Received: 28 September 2018 | Revised: 17 February 2019 | Accepted: 14 March 2019

DOI: 10.1002/mgg3.693

ORIGINAL ARTICLE

Molecular Genetics & Genomic Medicine Open Access WILEY

Transcriptome analysis identifies metallothionein as biomarkers to predict recurrence in hepatocellular cacinoma

Sufang Wang[1,2] | **Michael Gribskov**[3,4]

[1]School of Life Sciences, Northwestern Polytechnical University, Xi'an, Shaanxi, China

[2]Center of Special Environmental Biomechanics & Biomedical Engineering, Northwestern Polytechnical University, Xi'an, Shaanxi, China

[3]Department of Biological Sciences, Purdue University, West Lafayette, Indiana, USA

[4]Department of Computer Sciences, Purdue University, West Lafayette, Indiana, USA

Correspondence
Michael Gribskov, Department of Biological Sciences, Purdue University, West Lafayette, IN, USA.
Email: gribskov@purdue.edu

Funding information
National Natural Science Foundation of China, Grant/Award Number: 31800781 and 31741044; China Postdoctoral Science Foundation, Grant/Award Number: 2018M631198

Abstract

Background: Liver cancer is the fifth most common cancer, and hepatocellular carcinoma (HCC) is the major liver tumor type seen in adults. HCC is usually caused by chronic liver disease such as hepatitis B virus or hepatitis C virus infection. One of the promising treatments for HCC is liver transplantation, in which a diseased liver is replaced with a healthy liver from another person. However, recurrence of HCC after surgery is a significant problem. Therefore, it is important to discover reliable cellular biomarkers that can predict recurrence in HCC.

Methods: We analyzed previously published HCC RNA-Seq data that includes 21 paired tumor and normal samples, in which nine tumors were recurrent after orthotopic liver transplantation and 12 were nonrecurrent tumors with their paired normal samples. We used both the reference genome and de novo transcriptome assembly based analyses to identify differentially expressed genes (DEG) and used RandomForest to discover biomarkers.

Results: We obtained 398 DEG using the Reference approach and 412 DEG using de novo assembly approach. Among these DEG, 258 genes were identified by both approaches. We further identified 30 biomarkers that could predict the recurrence. We used another independent HCC study that includes 50 patients normal and tumor samples. By using these 30 biomarkers, the prediction accuracy was 100% for normal condition and 98% for tumor condition. A group of Metallothionein was specifically discovered as biomarkers in both reference and de novo assembly approaches.

Conclusion: We identified a group of Metallothionein genes as biomarkers to predict recurrence. The metallothionein genes were all down-regulated in tumor samples, suggesting that low metallothionein expression may be a promoter of tumor growth. In addition, using de novo assembly identified some unique biomarkers, further confirmed the necessity of conducting a de novo assembly in human cancer study.

KEYWORDS
biomarker, de novo transcriptome assembly, hepatocellular carcinoma, metallothionein, recurrence

1 | INTRODUCTION

Liver cancer is the fifth most common cancer, and the third leading cause of cancer-related death worldwide. Hepatocellular carcinoma (HCC) is the major liver tumor type seen in adults (Bosch, Ribes, Díaz, & Cléries, 2004; Shibata & Aburatani, 2014; Thomas et al., 2010). HCC is usually caused by chronic liver disease such as hepatitis B virus (HBV) or hepatitis C virus (HCV) infection, which accounts for 75%–80% of the cases (Arzumanyan, Reis, & Feitelson, 2013; Bosch et al., 2004). Abuse of alcohol and exposure to aflatoxin are also risk factors for HCC.

The pathogenic mechanisms of hepatitis B or C associated HCC have been heavily investigated. Alterations in the activities and expression levels of several signaling pathways have been identified. For example, inactivation of tumor suppressor genes *p53* (Christofori, Naik, & Douglas, 1995), *RAS* (Oishi et al., 2007), *PI3K* (Zender et al., 2008), overexpression of β-catenin in the Wnt signaling pathways (Edamoto et al., 2003; Peng et al., 2004), overexpression of epidermal growth factor receptor family members (Blivet-Van Eggelpoël et al., 2012; Ito et al., 2001), overexpression of *MET* and its ligand hepatocyte growth factor (Daveau et al., 2003) and overexpression of insulin-like growth factor (Sedlaczek, Hasilik, Neuhaus, Schuppan, & Herbst, 2003). In addition, methylation of cancer relevant genes have been also identified (Kubo et al., 2004; Lee et al., 2003; Liew et al., 1999; Matsuda, Ichida, Matsuzawa, Sugimura, & Asakura, 1999; Murata et al., 2004; C. Wong, Lee, Ching, Jin, & Ng, 2003; I. H. N. Wong et al., 1999), including *APC*, *p16*, E-cadherin, *GSTP1*, *COX2*, apoptosis-associated speck-like protein (*ASC*) and deleted in liver cancer 1, and allelic gains or losses on chromosomes (Kuroki et al., 1995; Maggioni, 2000; Wilkens et al., 2001). However, due to the heterogeneity of HCC, it is not yet clear what early biomarkers could be used for detection of HBV or HCV-mediated HCC (Arzumanyan et al., 2013).

The treatment for HCC includes liver resection, liver transplantation, chemotherapy, and radiation. Liver transplantation is one of the promising treatments, in which a diseased liver is replaced with a healthy liver from another person. However, recurrence of HCC after surgery is a significant problem. It has been reported that the recurrence after liver transplantation ranges from 6% to 40% (Cheng et al., 2011; Marsh et al., 1997; Shimoda et al., 2004). Although, many studies have attempted to identify biomarkers, in order to predict recurrence in patients with HCC, the early detection of recurrence still remains challenge. The current biomarkers for HCC are mostly serum markers, which show low sensitivity (Tsuchiya et al., 2015). Three recent studies identified some novel markers. One was using DNA markers from urine, but the study was based on only 10 individual patients and did not show wide applicability (Hann et al., 2017); a meta-analysis showed a possibility of using circRNA as biomarkers, but they did not talk about the recurrence problem in HCC (M. Wang et al., 2018); marker for the prediction of sorafenib response has been proposed, but its relevance to the recurrence problem is unclear(Kim et al., 2018). Therefore, it is important to discover reliable cellular biomarkers that can predict recurrence in HCC.

