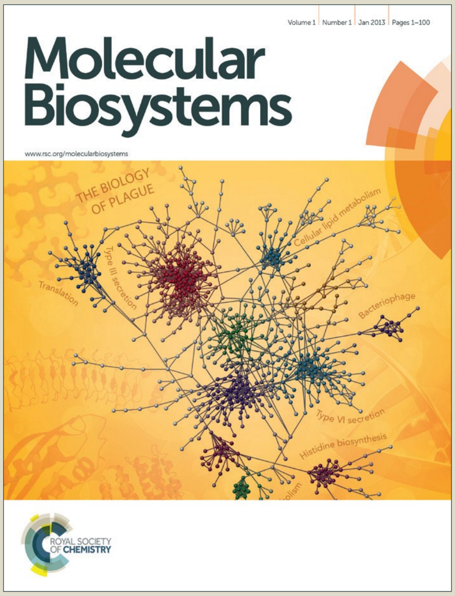
Molecular BioSystems


Volume 1 | Number 1 | Jan 2013 | Pages 1–100
Molecular Biosystems
www.rsc.org/molecularbiosystems
THE BIOLOGY OF PLAGUE
Cellular lipid metabolism
Type III secretion
Translation
Bacteriophage
Type VI secretion
Histidine biosynthesis
ROYAL SOCIETY OF CHEMISTRY

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems

Molecular BioSystems

PAPER

The relationships among host transcriptional responses reveal distinct signatures underlying viral infection-disease associations

Received 00th October 2015,
Accepted 00th October 2015

DOI: 10.1039/x0xx00000x

www.rsc.org/

Lu Han,‡[a,b] Haochen He,‡[b] Xinyan Qu,‡[b] Yang Liu,[b] Song He,[b] Xiaofei Zheng,[c] Fuchu He,[b] Hui Bai*[bd] and Xiaochen Bo*[b]

Genome-scale DNA microarray and computational biology facilitate new understanding of viral infections at system level. Recent years have witnessed a major shift from microorganism-centric toward host-oriented characterization and categorization of viral infections and infection related diseases. We established host transcriptional response (HTR) relationships among 23 different types of human viral pathogens based on calculating HTR similarities using computationally integrated 587 public available gene expression profiles. We further identified five virus clusters that show consensus internal HTRs and defined cluster signatures using common differentially regulated genes. Individual cluster signature genes distinguish from each other, and functional analysis revealed common and specific host cellular bioprocesses and signaling pathways involved in confrontation to viral infections. Through literature investigation and support from epidemiological studies, they were confirmed to be important gene factors associating viral infections with cluster-common and -specific non-infectious human disease(s). Our analyses were the first to feature differential HTRs to viral infections as clusters, and suggest new perspective of understanding infection-disease associations and the underlying pathogenesis.

Introduction

Viruses are etiology of many diseases, associating with leading morbidity and mortality worldwide[1]. It is well acknowledged that the complex and dynamic interaction between virus and host determines the progress and prognosis of related diseases. Importantly, infected hosts recognize the presence of viruses and mobilize specific defense mechanisms, while viruses in turn can actively modulate host-signaling pathways to enhance their persistence and survival. Thus, understanding viral infections from the perspective of host response provides insights into the molecular pathogenesis of infectious diseases[2,3], and may finally contribute to identification of novel biomarkers[4], invention of new antiviral therapies[5], discovery of potential drug targets[6], and improvement in diagnostic accuracy[7].

As a powerful high-throughput tool, genome-wide DNA microarray technology permits simultaneous interrogation of the transcriptional status of thousands of genes, which can provide a holistic picture of host transcriptional events underlying viral infections[8,9]. Along with statistical tools and gene functional annotation, comprehensive analysis based on transcriptional profiling of infection model systems (e.g., cells, tissues, and in vivo) have facilitated identifying new gene factors and deciphering exquisite host defense mechanisms[10–12]. With the publication that document gene expression changes of different host cell types upon infections of different viral pathogens increasing in recent years, researchers have begun to explore the molecular mechanism of infection in a systematic manner[13], on the basis that a set of consistently dysregulated genes (i.e., signature) may tailor cellular host response to individual pathogen[14]. However, most analysis that systematically compare host transcriptional responses (HTRs) to different pathogens from heterogeneous experiments focus on defining a common HTR gene set[6,15], whereas landscape relationships of HTRs to diverse pathogens indicating specific infection patterns have not been established.

An obstruction to such large-scale comparison of HTR results has been the wide-distribution nature of data generated by independent and sporadic transcriptional profiling experiments that utilize diverse cell types as infection models. Gene Set Enrichment Analysis (GSEA) algorithm based on the relative expression changes is a successful strategy to excavate transcriptional data applied in dozens of drug reposition studies[16–19] and thus allows comparison of expression profiles

[a] *Department of Traditional Chinese Medicine and Neuroimmunopharmacology, Beijing Institute of Pharmacology and Toxicology, Beijing, China*
[b] *Department of Biotechnology, Beijing Institute of Radiation Medicine, Beijing, China. E-mail: boxc@bmi.ac.cn, huibai13@hotmail.com; Tel: +86-10-66931207*
[c] *Department of Biochemistry and Molecular Biology, Beijing Institute of Radiation Medicine, Beijing, China*
[d] *No. 451 Hospital of Chinese People's Liberation Army, Xi'an, China*

† Footnotes relating to the title and/or authors should appear here.
Electronic Information (ESI) available. See DOI: 10.1039/x0xx00000x
‡ These authors contributed equally to this work
* Correspondence authors.

from different infection models. Here, by computational integrating 587 standard-format host cellular expression profiles upon virus inections from Gene Expression Omnibus (GEO) based on GSEA, we provide a virus-virus HTR network that presents the HTR similarities among 23 human viral pathogens. Through further dissection and comprehensive analysis, we classify and characterize HTR network of 23 viral pathogens by five clusters that show significantly consensus internal similarity, and their common dysregulated genes form differential infection patterns. Notably, we address the biological importance of cluster signature genes with respect to uncovering the associations between viral infection and human diseases and understanding the underlying pathogenesis.

