Published in final edited form as:

Int J Comput Biol Drug Des. 2020 ; 13(5-6): 454–474. doi:10.1504/ijcbdd.2020.10036395.

Rapid Evolution of Expression Levels in Hepatocellular Carcinoma

Fan Zhang*,

Vermont Genetics Network, University of Vermont, Burlington, Vermont 05405 USA

Department of Biology, University of Vermont, Burlington, Vermont 05405 USA

Michael D. Kuo*

Department of Diagnostic Radiology, University of Hong Kong, Room 406, Block K, Queen Mary Hospital, 102 Pok Fu Lam Rd, Hong Kong

Abstract

The human evolution and cancer evolution have been researched for several years, but little is known about the molecular similarities between human and cancer evolution. One interesting and important question when comparing and analyzing human evolution and cancer evolution is whether cancer susceptibility is related to human evolution. There are a few microarray studies on human evolution or cancer development. Yet, to date, no microarray studies have been performed with both. Since cancer is an evolution on a small time and space scale, we compared and analyzed liver gene expression data among orangutan, chimpanzee, human, nontumor tissue, and primary cancer using linear mixed model, Analysis of Variance (ANOVA), Gene Ontology (GO), and Human Evolution Based Cancer Gene Expression Analysis. Our results revealed not only rapid evolution of expression levels in hepatocellular carcinoma relative to the gene expression evolution rate of human, but also the correlation between human specific gene expression and cancer specific gene expression. Further gene ontology analysis also suggested statistical relationship between gene function and expression pattern might help understanding the relationship between human evolution and cancer development.

Keywords

cancer evolution; gene expression analysis; pathway analysis; Hepatocellular Carcinoma

1 Introduction

Human evolution is strongly driven by natural selection during the evolution process from primates to Hominins to human. The evolution of cancer differs in some ways from the human evolution. The mutations of human evolution are carried down in sperm and eggs from one generation to the next. Cancer, however, is about somatic evolution. Moreover, human evolution is a process of billions of years, and cancer evolution occurs on a timescale of a human lifetime not millennia (Merlo et al., 2006).

*Corresponding author: fan.zhang@uvm.edu. *Corresponding author: mikedkuo@gmail.com.

Though the human evolution and cancer evolution are not completely identical, we could compare and analyze them if we consider cancer as an evolution on a small time and space scale. For example, survival of the fittest favored by selective pressure which cancer cells face and is exerted by the surrounding environment, would be linked to metastatic potential (Castelli et al., 2017). The dynamics of evolution are fully in play not only in human evolution but also in cancer evolution. Carlo C. Maley believed that “A tumor cell population is constantly evolving through natural selection”(Merlo et al., 2006).

One interesting and important question when comparing and analyzing human evolution and cancer evolution is whether cancer susceptibility is related to human evolution. In order to answer the question, we first studied the correlation between human evolution’s gene expression and cancer gene expression.

There are a few microarray studies on human evolution or cancer evolution. For example, DNA microarrays have been used to compare patterns of gene expression profiles between primates, for a number of tissues, including brain, liver, and heart (Enard et al., 2002; Gilad et al., 2006; Khaitovich et al., 2006). However, none of these studies compare both together.

Gilad et al. compared and analyzed the similarities and differences in gene-expression data between humans, chimpanzees, orangutans and rhesus monkeys. Their results demonstrated that most genes were under a natural selection pressure to maintain a constant level of expression, but a few genes showed species-specific changes. They argued that the big differences between chimpanzees and humans are more due to changes in gene regulation than the individual differences in genes(Gilad et al., 2006). However, the comparison of gene expression in the paper was limited to primates and did not provide information about whether similar conclusions exist in cancer evolution.

Enard et al. used Affymetrix U95A arrays to compare mRNA levels in brains and livers of chimpanzees, orangutans and humans and found that gene expression patterns in the brain had changed more than in the liver during recent human evolution (Enard et al., 2002). Again, no possible relationship between human evolution and cancer evolution had been discussed.

Chen et al. used microarrays to characterize patterns of gene expression in primary liver cancer, and found consistent differences in the expression patterns between hepatocellular carcinoma (HCC) and nontumor liver tissues (Chen et al., 2002). Lee et al. analyzed 91 gene expression profile of human hepatocellular carcinoma and found two subclasses of HCC highly associated with patient survival (Lee et al., 2004). However, neither study provided information about possible differences between liver cancer evolution and human evolution.

In this work, we present the comparative analysis of gene expression data from human evolution and hepatocellular carcinoma evolution using Human Evolution Based Cancer Gene Expression Analysis.

2 Materials and Methods

2.1 Microarray and RNA-Seq

We downloaded five data sets from public repositories including two data sets for primate microarray and three data sets for human liver cancer data. Table 1 shows the summary of the five data sets. Primates microarray data was obtained from GSE2569 (Gilad et al., 2006) and <http://www.sciencemag.org/cgi/content/full/296/5566/340/DC1> (Enard et al., 2002). The first cDNA array contains four species: rhesus macaques, orangutans, chimpanzee, and human, four probes for 1045 orthologous genes from the four species, five liver samples of adult males from each of the four species, and four technical replicates of each comparison. In total, there were 80 samples and 320 hybridizations. The second GeneChip Human Genome U95A array contains three species: orangutans, chimpanzee, and human for approximately 12,000 full-length genes from the three species, three live samples of individuals from human and chimpanzee and one live sample from orangutans, and two technical replicates of each comparison. Enard's data set has 12625 Affymetrix IDs, and we used annotation data package (hgu95av2) in R to annotate Enard's Affymetrix IDs to 9045 gene symbols and used only the liver data from this dataset.

