Affinity analysis of differentially expressed genes in hepatocytes expressing HCV core genotype 1b or 3a

Tommaso Mazza, Stefano Castellana, Angelo Andriulli, Charles Auffray, Manlio Vinciguerra and Valerio Pazienza

1) Bioinformatics Unit, IRCCS “Casa Sollievo della Sofferenza” - Mendel Laboratory, Rome, Italy; 2) Gastroenterology Unit, IRCCS “Casa Sollievo della Sofferenza” Hospital, San Giovanni Rotondo, (FG) Italy; 3) European Institute for Systems Biology & Medicine, CNRS-UCBL-ENS, Université de Lyon, France 4) UCL – Institute for Liver and Digestive Health, Division of Medicine, Royal Free Hospital, London, United Kingdom and 5) Istituto EuroMediterraneo di Scienza e Tecnologia (IEMEST), Palermo, Italy

*: Corresponding author: Dr. Pazienza Valerio
I.R.C.C.S. "Casa Sollievo della Sofferenza" Hospital,
Viale Cappuccini n.1,
71013, San Giovanni Rotondo (FG) Italy.
Phone: +39-0882.416281
Fax: +39-0882.410271
e-mail: pazienza_valerio@yahoo.it

Key words: System Biology, HCV core protein
Abstract

Chronic hepatitis C patients display many genotype-specific clinical features of HCV infection. The core proteins encoded by different genotypes dysregulate numerous sets of distinct host genes. In this study we tested the hypothesis that HCV core proteins 1b and 3a would actually act on a limited number of independent cellular players, as well as on several functionally linked gene products. Structural and functional tests identified a core set of host genes dysregulated by HCV core genotypes 1b and 3a. The core proteins of HCV genotypes 1b and 3a target specifically limited sets of functionally related gene products, which may be responsible for the variations in the clinical spectra associated with HCV infection.
1. Introduction

Hepatitis C Virus (HCV) infects approximately 170 million people worldwide (Shepard et al, 2005). HCV is an enveloped, positive strand RNA virus, belonging to the Flaviviridae family genus Hepacivirus as its natural targets are hepatocytes. Its genome encodes a polyprotein of more than 3,000 amino acids that is cleaved post-translationally by host and viral proteases yielding three structural (Core, E1 and E2) and seven non-structural proteins (p7, NS2 to NS5B) (Grakoui et al, 1993).

The mechanisms through which HCV infection promotes the development of liver diseases remain poorly understood. An influence of specific viral proteins on the pathways played by the host cell would explain some physiopathological features observed in HCV-infected individuals. In particular, the core protein, a key component of the mature virion, has been reported to interact with host proteins involved in apoptotic pathway(s) (Dumoulin et al, 1999; Honda et al 2000; Ray et al 1996), the hepatocyte lipid metabolism (Abid et al, 2005; Moriya et al, 1997; Hourioux et al, 2007; Kim et al, 2007) and insulin signalling (Kawaguchi et al, 2004; Banerjee et al, 2007).

In addition, some of these interactions have been shown to vary according to the viral genotype, both quantitatively and qualitatively (Abid et al, 2005; Pazienza et al, 2007). In the present study, two different HCV genotypes (1b and 3a), which appear to be relevant in the main pathological aspects of HCV infection, were considered. Insulin resistance, steatosis, responsiveness to treatment, progression towards cirrhosis, fibrosis and hepatocellular carcinoma establish and develop following genotype-specific mechanisms (Abid et al, 2005, Pazienza et al, 2007; Ripoli and Pazienza 2011). Supporting evidence also suggests that different HCV genotypes show distinct cell entry properties (Lavillette et al 2005) and modes of transmission.
The viral genotype is among the factors affecting responsiveness to interferon therapy, together with interferon dose (Causse et al, 1991), duration of treatment (Karino et al, 1991) and viral load (Weiland et al, 1993). Despite the small sequence divergence, their core proteins dysregulate two large sets of distinct host gene products involved in lipid transport and metabolism, cell cycle, immune response and insulin signalling (Pazienza et al, 2009). These results point to differential gene expression programs activated by the core proteins of HCV genotypes 1b and 3a. The functional interactions and impact of these altered patterns of expression on the hepatocyte cell homeostasis and dynamics are unknown. In order to decipher them, we developed and used a model based on a semantic protein-to-protein (P2P) interaction network for each of the two genotypes. The interactions included in the model were selected because of physical interaction, co-expression, coexistence in literature, shared drug response and functional and biological processes, then used to define a set of gene products and their functional relationships.

