR	+	part of Wnt signaling pathway
dishevelled homolog 3	DVL3	BUD GR	++	part of Wnt signaling pathway
E74-like factor 5	ELF5	BUD GR	-	Transcription factor activated by HGF-beta-catenin nuclear translocation
Epithelial cell adhesion molecule	EPCAM	BUD GR	-	Required for hepatoblast proliferation
Epidermal Growth Factor Receptor 2	ERBB2	HEP DIFF	+	Epidermal growth factor receptor
Fibroblast Growth Factor 1	FGF1	HEPB MIGR	+	maintain hepatic progenitors in undifferentiated state
Fibroblast growth factor	FGF2	HEP DIFF	+	promote hepatoblast differentiation to biliary lineage
Fibroblast growth factor 7	FGF7	HEP DIFF	-	promote hepatoblast differentiation to biliary lineage; induces branching of hepatic epithelium
Fibroblast Growth Factor 8	FGF8	HEPB MIGR	-	maintain hepatic progenitors in undifferentiated state
Fibroblast Growth Factor Receptor 1	FGFR1	FATE SPEC, HEP DIFF	+	FGF receptor

Fibroblast Growth Factor Receptor 2	FGFR2	FATE SPEC, HEP DIFF	++	FGF receptor
Fibronectin	FN1	HEPB MIGR, HEP DIFF	+++	forms basement membrane of hepatic endoderm
Forkhead homeobox A1	FOXA1	FATE SPEC	++	regulates endoderm differentiation; de-compacts chromatin around Albumin
Forkhead homeobox A2	FOXA2	FATE SPEC	++	regulates endoderm differentiation; de-compacts chromatin; regulates hepatocyte maturation
Forkhead box M1	FOXM1	BUD GR	+	activates regulators of the G2/M phase of the cell cycle during hepatoblast proliferation
Follistatin	FST	FATE SPEC	+	BMP Inhibitor
Follistatin-like protein 3	FSTL3	FATE SPEC	+	BMP Inhibitor
Frizzled 1	FZD1	MULTIPLE	-	Wnt receptors
Frizzled 2	FZD2	MULTIPLE	-	Wnt receptors
Frizzled 3	FZD3	MULTIPLE	+	Wnt receptors
Frizzled 4	FZD4	MULTIPLE	++	Wnt receptors
Frizzled 5	FZD5	MULTIPLE	++	Wnt receptors
Frizzled 6	FZD6	MULTIPLE	+	Wnt receptors
Frizzled 7	FZD7	MULTIPLE	+	Wnt receptors
glucose-6-phophatase	G6PC	LVR MAT	+++	Marker of terminal hepatocyte differentiation
GATA binding protein 4	GATA4	MULTIPLE	++	de-compacts chromatin; binds albumin promoter; maintains STM during hepatoblast migration
GATA binding protein 6	GATA6	MULTIPLE	+	maintains hepatoblast differentiation
GDNF family receptor alpha 2	GFRA2	HEPB MIGR	-	Neurturn receptor
Glypican 3	GPC3	FATE SPEC	-	BMP Inhibitor
Growth factor receptor-bound protein 2	GRB2	MULTIPLE	++	signal transduction for MET, ERBB2, MST1R, and other receptors
Gremlin	GREM1	FATE SPEC	+	BMP Inhibitor
Heart- and neural crest derivatives-expressed protein 2	HAND2	BUD GR	+	Regulates remodeling of ECM to form the gut loop at the beginning of bud formation
Hepatoma-derived growth factor	HDGF	BUD GR	++	stimulates hepatoblast proliferation
Hepatocyte Growth Factor	HGF	MULTIPLE	+	promote hepatoblast proliferation via SEK1/MKK4
Hairy/enhancer-of-split related	HEYL	HEP DIFF	+	NOTCH signaling protein
Hematopoietically expressed homeobox	HHEX	FATE SPEC, BUD GR	++	promotes hepatoblast migration into STM
Hypoxia-inducible factor 1a	HIF1A	BUD GR	++	induces angiogenesis to growing liver bud
H2.0-like homeobox	HLX	BUD GR	+	promote hepatoblast proliferation, inhibit apoptosis