Two human liver cancer microarray datasets and one human liver cancer RNA-Seq dataset were selected from GEO Database GSE3500 (Chen et al., 2002), GSE1898 (Lee et al., 2004), and GSE77314 (Liu et al., 2016). The three liver cancer data sets contain two classes: primary liver cancer (HCC) and nontumor tissue. In Chen's dataset, primary liver cancer (HCC) has 104 samples with 82 patients, nontumor tissue has 76 samples with 73 patients, and the total number of clone IDs for primary liver cancer and nontumor tissue is 40725. We used the Stanford's Batch Source to convert the clone IDs of Chen's data set to 14475 Unigene gene names. Lee's dataset includes 21794 NIH mAdb plate-well identifiers for oligo samples which was converted to 16711 genes. Liu's dataset is RNA-seq of Illumina sequencing of 50 paired normal and tumor samples, including 30926 genes after reads alignment.

In merging all of the data, we discarded genes for which expression was not found common in primate and HCC dataset. Genes for which expression are not in common in samples of each dataset are also removed. We have 801 genes in common in Gilad's and Chen's, 7484 genes in Enard's and Chen's, 882 genes in Gilad's and Lee's, 8214 genes in Enard's and Lee's, 847 genes in Gilad's and Liu's, and 7584 genes in Enard's and Liu's.

2.2 Normalization

We first normalized the expression of each gene in the five datasets separately. The original five datasets are Affymetrix for Enard's, cDNA chips for Gilad's, cDNA chips for Chen's, spotted oligonucleotides for Lee's, and gene expression quantification data from RNA-Seq for Liu's. Enard's array came with data carried out with Microarray Suite by using default setting and global normalization method normalized to the same value using all probe sets without log transformation. We later applied a log base 2 transformation to Enard's array data. Gilad's and Chen's array data, have been originally processed with log ratio base 2 transformation, and Lowess normalization and quantile normalization, respectively.

For Lee's dataset, we used the dye swap normalization method to remove dye bias from each pair of samples. The relative gene expression levels in Liu's dataset were originally normalized with "cuffnorm" method.

2.3 Linear mixed model

System variability in a general microarray experiment can be caused by a number of factors including: gene effects, treatment effects, array effects, dye effects, individual effects, and interactions of these factors. Linear mixed models are used to describe each of these main effects and their interactions. The variance components in linear mixed model are estimated using the residual maximum likelihood (REML) approach (Pinheiro and D.M.Bates, 2002). We used R (<http://www.r-project.org>, version 3.4.4) to fit linear mixed models to each dataset.

For Gilad's data set, we used a full factorial model to represent the fixed effects and three random effects with their interactions, accounting for individual, species, and probe, with the last two random effects nested within Individual, but crossed with each other. The expression log ratio value y_{ijk} for each gene in the j th species, k th probe of the i th individual is expressed as

$$y_{ijk} = \mu + S_j + P_k + SP_{jk} + I_i + SI_{ij} + PI_{ik} + \varepsilon_{ijk}, \quad (1)$$

where $I_i \sim N(0, \sigma_1^2)$, $SI_{ij} \sim N(0, \sigma_2^2)$, $PI_{ik} \sim N(0, \sigma_3^2)$, and $\varepsilon_{ijk} \sim N(0, \sigma^2)$. The fixed effects in equation (1) are μ , the mean, S_j and P_k , the species and probe main effects, and SP_{jk} , the species-probe interaction. The random effects in equation (1) are I_i , the individual random effect, SI_{ij} , the species within individual random effect, and PI_{ik} , the probe within individual random effect. The within-groups errors ε_{ijk} is assumed to be independent of all the random effects. $j = 1, 2, 3, 4$ stands for the four species: rhesus macaques, orangutans, chimpanzee, and human. For each species, there are four types of probe, therefore $k = 1, 2, 3, 4$. Each species contains five individuals, therefore $i = 1, 2, 3, 4, 5$.

For Enard's data set, we used a full factorial model to represent the fixed effect and two random effects accounting for individual, species and their interaction. The expression log ratios value y_{ijk} for each gene in the j th species of the i th individual is expressed as

$$y_{ijk} = \mu + S_j + I_i + SI_{ij} + \varepsilon_{ijk}, \quad (2)$$

where $I_i \sim N(0, \sigma_1^2)$, $SI_{ij} \sim N(0, \sigma_2^2)$, and $\varepsilon_{ijk} \sim N(0, \sigma^2)$. The fixed effects in equation (2) are μ , the mean, and S_j , the species. The random effects in equation (2) are I_i , the individual random effect, and SI_{ij} , the species within individual random effect. The within-groups errors ε_{ijk} is assumed to be independent of all the random effects. $j = 1, 2, 3$ stands for the three species: orangutans, chimpanzee, and human. Enard's data contains three live sample of individuals from human and chimpanzee and one live sample from orangutans. Therefore, for $j = 1$, $i_1 = 1$. For $j = 2$, $i_2 = 1, 2, 3$. For $j = 3$, $i_3 = 1, 2, 3$. $k = 1, 2$ means the two technical replicates of each comparison.