Results

Virus-virus HTR network and clusters

To globally quantify the degree of HTR similarities to different viral pathogens, We exploited a repository of 50 datasets of 587 expression profiles representing *in vitro* and/or *in vivo* infected host (cell or tissue) transcriptional responses to 23 different types of viruses (Each type of virus contain multiple strains or subtypes, Table S1 for detailed information). For each virus type, we considered all the transcriptional responses following infections across different cell or tissue types as control and infected sample, thereafter we obtained a total of 587 paired samples (Methods). We ranked the 22160 validated probes of each paired sample to generate Probe Rank Lists (PRLs) based on expression changes, which represented HTRs to individual viral infection across different cell types. We combined the PRLs for the same virus by a hierarchical majority-voting scheme to a single merged PRL (mPRL)[20], representing HTR to a specific virus type. We represented the HTR similarity between two virus types as a “distance” between their mPRLs and computed it with GSEA[21–23]. *P* values and FDR values of the distances were estimated according to the distance distribution generated from one million paired random permutated lists of the same size. The distances between HTRs to different viral infections were shown as a heat map (Fig. 1).

Further, we tried to evaluate if certain viruses present more consensus internal HTRs. To assure rational dissection, we identified HTR clusters on basis of the calculated virus distance via an automatic and parameter-independent clustering algorithm, the affinity propagation cluster method[24](Methods). An “HTR cluster” is defined as a group of viruses densely interconnected with each other indicating more significant HTR similarities. As shown in Fig. 2a, we identified five HTR clusters for 23 viral pathogens with sizes ranging from 4 to 7.

Of note, despite that the expression data were collected from various infection models, viruses classified in an HTR cluster share certain infection attributes. For example, small RNA viruses EV71 and HAV are both classified in HTR cluster 2, and Flavivirus HCV and DENV are both classified in HTR cluster 4. These are demonstrations of close phylogenetic relationship among viruses within specific HTR cluster. Meanwhile, we found that viruses of the same genus may be classified into different HTR clusters, e.g. Herpesviruses clustered into HTR cluster 3 and 5. This highly indicates that infection attributes other than close phylogenetic relationship are correlated with significant HTR similarities within an HTR cluster. For example, respiratory viruses DHOV, RSV and SARS-CoV are classified in HTR cluster 1, which is a demonstration of same infection implicated organisms; Digestive tract viruses EV71, HAV and NV are classified in HTR cluster 2, which is a demonstration of same transmission route; HHV-1, HRV, and VV that cause latent infections are classified in HTR cluster 3, which is a demonstration of common infection manifestations; Tumor Virus HHV4，HHV5 and HPV are classified in HTR cluster 5, which is a demonstration of shared infection mechanisms underlying carcinogenesis. l These findings indicated that HTR-based clustering is of certain rationality, the mechanisms behind which rely on the annotation of common dysregulated host cellular genes and elucidation of infection patterns within clusters.

To characterize the general HTR features within HTR clusters, we merged the PRLs of pathogens within a cluster using Borda Merging Method to get a “cluster PRL” (Methods). The probes with the highest or lowest 250 rankings, representing a total of

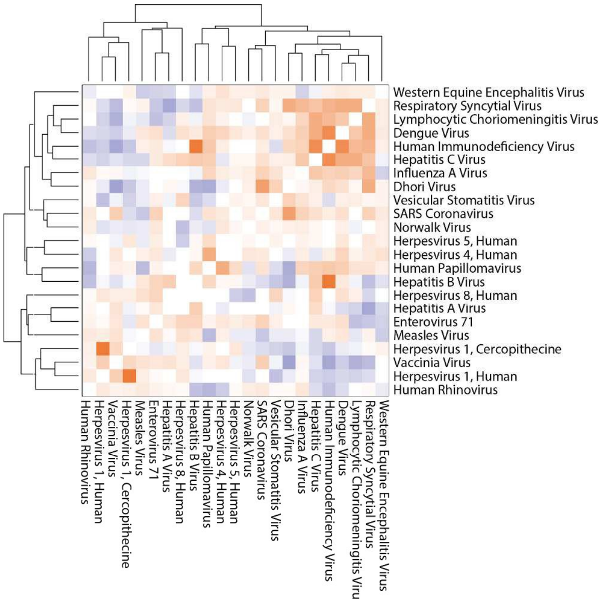


Fig. 1 Heat map presentation of virus-virus host transcriptional response (HTR) similarity. The squares in the heat map were colored according to the distances between HTRs to different virus infections estimated based on Gene Set Enrichment Analysis (Methods). The distance value between each two viruses varied from 0 to 2. And distances below 1 (orange squares) indicated similar HTRs, while those above 1 (blue squares) indicated opposite HTRs. The darkness of colored squares indicates the distances’ offset to 1. All distances and their significance information were provided in Table S2.