Next-generation sequencing, which identifies genomic alterations and somatic mutations at the nucleotide base level, is providing insights into the etiology of cancer and corresponding diagnostics (Chin et al., 2012; Davey et al., 2011; Meyerson, Gabriel, & Getz, 2010; Schuster, 2008). Scientists have started to sequence patients' DNA or mRNA to obtain their genome or transcriptome, but gene expression is usually measured based on the annotation of the human reference genome. Recently, it has been suggested that de novo assembly is valuable even when a reference genome is available (S. Wang & Gribskov, 2017). Although the human reference genome is available, in the case of tumors, where mutation and chromosomal rearrangement may have altered gene/transcript structure, incorporation of de novo assembly is even more important.

In this project, we analyzed previously published HCC RNA-Seq data (Xue et al., 2015) that includes 21 paired tumor and normal samples, in which nine tumors were recurrent after orthotopic liver transplantation and 12 were nonrecurrent tumors with their paired normal samples. In this study, we compared the results of reference and de novo transcriptome assembly based analyses, in order to identify biomarkers that predict recurrence of tumors in HCC.

2 | MATERIALS AND METHODS

2.1 | Data description

The RNA-seq data were directly downloaded from the NCBI sequence read archive (SRP040998). There were nine recurrent tumor with paired adjacent normal samples, and 12 nonrecurrent tumor with paired adjacent normal samples. In total, there were $9 \times 2 + 12 \times 2 = 42$ samples. Details of library construction and patient information are described in Xue et al. (2015).

2.2 | Quality control of raw data

Adapter sequences and low-quality portions of reads were removed using Trimmomatic (version 0.32) (Bolger, Lohse, & Usadel, 2014). Adapters and low-quality read regions with average quality below 13 (phred score) over a four base window were removed. Low-quality sequences at the 5′ and 3′ end, with quality score <10 were also removed. Only reads with a trimmed length over 30 bases were used in further analysis. The number of paired-end reads in each sample is shown in Table S1.

2.3 | Transcriptome assembly of cleaned data

We pooled all left cleaned reads, right cleaned reads, and unpaired cleaned reads from all 42 samples together for de novo transcriptome assembly. Cleaned reads were assembled using Trinity (Grabherr et al., 2011) (version 2.0.6) with the parameters recommended by the authors.

2.4 | Alignment and quantification

Bowtie2 was used to align cleaned reads to both human reference genome (GRCH38) and de novo transcriptome assembly. Then RSEM (version 1.2.30) program (Li & Dewey, 2011; Li, Ruotti, Stewart, Thomson, & Dewey, 2010) was used to quantify gene expression level.

2.5 | Differential expression gene analysis

The DeSeq2 package (Love, Huber, & Anders, 2014) was used to determine differential expression. The integrated statistical model is:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk},$$

where i is group (recurrent or nonrecurrent), j is condition (normal or tumor), k is individual (patient). In this model, we integrated the sample type (tumor or normal) and recurrence type (yes or no), which identified genes that were both differentially expressed in these conditions. Only genes with observed counts >100 (summed over all conditions) were analyzed.

2.6 | Blast search

We compared the de novo assembly to the human reference genome (GRCH38) using BlastN with default settings (Blast version 2.2.29+, National Center for Biotechnology Information, National Library of Medicine, National Institues of Health, Bethesda, MD, USA). We filtered hits by two criteria: identity score ≥95%; and aligned length ≥100 bases.

2.7 | Biomarker identification and confirmation

The RandomForest package (Liaw & Wiener, 2002) was used to identify biomarkers from recurrent and nonrecurrent patients' gene expression levels. Another independent data set was downloaded from the NCBI sequence read archive (SRP068976) for use as confirmation data, to predict the patient outcome using the biomarkers identified in the RandomForest analysis. The confirmation data included 50 patients paired normal and tumor RNA-Seq data. Details of library construction and patient information are described in Liu et al. (2016).

3 | RESULTS

3.1 | De novo transcriptome assembly

We pooled all patient reads together to assemble the transcriptome using the Trinity program ($k\text{-}mer = 25$). In total, there were 1,036,270 predicted transcripts in the assembly with minimum length 224 bp and maximum length 31,736 bases. Trinity labels predicted transcripts systematically with designated form such as $TR_i|c_j\_g_k\_i_l$ (e.g., $TR_{101}|c_0\_g_2\_i_2$), where i, j, k, and l are integers indicate the transcripts, component, group, and isoform, respectively. We determined that sequences with the same component (e.g., c_0, c_1) but different groups such as $TR_1|c_0\_g_1$ and $TR_1|c_0\_g_2$, usually match the same gene. Therefore, we used transcripts number for example, TR_1 as the gene unit. The number of genes when pooled at this level in de novo transcriptome assembly was 797,713, substantially more than the number of genes annotated in the human reference genome. This is possibly due to two effects. (a) many of the predicted transcripts are similar or duplicated; (b) many of them are expressed at low levels which leads to incomplete transcripts.

3.2 | Differentially expressed genes on recurrent/nonrecurrent tumor analysis

We analyzed nine recurrent tumors (after orthotopic liver transplantation) and 12 nonrecurrent tumors, each with a paired normal sample. We did analysis using two approaches, (a) all samples were analyzed with respect to the human Reference genome; (b) all samples were analyzed with respect to de novo assembly (Trinity). Thus, we obtained two gene expression profiles (one using the reference genome, the other using the assembly), and two differentially expressed genes (DEG) lists (see Materials and Methods2 for details).

First, we used the gene expression level to do a principal component analysis. Gene counts were $\log_{10}$ transformed. All normal samples clustered closely, while the tumor samples were distributed widely in both the Reference and assembly cases (Figure 1a,b). However, data from one patient (both normal and tumor) showed very large reciprocal deviations from the expected clusters (Figure 1a,b), suggesting that the tumor and normal samples may be mislabeled. If this is the case, it would cause large errors in estimates of variance, suggesting this sample should be omitted. Without this patient, the paired normal and tumor samples, were better separated (Figure 1c,d). In order to gain more confidence in identifying DEG, we excluded this patient from all analysis.