The expression log ratios y_{ij} of each gene in Chen's data set was analyzed using the linear mixed model:

$$y_{ij} = \mu + T_j + I_i + \varepsilon_{ij}, \quad (3)$$

where $I_i \sim N(0, \sigma_1^2)$, and $\varepsilon_{ij} \sim N(0, \sigma^2)$. The fixed effects in equation (3) are μ , the mean, T_j the nontumor and HCC effects. The random effects in equation (3) is I_i , the individual random effect. The within-groups errors ε_{ij} is independent of the random effect. $j = 1, 2$ stands for the two disease states: nontumor and HCC. $i = 1, 2, \dots, 97$ stands for the 97 individuals.

2.4 Gene expression diversity index

We define Simpson's gene expression diversity index (D) as a simple mathematical measurement that characterizes gene expression diversity in a sample (Zhang et al., 2015).

$$D = 1 - \sum_{i=1}^S p_i^2, \quad (4)$$

The proportion of gene i expression value relative to the total expression value of genes p_i is calculated as followed

$$p_i = \frac{g_i}{\sum_{i=1}^S g_i}, \quad (5)$$

where g_i is the i th gene's expression value.

2.5 ANOVA

We used the ANOVA function in R package to test the **null hypothesis** that all the **population** means are equal. In order to compare gene expression between human evolution and cancer development, we classified genes into four categories, human conserved expression (H_0 : $\mu_{nh} = \mu_h$ (nh , nonhuman primates; h , human), no significant differences between human and non-human primates); human specific expression (H_0 : $\mu_{nh} \neq \mu_h$, significant differences between human and non-human primates); cancer conserved expression (H_0 : $\mu_{non} = \mu_{hcc}$ (non , nontumor; hcc , HCC), no significant differences between nontumor and tumor); and cancer specific expression (H_0 : $\mu_{non} \neq \mu_{hcc}$, significant differences between nontumor and tumor). Combined with the linear mixed models (1), (2) and (3), we made a little bit modification to test the null hypothesis for human conserved, human specific, cancer conserved, and cancer specific expression. In Gilad's dataset, we redefined the species as $j = 1$ for rhesus macaques, orangutans, or chimpanzee, and $j = 2$ for human. In Enard's dataset, we redefined the species as $j = 1$ for orangutans, or chimpanzee, and $j = 2$ for human. In Chen's dataset, the null hypothesis actually became H_0 : $T_1 = T_2$. The cancer specific and conserved expression in Lee's dye swap microarray were calculated by t-test method in R package. All p-values were adjusted using p.adjust function and Benjamini correction in R.

2.6 Human evolution based cancer gene expression analysis

Although much is known about genetic variation in human evolution genomes and cancer genomes, little is known about the gene expression profiles across human evolution and cancer development. This information is necessary for identifying transcriptional regulatory networks that are unique to the evolution of humans and cancer development.

Human Evolution Based Cancer Gene Expression Analysis is used to investigate the gene expression patterns across human evolution and cancer development using gene expression data analysis methods in order to study on the relationship between human evolution and cancer development. A specific expression refers to the expression pattern where levels of gene expression are significantly different across samples. A conserved expression refers to the expression pattern where levels of gene expression are not significantly different across samples. The human evolution based cancer gene expression patterns might represent molecular adaptations that influenced the development of human-conserved and human-specific expression traits in human evolution, and cancer-conserved and cancer-specific ones in cancer development.

Here we used the method to investigate gene expression profiles across human evolution (orangutan, chimpanzee and human) and HCC development by comparing the profiles from orangutan, chimpanzee and human to nontumor and HCC. We classified the gene expression pattern across human evolution and cancer development as Human Conserved and HCC Conserved, Human Conserved but HCC Up-Regulated, Human Conserved but HCC Down-Regulated, Human Specific and HCC Conserved, Human Specific but HCC Up-Regulated, or Human Specific but HCC Down-Regulated expression according to the definitions of specific and conserved expression.

2.7 Gene ontology and pathway maps

We used Database for Annotation, Visualization and Integrated Discovery (DAVID) 2018 to perform gene ontology analysis, and explore genes on BioCarta & KEGG pathway maps (Dennis et al., 2003; Huang da et al., 2009; Jiao et al., 2012). We also used Integrated Pathway Analysis Database (IPAD) (<http://fzhang.w3.uvm.edu/ipad/>) (Zhang and Drabier, 2012).

3 Results

We primarily analyzed Enard's and Chen's datasets which have 7484 genes in common. After removing missing data, we obtained 6314 genes. Enard's dataset has three species: orangutan, chimpanzee and human, while Chen's dataset has two disease states: nontumor and HCC. After normalization of the five species (three species and two cancer states), Enard's and Chen's array data have nearly identical median values; the extreme values are closer to the median in the three human species and are further away from the median in the nontumor and HCC; and, the variability of expressions in the HCC decreased a little relative to that in the nontumor (Figure 1).

3.1 Rapid evolution of expression levels in HCC

We considered humans who had cancer as a distinct kind of species and compared the rate of evolutionary change of gene expression levels among the five species (orangutan, chimpanzee, human, nontumor, and HCC). We calculated the Pearson correlation coefficient matrix among the five species. Then, the gene expression distance between species was calculated by one minus the Pearson correlation coefficient.