High-mobility group protein A2	HMGA2	BUD GR	-	regulates proliferation genes and maintains hepatoblasts in undifferentiated state
High-mobility group protein B2	HMGB2	BUD GR	+	regulates proliferation genes and maintains hepatoblasts in undifferentiated state
Hepatocyte nuclear factor 1 homeobox A	HNF1A	LVR MAT	+	regulates hepatocyte maturation
Hepatic Nuclear Factor 1 beta	HNF1B	FATE SPEC, LVR MAT	+	stimulates expression of FOXA1 and FOXA2 in pre-hepatic endoderm; later regulates hepatocyte maturation
Hepatic Nuclear Factor 4 alpha	HNF4A	HEP DIFF, LVR MAT	++	specifies hepatoblast differentiation into hepatocyte; regulates liver zonation post-natally
Homeobox A7	HOXA7	MULTIPLE	-	regulated nuclear export of c-MYC, FGF2, CCND1
Heparin sulfate proteoglycan	HSPG2	HEPB MIGR	++	forms basement membrane of hepatic endoderm
isoprenylcysteine carboxyl methyltransferase	ICMT	HEPB MIGR	++	remodels basement membrane to allow hepatoblast migration
Inhibitor of differentiation 3	ID3	BUD GR	+	inhibits TCF3 to enhance hepatoblast proliferation
Insulin-like Growth Factor 2	IGF2	BUD GR	+++	promotes proliferation of hematopoietic cells in the liver
Indian hedgehog	IHH	BUD GR	-	Inhibits hepatoblast differentiation during bud growth
Interleukin 6 signal transducer (gp130, oncostatin M receptor)	IL6ST	LVR MAT	+++	OSM receptor
Inhibin, alpha	INHA	FATE SPEC	-	negative regulator of activing
Activin	INHBA	FATE SPEC	+	initiates endoderm/mesoderm formation
Inhibin, beta B	INHBB	FATE SPEC	+	subunit of both inhibin and activin
Inhibin, beta C	INHBC	FATE SPEC	++	subunit of both inhibin and activin
Inhibin, beta E	INHBE	FATE SPEC	++	subunit of both inhibin and activin
insulin receptor substrate 2	IRS2	BUD GR	++	enhances hepatoblast survival during proliferation
Integrin alpha 3	ITGA3	BUD GR	+	receptor for fibronectin, collagens, laminins, and cadherins
Integrin alpha 5	ITGA5	BUD GR	++	receptor for fibronectin, collagens, laminins, and cadherins
Integrin alpha 6	ITGA6	BUD GR	++	receptor for fibronectin, collagens, laminins, and cadherins
Integrin beta 1	ITGB1	BUD GR	++	receptor for fibronectin, collagens, laminins, and cadherins
Integrin beta 4	ITGB4	BUD GR	-	receptor for fibronectin, collagens, laminins, and cadherins
Jagged 1	JAG1	HEP DIFF	+	NOTCH pathway ligand; induces expression of HNF1B and SOX9
Jumonji	JARID2	LVR MAT	+	activates OSM; promotes morphological maturation
Jun protoncogene (c-JUN)	JUN	BUD GR	+	required for proliferation
kinase insert domain receptor (vegfr2)	KDR	HEPB MIGR	++	required for blood vessel formation as hepatoblasts migrate into STM
c-kit	KIT	HEPB MIGR	+	cytokine receptor expressed in undifferentiated hepatic progenitors

Kruppel-like factor 6	KLF6	BUD GR	++	Required for hepatocyte proliferation
k-RAS	KRAS	BUD GR	++	regulates proliferation and survival
cytokeratin-19	KRT19	HEP DIFF	+	specifies hepatoblast differentiation into biliary epithelial cell
Laminin alpha 2	LAMA2	HEPB MIGR	+	Structural component of the basement membrane
Laminin alpha 3	LAMA3	HEPB MIGR	+	Structural component of the basement membrane
Laminin alpha 4	LAMA4	HEPB MIGR	+	Structural component of the basement membrane
Laminin beta 1	LAMB1	HEPB MIGR	++	Structural component of the basement membrane
Laminin beta 2	LAMB2	HEPB MIGR	++	Structural component of the basement membrane
Laminin beta 3	LAMB3	HEPB MIGR	+	Structural component of the basement membrane
Laminin beta 4	LAMB4	HEPB MIGR	-	Structural component of the basement membrane
Laminin gamma 1	LAMC1	HEPB MIGR	++	Structural component of the basement membrane
Laminin gamma 2	LAMC2	HEPB MIGR	-	Structural component of the basement membrane
Laminin gamma 3	LAMC3	HEPB MIGR	+	Structural component of the basement membrane
Lymphoid enhancer-binding factor	LEF1	LVR MAT	-	HNF4A cofactor
LIM/homeobox protein	LHX2	BUD GR	+	promote hepatoblast proliferation, inhibit apoptosis
Mitogen-activated protein kinase kinase 4 (SEK1)	MAP2K4	BUD GR	++	direct activator of MAP kinases including MAPK8
p38	MAPK14	BUD GR	++	Activate ATF2/7 in response to HGF signaling
Mitogen-activfated protein kinase 8 (JNK)	MAPK8	BUD GR	+	Required for differentiation
Mitogen-activated protein kinase kinase kinase kinase 4	MAP4K4	BUD GR	++	activates JNK/MAPK8
Midkine	MDK	BUD GR	-	regulates PTN expression in developing catecholemine and rennin-angiotensin pathways
Met proto-oncogene (c-MET)	MET	BUD GR	++	HGF receptor
matrix metalloproteinase 1	MMP1	HEPB MIGR	-	Promotes invasion through the basement membrane
matrix metalloproteinase 11	MMP11	HEPB MIGR	-	Promotes invasion through the basement membrane
matrix metalloproteinase 12	MMP12	HEPB MIGR	-	degrades elastin
matrix metalloproteinase 13	MMP13	HEPB MIGR	-	Promotes invasion through the basement membrane
matrix metalloproteinase-14	MMP14	HEPB MIGR	+	activates MMP2; required for hepatoblast migration
matrix metalloproteinase 15	MMP15	HEPB MIGR	++	Degrades fibronectin, nidogen, and laminin and activates MMP2
matrix metalloproteinase 16	MMP16	HEPB MIGR	-	Degrades fibronectin, nidogen, and laminin and activates MMP2
matrix metalloproteinase 17	MMP17	HEPB MIGR	-	Expressed at low levels but activity is unclear. May activate pro-form of MMP2
matrix metalloproteinase 19	MMP19	HEPB MIGR	+	Degrades Collagen IV, fibronectin, nidogen, laminin to disrupt the basement membrane
matrix metalloproteinase-2	MMP2	HEPB MIGR	+	required for hepatoblast migration; specific for Collagen IV