1.2 Material and methods

Two lists of genes, reported in Table 1 of a previous work (Pazienza et al, 2009), dysregulated by the HCV core proteins genotypes 1b and 3a, were used as a starting point. One list was used to train an affinity model based on co-occurrence relationships detected between its members, and the other list to test the existence of any functional affinity with the genes making the model. Each model consists of sub-models, which were built on evidence derived from structural and functional properties of the corresponding group of gene products. Information was gathered from several data sources: MEDLINE abstracts and LocusLink textual descriptions, sequence similarity, motifs, gene co-expression, drug response, and gene ontology.
terms. The degree of affinity of each test gene was assessed against each sub-model, and a significance score was estimated using a 0.05 cutoff value with a Bonferroni correction for multiple testing. A random sampling of the background gene set from the genome was computed to calculate the p-values. A value of 6% of the genome (or 1500 genes from a total of 25000 genes) for random sampling size was considered. Multiple p-values computed for the same test gene were combined with the Fisher's combined probability test. The presented results associate the outputs of DAVID 6.7 (Huang et al, 2009), ToppGene (Chen et al, 2009) and Endeavour 2.44 (Aerts et al, 2009) web services (accessed in January 2013).

The extent of functional homogeneity among the nonspecific genes, was assessed using GeneMANIA (Mostafavi et al, 2008), based on the reconstruction of a weighted interaction network, in which each pair of genes is associated with a positive number reflecting the strength of the interaction or the reliability of the observation that they interact. Strengths correspond to coverage ratios for the number of annotated genes in the network, the number of genes with that annotation in the genome. Their significance was measured in terms of Q-values from FDR corrected hypergeometric tests. Q-values were estimated using the Benjamini-Hochberg procedure. Relationships were determined up to a Q-value cutoff of 0.05 and were calculated on a variety of genomics and proteomics data from several sources, including data from gene and protein expression profiling studies and primary and curated molecular interaction networks, protein domains and phenotypic screening profiles. The more the resulting network is wired, the more genes are functionally related. Therefore, some well-known topological indices were calculated to quantify its degree of connectivity. The number of connected components corresponds to the number of groups of nodes that are linked by at least one pairwise connection. A lower number of connected components suggests a stronger connectivity. The average number of
neighbors is counted as an index of the average connectivity of a node in the network. A normalized version of this parameter is the network density. The density is a value between 0 and 1. It shows how densely the network is populated with edges. A network containing no edges and solely isolated nodes has a density of 0. In contrast, the density of a clique (i.e., a fully connected component) is 1. Finally, a significant expression of the selected genes is monitored in any pathway deposited in the following databases: BioCyc, GenMAPP, BioCarta, Reactome, SigmaAldrich, Signaling Gateway, Signaling Transduction KE, SuperArray, NCI-Nature Curated, PantherDB, Pathway Ontology, SMPDB, WikiPathways, and Bonferroni corrected p values quantify the degree of significance of the outcome (Figure 1).

2. Results

2.1 Affinity Analysis

In a first analysis round, an affinity model was trained using the genes found to be dysregulated by the core protein of HCV genotype 1b, and tested against the genes dysregulated by the core protein of HCV genotype 3a. Affinity was firstly measured on the basis of characteristics of the training set for all data coming from interaction (Bind, Hprd, BioGrid, Intact, Mint, String) and expression (HomoInp, HomoMart, Homologene, InNetDb, and Chdma) resources as well as from literature text (MEDLINE abstracts and LocusLink textual descriptions). The combined p-values, calculated for the test set genes by Endeavour, were significant for the first four genes reported in Table 1: ABCA1 (p=1.384E-6), CYP26B1 (p=5.080E-6), KHDRBS1 (p=0.004) and YWHAB (p=0.038). ToppGene was then used to enrich the model with data coming from drug response (Drug Bank, CTD Marker and Therapeutic, Stitch), and gene ontology databases. Similarly, the most affine genes to the new training set
genes were: APOC1 (p=3.014E-7), EDN1 (p=3.475E-4), PHF15 (p=9.126E-3), SOD2 (p=1.599E-2), ACER3 (p=1.909E-2), MAP4K4 (p=1.911E-2), TNPO1 (p=2.483E-2), SCD (p=2.736E-2), and FNDC3B (p=4.728E-2) Table 1 (lines 5 to 13). There is no overlap between the results of both analyses. Overall, the core HCV protein of genotype 3a was found to dysregulate 13 out of the 88 genes in the list.