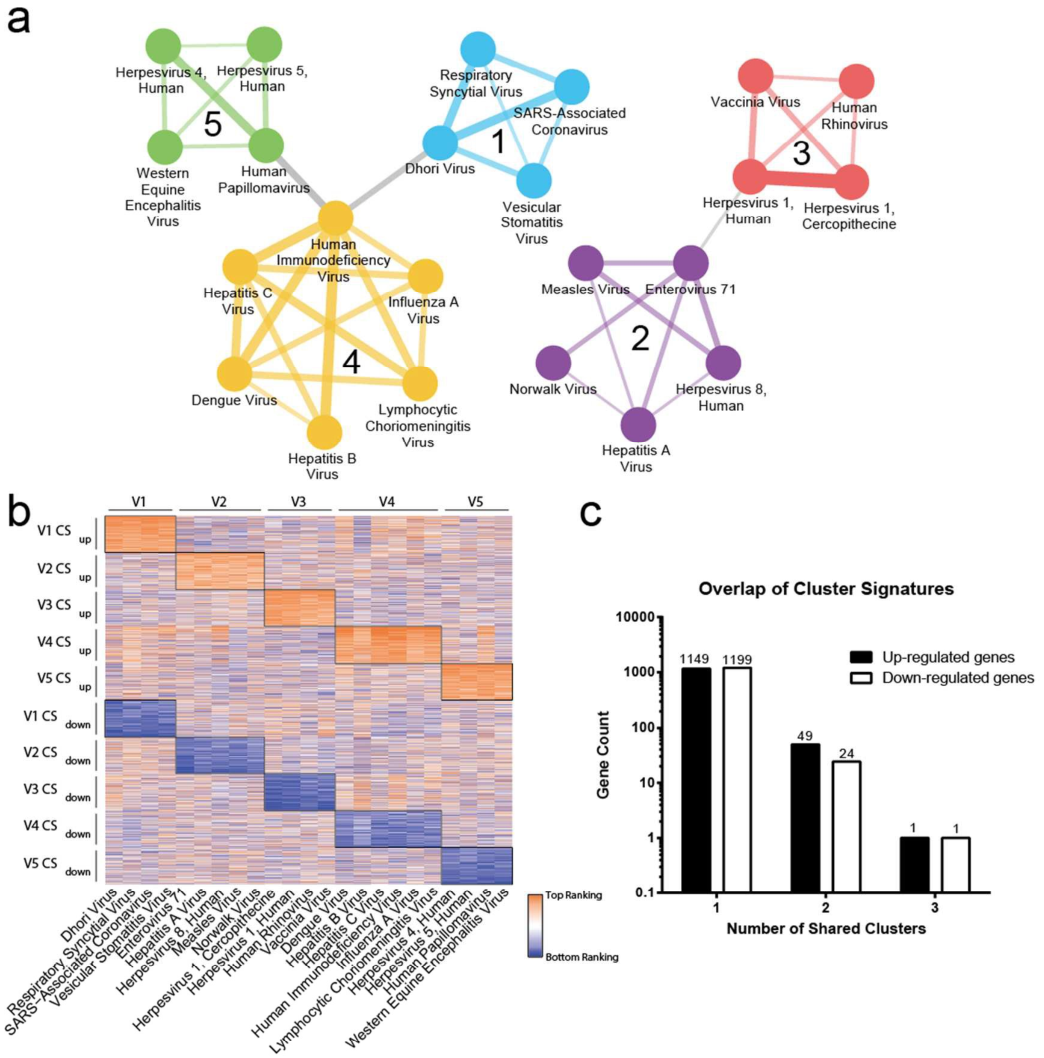


Fig. 2 HTR network and clusters. (a) Clusters are identified based on the calculated virus distance (Methods), and numerically labelled according to the alphabetical precedence of the exemplar (i.e., the virus whose HTR best represents the HTRs of the other viruses within a cluster). Each node shows a viral species, and thick edge connecting two viruses represents significant HTR similarity between two viruses (virus distance < 0.01), while thin edge shows HTR similarity between two viruses (distance < 1). Nodes and edges within a cluster are differentially coloured, whereas edges connecting the "exemplar" viruses in different clusters are colored grey. (b) Cluster signatures (CSs). Synthetic heat map of specific 500 signature genes of each cluster (whose probes with the highest or lowest 250 rankings in cluster PRL). The up- and down- regulated signature genes for each cluster (highlighted in solid-line frame) are depicted on the left. To allow better visualization of the representative characteristics of CS, 500 signature genes for each component virus within a cluster were separately listed and integrated into an amount of 2500 genes (vertically listed as squares). Orange squares indicate the top ranking genes and blue squares indicate the bottom ranking genes. Ranking values for 2500 genes are reproduced in Table S3. (c) Gene counts in the up and down-regulated signature genes with respect to the number of clusters (see Table S4 for detailed information).

500 most consistently dysregulated genes for each cluster PRL, were selected to be cluster signature, i.e., a common set of genes that was most consistently up- (or down-) regulated (Table S3). A heat map representation of the rankings of 500 dysregulated signature genes of each cluster as compared to those of component HTR to individual virus infection was depicted (Fig. 2b).

Obviously, cluster signatures genes could in general represent the consistent expression level of the same gene sets in the HTR to individual component viral infections, while distinguish from those to external viral infections. Through further statistical analysis, we found that the signature genes across five HTR clusters showed limited overlap (Fig. 2c and Table S4). Notably, there was no signature gene identified as being universal to all HTR clusters. These results highlighted the importance of HTR cluster signature genes in revealing the differentiating characters of individual HTR cluster.