matrix metalloproteinase 23	MMP23A	HEPB MIGR	-	Expressed at low levels but activity is unclear.
matrix metalloproteinase 24	MMP24	FATE SPEC	+	Expressed in mesoderm and induce FOX and GATA expression
matrix metalloproteinase 25	MMP25	HEPB MIGR	+	Expressed at low levels but activity is unclear.
matrix metalloproteinase 7	MMP7	HEPB MIGR	-	degrades Collagen IV, laminin 1, fibronectin to disrupt the basement membrane
matrix metalloproteinase 9	MMP9	HEPB MIGR	-	degrades Collagen IV AND v
Macriogage stimulating 1 (hepatocyte growth factor-like)	MST1	LVR MAT	++	regulates YAP; maintains hepatocyte quiesence in adult liver
Metal-regulatory transcription factor	MTF1	BUD GR	+	regulates proliferation and survival
c-Myc	MYC	LVR MAT	++	regulated hepatocyte size and morphology
N-myc	MYCN	BUD GR	-	promote hepatoblast proliferation, inhibit apoptosis
Neuroblastoma, suppression of tumorigenicity 1	NBL1	FATE SPEC	+	BMP inhibitor
Necdin	NDN	BUD GR	+	regulates hematopoietic stem cells
Nuclear factor 1	NF1	FATE SPEC	++	initiates albumin transcription
Nuclear factor of kappa light polypeptide gene enhancer in B-cells 1	NFKB1	BUD GR	+	protects hepatoblasts against TNF-induced apoptosis
Nuclear factor of kappa light polypeptide gene enhancer in B-cells 2	NFKB2	BUD GR	+	protects hepatoblasts against TNF-induced apoptosis
Nidogen	NID1	HEPB MIGR	++	forms basement membrane of hepatic endoderm
NK2 homeobox 8	NKX2-8	FATE SPEC	-	promotes AFP expression
Nodal	NODAL	FATE SPEC	-	initiates endoderm/mesoderm formation; induces GATA4/6 and SOX17
Noggin	NOG	FATE SPEC	-	BMP Inhibitor
Neurogenic locus notch homolog protein 2	NOTCH2	HEP DIFF	++	JAG1 receptor
Neurogenic locus notch homolog protein 3	NOTCH3	HEP DIFF	+	JAG1 receptor
liver receptor homolog 1	NR5A2	BUD GR, LVR MAT	++	antagonizes PROX1-promoted hepatoblast proliferation; regulates hepatocyte maturation
Nuclear respiratory factor 1	NRF1	BUD GR	+	regulates proliferation and survival
Neurturin	NRTN	HEPB MIGR	+	hepatoblast chemoattractant
Onecut 1	ONECUT1	MULTIPLE	+	regulate expression of ECM and MMP genes; modulates gradient of TGFβ signaling during hepatoblast differentiation; regulates hepatocyte maturation
Onecut 2	ONECUT2	MULTIPLE	++	regulate expression of ECM and MMP genes; modulates gradient of TGFβ signaling during hepatoblast differentiation
Oncostatin M	OSM	LVR MAT	-	promotes terminal hepatocyte differentiation
Proliferation-assocaited 2G4	PA2G4	BUD GR	++	negative regulator of PROX1