Figure 2 shows that the distance between nontumor and HCC is 0.06884353, which is 2.8 times the distance within the 3 primate species (0.02454262). By splitting the data into autosome genes and allosome genes, we found that allosome genes ($0.08682509/0.02410024 = 3.6$) could yield higher ratios of the distance between nontumor and HCC to the distance within the 3 primate species than autosome genes ($0.06697080/0.02883525 = 2.32$). The results are consistent with analysis from other combinations of the two primate datasets and three cancer datasets. For example, using Enard's dataset and Liu's dataset, we found that the distance between nontumor and HCC is 0.09059189, which is 3.28 times the distance within the 3 primate species (0.02762079). For allosome genes, the ratios of the distance between nontumor and HCC to the distance within the 3 primate species is $0.11029500/0.02433740 = 4.531914$, which is higher than for autosome genes ($0.08653313/0.02834891 = 3.052432$). This increased distance between nontumor and HCC might suggest that the rate of evolutionary change of gene expression levels may be accelerated from nontumor to HCC compared to the normal human evolution, especially for allosome genes. We also observed the rapid evolution of expression levels in HCC based on the gene expression diversity index (Zhang et al., 2015). From nontumor to HCC, the gene expression diversity index was increased from 0.9894 to 0.9918 with the p-value of 0.0178 and the increase rate of 0.247%. We did not observe significant change in the gene expression diversity index with the human evolution (p-value = 0.7177).

3.2 Human evolution based cancer gene expression analysis

According to the definition of specific and conserved expression, we classified all genes among primate species and two HCC states into four classes: Human Conserved and Cancer (HCC in the paper) Conserved and (HCCC), Human Specific and Cancer Conserved (HSCC), Human Conserved and Cancer Specific (HCCS), and Human Specific and Cancer Specific (HSCS), shown in the Figure 3 A and B for Enard's and Chen's dataset and Gilad's and Chen's dataset, respectively.

Expression for 2209 of 6314 (35%) genes are human conserved and HCC conserved; whereas expression for 2990 of 6314 (47%) genes are human conserved and HCC Specific. Expression for 413 of 6314 (7%) genes are human specific and HCC conserved; whereas expression for 702 of 6314 (11%) genes are human specific and HCC Specific. It shows that the ratio of the number of human specific and cancer specific gene expression to the number of human conserved and cancer specific gene expression is greater than the ratio of the number of the total human specific gene expression to the number of the total human conserved gene expression ($p = 0.0008$ Chi Square Test; $p = 0.0004$ Fisher's Exact Test)

Figure 4 showed mean gene expression profile across human evolution and HCC development. Whether HCC up-regulated or HCC down-regulated expression, HCC shows a pattern to get closer to primates, which is consistent with gene expression distance between HCC and primates in the phylogenetic tree above.

We used a one sample t-test to test our null hypothesis that there is no difference between the ratio of the number of genes with human specific and cancer specific expression to the number of genes with human conserved and cancer specific expression and the ratio of the total number of genes with human specific expression to the total number of genes with human conserved expression. The alternative hypothesis is the ratio of the number of genes with human specific and cancer specific gene expression to the number of genes with human conserved and cancer specific expression is greater than the ratio of the total number of genes with human specific expression to the total number of genes with human conserved expression. The equation is shown below,

$$H_0: \frac{N(h_s | c_s)}{N(h_c | c_s)} = \frac{N(h_s)}{N(h_c)} H_1: \frac{N(h_s | c_s)}{N(h_c | c_s)} > \frac{N(h_s)}{N(h_c)}, \quad (6)$$

where $N(h_s)$ is the total number of genes with human specific expression, $N(h_c)$ the total number of genes with human conserved expression, $N(h_s|c_s)$ the number of genes with human specific and cancer specific expression, and $N(h_c|c_s)$ the number of genes with human conserved and cancer specific expression. Both $N(h_s|c_s)$ and $N(h_c|c_s)$ considers all reasonable thresholds (we set the thresholds for cancer specific expression as $0.05 * (0.8)^i$ $i = 0, 1, \dots, 150$). Various thresholds result in different $\frac{N(h_s | c_s)}{N(h_c | c_s)}$ ratios.

Since the p-value ($8.860576e-88$) is less than the significance level (0.05), we rejected the null hypothesis. Therefore, we conclude that the ratio of the number of genes with human specific and cancer specific expression to the number of genes with human conserved and cancer specific expression is significantly greater than the ratio of the total number of genes with human specific expression to the total number of genes with human conserved expression. The mean of the human specific/human conserved cancer-specific expression number ratio is 0.2795241. It is greater than the ratio of the total number of genes with human specific expression to the total number of genes with human conserved expression (1115/5199).

Similarly, the null hypothesis for the human conserved/human specific cancer-conserved number ratio is that there is no difference between the ratio of the number of genes with human conserved and cancer conserved expression to the number of genes with human specific and cancer conserved expression and the ratio of the total number of genes with human conserved expression to the total number of genes with human specific expression. The alternative hypothesis is the ratio of the number of genes with human conserved and cancer conserved expression to the number of genes with human specific and cancer conserved expression is greater than the ratio of the total number of genes with human conserved expression to the total number of genes with human specific expression. The equation is shown below,

$$H_0: \frac{N(h_c | c_c)}{N(h_s | c_c)} = \frac{N(h_c)}{N(h_s)} H_1: \frac{N(h_c | c_c)}{N(h_s | c_c)} > \frac{N(h_c)}{N(h_s)}, \quad (7)$$

where $N(h_s)$ is the total number of genes with human specific expression, $N(h_c)$ the total number of genes with human conserved expression, $N(h_c|c_c)$ the number of genes with human conserved and cancer conserved expression, and $N(h_s|c_c)$ the number of genes with human specific and cancer conserved expression. Both $N(h_c|c_c)$ and $N(h_s|c_c)$ considers all reasonable thresholds (we set the thresholds for cancer conserved expression as $0.05 * (0.8)^i$ $i = 0, 1, \dots, 150$). Various thresholds result in different $\frac{N(h_c | c_c)}{N(h_s | c_c)}$ ratios.