HTR cluster based infection patterns

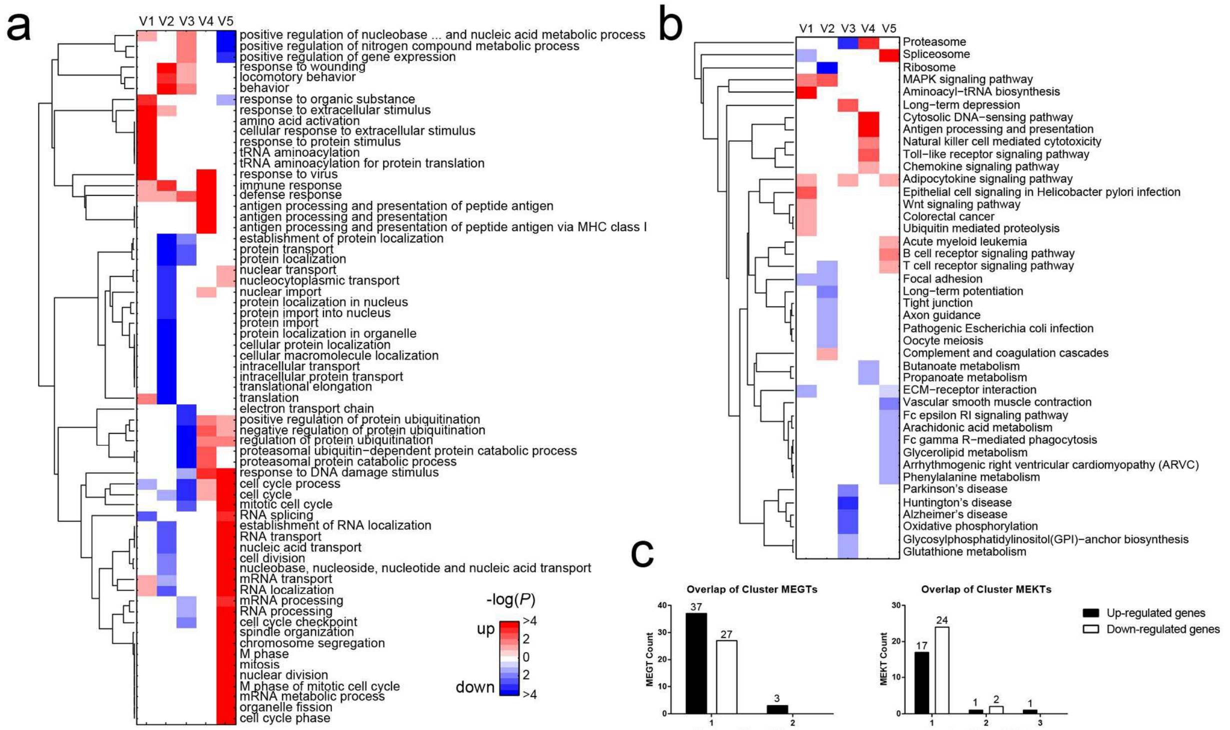


Fig. 3 Functional annotation and clustering of cluster signature (CS) genes. Heat map presentation of (a) most enriched GO biological processes (BP) terms (MEGTs, p < 0.001) and (b) the most enriched KEGG pathways (MEKTs, p < 0.01) according to their enrichment p values. The color scale indicates the significance of the enrichment, with red representing that in up-regulated signature genes and blue representing that in down-regulated signature genes. The order of gene families is determined by hierarchical clustering (Methods). Annotation is given according to GO BP and KEGG nomenclatures. (c) MEGTs and MEKTs count with respect to the number of clusters (see Table S5 and S6 for detailed information). Note: As no up- and down-regulated signature genes are both enriched in certain terms in this figure, the up- and down MEGTs or MEKTs are presented in the same heat map.

The distinct nature of cluster signature genes is highly possible to unravel differential expression patterns that host cells utilize to confront infections of component viruses within an HTR cluster. To this end, we mapped signature genes to Gene Ontology (GO) Biological Processes (BP) and Kyoto Encyclopedia of Genes and Genomes (KEGG) for functional annotations, which were further hierarchically clustered according to enrichment scores (-$\log_{10}$ of *p*). The most enriched GO BP terms (MEGTs, $p < 0.001$) and KEGG pathway terms (MEKTs, $p < 0.\ 1$) clustered from up- and down-regulated signature genes were shown in a heat map representation of Fig. 3 (see Table S5 and S6 for detailed information). Most MEGTs and MEKTs were HTR cluster specific, while BPs or pathways closely related to virus replication and host defense tended to be common among different HTR clusters. For example, the defense and immune responses were enriched in the up-signatures in HTR cluster V1, V2, V3 and V4. The cell cycle and nucleic acid metabolic processes were also enriched in multiple HTR cluster signatures, but differed in regulation orientation.

Cluster-specific BPs or pathways further explicit the distinct infection mechanisms for individual HTR cluster. HTR Cluster V1 contains three viruses that usually cause respiratory tract infections and acute inflammations[25–27]. Accordingly, BPs and pathways specifically dysregulated in HTR cluster V1 were closely related to epithelial cell responses to stimulus such as responses to extracellular stimulus, epithelial cell signaling in Helicobacter pylori infection, and Wnt Signaling pathway. HTR Cluster V2 that includes two small RNA viruses, HAV and EV71, showed specific down regulation in the inhibition of protein import and localization. Four viruses that cause subclinical or latent infections were classified in HTR cluster V3, which showed specifically significant down regulation in protein catabolic and cell cycle processes. Moreover, neurodegenerative diseases related pathways were also specifically dysregulated in HTR cluster V3, in spite of the fact that none of the source cells used to generate expression profiles upon infections of V3 component viruses was neurocytes. Intriguingly, HHV-1 is one of the leading infectious agents strongly proposed as potential cause of Alzheimer's disease for its neurotropic feature and lifelong latency[28]. Viruses in HTR cluster V4, including HIV, HBV, HCV and LCMV that usually lead to chronic infections causing long-term inflammation and immune-response mediated cell injury in the host, as well as DENV and IAV that usually causes acute inflammation and high fever, showed strongest internal HTR similarities. As shown in Fig. 3a, activation of antigen processing and presentation was a specifically up regulated signaling pathway for HTR cluster V4. This feature helps component viruses distinguish from viruses in other clusters