PHD finger protein 2	PHF2	LVR MAT	+	coActivator of HNF4A
Phosphoinositide-3-kinase, catalytic, alpha polypeptide (Pi3K)	PIK3CA	BUD GR	+	Activates signaling cascades
Phophoinositide-3-kinase, regulatory subunit 1 (alpha)	PIK3R1	BUD GR	++	regulates proliferation and survival
prospero homeobox 1	PROX1	FATE SPEC, BUD GR	+	Promotes hepatoblast delamination from basement membrane; promotes hepatoblast proliferation
pleiotrophin	PTN	BUD GR	-	required for development of catecholemine and rennin-angiotensin pathways
v-raf-1 murine leukemia viral oncogene homolog 1	RAF1	BUD GR	++	decrease hepatoblast sensitivity to FasL apoptotic signals
Retinoic Acid Receptor alpha	RXRA	FATE SPEC, BUD GR	+++	Retinoic Acid Receptor
SET domain bifurcated 1	SETDB1	BUD GR	+	chromatin remodeling
secreted frizzled-related protein 5	SFRP5	FATE SPEC	+	Inhibits Wnt signaling to establish foregut identity
Sonic hedgehog	SHH	BUD GR, HEP DIFF	-	Inhibits hepatoblast differentiation during bud growth
Smad2	SMAD2	FATE SPEC	++	following Nodal stimulation, initiates transcription of Sox17 and FoxA1-3; later promote hepatoblast proliferation
Smad3	SMAD3	FATE SPEC	+	following Nodal stimulation, initiates transcription of Sox17 and FoxA1-3; later promote hepatoblast proliferation
Smad4	SMAD4	FATE SPEC	++	following Nodal stimulation, initiates transcription of Sox17 and FoxA1-3
SMAD family member 5	SMAD5	HEP DIFF	++	Transduces BMP signals
Smad6	SMAD6	FATE SPEC	+	antagonist of Smad signaling
Smad7	SMAD7	FATE SPEC	+	antagonist of Smad signaling
Sox17	SOX17	FATE SPEC	+	regulates endoderm differentiation
SRY-box 9	SOX9	HEP DIFF	+	specifies hepatoblast differentiation into biliary epithelial cell
Osteopontin	SPP1	HEP DIFF	++	Mediates integrins and CD44 signaling
Serine/threonine-protein kinase	SRPK1	BUD GR	++	regulates alternative splicing
Signal transducer and activator of transcription 3	STAT3	LVR MAT	+++	activated by IL6ST to promote terminal hepatocyte differentiation
STEAP family member 3, metalloreductase	STEAP3	LVR MAT	+++	inhibits apoptosis during rapid growth
T-box transcription factor 3	TBX3	HEPB MIGR, HEP DIFF	++	stimulates expression of PROX1 during hepatoblast migration; may determine timing of hepatoblast differentiation
Transcription factor 3	TCF3	BUD GR	+	Associates with LEF1 in the Wnt pathway; inhibits proliferation
Transforming growth factor, beta 1	TGFB1	BUD GR	+	promote hepatoblast proliferation; promote hepatoblast differentiation to biliary cells
Transforming Growth Factor beta 2	TGFB2	HEP DIFF	-	specifies hepatoblast differentiation into biliary epithelial cell

Transforming Growth Factor beta 3	TGFB3	HEP DIFF	+	specifies hepatoblast differentiation into biliary epithelial cell
Transforming growth factor receptor	TGFBR1-3	BUD GR	++	TGF-beta receptors
Transforming Growth Factor beta receptor 3	TGFBRIII	HEP DIFF	++	TGFb receptor critical to hepatoblast differentiation into biliary cells
TIMP metallopeptidase inhibitor 2	TIMP2	HEPB MIGR	++	MMP inihibitor
TIMP metallopeptidase inhibitor 4	TIMP4	HEPB MIGR	-	MMP inihibitor
Tumor Necrosis Factor	TNF	BUD GR	-	negative regulator of hepatoblast proliferation; maintains proliferative capacity of fetal hepatocytes
regulator of nonsense transcripts homolog (yeast)	UPF2	MULTIPLE	++	loss leads to activation of DNA damage response
Vimentin	VIM	HEP DIFF	++	BEC marker; intermediate filament
Wnt 5A	WNT5A	HEP SPEC	+	may inhibit Wnt signaling in the anterior endoderm to allow foregut identity to be established
Wilms tumor protein	WT1	BUD GR	-	controls retinoic acid signaling during liver bud growth
X-box binding protein 1	XBP1	BUD GR	+++	controls expansion of the endoplasmic reticulum in proliferating hepatoblasts
Yes-associated protein 1	YAP1	LVR MAT	++	regulates organ size via cell contact inhibition of cell proliferation
Zinc finger factor	ZBTB20	LVR MAT	+	represses AFP and GPC3 post-natally
Zinc finger and homeoboxes factor 2	ZHX2	LVR MAT	+	represses AFP and GPC3 post-natally
thyroid hormone receptor interacting protein 3	ZNHIT3	LVR MAT	+	co-factor for HNF4A in maturing hepatocytes

Appendix B. Quality Assessment results

The seventy-three chips (listed below) were excluded for failing at least one of the following criteria: Nbr Corr >40%, Row Corr <70%, or log(PM/MM) >50% different from average (log(PM/MM)).

Abbreviations: Nbr Corr = probe-neighbor correlation (ideally = 0); Row Corr = correlation between adjacent rows of probes (ideally = 1); log(PM/MM) avg = average log(PM/MM) for all probes on the chip (ideally should be nearly the same for all chips)