Since the p-value ($6.965877e-39$) is less than the significance level (0.05), we rejected the null hypothesis. Therefore we conclude that the ratio of the number of genes with human conserved and cancer conserved expression to the number of genes with human specific and cancer conserved expression is significantly greater than the ratio of the total number of genes with human conserved expression to the total number of genes with human specific expression. The mean of the human conserved/human specific cancer-conserved number ratio is 4.959243. It is greater than the ratio of the total number of genes with human conserved expression to the total number of genes with human specific expression (5199/1115).

We found same conclusion for other combinations of the two primate datasets and the three cancer datasets. For example, for Gilad's and Chen's dataset (Figure 3 B), the mean of the human specific/human conserved cancer-specific expression number ratio is 0.2346383. It is greater than the ratio of the total number of genes with human specific expression to the total number of genes with human conserved expression (127/600) (p-value = $1.676708e-22$). And the mean of the human conserved/human specific cancer-conserved number ratio is 4.918385. It is greater than the ratio of the total number of genes with human conserved expression to the total number of genes with human specific expression (600/127) (p-value = $1.292519e-24$).

3.3 Correlation of gene function with gene expression pattern

Statistically significant associations between gene function and each cluster were identified using DAVID 2017. The significant associations are listed in the Table 2. Statistically significant associations between pathway and each cluster were identified using DAVID 2017 and the IPAD (Zhang and Drabier, 2012). The significant associations with Benjamini adjusted p-value less than 0.01 are listed in the Table 3. The relationship between gene function and gene expression pattern was shown in the Figure 5. We defined Human Specific and Cancer Specific (HSCS) as the combination of the Human Specific but HCC Up-Regulated and the Human Specific but HCC Down-Regulated, and the Human Conserved and Cancer Specific (HCCS) as the combination of the Human Conserved but HCC Up-Regulated and the Human Conserved but HCC Down-Regulated. We found a strong statistical relationship between gene function and expression pattern (Benjamini adjusted p-value < 0.01). The difference between cancer specific and cancer conserved expression pattern is in response to stimulus, regulation of biological process, reproduction,

cell proliferation, cell death, cell cycle, and development. The HCCS is more associated with the following functions than the HSCS: cell communication, cell growth, cell differentiation, cell proliferation, cell death, cell cycle, and development. It may suggest that human conserved expression plays important role in intermediate phase of cancer development whereas human specific expression plays important role in initial phase of cancer development by metabolism and response to stimulus. The difference between HCCC and other three patterns is in the following functions: metabolism, cell growth, response to stimulus, regulation of biological process, reproduction and development. It may suggest that human conserved and cancer conserved expression plays important role in the last phase of cancer evolution.

4 Discussion

4.1 Samples

To date there are only a few microarray and RNA-Seq studies relative to HCC and human evolution and very few of them have the same type of tissue. They are Gilad's and Chen's (801 genes), Enard's and Chen's (7484 genes), Gilad's and Lee's (882 genes), Enard's and Lee's (8214 genes), Gilad's and Liu's (847 genes), and Enard's and Liu's (7584 genes). Although we showed the result of Enard's and Chen's dataset in the paper due to space limit, the other five paired datasets we also analyzed showed similar conclusions.

4.2 HCC rapidly evolved

Cancer evolution differs from the evolution of humans. First, the evolution in a cancer cell occurs on a timescale of years, whereas human evolution occurs across a millennia. Second, cancer cells are asexual entities whereas human evolution concerns a sexually reproducing species. However, from an evolutionary perspective, cancer development can also be viewed as a large, genetically and epigenetically heterogeneous population of individual cells (Merlo et al., 2006). Gilad et al. (Gilad et al., 2006) described a rapid evolution of expression levels and human transcription factors in human evolution. Enard et al. (Enard et al., 2002) found that human evolution showed more rapid gene expression evolution in brain tissue than in other tissue. Khaitovich et al. found that the ratio of expression divergence between species to diversity within species in testis is highest of all tissues (Khaitovich et al., 2005). They found a positive association between LD and genes expression divergence in the brain (Khaitovich et al., 2006). In addition to these rapid evolutions of human gene expression, (Stevenson et al., 2007) demonstrated Rapid evolution of cancer genes on the X chromosome, and (de Gramont, 2005) illustrated a rapid evolution in colorectal cancer.

As Figure 2 and Figure 4 illustrate, we found the rapid evolution of HCC gene expression compared to the rate of human evolution gene expression, which was consistent with the results from the gene expression diversity index analysis. We also found that: 1) majority of genes are differentially expressed in HCC development; 2) majority of genes are not differentially expressed in human evolution; 3) majority of differentially expressed genes in HCC development are human-conserved gene expression; and 4) majority of similarly expressed genes in human evolution are differentially expressed in HCC development.

One interesting finding is that for both human specific and conserved gene expression, HCC is closer to primate than nontumor is. In the Figure 2, the distance between 3 primate species and nontumor is 0.7432036, the distance between 3 primate species and HCC 0.741498, the distance between 2 nonhuman species and nontumor 0.7542441, and the distance between 2 nonhuman species and HCC 0.7494748. Gene expression profile across human evolution and cancer development and its average profile showed in the Figure 4 also suggest that HCC might try to return to normal states under the natural selection's pressure but becomes more closer to nonhuman than nontumor does.