The same data resources were used to build and train an affinity model using the genes dysregulated by the core protein genotype 3a and tested again the genes dysregulated by the core protein of HCV genotype 1b. Using the same analytic procedure and according to interaction, expression and literature text resources, seven genes ABCA1 (p=6.990E-4), PSMD8 (p=0.036), PCTK2 (p=0.040), ABI1 (p=0.041), YES1 (p=0.042), CREB1 (p=0.044) and SFRS7 (p=0.046) (Table 1, lines 14-20) were found to be dysregulated. In contrast, 19 genes SIRT1 (p=1.54E-4), CREB1 (p=3.34E-4), YES1 (p=0.002), CLK1 (p=0.004), IL8 (p=0.005), ABI1 (p=0.005), RHOU (p=0.007), LPGAT1 (p=0.008), TAGLN (p=0.009), PNN (p=0.012), CSPP1 (p=0.013), TTR (p=0.014), DLX2 (p=0.016), NR2F2 (p=0.017), CDK17 (p=0.021), ZFP36L1 (p=0.022), TANC1 (p=0.028), CDH17 (p=0.030), PTBP2 (p=0.044), (Table 1, lines 21-39) were found to be significantly associated to the genes dysregulated by the core protein of HCV genotype 3a, when also considering information about drug response and the Gene Ontology. CREB1, YES1, ABI1 and PCTK2 (alias of CDK17) were detected in both analyses. Overall, the core HCV protein genotype 1b was found to dysregulate 22 genes.

2.2 Functional Analysis

A total of 34 genes identified from the analysis of both sets were wired up with 56 links in an interaction network (Figure 2). Among the interactions, 66% were inferred
from co-expression data, 16% were relationships based on co-localization, with 14% linked gene products belonging to the same pathways. In addition, 4% of the total links were predicted and only 0.09% were experimentally verified physical interactions. All but six genes (CYP26B1, CREB1, CDH17, TANC1, OHF15 and DLX2) were connected in the network (Figure 2), which exhibited two connected components, meaning that most genes are closely related. The average number of neighbor nodes is 3.8, indicating that each gene was associated to at least three genes, on average. The resulting network density is 0.140, corresponding to a moderate association rate among genes (Mazza et al, 2010; Piepoli et al, 2012; Mazza et al, 2012).

The resulting significantly enriched biological processes were the lipid metabolic process (GO:0006629, p=7.414E-3), regulation of fat cell differentiation (GO:0045598, p=9.689E-3), cellular lipid metabolic process (GO:0044255, p=1.886E-2), regulation of sterol transport (GO:0032371, p=4.358E-2), and regulation of cholesterol transport (GO:0032374, p=4.358E-2). For each of these semantic terms, genes dysregulated by one genotype were not found to be more representative for one term compared to those dysregulated by the other genotype. The hypothesis according to which the proportions of genes dysregulated by the two genotypes are significantly different for each term was tested using Fisher's exact test (GO:0006629, p=0.2494; GO:0045598, p=0.6307; GO:0044255, p=0.7482; GO:0032371, p=0.6307; GO:0032374, p=0.6307). This showed that neither genotype is exclusive for, or represents better than the other, these biological processes.