Chip	Nbr Corr	Row Corr	log(PM/MM) avg	Chip	Nbr Corr	Row Corr	log(PM/MM) avg
8-D-401	0.290	0.940	0.404	D712_T	0.572	-0.372	0.443
9-D-310	0.413	0.908	0.262	D-728T.1B	0.362	0.277	0.400
B-290	0.714	0.928	0.489	D787_A1	0.104	0.951	0.698
CIR122	0.204	0.811	0.330	D787_A7	0.128	0.938	0.733
CIR123	0.215	0.799	0.418	D-796T.A1	0.180	0.450	0.096
CIR128	0.212	0.800	0.326	D-817_N	0.504	-0.232	0.515
CIR129	0.161	0.939	0.234	D-819_N	0.338	0.229	0.515
CIR283	0.295	0.602	0.259	D-833	0.091	0.974	0.450
D-260	0.375	0.940	0.503	D834_T	0.461	-0.130	0.566
D-264	0.420	0.923	0.496	DC-679	0.167	0.943	0.531
D-265	0.362	0.845	0.469	HCC-I.125	0.222	0.801	0.368
D-269	0.286	0.800	0.359	R2858T	0.359	0.628	0.377
D-278	0.493	0.930	0.423	R2925T	0.269	0.254	0.228
D-345T	0.592	-0.213	0.232	R2926	0.202	0.888	0.411
D-357	0.186	0.941	0.292	R3394_T_III	0.195	0.892	0.522
D-363	0.285	0.933	0.506	R3399_T_II	0.514	-0.138	0.527
D-364	0.469	0.932	0.493	R3400_T	0.159	0.969	0.612
D-374	0.185	0.928	0.370	R3465.V.T	0.149	0.879	0.403
D-410	0.590	-0.265	0.270	R3465_T_VI	0.114	0.951	0.620
D-448	0.421	-0.058	0.160	R3508_T_VIII	0.107	0.969	0.633
D513_TC1	0.547	-0.291	0.566	R3517_T_VII	0.184	0.933	0.689
D520_T2B	0.537	-0.148	0.485	R3520_T	0.140	0.951	0.732
D528_T1A	0.492	-0.124	0.491	R3548_T	0.176	0.967	0.678
D582_T	0.483	-0.113	0.488	R3551_T	0.179	0.970	0.622
D599_T2A	0.175	0.961	0.640	R3552_T	0.203	0.956	0.680
D-69	0.373	0.913	0.274	R3658_TA2	0.169	0.954	0.660
D691_T	0.367	0.004	0.545	R3659_T_III	0.280	0.590	0.545

Appendix C. Differentially expressed liver development genes.

Names	Mean FC HCV-CIR	q-value HCV-CIR	Mean FC Early HCC	q-value Early HCC	Mean FC late HCC	q-value Late HCC
ACVR2B	0.49*	3.46E-11	0.57*	0.000182	0.59	0.038189
AFP	0.87	0.086872	1.79*	4.48E-20	1.57	0.324027
ARID5B	3.44*	1.50E-07	1.77*	4.61E-08	1.64	0.745184
ATF2	0.91	0.176309	1.24*	1.06E-08	1.35	0.243972
BMP2	1.53*	3.58E-06	1.23*	0.002834	1.19	0.935888
BMP4	0.76*	0.001841	0.88*	2.52E-06	1.02	0.249444
CADM1	1.55*	5.92E-06	1.51[V]	0.001003	1.73	0.502831
CCNE2	1.16	0.003566	1.4[V]	2.53E-14	1.7	0.285766
CDH1	1.71*	3.42E-05	1.35[V]	1.21E-06	1.29	0.85288
CEBPA	0.57*	0.008095	0.79*	0.008635	0.86	0.598495
CITED2	2.48*	0.000259	1.56*	0.000295	1.48	0.116526
COL4A1	4.94*	2.53E-08	3.82	0.070348	3.84	0.548252
COL4A2	3.56*	5.24E-11	2.32*	0.002803	2.19	0.255501
COL4A4	1.47*	3.21E-11	1.13*	0.000266	1.16	0.241786
COL4A5	1.36*	7.62E-07	1.36	1.13E-07	1.35	0.064255
CP	0.85	0.335561	0.8	0.253454	0.54[V]	0.000669
CSNK1D	1.42*	9.79E-08	1.02*	5.55E-05	1.06	0.820975
DKK1	1.05	0.289909	1.34*	5.74E-15	1.34	0.382324
DKK3	5.19*	5.88E-22	3.19*	1.66E-08	2.39	0.251674
DKK4	1.0	0.041729	1.16[V]	1.06E-09	1.44	0.026289
EPCAM	13.99*	3.88E-18	7.08[V]	3.97E-09	2.85	0.065963
ERBB2	1.36*	1.81E-13	1.26	0.042029	1.31	0.106232
FGF7	1.48*	2.98E-10	1.24[V]	0.000328	1.14	0.521088
FGFR2	4.09*	5.32E-18	2.01*	1.42E-09	1.34	0.127774
FOXA1	0.40*	3.74E-08	0.46[V]	0.000409	0.37	0.404135
FOXA2	0.58*	2.01E-05	0.63	0.348996	0.63	0.958847
FOXM1	0.87*	0.006229	1.02*	1.55E-05	1.2[V]	3.73E-05
FSTL3	2.94*	9.41E-13	1.54*	3.40E-07	1.47	0.857174
GATA4	0.66*	1.72E-06	0.74	0.02558	0.79	0.181081
GATA6	2.73*	2.66E-15	1.19*	7.71E-13	0.72	0.044615
GPC3	1.93*	7.19E-06	3.98*	2.23E-10	4.49	0.713219
GREM1	1.16	0.002514	1.87[V]	3.10E-08	1.49	0.192348
GSK3B	0.78	0.031171	0.58*	7.55E-05	0.62	0.576869
HAND2	2.06*	8.92E-10	1.56	0.022322	1.27	0.152358
HMGB2	2.37*	8.06E-05	2.8	0.016347	3.21	0.192274
HNF1B	1.28*	0.000124	0.95[V]	0.00099	1.07	0.41923