We hypothesize the following although this conclusion needs further experimental validation. Nontumor liver (cirrhosis) represents a large fundamental change and that HCC is the liver cells' attempts to adapt to that environment (nontumor liver). Thus, HCC is a natural selection process where hepatic cells are trying to adapt to that new environment (cirrhosis) which is represented by activation of new genes and repression of others. This appears to be generally supported by the fact that evolution of gene expression is relatively stable over 70 million years (3 species) but the largest change in gene expression is from 3 species to non tumor, and this change is quite large. There is a reduction in change from nontumor to HCC which could suggest that HCCs are attempting to return to their normal state but somehow have shifted to a new "stable" species with different properties that allows them survive in that environment.

4.3 Human specific gene expression is more likely to be cancer specific

We calculated the eight variables over p-value exponent: 1) human specific/human conserved cancer-specific number ratio, 2) human conserved/human specific cancer-conserved number ratio, 3) Difference between Mean P-Values for HSCS and HCCS, 4) Difference between Mean P-Values for HCCC and HSCC, 5) Pearson chi-square p-value, 6) Fisher's exact test p-value, 7) Conditional Probability of Cancer Specific expression Given Human Specific or Human Conserved expression, and 8) Conditional probability of cancer conserved expression given human conserved or human specific expression to test if the proportion of human conserved gene expression among cancer specific gene expression or cancer conserved gene expression is more than random chance comparing to the proportion of the total number of human conserved gene expression to the total number of human specific gene expression, and if the proportion of human specific gene expression among cancer specific gene expression or cancer conserved gene expression is more than random chance comparing to the proportion of the total number of human specific gene expression to the total number of human conserved gene expression.

Although the proportion of human conserved gene expression is several times greater than the proportion of human specific gene expression, among cancer specific gene expression or cancer conserved gene expression, our results show that the proportion of human specific gene expression is statistically significantly higher among cancer specific gene expression than that among cancer conserved gene expression, and the proportion of human conserved gene expression is statistically significantly higher among cancer conserved gene expression than that among cancer specific gene expression.

This suggests that human conserved gene expression is significantly associated with cancer conserved gene expression more than random chance, and human specific gene expression is significantly associated with cancer specific gene expression more than random chance. In other words, cancer specific gene expression is correlated with human specific gene expression and cancer conserved gene expression is correlated with human conserved gene expression.

Therefore, we concluded that human specific gene expression may be significantly more likely to be cancer specific than human conserved gene expression, and that human conserved gene expression may be significantly more likely to be cancer conserved than human specific gene expression. That is, human conserved gene expression that are important to primate function and are under negative selection during human evolution may be more likely to remain invariant under negative selection during cancer development, whereas human specific gene expression that are crucial to normal human function and are under positive selection during human evolution may be more likely to continue to change under positive selection during cancer development.

5 Conclusion

We performed a human evolution based cancer gene expression analysis and found 1) rapid evolution of expression levels in hepatocellular carcinoma relative to the gene expression evolution rate of human, 2) the correlation between human specific gene expression and cancer specific gene expression, and 3) statistical relationship between gene function and expression pattern. These findings might be able to help us understand the relationship between cancer and human evolution while comparing and analyzing human evolution and cancer evolution.

Acknowledgements

This work was supported by grants from University California, San Diego(M.D.K.). This work was also supported by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under grant number P20GM103449. We especially thank Catrina Hood from Vermont Genetics Network at University of Vermont for thoroughly proofreading the manuscript.

Biography

Biographical notes: Fan Zhang is a research associate professor in the department of biology University of Vermont and director of bioinformatics in the Vermont Genetics Network. Dr. Michael D. Kuo is a professor in the department of diagnostic radiology Hong Kong University and a Director of Radiogenomics and Rad-Path.

References

- Castelli G, Pelosi E, Testa U (2017). Liver Cancer: Molecular Characterization, Clonal Evolution and Cancer Stem Cells. *Cancers* (Basel) 9.
- Chen X, Cheung ST, So S, Fan ST, Barry C, Higgins J, Lai KM, Ji J, Dudoit S, Ng IO, Van De Rijn M, Botstein D, Brown PO (2002). Gene expression patterns in human liver cancers. *Mol Biol Cell* 13: 1929–1939. [PubMed: 12058060]