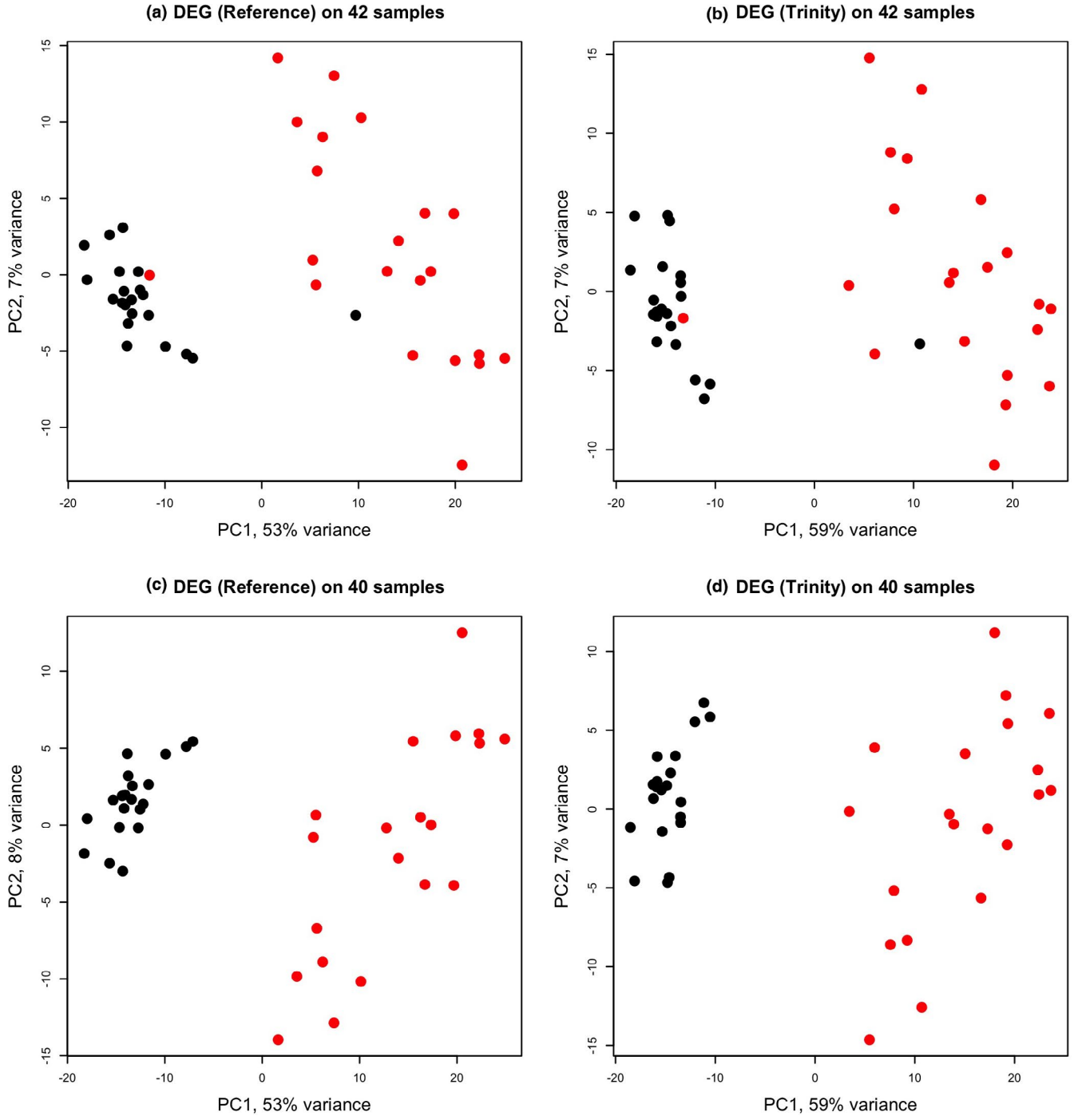

FIGURE 1 Principal component analyses on normal (black dots) and tumor (red dots) samples. (a) Principal component analyses for Reference approach. (b) Principal component analyses for de novo assembly (Trinity) approach. (c) Principal component analyses with one patient excluded in Reference approach. (d) Principal component analyses with one patient excluded in de novo assembly (Trinity) approach. DEG, differentially expressed gene

DEG were identified using the DeSeq2 package with the integrated model where takes group (recurrent or nonrecurrent) and condition (normal or tumor) into account. The significance level was defined as a false discovery rate < 0.0001, and $\log_2$ fold-change ($\log_2$FC) larger than ± 3 (i.e., eightfold change). In total, we obtained 398 DEG using the Reference approach (Table S2) and 412 DEG using de novo assembly approach (Table S3). We further compared the DEG between these two methods and found that 258 DEG were identified by both approaches.

After identifying the DEG, we first checked DEG expression. In order to provide a more straightforward and detailed perspective on gene expression, up- and down-regulated genes were displayed as a heatmap. We chose the top 100 DEG (for a better viewing) exhibiting the largest fold change

and used hierarchical clustering with Euclidean distance and complete linkage method. Gene counts were $\log_{10}$ transformed and normalized as *Z*-score. From the heatmap, there were clearly two clusters (Figure 2); all 20 normal samples were in one cluster and the 20 tumor samples in another cluster. The tumor and normal samples separated very well in both the Reference and de novo assembly (Figure 2), which gives the confidence that these DEG express differentially between tumor and normal samples.

3.3 | Comparison with known cancer genes

In order to validate the involvement of the identified genes in cancer etiology, we compared the DEG to two cancer gene databases: tumor suppressor genes (Zhao, Kim, Mitra, Zhao, & Zhao, 2016; Zhao, Sun, & Zhao, 2013) and oncogenes (Y. Liu, Sun, & Zhao, 2017). We combined the two databases to produce a list of 1616 cancer oncogenes and tumor suppressor genes. We refer this combined database as known cancer genes and it serves as an internal positive control because we expect some of these known cancer genes to be identified in the analysis. We matched our DEG to this list and found 41 known cancer genes in Reference approach and 39 known cancer genes in de novo assembly approach. Twenty-two known cancer genes were found using both methods (Table 1). Some of these cancer genes have previously identified in liver cancer. For example, the overexpression of insulin like growth factor 2 was found to be associated with HCC (Sedlaczek et al., 2003). This gives high confidence that the DEG represent genes involving in regulating pathways in HCC.

3.4 | Biomarker identification and confirmation

We used RandomForest (Liaw & Wiener, 2002), a decision tree-based classification method, to identify biomarkers. A decision tree uses a tree-like graph, which each branch represents a "test" on an attribute (e.g., whether a gene turned on or not, or if the expression level >20), and each leaf node represents the outcome of the test, usually it is a class label (e.g., "Yes" or "No," "tumor" or "nontumor"). RandomForest builds a forest of decision trees to make classifications and rank the importance of attributes (e.g., genes).