ID3	2.27*	1.23E-11	1.54*	0.001166	1.22	0.058117
IGF2	0.92	0.245092	0.75	0.008089	0.40*	0.000234
ITGA3	1.35*	1.46E-05	1.17	0.019953	1.23[V]	0.0038
ITGA6	1.41*	0.002155	1.71*	0.003043	1.69	0.573798
JAG1	2.20*	1.02E-14	1.96	0.026299	1.82	0.742372
JUN	2.32*	1.12E-05	1.41*	4.26E-07	1.15	0.005778
KIT	1.88*	7.28E-12	1.46	0.072598	1.39	0.860014
KLF6	3.9*	2.54E-09	2.39*	4.65E-08	1.61	0.007397
KRAS	0.91	0.155308	0.73*	0.000112	0.66	0.643364
KRT19	5.01*	3.74E-13	2.12*	5.45E-05	1.6	0.049275
LAMA2	4.16*	1.34E-14	1.91*	1.34E-09	1.44	0.363741
LAMA3	1.32*	0.002967	1.25[V]	0.003373	1.53	0.208111
LAMB1	3.21*	4.72E-06	1.83[V]	1.41E-06	1.5	0.587399
LAMC1	1.42*	2.13E-05	1.5[V]	0.000314	1.61	0.422084
LAMC3	2.18*	5.15E-13	1.32*	1.36E-05	1.24	0.603517
LRP5	0.67*	0.002842	0.61	0.690185	0.54	0.601217
LRP6	0.74*	0.000792	0.82	0.275806	0.81	0.985316
MDK	1.65*	4.89E-09	2.18[V]	1.08E-05	2.24	0.239585
MET	0.79*	0.000278	1.08*	2.62E-10	1.22	0.321991
MMP2	4.9	0.027847	3.29[V]	0.003952	2.69	0.001829
MMP7	6.33*	2.33E-16	2.99*	0.000111	3.58	0.533479
MMP9	1.1[V]	0.000558	1.44	0.012296	2.06	0.058299
MMP12	1.07	0.001873	2.02*	8.82E-16	3.07	0.385854
MMP15	0.71*	0.002287	0.7	0.954317	0.71	0.86459
MMP19	1.43*	2.86E-06	1.14	0.010725	1.18	0.670128
MST1	0.89	0.062756	0.57*	2.43E-06	0.47	0.292879
MYC	1.77	0.032188	1.04*	0.002319	0.75	0.122622
NDN	2.11*	3.14E-10	1.16*	8.35E-09	1.13	0.931531
NID1	1.08	0.385875	0.81[V]	4.49E-08	0.82	0.18698
NOTCH2	1.05	0.585906	0.78 *	6.89E-08	0.83	0.753159
NR5A2	0.56*	2.83E-06	0.85*	1.57E-09	0.7	0.309775
NRTN	0.33*	2.87E-15	0.38	8.85E-05	0.35	0.398713
PTN	2.02*	1.38E-18	1.48*	0.000376	1.31	0.464926
REL	1.39*	0.006132	1.37	0.033825	1.15	0.066352
RXRA	0.50*	1.03E-12	0.51	0.528157	0.52	0.630088
SFRP5	2.4*	1.36E-09	1.56*	4.62E-06	1.3	0.401281
SMAD2	1.35*	2.86E-07	1.4	0.017922	1.43	0.92141
SMAD7	2.76*	1.36E-05	1.47*	3.75E-06	1.11	0.116017
SOX9	4.66*	8.08E-11	2.51*	7.30E-08	2.8	0.394532
SPP1	7.63*	3.28E-06	5.63	0.042465	11.7	0.031642
SRPK1	0.75*	0.00073	0.85[V]	0.000124	0.97	0.528191

STAT3	0.81	0.066163	0.54*	4.96E-07	.48	0.558456
STK3	0.85*	0.000194	0.97*	2.02E-06	1.15[V]	0.000737
TBX3	0.73	0.048728	1.0[V]	3.71E-06	1.05	0.698753
TCF4	2.57*	1.72E-08	1.88*	0.000575	1.56	0.489453
TGFB1	2.67*	3.94E-11	1.64	0.007678	1.57	0.746376
TGFBR2	1.48*	2.93E-05	1.35	0.027346	1.14	0.08455
TGFBR3	1.09	0.503817	0.60*	7.04E-06	0.50	0.461152
TIMP2	1.39*	2.58E-06	1.39	0.109651	1.28	0.232708
VIM	5.29*	8.28E-09	4.09	0.002167	4.61	0.504156
WNT5A	1.04	0.518411	1.25[V]	1.04E-09	1.57[V]	0.001328
XBP1	0.65*	1.95E-05	0.82*	2.41E-05	0.69	0.185357
YAP1	1.03	0.051737	0.9	0.023557	0.73**	0.006019
ZBTB20	1.99*	1.27E-06	1.25*	3.56E-10	1.14	0.749199

Appendix D. Differential Co-expression modules

Table 1. Gene sets defined by >.5 correlation to the seed gene within cirrhosis, which have significantly different correlation patterns in early HCC.