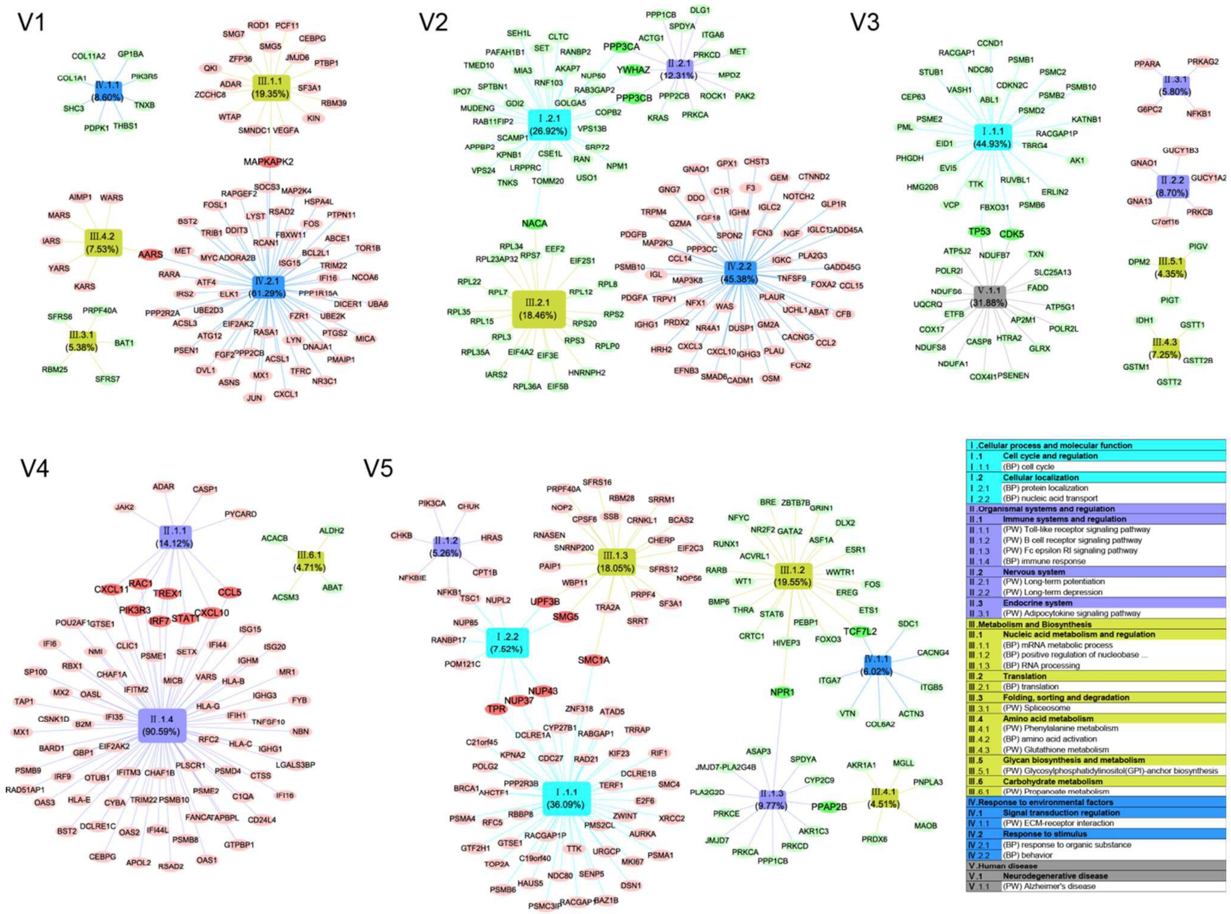

V1
IV.1.1 (8.60%)
III.1.1 (19.35%)
III.4.2 (7.53%)
IV.2.1 (61.29%)
III.3.1 (5.38%)
V2
I.2.1 (26.92%)
II.2.1 (12.31%)
III.2.1 (18.46%)
IV.2.2 (45.38%)
V3
I.1.1 (44.93%)
II.3.1 (5.80%)
II.2.2 (8.70%)
III.5.1 (4.35%)
V.1.1 (31.88%)
III.4.3 (7.25%)
V4
II.1.1 (14.12%)
III.6.1 (4.71%)
II.1.4 (90.59%)
V5
II.1.2 (5.26%)
III.1.3 (18.05%)
III.1.2 (19.55%)
I.2.2 (7.52%)
IV.1.1 (6.02%)
I.1.1 (36.09%)
II.1.3 (9.77%)
III.4.1 (4.51%)
I .Cellular process and molecular function
I .1 Cell cycle and regulation
I .1.1 (BP) cell cycle
I .2 Cellular localization
I .2.1 (BP) protein localization
I .2.2 (BP) nucleic acid transport
II .Organismal systems and regulation
II .1 Immune systems and regulation
II .1.1 (PW) Toll-like receptor signaling pathway
II .1.2 (PW) B cell receptor signaling pathway
II .1.3 (PW) Fc epsilon RI signaling pathway
II .1.4 (BP) immune response
II .2 Nervous system
II .2.1 (PW) Long-term potentiation
II .2.2 (PW) Long-term depression
II .3 Endocrine system
II .3.1 (PW) Adipocytokine signaling pathway
III .Metabolism and Biosynthesis
III .1 Nucleic acid metabolism and regulation
III .1.1 (BP) mRNA metabolic process
III .1.2 (BP) positive regulation of nucleobase ...
III .1.3 (BP) RNA processing
III .2 Translation
III .2.1 (BP) translation
III .3 Folding, sorting and degradation
III .3.1 (PW) Spliceosome
III .4 Amino acid metabolism
III .4.1 (PW) Phenylalanine metabolism
III .4.2 (BP) amino acid activation
III .4.3 (PW) Glutathione metabolism
III .5 Glycan biosynthesis and metabolism
III .5.1 (PW) Glycosylphosphatidylinositol(GPI)-anchor biosynthesis
III .6 Carbohydrate metabolism
III .6.1 (PW) Propanoate metabolism
IV.Response to environmental factors
IV.1 Signal transduction regulation
IV.1.1 (PW) ECM-receptor interaction
IV.2 Response to stimulus
IV.2.1 (BP) response to organic substance
IV.2.2 (BP) behavior
V.Human disease
V.1 Neurodegenerative disease
V.1.1 (PW) Alzheimer's disease