- de Gramont A (2005). Rapid evolution in colorectal cancer: therapy now and over the next five years. *Oncologist* 10 Suppl 2: 4–8. [PubMed: 16272453]
- Dennis G Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* 4: P3. [PubMed: 12734009]
- Enard W, Khaitovich P, Klose J, Zollner S, Heissig F, Giavalisco P, Nieselt-Struwe K, Muchmore E, Varki A, Ravid R, Doxiadis GM, Bontrop RE, Paabo S (2002). Intra- and interspecific variation in primate gene expression patterns. *Science* 296: 340–343. [PubMed: 11951044]
- Gilad Y, Oshlack A, Smyth GK, Speed TP, White KP (2006). Expression profiling in primates reveals a rapid evolution of human transcription factors. *Nature* 440: 242–245. [PubMed: 16525476]
- Huang da W, Sherman BT, Lempicki RA (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 4: 44–57. [PubMed: 19131956]
- Jiao X, Sherman BT, Huang da W, Stephens R, Baseler MW, Lane HC, Lempicki RA (2012). DAVID-WS: a stateful web service to facilitate gene/protein list analysis. *Bioinformatics* 28: 1805–1806. [PubMed: 22543366]
- Khaitovich P, Hellmann I, Enard W, Nowick K, Leinweber M, Franz H, Weiss G, Lachmann M, Paabo S (2005). Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science* 309: 1850–1854. [PubMed: 16141373]
- Khaitovich P, Tang K, Franz H, Kelso J, Hellmann I, Enard W, Lachmann M, Paabo S (2006). Positive selection on gene expression in the human brain. *Curr Biol* 16: R356–358. [PubMed: 16618540]
- Lee JS, Chu IS, Heo J, Calvisi DF, Sun Z, Roskams T, Durnez A, Demetris AJ, Thorgeirsson SS (2004). Classification and prediction of survival in hepatocellular carcinoma by gene expression profiling. *Hepatology* 40: 667–676. [PubMed: 15349906]
- 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. [PubMed: 27119355]
- Merlo LM, Pepper JW, Reid BJ, Maley CC (2006). Cancer as an evolutionary and ecological process. *Nat Rev Cancer* 6: 924–935. [PubMed: 17109012]
- Pinheiro JC, Bates DM (2002). *Mixed-Effects Models in S and S-PLUS*. Springer.
- Stevenson BJ, Iseli C, Panji S, Zahn-Zabal M, Hide W, Old LJ, Simpson AJ, Jongeneel CV (2007). Rapid evolution of cancer/testis genes on the X chromosome. *BMC Genomics* 8: 129. [PubMed: 17521433]
- Zhang F, Drabier R (2012). IPAD: the Integrated Pathway Analysis Database for Systematic Enrichment Analysis. *BMC Bioinformatics* 13.
- Zhang F, Cui L, Kuo MD (2015). Diversity of Gene Expression in Hepatocellular Carcinoma Cells. *Genomics Proteomics Bioinformatics* 13: 377–382. [PubMed: 26779818]

Figure 1.

Boxplots represent expression value for three species and two HCC states. The five boxplots have similar median values. The inter-quartile range is increasing from three human species to nontumor and HCC, indicating increased variability of expressions in the nontumor and HCC.

Figure 2.

Phylogenetic tree representing the relative extent of expression changes among cancer evolution and human evolution. We used $d=1-\text{corr}$, where corr indicates the sample correlation between two data points, to describe the rate of evolutionary change of gene expression levels. The rate of evolutionary change of gene expression levels is 0.03 and 0.07 for human evolution and cancer evolution, respectively.

A

B

Figure 3.

Pie charts showing the intersections between HCC and primates microarray data. (A) for Enard's and Chen's dataset; (B) for Gilad's and Chen's dataset.

Figure 4.

Average gene expression profile across human evolution and HCC development. X axis is the five species: orangutan, chimpanzee, human, nontumor and HCC. Y axis is the mean log(2)-transformed gene expression value.

Figure 5. Correlation of gene function with gene expression pattern. The statistical association between gene function and expression pattern are filtered by Benjamini adjusted p-value < 0.01.

Table 1

Summary of data set

	Gilad's	Enard's	Chen's	Lee's	Liu's
Source	GSE2569	http://www.sciencemag.org/cgi/content/full/296/5566/340/DC1	GSE3500	GSE1898	GSE77314
Species	rhesus macaques, orangutans, chimpanzee, and human	orangutans, chimpanzee, and human	Nontumor and HCC	Nontumor and HCC	Nontumor and HCC
Sample size	80 (20 each species)	14 (2 orangutans samples, 6 chimpanzee samples and 6 human samples)	180 (104 HCC, 76 nontumor)	182 (channel 1 for nontumor, channel 2 for HCC)	100 (50 HCC, 50 nontumor)
Number of Individuals	4 rhesus macaques, 4 orangutans, 4 chimpanzee, and 4 human	1 orangutans, 3 chimpanzee and 3 human	97 patients	91 patients	50 patients
probe set	4608	12625	40725	21794	RNA-seq
Gene number	1045	9045	14475	16711	30926

Table 2

Relationship between gene expression pattern and gene function among human evolution and cancer evolution

Category	Biological Process	Molecular Function
Human conserved and HCC Conserved 2209	regulation of biological process 475 regulation of cellular process 446 cell communication 426 signal transduction 378 regulation of metabolism 303 regulation of cellular metabolism 294 transcription 283 regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism 276	ion binding 414 metal ion binding 414 cation binding 371 nucleic acid binding 356 signal transducer activity 315 DNA binding 260
Human Conserved but HCC Up-Regulated 1595	cellular process 1025 physiological process 980 cellular physiological process 932 metabolism 719 cellular metabolism 686 primary metabolism 682 macromolecule metabolism 461 regulation of biological process 373 nucleobase, nucleoside, nucleotide and nucleic acid metabolism 358	binding 894 protein binding 493 nucleic acid binding 317 nucleotide binding 229 DNA binding 215 purine nucleotide binding 200 transferase activity 174 adenyl nucleotide binding 166 ATP binding 165
Human Conserved but HCC Down-Regulated 851	cell communication 302 signal transduction 276 response to stimulus 245 organismal physiological process 236 development 203 response to biotic stimulus 148	catalytic activity 441 cation binding 253 signal transducer activity hydrolase activity 188 receptor activity 146
Human Specific and HCC Conserved 413	physiological process 263 cellular physiological process 245 macromolecule metabolism 122 establishment of localization 78 localization 78	binding 242 protein binding 136 oxidoreductase activity RNA binding 23 cytoskeletal protein binding 18
Human Specific but HCC Up-Regulated 389	cellular physiological process 226 cellular metabolism 164 macromolecule metabolism 119 biopolymer metabolism 86	nucleotide binding 64 purine nucleotide binding 50
Human Specific but HCC Down-Regulated 313	response to stimulus 61 organismal physiological process 58 response to stress 43	catalytic activity 113 oxidoreductase activity 26 calcium ion binding 25