A. BMP2, cyclin D1 (CCND1), E-cadherin (CDH1)

P=0.0078	BMP2		0.0003	CCND1		P=0.0001	CDH1
BMP2	1.0		PROX1	0.65		CDH1	1.00
NFKB1	0.58		ONECUT1	0.60		MET	0.65
MMP19	0.57		CDH1	0.58		CCND1	0.58
TGFB3	0.57		KRAS	0.56		GPC3	0.52
CSNK1D	0.53		GPC3	0.51		ID2	0.51
LAMC3	0.53		UPF2	-0.52		TGFBR3	-0.56
HAND2	-0.51		TGFBR2	-0.53		PROX1	1.00
ATF2	-0.61		LAMA2	-0.65			
SFRP5							
PA2G4							
ZHX2							

B. c-Met (MET), SMAD2, SMAD5, thyroid hormone receptor interactor 3 (ZNHIT3)

P=0.0001	MET		P=0.0003	SMAD2		P=0.0001	SMAD5		P=0.004	ZNHIT3
CDH1	0.65		UPF2	0.65		SMAD5	1.00		ZNHIT3	1.00
ID2	0.63		LAMA2	0.63		PIK3R1	0.67		NDN	0.66
CP	0.59		NDN	0.62		OC1	0.66		FGF7	0.64
FN1	0.56		SMAD7	0.56		PROX1	0.62		ID3	0.58
NR5A2	0.56		ZNHIT3	0.56		IL6ST	0.52		SMAD2	0.56
CADM1	0.51		LAMB1	0.52		SRPK1	-0.50		TIMP2	0.55
ERBB2	0.50		YAP1	0.50		CSNK1D	-0.51		FN1	-0.51
ARID5B	-0.53		ACVR2B	-0.50		STAT3	-0.56		ATF2	-0.55
GATA6	-0.54		FN1	-0.54						
TGFBR3	-0.57		ID2	-0.57						

C. Vimentin (VIM), Follistatin-like-3 (FST), ceruloplasmin (CP), proliferation-associated 2G4 (PA2G4), laminin B1 (LAMB1)

P=0.012	**VIM**		**P=0.0037**	**FSTL3**		**P=0.109**	**LAMB1**
VIM	1.00		FSTL3	1.00		LAMB1	1.00
MMP2	0.75		NFKB1	0.74		LAMC1	0.73
MMP9	0.73		CSNK1D	0.72		LAMA2	0.71
KIT	0.71		COL4A2	0.72		MMP2	0.66
TIMP2	0.70		TGFB1	0.66		COL4A1	0.64
TGFBR2	0.68		COL4A1	0.64		JAG1	0.64
TGFB1	0.66		PA2G4	0.64		TGFB3	0.61
COL4A2	0.64		MMP19	0.61		COL4A2	0.61
LAMA2	0.64		SMAD7	0.59		VIM	0.60
LAMB1	0.60		VIM	0.58		KIT	0.60
LAMC3	0.59		STAT3	0.58		LAMC3	0.54
NDN	0.59		LAMA3	0.55		SMAD2	0.52
SMAD7	0.58		SRPK1	0.52		ITGA3	0.51
FSTL3	0.58		ITGA5	0.52		TGFBR3	0.50
GREM1	0.57		LAMC3	0.51		KRT19	0.50
ITGA3	0.57		ARF6	0.51		FSTL3	0.50
LAMB2	0.56		TGFB3	0.51		ITGA6	0.50
CITED2	0.56		KRT19	0.51		IRS1	-0.53
KRT19	0.55		MMP2	0.51		CP	-0.55
ITGA6	0.53		ITGA6	0.50		MST1	-0.56
LAMC1	0.50		LAMB1	0.50		CEBPA	-0.58
HMGB2	0.50		CADM1	-0.51		FN1	-0.58
PTN	0.50		OC1	-0.52		PROX1	-0.59
IRS1	-0.50		FST	-0.55		RXRA	-0.63
CADM1	-0.61		CP	-0.61			
CP	-0.62		IRS1	-0.71			
FN1	-0.62		G6PC	-0.73			

Table 2. Differentially co-expressed genes in early HCC. (A) MET, CDH1, STAT3, LAMB2