In our analysis, we splitted the 40 patients data into two data sets. One set was the training data to train the decision

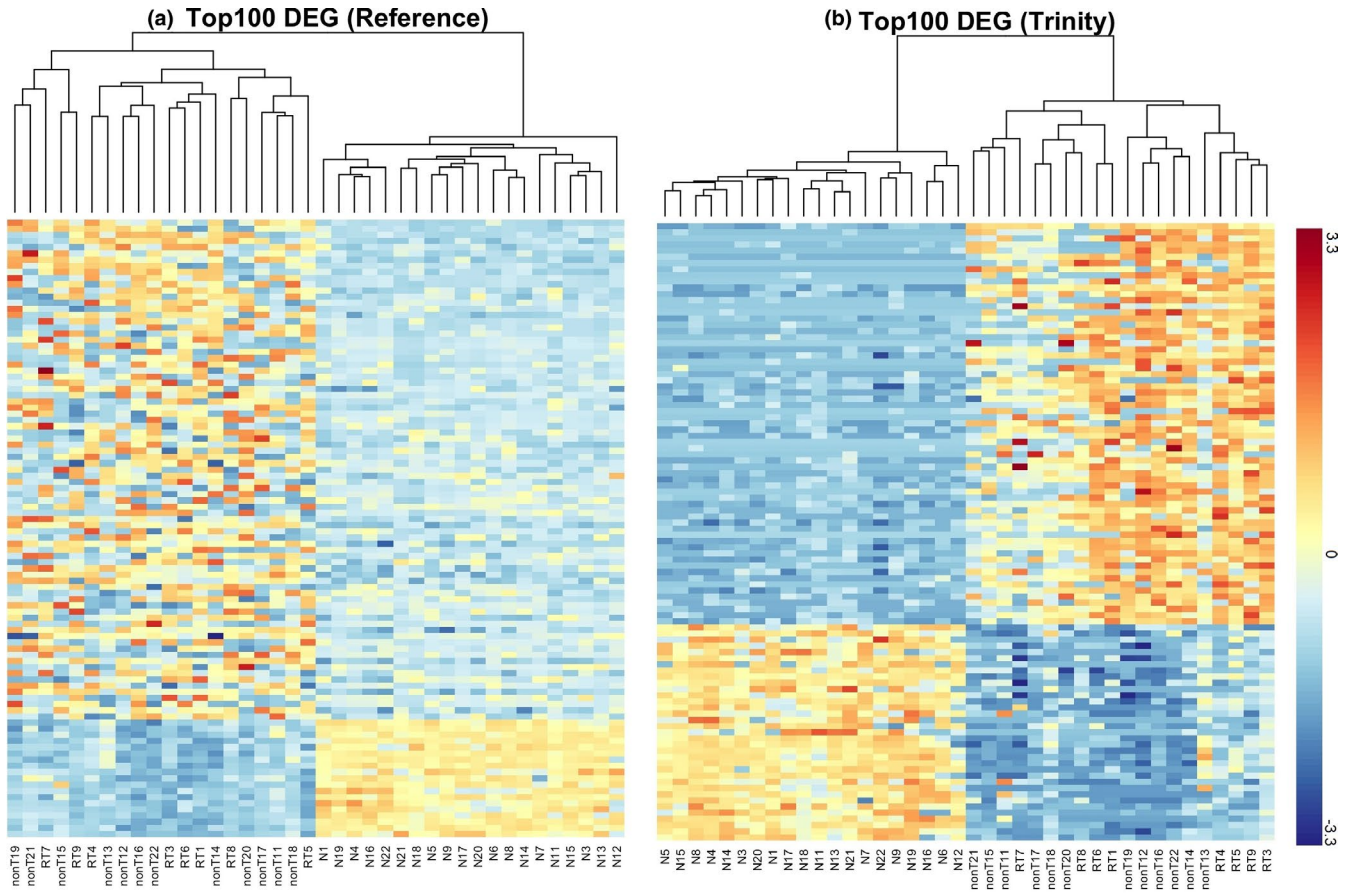

FIGURE 2 Heatmaps of top 100 DEG with largest fold change. The hierarchical clusters were based on *Z*-score. (a) Heatmap of top 100 DEG identified using Reference approach (b) Heatmap of 100 DEG identified using de novo assembly (Trinity). DEG, differentially expressed gene

TABLE 1 Known cancer genes found using both methods

		Reference		De novo assembly	
Name	**Description**	**log_2FC**	**FDR**	**log_2FC**	**FDR**
MT1G	Metallothionein 1G	−4.1	3.96E-18	−4.15	4.40E-19
MT1F	Metallothionein 1F	−3.85	2.24E-24	−4.84	1.47E-22
CXCL14	C-X-C motif chemokine ligand 14	−3.61	1.00E-11	−3.53	1.73E-10
RAB25	Member of RAS oncogene family	−3.57	8.91E-10	−4.27	2.12E-09
BMP10	Bone morphogenetic protein 10	−3.31	4.66E-09	−3.43	1.52E-09
SOX2	SRY-box 2	3.05	4.67E-06	3.39	1.97E-08
CCNE1	Cyclin E1	3.18	1.07E-13	3.1	2.10E-06
CCNB1	Cyclin B1	3.19	6.98E-22	3.19	6.43E-18
MAFA	MAF bZIP transcription factor A	3.27	4.21E-06	3.66	4.61E-09
PTTG1	Pituitary tumor-transforming 1	3.27	3.83E-18	3.74	7.13E-10
KIF14	Kinesin family member 14	3.36	3.86E-32	3.52	9.63E-28
CDK1	Cyclin dependent kinase 1	3.38	2.18E-22	3.33	1.94E-19
MYO18B	Myosin XVIIIB	3.4	9.81E-07	3.68	3.57E-06
CDKN3	Cyclin dependent kinase inhibitor 3	3.64	3.63E-23	3.43	6.51E-19
E2F1	E2F transcription factor 1	3.76	3.63E-23	3.74	7.42E-21
MYO1A	Myosin IA	3.83	9.30E-15	3.28	1.73E-05
CDC25C	Cell division cycle 25C	3.9	5.37E-24	3.84	4.30E-20
GPC3	Glypican 3	4.28	2.79E-12	3.38	1.91E-08
CSMD1	CUB and Sushi multiple domains 1	5.26	2.78E-12	4.25	1.23E-07
SIX1	SIX homeobox 1	5.27	1.52E-12	4.59	2.09E-18
IGF2BP1	Insulin like growth factor 2 mRNA binding protein 1	6.34	3.80E-28	4.46	1.11E-08
ZIC2	Zic family member 2	6.56	6.91E-22	5.64	1.03E-14

Abbreviations: log_2FC, log_2 fold-change; FDR, false discovery rate.

tree; the other used as validation data. We used 80% of original data as training data, after training, we used the trees to predict the patient's condition (normal, recurrent tumor, or nonrecurrent tumor) in the validation data set (20% of original data) and compared the prediction with the patient's real condition. If the accuracy was over 80%, we kept the trees and listed the top 30 important genes in the trees according to the importance plot (data not shown). The top 30 genes for reference and de novo assembly approaches are listed in Table 2.