Fig. 4. Functional annotation clustering network (FACN) of HTR cluster signature genes. Network represents "signature gene-function exemplar" associations visualized by network edges that are colored according to the manually categorized first-level classification term. Exemplar is the best representative of the grouped functional terms, i.e., most enriched GO biological processes (BPs) terms (MEGTs, $p < 0.001$) or enriched KEGG pathway terms (MEKTs, $p < 0.1$) elected automatically by the affinity propagation cluster algorithm (see Table S7 for detailed information). The size of exemplar indicates the average enrichment score of all clustered MEGTs and/or MEKTs. The percentage of clustered gene count in total gene count of a functional term cluster is denoted under each exemplar. Red and green nodes indicate up- and down- regulated signature genes, respectively. The color scale indicates the number of FACs that individual signature gene participated.

that also induce immune response, e.g., viruses in HTR cluster V1 prominently activate signaling pathways of cell responses. HTR cluster V5 including three DNA tumor viruses showed that cell cycle related processes, mitosis, and nuclear division were specifically activated, while transcription and several signaling pathways were specifically inhibited. These were consistent with the fact that infections of the three DNA tumor viruses increase cell proliferation.

To better characterize the infection patterns in individual cluster, we extracted the representative feature of functionally annotated cluster-enriched BPs and pathways through further clustering. To this end, we built each cluster an exemplar-centric "functional annotation clustering network" (FACN), in which a total of 277 MEGTs ($p < 0.001$) and 83 MEKTs ($p < 0.1$) were further clustered using the shared signature gene numbers between enriched terms as clustering weight[24] (see Methods and Table S7). The functional term cluster was composed of genes involved in functional categories of similar gene composition, and the algorithm automatically elected a functional term as the cluster exemplar, defined as the best representative term within the group. For better annotation, we manually classified all exemplars into a three-level category (see Methods). As illustrated in Fig. 4, the gene expression pattern of each HTR cluster signature can be differentially characterized by several exemplar-named enriched functional annotation clusters (EFACs) of varied quantitative preponderance. The same orientation of host cellular gene regulation upon cluster virus infections could also be reflected. In this way, the number of signature genes involved in each cluster-enriched BP or signaling pathway and their mutual correspondences can be depicted, especially those genes involved in different BPs or signaling pathways.

In general, HTR cluster signatures were characterized by limited EFACs in which specific functional gene clusters (BPs and/or signaling pathways) were preferentially elicited or subverted by cluster viruses (Fig. 4). HTR Cluster V1 was of predominant EFAC genes up-regulated in cellular antiviral response (61.29%) and molecular metabolism such as mRNA metabolic process and amino acid activation (a total of 26.88%), whereas a few genes down-regulated in extracellular signal transduction pathways (8.6%) and spliceosome (5.38%). Notably, MAPKAPK2 and AARS are the two significantly activated signature genes associating three major EFACs. Cluster V2 was of EFAC genes down-regulated in translation and protein localization related BPs (a total of 57.69%), as well as up-regulated in immune response related BPs and pathways (45.38%). PPP3CA, PPP3CB, YWHAZ, and NACA are the highly inhibited signature genes associating three minor EFACs. Consistently, HTR cluster V3 was of predominant EFAC genes down-regulated in cellular processes such as cell cycle (44.93%) and neurodegenerative disease related pathways (31.88%), while the up-regulated EFAC genes were involved in nervous and endocrine systems regulation pathways of long-term depression and adipocytokine signaling pathway (a total of 14.5%). Notably, TP53 and CDK5 are functionally activated associating two major EFACs. HTR cluster V4 was of EFAC genes overwhelmingly up-regulated in innate immune response related BPs and pathways (a total of 104.71%), in which CXCL10, CXCL11, CCL5, RAC1, TREX1, IRF7, and STAT1 coordinately associate with EFAC representing Toll-like receptor signaling pathway. However, down-regulated EFAC genes of HTR cluster V4 were only 4.71%, involved in propanoate metabolism pathway. Approximately 60% of EFAC genes of HTR cluster V5 were up-regulated in cellular processes such as cell cycle and RNA localization (a total of 46.62%) plus molecular metabolism such as RNA processing (20.3%) and adaptive immune response (5.26%). And the three major up-regulated EFACs use UPF3B, SMG5, SMC1A, NUP43, NUP37, and TPR to be functionally associated. The left 40% signature genes in V5 were down-regulated in nucleic acid metabolic process (19.55%), environmental signal transduction and inflammatory response activation, with NPR1, PPAP2B and TCF7L2 associating each other.

Taken together, these results facilitated the demonstration of distinguished feature of HTRs to viral infections as cluster components, and highlighted that viral infections are indeed of differential patterns, in which several genes play important roles in functional associating diverse gene clusters (BPs and/or signaling pathways).