2.3 Pathway analysis

Assessment of common biological processes has identified a small but significant (p=1.668E-2) proportion of genes that were involved in the pathway where SREBF
and miR-33 govern cholesterol and lipid homeostasis (Yang et al, 2006). Among the various genes central to this pathway and dysregulated by both genotypes were ABCA1, SCD and SIRT1, controlling PPARGC1A, PPARA, NR1H3 and SIRT6 (Figure 3). MiR-33a/b play a capital role in this pathway. By querying AURA (Dassi et al, 2012), we could confirm that their regulatory activities are positively exerted on the 3’ UTR of ABCA1, PRKAA1 and negatively on that of SIRT6, which in turn cooperates with SIRT1 to inhibit the biological processes mediated by SREBF1/2. Moreover, the direct and positive regulation of SIRT1 on SIRT6 strengthen their cooperative inhibition over SREBF1/2, thereby making a central regulative cluster within the pathway. It is evident both and interesting the opposing functions played by miR-33a/b and SIRT1 over the regulation of the expression of SREBF1/2 and, consequently, of that of its target genes.

3. Discussion.

In recent years, systems biology approaches combining high-dimensional experiments and network modelling have been increasingly applied to biomedical research (Auffrey et al, 2009). In the present work, we aimed at investigating, using affinity and network analyses, the relationship between the genes that were found to be dysregulated by the HCV core protein genotype 1b or 3a and the biological processes influenced by the encoded gene products. Previous works agreed on the fact that each core protein triggers its own specific gene expression profile (Dou et al, 2006). In fact, despite the limited sequence divergence (the HCV core gene contains the most conserved sequence in the coding region of most HCV genotypes), HCV affects biological pathways in a genotype-dependent manner. Genotype 1-infected patients are exposed to a more aggressive liver disease with the worst response to therapy, higher risk of cirrhosis and hepatocellular carcinoma development (Ripoli
and Pazienza 2011). This worst-case prognosis is probably due to the multiple oncogenic properties of the genotype 1 HCV core protein whilst increased steatosis and fibrosis are associated with genotype 3.

Pathway analysis of the common biological processes in which the genes dysregulated by both HCV genotypes 1b and 3a are involved revealed that a significant proportion of the gene products act in cholesterol and lipid homeostasis pathways and are influenced by the sterol regulatory element-binding protein transcription factor (SREBF1/2) and miR-33a/b (Figure 3). The two molecules are located in the same evolutionary conserved genetic locus and act in concert to control cholesterol homeostasis. The expression of miR-33a/b from an SREBF1/2 intron inhibits cholesterol export and fatty acid oxidation (Gerin et al, 2010). The miR-33a/b gene embedded within introns of the SREBF1/2 genes, in turn, target the adenosine triphosphate-binding cassette transporter A1 (ABCA1), an important regulator of high-density lipoprotein (HDL) synthesis and reverse cholesterol transport (Najafi-Shoushtari et al, 2010) (Figure 3). Inhibition of miR-33 in cell lines and mice causes up-regulation of ABCA1 expression and increased cholesterol efflux (Najafi-Shoushtari et al, 2010). A central node in the pathway analysis of the selected genes is the NAD-dependent deacetylase SIRT1, which we have previously directly associated to the HCV core genotype 1b (Ripoli et al, 2011) but is now proven to be also indirectly associated to the genotype 3a. The role of SIRT1 in HCV life cycle has not been fully investigated. However an increase in its expression could be interpreted either as an indirect effect on the Wnt pathway within hepatocytes, since SIRT1 inhibition reverses the E-cadherin downregulation (Ripoli et al, 2011) and the DNMT3b expression (Benegiamo et al, 2012) due to HCV core protein, or as a compensatory mechanism of the hepatocyte, since high levels of SIRT1 protect against lipid accumulation (Podrini et al, 2012; Pfluger et al, 2008). SIRT1 can
directly deacetylate and inhibit SREBF1 (Figure 3) (Ponugoti et al, 2010). In turn, gene expression of Stearoyl-CoA Desaturase 1 (SCD), which synthesizes oleate necessary for the biosynthesis of triglycerides and other lipids is mediated by SREBF1 (Figure 3) (Myazaki et al, 2004).