Then we used these top 30 genes (Table 2) to predict the patient condition in the confirmation data set, which had 50 HCC patients' tumor and normal samples. With these 30 biomarkers, the accuracy for predicting the normal and tumor condition was 100% and 98%, respectively, suggesting these genes might be used for potential biomarkers to predict HCC.

Interestingly, we identified a group of Metallothionein genes as biomarkers (down-regulated in tumor samples), including metallothionein 1E,1F,1G,1H,1J,1M,1X (Table 3). Metallothioneins (*MT*), are a groups of cysteine-rich, low molecular weight proteins that bind to heavy metals. Their major function is protection against DNA damage, oxidative stress, and apoptosis, and they play an important role in transcription factor regulation. Therefore, defects in *MT* expression may lead to malignant transformation of cells and ultimately cancer. It has previously been reported that metallothionein is associated with tumors (Arriaga, Bravo, Mordoh, & Bianchini, 2017; Cherian, Jayasurya, & Bay, 2003; Han et al., 2013; Zheng et al., 2017). Here, *MT* were all down-regulated in tumor samples, suggesting that low *MT* expression may be a promoter of tumor growth.

TABLE 2 Top 30 biomarkers that are used to predict recurrence

Reference			De novo assembly		
	Name	**Description**		**Name**	**Description**
1	***PLP1***	**Proteolipid protein 1**	**1**	***MT1JP***	**Metallothionein 1J**
2	*MT1M*	Metallothionein 1M	**2**	***SYT9***	**Synaptotagmin 9**
3	***SYT9***	**Synaptotagmin 9**	3	*TH*	Tyrosine hydroxylase
4	NA	LincRNA	4	*MT1F*	Metallothionein 1F
5	*KRT16P2*	Keratin 16 pseudogene 2	**5**	***CYP1A2***	**Cytochrome P450 family 1 subfamily A member 2**
6	*KRT16P1*	Keratin 16 pseudogene 1	6	*CLEC4M*	C-type lectin domain family 4 member M
7	***MT1JP***	**Metallothionein 1J**	7	*PLIN2*	Perilipin 2
8	*KRT16P3*	Keratin 16 pseudogene 3	**8**	***HAMP***	**Hepcidin antimicrobial peptide**
9	NA	LincRNA	9	*SYT10*	Synaptotagmin 10
10	***CYP1A2***	**Cytochrome P450 family 1 subfamily A member 2**	**10**	***CLEC4G***	**C-type lectin domain family 4 member G**
11	*ZAN*	Zonadhesin	11	*CD209*	CD209 molecule
12	*CLEC1B*	C-type lectin domain family 1 member B	12	*RAB25*	RAB25 member RAS oncogene family
13	***HAMP***	**Hepcidin antimicrobial peptide**	13	*FAM83F*	Family with sequence similarity 83 member F
14	***CLEC4G***	**C-type lectin domain family 4 member G**	14	*CD5L*	CD5 molecule like
15	*LYPD2*	LY6/PLAUR domain containing 2	**15**	***MARCO***	**Macrophage receptor with collagenous structure**
16	***MARCO***	**Macrophage receptor with collagenous structure**	**16**	***MT1G***	**Metallothionein 1G**
17	*MT1X*	Metallothionein 1X	**17**	***PLP1***	**Proteolipid protein 1**
18	***MT1G***	**Metallothionein 1G**	18	*MT1E*	Metallothionein 1E
19	*MT1H*	Metallothionein 1H	19	*GDF2*	Growth differentiation factor 2
20	NA	LincRNA	20	*NRDE2*	NRDE necessary for RNA interference domain containing
21	*DHAP8*	Double Homeobox A Pseudogene 8	21	*SLITRK6*	SLIT and NTRK like family member 6
22	*IBSP*	Integrin binding sialoprotein	22	*GPM6A*	Glycoprotein M6A
23	*CLEC2L*	C-type lectin domain family 2 member L	23	*ADGRA1*	Adhesion G protein-coupled receptor A1
24	*TOP2A*	Topoisomerase (DNA) II alpha	24	*SFRP5*	Secreted frizzled related protein 5
25	NA	RNA gene	25	*AGBL4*	ATP/GTP binding protein-like 4
26	*UCHL1*	Ubiquitin C-terminal hydrolase L1	26	*B3GNT5*	Beta-1,3-N-Acetylglucosaminyltransferase 5
27	*CDCA7*	Cell division cycle associated 7	27	*CKAP2L*	Cytoskeleton associated protein 2 like
28	NA	RNA gene	28	*MCF2L2*	MCF2 cell line derived transforming sequence-like 2
29	*EXO1*	Exonuclease 1	29	*DUXAP9*	Double homeobox A pseudogene 9
30	NA	RNA gene	30	*TLX1*	T-cell leukemia homeobox 1

Note: The biomarkers identified by both methods were in bold.