Cluster signatures associate viral infection with disease

The above results highlighted a bunch of signature genes as key modulators for each cluster, the dysregulation of which upon infection may perturb multiple host cellular functions of biological essentiality and thereby affect the consequences where infection may lead to. In addition, certain cluster signatures showed several disease-related signaling pathways specifically induced or inhibited, highly indicating the etiological role that cluster viruses may play in infection-associated human diseases. For example, signaling pathways of colorectal cancer and acute myeloid leukemia were significantly up regulated in viral infections of HTR cluster V1 and V5, respectively; several signaling pathways in neurodegenerative diseases, including Parkinson's disease, Huntington's disease, and Alzheimer's disease, were down regulated for viral infections in HTR cluster V3.

Of the many databases storing human diseases related genes, Genome-Wide Association Studies (GWAS) catalog (http://www.genome.gov/gwastudies, accessed July 15, 2013) was advantageous in way of large-population based experiment design and rigorous statistics with less research bias. Therefore, we used data in GWAS as reference disease-related genes and the plain guilt by association strategy to build infection-disease relations. In specific, we assessed the relevance of cluster signatures to human disease(s) by mapping them to candidate genes adjacent to single nucleotide polymorphisms (SNPs) that have been annotated as being associated with various diseases or disease-related intermediate phenotypes in GWAS catalog.

To explore the possible associations between viral infections and human diseases, we built a viral infection-disease network (Fig. 5a), in which HTR cluster was connected by an edge to a well-defined human disease if the mapped genes of 3 different disease-associated SNPs were shared in viral cluster signature (Methods and Table S9). In general, 57 virus cluster infection-disease associations were identified to designate a total of 25 human diseases to all HTR clusters, and 9 diseases were specific for individual HTR cluster. Notably, 12 out of 25 diseases were found to be of well-acknowledged relationships with chronic inflammation and/or immune dysfunction (highlighted as hexagons with black circle in Fig. 5a). And 4 of them were among the top ranking of the 5 most common diseases (with connection to at least four virus clusters) identified to associate with cluster viruses. This highlighted a shared inflammation and/or immune related background for viral infection-associated diseases. Overall, virus clusters are mainly associated with mental and behavior disorders (25%), diseases of the digestive system (25%) and of the musculoskeletal system and connective tissue (12%), as well as malignant neoplasms (10%) (Fig. 5b).

To further validate the reliability of associations established for cluster viral infections and human diseases, we searched literature support in PubMed for a total of 256 viral infection-disease associations with respect to individual cluster virus. Generally, the keywords for searching literatures related specific viral infection-disease associations are in the form of "xxx virus AND xxx (i.e., GWAS disease name). We have virologist perform the expert interpretation of Abstracts and full articles when necessary to identify the association types and designate publications that support the identified association.

To be noted, most reports of infection-disease associations have been epidemiological studies. Therefore, statements of significant correlations between viral infection and non-infectious disease concluded from large sample studies as well as those from small-sample studies suggesting viral infection as risk factors for certain disease, have both been adopted.