SIRT1 can increase the hepatic expression of another member of the sirtuin family, SIRT6, by forming a complex with FOXO3a and NRF1 on the SIRT6 promoter. This process can, in turn, negatively regulate glycolysis, triglyceride synthesis and fat metabolism, by blocking SREBF1 (Kim et al, 2010). Consistently, SIRT1 activates also PPARGC1A and PPARA, nuclear factors that mediate the adaptive liver response to fasting and starvation (Purushotham et al, 2009) (Figure 3), and NR1H3 [also known as liver X receptor (LXR)], a nuclear receptor that functions as a cholesterol sensor and regulates the whole-body cholesterol (Li et al, 2007). Upregulation of LXR leads to an increased expression of the regulator of cholesterol efflux ABCA1 (Li et al, 2007). Not surprisingly, the dysregulation of miR33a/b observed in this study could also lead to an increase in ABCA1 and SIRT6 (Davalos et al 2011; Gerin et al, 2010) (Figure 3). Interestingly, among the various genes central to this pathway and dysregulated by both HCV genotypes are SIRT1, SCD and ABCA1 (Figure 3).

Thus, our analyses demonstrated that, although the two clusters of genes found dysregulated by HCV of different genotypes 1b and 3a are dissimilar, from a functional point of view they contain a cohort of genes involved in the same biological process.

System approaches to disease are grounded in the idea that disease-perturbed protein and gene regulatory networks differ from their normal counterparts and they open new perspectives toward systemic understanding of perturbations induced e.g. during viral infections. Deciphering pathogen-host interactions also relies on the
analytical and predictive power of systems biology computer models taking into account the whole picture of the molecules and their interactions. Our study represents the first step of a systems biology approach aimed at the identification of a limited number of molecular entities explaining the functional differences between the responses triggered by HCV genotype 1b and 3a infection in hepatocytes. Further experimental validation studies will be required before they can be considered as useful targets for the treatment of HCV infections.

Acknowledgments

This research was supported by the “Ricerca Corrente” RC1103GA48 to VP and RC1201MD14 to TM funding granted by the Italian Ministry of Health and by the “5x1000” voluntary contributions to our Hospital. PV and MV are supported by the Italian Ministry of Health (GR program 2010-2311017). MV is supported by a My First AIRC Grant. No conflict of interest is declared.

Figures legends

Table 1: List of nonspecific genes identified by Endeavou and ToppGene (5th column) on a semantic model built on the lists of genes deregulated by both core protein genotypes (4th column). Bonferroni corrected p-values are shown in column 3. In red, genes that are classified as unspecific by both software on the same model. In blue the gene in common between the two lists of genes.

Figure 1

Several heterogeneous data sources (the furthest right) are used to build semantic models around the two gene lists (the furthest left). Functional proximity is measured between each genes list and the model built around the other list. The closest genes populate an interaction network and are deemed nonspecific.
Figure 2
The diagram summarizes the interactions among the nonspecific genes resulting from the analysis. Violet edges correspond to co-expression evidence of pairs of genes. Cyan edges associate genes participating to the same pathways. Red and orange edges denote physical and predicted interactions, respectively. Thickness of an edge measures its reliability.

Figure 3
Picture drawn with the Systems Biology Graphical Notation (SBGN). SIRT-centered pathway involving the HCV core proteins responsible of cholesterol and lipid homeostasis. Arrows stand for activation, while bar-headed lines mean inhibition of the activities of the target species.