4 | DISCUSSION

In this research, we used both the Reference and de novo assembly approaches to identify genes that could be used as biomarkers to predict recurrence in HCC. We analyzed one RNA-Seq dataset that with the recurrent tumors after orthotopic liver transplantation (and their paired normal samples) and the nonrecurrent tumor after orthotopic liver transplantation (and their paired normal samples). We did both de novo transcriptome assembly and reference-based analysis because through our previous research, we discovered that de novo assembly is valuable even when a reference genome available (S. Wang & Gribskov, 2017). And we indeed identified some unique and interesting biomarkers that were not showed in

TABLE 3 Biomarker-Metallothionein expression and significance level identified using references and de novo assembly

	Name	log_2FC	FDR
Genes from reference			
ENSG00000205364	metallothionein 1M	−5.33	9.04E-34
ENSG00000255986	metallothionein 1J	−4.72	5.95E-26
ENSG00000125144	metallothionein 1G	−4.10	3.96E-18
ENSG00000187193	metallothionein 1X	−4.10	1.92E-20
ENSG00000205358	metallothionein 1H	−4.07	2.67E-13
ENSG00000198417	metallothionein 1F	−3.85	2.24E-24
ENSG00000169715	metallothionein 1E	−3.57	3.09E-16
ENSG00000125148	metallothionein 2A	−3.35	4.09E-17
ENSG00000205361	metallothionein 1D	−3.13	3.46E-08
ENSG00000260549	metallothionein 1L	−3.10	5.87E-19
Genes from trinity			
ENSG00000255986	metallothionein 1J	−5.53	4.70E-26
ENSG00000198417	metallothionein 1F	−4.84	1.47E-22
ENSG00000125144	metallothionein 1G	−4.15	4.40E-19
ENSG00000169715	metallothionein 1E	−4.11	3.71E-16
ENSG00000205358	metallothionein 1H	−3.75	3.25E-12
ENSG00000260549	metallothionein 1L	−3.59	9.01E-17
ENSG00000205361	metallothionein 1D	−3.24	9.65E-15

Abbreviations: log_2FC, log_2 fold-change; FDR, false discovery rate.

reference method. For example, *CLEC4M*, a protein encodes a transmembrane receptor and expressed in the endothelial cells of the lymph nodes and liver, together with *CD209*, mediate transinfection of liver cells by HCV (Cormier et al., 2004). Another example is *PLIN2*, belonging to the perilipin family, members of intracellular lipid storage droplets. This protein is found in hepatocytes in alcoholic liver cirrhosis, suggesting that it may serve as a marker of lipid accumulation in liver diseases (Graffmann, Ring, Kawala, Wruck, & Ncube, 2016). And *CD5L*, a key regulator of lipid synthesis, was also identified as a possible marker in liver disease (Gangadharan, Antrobus, Dwek, & Zitzmann, 2007). Therefore, these results further confirmed the necessity of conducting a de novo assembly.

In addition to some unique biomarkers identified in de novo method, it was also very interesting that we identified some long noncoding RNA in reference method. Since long noncoding RNA plays important roles in regulating gene expression, these long-noncoding RNA may be promising biomarkers in HCC diagnostics. However, because we used BLAST program to match the assembled Trinity transcripts to known cDNA gene file, long noncoding RNA was not in the cDNA gene file, therefore, no long noncoding RNA identified in de novo assembly. Furthermore, lower expression of metallthionein protein in HCC tumor has been found before (Cherian et al., 2003), but through our analysis, we systematically pointed out that these genes may use as biomarkers in HCC. However, we recommend more analyses and molecular experiments are needed to confirm the utility of these biomarkers.

In terms of de novo transcriptome assembly programs, it was suggested that SOAPdenovo-trans and Trinity were the best in case of Arabidopsis study (S. Wang & Gribskov, 2017). In this study, we used Trinity, but we found that Trinity produced many redundant or duplicated transcripts when compared with human reference gene annotation. Therefore it may be advantageous using more transcriptome assembly programs in de novo assembly. And for the bioinformatics analysis, DEG were usually discovered by comparing two conditions at one time. But in our analysis, we compared four conditions simultaneously, taking into account the group (recurrent or nonrecurrent) and condition (normal or tumor) information into the integrated statistical model, therefore improves the accuracy of identifying the significant DEG.

ACKNOWLEDGMENTS

This work was funded by the National Natural Science Foundation of China (31741044, 31800781) and China Postdoctoral Science Foundation (2018M631198).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ORCID

Sufang Wang https://orcid.org/0000-0002-7691-9895

REFERENCE

Arriaga, J., Bravo, A., Mordoh, J., & Bianchini, M. (2017). Metallothionein 1G promotes the differentiation of HT-29 human colorectal cancer cells. *Oncology Reports*, *37*, 2633–2651. https://doi.org/10.3892/or.2017.5547

Arzumanyan, A., Reis, H. M., & Feitelson, M. A. (2013). Pathogenic mechanisms in HBV- and HCV-associated hepatocellular carcinoma. *Nature Reviews Cancer*, *13*, 123–135. https://doi.org/10.1038/nrc3449

Blivet-Van Eggelpoël, M.-J., Chettouh, H., Fartoux, L., Aoudjehane, L., Barbu, V., Rey, C., … Desbois-Mouthon, C. (2012). Epidermal growth factor receptor and HER-3 restrict cell response to sorafenib in hepatocellular carcinoma cells. *Journal of Hepatology*, *57*, 108–115. https://doi.org/10.1016/j.jhep.2012.02.019

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bosch, F. X., Ribes, J., Díaz, M., & Cléries, R. (2004). Primary liver cancer: Worldwide incidence and trends. *Gastroenterology*, *127*(5), S5–S16. https://doi.org/10.1053/j.gastro.2004.09.011

Cheng, J. W., Shi, Y. H., Fan, J., Huang, X. W., Qiu, S. J., Xiao, Y. S., … Zhou, J. (2011). An immune function assay predicts post-transplant recurrence in patients with hepatocellular carcinoma. *Journal of Cancer Research and Clinical Oncology*, *137*, 1445–1453. https://doi.org/10.1007/s00432-011-1014-0

Cherian, M. G., Jayasurya, A., & Bay, B. H. (2003). Metallothioneins in human tumors and potential roles in carcinogenesis. *Mutation Research*, *533*, 201–209. https://doi.org/10.1016/j.mrfmmm.2003.07.013

Chin, L., Hahn, W. C., Getz, G., Chin, L., Hahn, W. C., Getz, G., … Meyerson, M. (2012). Making sense of cancer genomic data. *Genes & Development*, *25*, 534–555. https://doi.org/10.1101/gad.2017311

Christofori, G., Naik, P., & Douglas, H. (1995). Functional inactivation but not structural mutation of p53 causes liver cancer. *Nature Genetics*, *10*, 196–201. https://doi.org/10.1038/ng0595-111

Cormier, E. G., Durso, R. J., Tsamis, F., Boussemart, L., Manix, C., Olson, W. C., … Dragic, T. (2004). L-SIGN (CD209L) and DC-SIGN (CD209) mediate transinfection of liver cells by hepatitis C virus. *Proceedings of the National Academy of Sciences of the United States of America*, *101*, 14067–14072. https://doi.org/10.1073/pnas.0405695101