Table 3

Relationship between gene expression pattern and BioCarta and KEGG pathways

Category	BioCarta pathway	KEGG pathway
Human conserved and HCC Conserved 2209	h_mTORPathway: mTOR Signaling Pathway 10 h_rhoPathway: Rho cell motility signaling pathway 8 h_pmlPathway: Regulation of transcriptional activity by PML 7 h_ctla4Pathway: The Co-Stimulatory Signal During T-cell Activation 7 h_telPathway: Telomeres 7 h_arfPathway: Tumor Suppressor Arf Inhibits Ribosomal Biogenesis 6	hsa04080: neuroactive ligand-receptor interaction 50 hsa04020: calcium signaling pathway 39 hsa04530: tight junction 31 hsa04720: long-term potentiation 21
Human Conserved but HCC Up-Regulated 1595	h_mapkPathway: MAPKinase Signaling Pathway 18 h_pyk2Pathway: Links between Pyk2 and Map Kinases 9 h_biopeptidesPathway: Bioactive Peptide Induced Signaling Pathway 9 h_vdrPathway: Control of Gene Expression by Vitamin D Receptor 9 h_smPathway: Spliceosomal Assembly 9 h_insulinPathway: Insulin Signaling Pathway 8 h_hSWI-SNFpathway: Chromatin Remodeling by hSWI/SNF ATP-dependent Complexes 8 h_TPOPathway: TPO Signaling Pathway 8 h_ghPathway: Growth Hormone Signaling Pathway 8 h_igf1Pathway: IGF-1 Signaling Pathway 8 h_rarrxrPathway: Nuclear receptors coordinate the activities of chromatin remodeling complexes and coactivators to facilitate initiation of transcription in carcinoma cells 6 h_epoPathway: EPO Signaling Pathway 6 h_il3Pathway: IL 3 signaling pathway 5	hsa04110: cell cycle 35 hsa04910: insulin signaling pathway 32 hsa04350: tgf-beta signaling pathway 18 hsa04120: ubiquitin mediated proteolysis 11 hsa03022: basal transcription factors 9 hsa00100: biosynthesis of steroids 5
Human Conserved but HCC Down-Regulated 1395	h_nuclearRsPathway: Nuclear Receptors in Lipid Metabolism and Toxicity 11 h_compPathway: Complement Pathway 8 h_alternativePathway: Alternative Complement Pathway 6 h_argininePathway: Catabolic Pathways for Arginine 4	hsa04060: cytokine-cytokine receptor interaction 44 hsa04610: complement and coagulation cascades 28 hsa04630: jak-stat signaling pathway 28 hsa04650: natural killer cell mediated cytotoxicity 24 hsa04620: toll-like receptor signaling pathway 19 hsa00980: metabolism of xenobiotics by cytochrome p450 17 hsa00330: arginine and proline metabolism 14 hsa00280: valine, leucine and isoleucine degradation 14 hsa00071: fatty acid metabolism 13 hsa00350: tyrosine metabolism 13 hsa00410: beta-alanine metabolism 9 hsa00360: phenylalanine metabolism 8
Human Specific and HCC Conserved 413	h_deathPathway: Induction of apoptosis through DR3 and DR4/5 Death Receptors 4 h_longevityPathway: The IGF-1 Receptor and Longevity 3	hsa04110: cell cycle 9 hsa00620: pyruvate metabolism 6 hsa00632: benzoate degradation via coa ligation 5
Human Specific but HCC Up-Regulated 389	h_arapPathway: ADP-Ribosylation Factor 5 h_gpcrPathway: Signaling Pathway from G-Protein Families 4 h_cell2cellPathway: Cell to Cell Adhesion Signaling 3 h_pgc1aPathway: Regulation of PGC-1a 3 h_mCalpainPathway: mCalpain and friends in Cell motility 3	hsa03050: proteasome 5
Human Specific but HCC Down-Regulated 313	h_pparaPathway: Mechanism of Gene Regulation by Peroxisome Proliferators via PPARa(alpha) 7 h_il1rPathway: Signal transduction through IL1R 5 h_amiPathway: Acute Myocardial Infarction 4 h_egfPathway: EGF Signaling Pathway 4 h_intrinsicPathway: Intrinsic Prothrombin Activation Pathway 4 h_extrinsicPathway: Extrinsic Prothrombin Activation Pathway 4 h_il6Pathway: IL 6 signaling pathway 4 h_cdMacPathway: Cadmium induces DNA synthesis and proliferation in macrophages 4 h_TPOPathway: TPO Signaling Pathway 4 h_tollPathway: Toll-Like Receptor Pathway 4 h_pdgfPathway: PDGF Signaling Pathway 4 h_4lbbPathway: The 4-1BB-dependent immune response 3	hsa00260: glycine, serine and threonine metabolism 8 hsa04610: complement and coagulation cascades 7 hsa00071: fatty acid metabolism 6 hsa00380: tryptophan metabolism 6 hsa00650: butanoate metabolism 5 hsa00310: lysine degradation 5 hsa00271: methionine metabolism 3