References


<table>
<thead>
<tr>
<th>Gene</th>
<th>P-Value</th>
<th>Genotype</th>
<th>Software</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABCA1</td>
<td>1.384E-6</td>
<td>1b</td>
<td>Endeavour</td>
</tr>
<tr>
<td>CYP26B1</td>
<td>5.080E-6</td>
<td>1b</td>
<td>Endeavour</td>
</tr>
<tr>
<td>KHDRBS1</td>
<td>0.004</td>
<td>1b</td>
<td>Endeavour</td>
</tr>
<tr>
<td>YWHAB</td>
<td>0.038</td>
<td>1b</td>
<td>Endeavour</td>
</tr>
<tr>
<td>APOC1</td>
<td>3.014E-7</td>
<td>1b</td>
<td>ToppGene</td>
</tr>
<tr>
<td>EDN1</td>
<td>3.475E-4</td>
<td>1b</td>
<td>ToppGene</td>
</tr>
<tr>
<td>PHF15</td>
<td>9.126E-3</td>
<td>1b</td>
<td>ToppGene</td>
</tr>
<tr>
<td>SOD2</td>
<td>1.599E-2</td>
<td>1b</td>
<td>ToppGene</td>
</tr>
<tr>
<td>ACER3</td>
<td>1.909E-2</td>
<td>1b</td>
<td>ToppGene</td>
</tr>
<tr>
<td>MAP4K4</td>
<td>1.911E-2</td>
<td>1b</td>
<td>ToppGene</td>
</tr>
<tr>
<td>TNPO1</td>
<td>2.483E-2</td>
<td>1b</td>
<td>ToppGene</td>
</tr>
<tr>
<td>SCD</td>
<td>2.736E-2</td>
<td>1b</td>
<td>ToppGene</td>
</tr>
<tr>
<td>FNDC3B</td>
<td>4.728E-2</td>
<td>1b</td>
<td>ToppGene</td>
</tr>
<tr>
<td>ABCA1</td>
<td>6.990E-4</td>
<td>3a</td>
<td>Endeavour</td>
</tr>
<tr>
<td>PSMD8</td>
<td>0.036</td>
<td>3a</td>
<td>Endeavour</td>
</tr>
<tr>
<td>PCTK2</td>
<td>0.040</td>
<td>3a</td>
<td>Endeavour</td>
</tr>
<tr>
<td>ABI1</td>
<td>0.041</td>
<td>3a</td>
<td>Endeavour</td>
</tr>
<tr>
<td>YES1</td>
<td>0.042</td>
<td>3a</td>
<td>Endeavour</td>
</tr>
<tr>
<td>CREB1</td>
<td>0.044</td>
<td>3a</td>
<td>Endeavour</td>
</tr>
<tr>
<td>SFRS7</td>
<td>0.046</td>
<td>3a</td>
<td>Endeavour</td>
</tr>
<tr>
<td>SIRT1</td>
<td>1.54E-4</td>
<td>3a</td>
<td>ToppGene</td>
</tr>
<tr>
<td>CREB1</td>
<td>3.34E-4</td>
<td>3a</td>
<td>ToppGene</td>
</tr>
<tr>
<td>YES1</td>
<td>0.002</td>
<td>3a</td>
<td>ToppGene</td>
</tr>
<tr>
<td>CLK1</td>
<td>0.004</td>
<td>3a</td>
<td>ToppGene</td>
</tr>
<tr>
<td>IL8</td>
<td>0.005</td>
<td>3a</td>
<td>ToppGene</td>
</tr>
<tr>
<td>ABI1</td>
<td>0.005</td>
<td>3a</td>
<td>ToppGene</td>
</tr>
<tr>
<td>RHOU</td>
<td>0.007</td>
<td>3a</td>
<td>ToppGene</td>
</tr>
<tr>
<td>LPGAT1</td>
<td>0.008</td>
<td>3a</td>
<td>ToppGene</td>
</tr>
<tr>
<td>TAGLN</td>
<td>0.009</td>
<td>3a</td>
<td>ToppGene</td>
</tr>
<tr>
<td>PNN</td>
<td>0.012</td>
<td>3a</td>
<td>ToppGene</td>
</tr>
<tr>
<td>CSPP1</td>
<td>0.013</td>
<td>3a</td>
<td>ToppGene</td>
</tr>
<tr>
<td>TTR</td>
<td>0.014</td>
<td>3a</td>
<td>ToppGene</td>
</tr>
<tr>
<td>DLX2</td>
<td>0.016</td>
<td>3a</td>
<td>ToppGene</td>
</tr>
<tr>
<td>NR2F2</td>
<td>0.017</td>
<td>3a</td>
<td>ToppGene</td>
</tr>
<tr>
<td>CDK17</td>
<td>0.021</td>
<td>3a</td>
<td>ToppGene</td>
</tr>
<tr>
<td>ZFP36L1</td>
<td>0.022</td>
<td>3a</td>
<td>ToppGene</td>
</tr>
<tr>
<td>TANC1</td>
<td>0.028</td>
<td>3a</td>
<td>ToppGene</td>
</tr>
<tr>
<td>CDH17</td>
<td>0.030</td>
<td>3a</td>
<td>ToppGene</td>
</tr>
<tr>
<td>PTBP2</td>
<td>0.044</td>
<td>3a</td>
<td>ToppGene</td>
</tr>
</tbody>
</table>

Table 1.