	MET			CDH1	STAT3			LAMB2
MET	1.00		CDH1	1.00	0.73		LAMB2	1.00
BMP4	0.63		STAT3	0.73	1.00		SMAD7	0.75
GPC3	0.59		ERBB2	0.71	0.59		ITGA3	0.74
CCNE2	0.58		FGFR2	0.69	0.51		GATA6	0.74
NR5A2	0.58		LHX2	0.68	0.70		NOTCH2	0.70
SRPK1	0.57		SFRP5	0.63	0.61		FSTL3	0.69
ATF2	0.56		GATA6	0.62	0.60		NDN	0.67
YAP1	0.55		SMAD7	0.61	0.53		ITGA5	0.67
ITGA6	0.53		LAMA2	0.60	0.50		MMP2	0.66
FOXM1	0.52		ZBTB20	0.59			STAT3	0.64
NOTCH2	-0.50		CITED2	0.59	0.64		TGFBR3	0.64
LAMC3	-0.50		NOTCH2	0.58	0.57		CITED2	0.63
IGF2	-0.50		NDN	0.58	0.54		LHX2	0.62
ITGA5	-0.50		COL4A2	0.58			KIT	0.61
EPCAM	-0.51		LAMC3	0.58			LAMC3	0.58
KIT	-0.51		HAND2	0.57			TGFB1	0.57
KLF6	-0.52		MMP15	0.54			LAMA2	0.56
COL4A2	-0.52		MST1	0.53	0.59		KLF6	0.56
LAMB1	-0.53		EPCAM	0.53			MMP15	0.55
LAMA2	-0.54		HNF1B	0.53			SFRP5	0.55
ERBB2	-0.54		COL4A4	0.53			COL4A2	0.54
SFRP5	-0.55		TGFBR3	0.50	0.56		ERBB2	0.54
TGFB1	-0.56		CCNE2	-0.50	-0.60		ATF2	-0.50
PTN	-0.56		BMP4	-0.51	-0.60		CCNE2	-0.52
NDN	-0.61		ITGA6	-0.55	-0.56		SRPK1	-0.54
MMP2	-0.61		GREM1	-0.59	-0.52		FOXM1	-0.61
ID2	-0.61		MAP4K4	-0.59	-0.56		MET	-0.74
CDH1	-0.62		MET	-0.62	-0.63			
GATA6	-0.62		CTNNB1	-0.64				
STAT3	-0.63		FOXM1	-0.67	-0.72			
TGFBR3	-0.64		ZNHIT3	-0.70	-0.55			
LHX2	-0.64		UPF2	-0.70				
ITGA3	-0.68		LAMB2		0.64			
LAMB2	-0.74		KLF6		0.63			
SMAD7	-0.79		ID2		0.61			
			CCND1		0.55			
			CP		0.55			
			GPC3		-0.50			

(B) KLF6, NR5A2, ITGA6

	KLF6			NR5A2			ITGA6
KLF6	1.00		NR5A2	1.00		ITGA6	1.00
GATA6	0.71		YAP1	0.66		MAP4K4	0.72
STAT3	0.63		IRS1	0.64		GPC3	0.71
ARID5B	0.62		TBX3	0.62		LAMC1	0.59
ARF6	0.60		HDGF	0.62		UPF2	0.58
LAMB1	0.58		MET	0.58		BMP4	0.58
LAMB2	0.56		ONECUT2	0.57		YAP1	0.55
SFRP5	0.56		BMP4	0.50		CCNE2	0.55
FSTL3	0.56		XBP1	0.50		TBX3	0.54
LHX2	0.55		COL4A2	-0.50		LAMA3	0.53
KIT	0.54		ARID5B	-0.50		MET	0.53
PIK3R1	0.53		LAMA2	-0.53		SRPK1	0.52
MMP2	0.53		VIM	-0.53		HAND2	-0.51
NDN	0.52		MMP2	-0.53		CDH1	-0.55
PTN	0.52		TIMP2	-0.58		STAT3	-0.56
ID2	0.50		MMP7	-0.58		CP	-0.56
MET	-0.52		LAMB1	-0.63		ID2	-0.61

(C) CCND1, CCNE2, MST1, YAP1

	CCND1			CCNE2			MST1			YAP1
CCND1	1.00		CCNE2	1.00		MST1	1.00		YAP1	1.00
MST1	0.63		FOXM1	0.77		CCND1	0.63		MDK	0.70
G6PC	0.61		SRPK1	0.70		CP	0.61		NR5A2	0.66
IRS2	0.59		MET	0.58		STAT3	0.59		TBX3	0.57
FOXA2	0.56		GPC3	0.58		MMP15	0.59		ITGA6	0.55
STAT3	0.55		ITGA6	0.55		ERBB2	0.59		MET	0.55
PIK3R1	0.55		ZNHIT3	0.51		CDH1	0.53		HDGF	0.55
MMP15	0.53		LAMC1	0.50		RXRA	0.53		SRPK1	0.52
ZBTB20	0.50		CDH1	-0.50		MMP12	-0.51		MAP4K4	0.51
GPC3	-0.51		ZBTB20	-0.51		SMAD2	-0.52		MMP7	-0.50
HMGB2	-0.51		LAMB2	-0.52		ATG7	-0.53			
SPP1	-0.52		TGFBR3	-0.55		FOXM1	-0.54			
SRPK1	-0.57		STAT3	-0.60		SPP1	-0.54			
GREM1	-0.58		CCND1	-0.60		ACVR2B	-0.60			
CCNE2	-0.60					LAMC1	-0.70			