Daveau, M., Scotte, M., François, A., Coulouarn, C., Ros, G., Tallet, Y., … Salier, J. (2003). Hepatocyte growth factor, transforming growth factor alpha, and their receptors as combined markers of prognosis in hepatocellular carcinoma. *Molecular Carcinogenesis*, *36*, 130–141. https://doi.org/10.1002/mc.10103

Davey, J., Hohenlohe, P., Etter, P., Boone, J., Catchen, J., & Blaxter, M. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, *12*, 499–510. https://doi.org/10.1038/nrg3012

Edamoto, Y., Hara, A., Biernat, W., Terracciano, L., Cathomas, G., Riehle, H. M., … Ohgaki, H. (2003). Alterations of RB1, p53 and Wnt pathways in hepatocellular carcinomas associated with hepatitis C, hepatitis B and alcoholic liver cirrhosis. *International Journal of Cancer*, *106*, 334–341. https://doi.org/10.1002/ijc.11254

Gangadharan, B., Antrobus, R., Dwek, R. A., & Zitzmann, N. (2007). Novel serum biomarker candidates for liver fibrosis in hepatitis C patients. *Clinical Chemistry*, *53*, 1792–1799. https://doi.org/10.1373/clinchem.2007.089144

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*, 644–652. https://doi.org/10.1038/nbt.1883

Graffmann, N., Ring, S., Kawala, M., Wruck, W., & Ncube, A. (2016). Modeling nonalcoholic fatty liver disease with human pluripotent stem cell-derived immature hepatocyte-like cells reveals activation of PLIN2 and confirms regulatory functions of Peroxisome Proliferator-Activated Receptor Alpha. *Stem Cells and Development*, *25*, 1119–1133. https://doi.org/10.1089/scd.2015.0383

Han, Y. C., Zheng, Z. L., Zuo, Z. H., Yu, Y. P., Chen, R., Tseng, G. C., … Luo, J. H. (2013). Metallothionein 1 h tumour suppressor activity in prostate cancer is mediated by euchromatin methyltransferase 1. *Journal of Pathology*, *230*, 184–193. https://doi.org/10.1002/path.4169

Hann, H.-W., Jain, S., Park, G., Steffen, J. D., Song, W., & Su, Y.-H. (2017). Detection of urine DNA markers for monitoring recurrent hepatocellular carcinoma. *Hepatoma Research*, *3*, 105–111. https://doi.org/10.20517/2394-5079.2017.15

Ito, Y., Takeda, T., Sakon, M., Tsujimoto, M., Higashiyama, S., Noda, K., … Matsuura, N. (2001). Expression and clinical significance of erb-B receptor family in hepatocellular carcinoma. *British Journal of Cancer*, *84*, 1377–1383. https://doi.org/10.1054/bjoc.2000.1580

Kim, H. Y., Lee, D. H., Cho, E. J., Yu, S. J., Kim, Y. J., Yoon, J.-H., & Lee, J.-H. (2018). A novel biomarker-Based model for the prediction of response to sorafenib and overall survival for advanced hepatocellular carcinoma: A prospective cohort study. *BMC Cancer*, *18*, 307. Retrieved from http://www.embase.com/search/results?subaction=viewrecord&from=export&xml:id=L612593836

Kubo, T., Yamamoto, J., Shikauchi, Y., Niwa, Y., Matsubara, K., & Yoshikawa, H. (2004). Apoptotic speck protein-like, a highly homologous protein to apoptotic speck protein in the pyrin domain, is silenced by DNA methylation and induces apoptosis in human hepatocellular carcinoma. *Cancer Research*, *64*, 5172–5177. https://doi.org/10.1158/0008-5472.CAN-03-3314

Kuroki, T., Fujiwara, Y., Nakamori, S., Imaoka, S., Kanematsu, T., & Nakamura, Y. (1995). Evidence for the presence of two tumour-suppressor genes for hepatocellular carcinoma on chromosome 13q. *British Journal of Cancer*, *72*, 383–385. https://doi.org/10.1038/bjc.1995.342

Lee, S., Lee, H. J., Kim, J.-H., Lee, H.-S., Jang, J. J., & Kang, G. H. (2003). Aberrant CpG island hypermethylation along multistep hepatocarcinogenesis. *The American Journal of Pathology*, *163*, 1371–1378. https://doi.org/10.1016/S0002-9440(10)63495-5

Li, B., & Dewey, C. N. (2011). RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, *12*, 323. https://doi.org/10.1186/1471-2105-12-323

Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A., & Dewey, C. N. (2010). RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, *26*, 493–500. https://doi.org/10.1093/bioinformatics/btp692

Liaw, A., & Wiener, M. (2002). Package "randomForest." *R News*, *2*(3), 18–22.

Liew, C. T., Li, H. M., Lo, K. W., Leow, C. K., Chan, J. Y., Hin, L. Y., … Lee, J. C. (1999). High frequency of p16INK4A gene alterations in hepatocellular carcinoma. *Oncogene*, *18*, 789–795. https://doi.org/10.1038/sj.onc.1202359

Liu, G., Hou, G., Li, L., Li, Y., Zhou, W., & Liu, L. (2016). Potential diagnostic and prognostic marker dimethylglycine dehydrogenase

(DMGDH) suppresses hepatocellular carcinoma metastasis in vitro and in vivo. *Oncotarget*, *7*, 32607–32616. https://doi.org/10.18632/oncotarget.8927

Liu, Y., Sun, J., & Zhao, M. (2017). ONGene: A literature-based database for human oncogenes. *Journal of Genetics and Genomics*, *44*, 119–121. https://doi.org/10.1016/j.jgg.2016.12.004

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*, 550. https://doi.org/10.1186/s13059-014-0550-8

Maggioni, M. (2000). Molecular changes in hepatocellular dysplastic nodules on microdissected liver biopsies. *Hepatology*, *32*, 942–946. https://doi.org/10.1053/jhep.2000.18425

Marsh, J. W., Dvorchik, I., Subotin, M., Balan, V., Rakela, J., Popechitelev, E. P., … Iwatsuki, S. (1997). The prediction of risk of recurrence and time to recurrence of hepatocellular carcinoma after orthotopic liver transplantation: A pilot study. *Hepatology*, *26*, 444–450. https://doi.org/10.1002/hep.510260227

Matsuda, Y., Ichida, T., Matsuzawa, J., Sugimura, K., & Asakura, H. (1999). p16(INK4) is inactivated by extensive CpG methylation in human hepatocellular carcinoma. *Gastroenterology*, *116*, 394–400. https://doi.org/10.1016/S0016-5085(99)70137-X

Meyerson, M., Gabriel, S., & Getz, G. (2010). Advances in understanding cancer genomes through second-generation sequencing. *Nature Reviews Genetics*, *11*, 685–696. https://doi.org/10.1038/nrg2841

Murata, H., Tsuji, S., Tsujii, M., Sakaguchi, Y., Fu, H. Y., Kawano, S., & Hori, M. (2004). Promoter hypermethylation silences cyclooxygenase-2 (Cox-2) and regulates growth of human hepatocellular carcinoma cells. *Laboratory Investigation*, *84*, 1050–1059. https://doi.org/10.1038/labinvest.3700118

Oishi, N., Shilagardi, K., Nakamoto, Y., Honda, M., Kaneko, S., & Murakami, S. (2007). Hepatitis B virus X protein overcomes oncogenic RAS-induced senescence in human immortalized cells. *Cancer Science*, *98*, 1540–1548. https://doi.org/10.1111/j.1349-7006.2007.00579.x

Peng, S. Y., Chen, W. J., Lai, P. L., Jeng, Y. M., Sheu, J. C., & Hsu, H. C. (2004). High alpha-fetoprotein level correlates with high stage, early recurrence and poor prognosis of hepatocellular carcinoma: Significance of hepatitis virus infection, age, p53 and beta-catenin mutations. *International Journal of Cancer*, *112*(1), 44–50. https://doi.org/10.1002/ijc.20279

Schuster, S. C. (2008). Next-generation sequencing transforms today's biology. *Nature Methods*, *5*(1), 16–18. https://doi.org/10.1038/nmeth1156

Sedlaczek, N., Hasilik, A., Neuhaus, P., Schuppan, D., & Herbst, H. (2003). Focal overexpression of insulin-like growth factor 2 by hepatocytes and cholangiocytes in viral liver cirrhosis. *British Journal of Cancer*, *88*, 733–739. https://doi.org/10.1038/sj.bjc.6600777

Shibata, T., & Aburatani, H. (2014). Exploration of liver cancer genomes. *Nature Reviews Gastroenterology & Hepatology*, *11*, 340–349. https://doi.org/10.1038/nrgastro.2014.6

Shimoda, M., Ghobrial, R. M., Carmody, I. C., Anselmo, D. M., Farmer, D. G., Yersiz, H., … Busuttil, R. W. (2004). Predictors of survival after liver transplantation for hepatocellular carcinoma associated with hepatitis C. *Liver Transplantation*, *10*, 1478–1486. https://doi.org/10.1012/lt.20303

Thomas, M. B., Jaffe, D., Choti, M. M., Belghiti, J., Curley, S., Fong, Y., … Venook, A. (2010). Hepatocellular carcinoma: Consensus recommendations of the national cancer institute clinical trials planning meeting. *Journal of Clinical Oncology*, *28*, 3994–4005. https://doi.org/10.1200/JCO.2010.28.7805

Tsuchiya, N., Sawada, Y., Endo, I., Saito, K., Uemura, Y., & Nakatsura, T. (2015). Biomarkers for the early diagnosis of hepatocellular carcinoma. *World Journal of Gastroenterology*, *21*, 10573–10583. https://doi.org/10.3748/wjg.v21.i37.10573

Wang, S., & Gribskov, M. (2017). Comprehensive evaluation of *de novo* transcriptome assembly programs and their effects on differential gene expression analysis. *Bioinformatics*, *33*, 327–333. https://doi.org/10.1093/bioinformatics/btw625

Wang, M., Yang, Y., Xu, J., Bai, W., Ren, X., & Wu, H. (2018). CircRNAs as biomarkers of cancer: A meta-analysis. *BMC Cancer*, *18*, 303. https://doi.org/10.1186/s12885-018-4213-0

Wilkens, L., Bredt, M., Flemming, P., Becker, T., Klempnauer, J., & Kreipe, H. H. (2001). Differentiation of liver cell adenomas from well-differentiated hepatocellular carcinomas by comparative genomic hybridization. *Journal of Pathology*, *193*, 476–482. https://doi.org/10.1002/path.825

Wong, C., Lee, J. M., Ching, Y., Jin, D., & Ng, I. O. (2003). Genetic and epigenetic alterations of DLC-1 gene in hepatocellular carcinoma. *Cancer Research*, *63*, 7646–7651.

Wong, I. H. N., Lo, Y. M. D., Zhang, J., Liew, C., Ng, M. H. L., Wong, N., … Johnson, P. J. (1999). Detection of aberrant p16 methylation in the plasma and serum of liver cancer patients. *Cancer Research*, *59*, 71–73.

Xue, F., Higgs, B. W., Huang, J., Morehouse, C., Zhu, W., Yao, X., … Yao, Y. (2015). HERC5 is a prognostic biomarker for post-liver transplant recurrent human hepatocellular carcinoma. *Journal of Translational Medicine*, *13*, 379. https://doi.org/10.1186/s12967-015-0743-2

Zender, L., Xue, W., Zuber, J., Semighini, C. P., Krasnitz, A., Ma, B., … Lowe, S. W. (2008). An oncogenomics-based In vivo RNAi screen identifies tumor suppressors in liver cancer. *Cell*, *135*, 852–864. https://doi.org/10.1016/j.cell.2008.09.061

Zhao, M., Kim, P., Mitra, R., Zhao, J., & Zhao, Z. (2016). TSGene 2.0: An updated literature-based knowledgebase for tumor suppressor genes. *Nucleic Acids Research*, *44*, 1023–1031. https://doi.org/10.1093/nar/gkv1268

Zhao, M., Sun, J., & Zhao, Z. (2013). TSGene: A web resource for tumor suppressor genes. *Nucleic Acids Research*, *41*, 970–976. https://doi.org/10.1093/nar/gks937

Zheng, Y., Jiang, L., Hu, Y., Xiao, C., Xu, N., Zhou, J., & Zhou, X. (2017). Metallothionein 1H (MT1H) functions as a tumor suppressor in hepatocellular carcinoma through regulating Wnt/β-catenin signaling pathway. *BMC Cancer*, *17*, 161. https://doi.org/10.1186/s12885-017-3139-2

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Wang S, Gribskov M. Transcriptome analysis identifies metallothionein as biomarkers to predict recurrence in hepatocellular cacinoma. *Mol Genet Genomic Med*. 2019;7:e693. https://doi.org/10.1002/mgg3.693