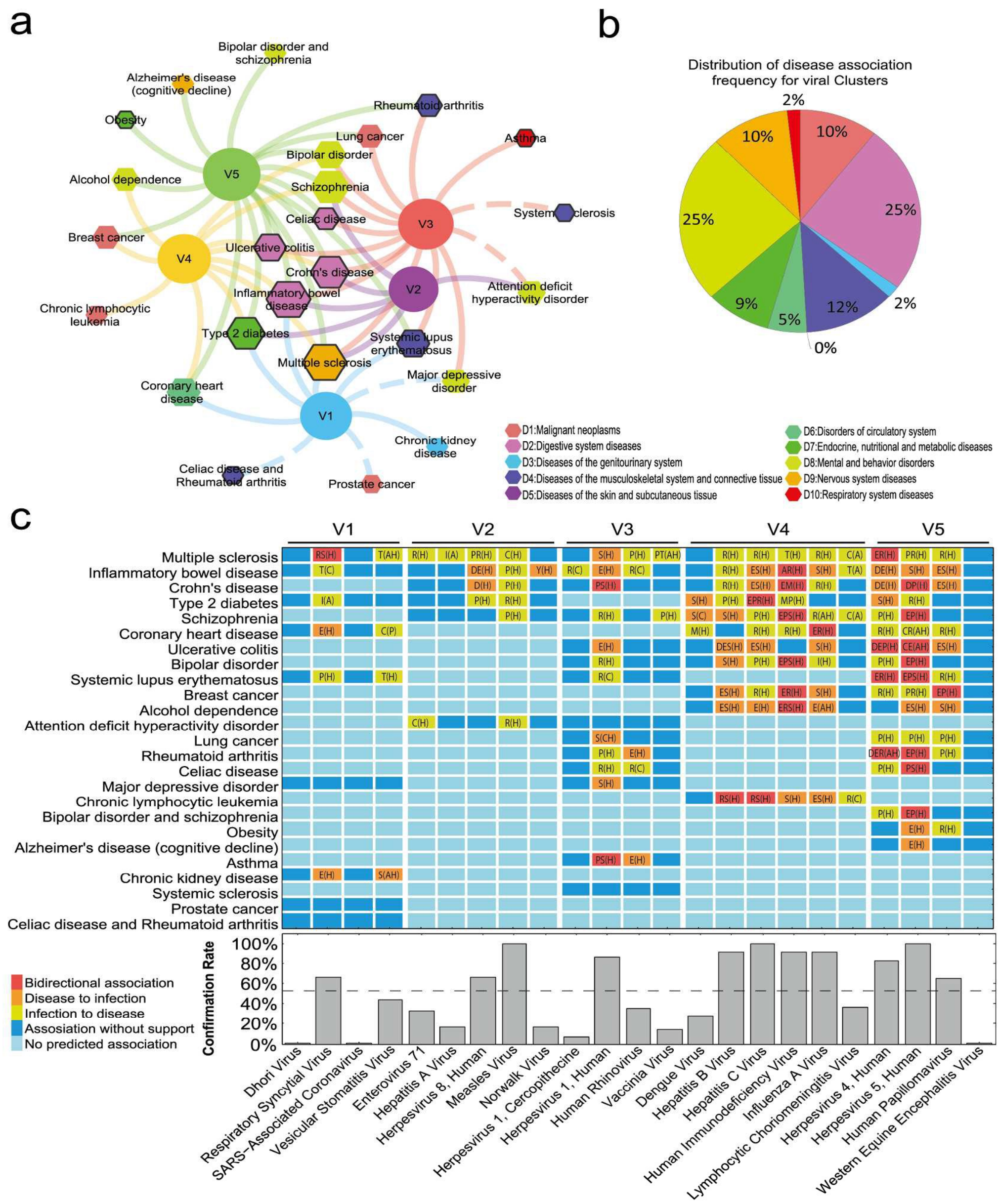


Fig. 5. Viral infection-disease associations. (a) Viral infection-disease network based on comparison of cluster-specific signature genes with Genome-Wide Association Studies (GWAS) diseases catalogue (see Methods and Table S9). Cluster infection-disease association(s) without current literature support for all composing viruses are highlighted by edges as dashed line(s). Diseases are classified into 10 categories according to the International Classification of Diseases (ICD) and are differentially colored. The size of the hexagon indicates the number of associated clusters. Diseases of well-acknowledged relationship with chronic inflammation and/or immune dysfunction are highlighted as hexagons with black circle. (b) Distribution pattern of associated disease(s) for virus clusters according to ICD classification. (c) Annotation of identified cluster infection-disease association based on literature investigation of individual virus and confirmation rate. Infection-disease association is shown as a matrix with rows representing human diseases (listed in a descending order according to the number of shared clusters) and columns representing composing viruses of individual cluster. The infection-disease association is generally represented by colored squares. Specific infection-disease association for individual cluster virus is represented by a combination of uppercase initial letters of key words (see Results). Subject types in studies supporting infection-disease association include Human (including mother to child transition), Animal model (mice, rats or rodents), Cell (in vitro), and Pig. Confirmation rate of literature support for cluster infection-disease associations with respect to individual component virus is illustrated under the heat map. The dashed red lines represent the average confirmation rates for cluster viruses.

With regard to the 57 cluster infection-disease associations, 52 can be correctly inferred from literatures for at least one component virus. Although literature support is overwhelmingly epidemiological studies on human subjects, it demonstrated a potentially causal nature of viral infection-disease association with bidirectional possibility, i.e., the "viral infection to disease” association and “disease to viral infection” association. In general, infection-disease associations (Fig. 5c) can be further summarized as 1) infectious virus is a **C**ausative agent of disease, 2) viral infection increases the **R**isk of disease, 3) disease is one of the clinical **M**anifestations of viral infection, 4) virus test shows **P**ositive or abnormal **P**revalence (usually higher) for subjects with disease, 5) viral infection **T**riggers disease though molecular mimicry mechanism, 6) viral infection **I**ndirectly leads to disease. The “disease to infection” associations include 7) subjects with disease are **S**usceptible to viral infection (including the condition after immunosuppressed **D**rug therapy), 8) viral infection **E**xacerbates disease, and 9) viral infection **A**meliorates disease (usually autoimmune diseases). Specifically, 26 viral infection-disease associations showed bidirectional causal possibilities/potentials for 9 viruses from four viral clusters. And 108 showed one-way causal possibilities/potentials, of which 41 were “disease to viral infection” and 67 were “viral infection to disease”. This provides implications of shared genetic background with respect to host cellular gene dysfunction under either condition of specific viral infection or human disease(s).

Besides, for the well-characterized viruses within a cluster, such as human immunodeficiency virus (HIV), hepatitis B virus (HBV), human herpesvirus 1, dengue virus (DENV), and Influenza A Virus (IAV), > 80% viral infection-disease associations were of literature support. Notably, viral infection-disease associations for measles virus, hepatitis C virus (HCV), and human herpesvirus 5 (also human cytomegalovirus, HCMV) were 100% supported by literature. Importantly, for viral infection-disease associations with solid footing, e.g., HIV associated with coronary heart disease[29–31] and Crohn’s disease[32], HHV-4 (also Epstein Barr virus, EBV) associated with multiple sclerosis[33], there also are pathological studies that show consistency of dysregulation of cluster-specific signature genes as shared underlying mechanism for viral infections and human diseases (Note S1). These further confirmed the critical role of signature genes in mediating multiple disorders and their credibility in identifying complex infection-disease associations as well as interpreting the underlying pathogenesis.

Discussion

The specificity and representativeness of identified cluster signatures concords with cluster identification results with respect to discrimination excellence. Meanwhile, our results show that identification of specific functional gene clusters is crucial in understanding the differential expression patterns of cluster signatures, whereas further annotation clustering analysis is capable of characterizing the general and common features of cluster component HTRs. More importantly, our results demonstrate the biological importance of signature genes in way of uncovering associations between human diseases and cluster viral infections. The confirmation rate (~65%) for cluster disease-infection associations on basis of individual component pathogen is obtained unitarily by literature investigation largely composed of epidemiological studies, and is partly affected by the limiting contribution of poorly-studied component pathogen(s).

Our findings are directly applicable to comparison of virus-virus HTR similarity and HTR cluster identification at larger scale, the reliability of which is ensured by our mixture use of methodologies. GSEA is a nonparametric, cross-platform, rank-based pattern-matching strategy capable of revealing hidden relations between perturbations of different characteristics on the same subject by comparing individual expression profile[16,18,34,35]. A hierarchical majority-voting scheme has been invented and implemented to integrate expression profiles from multiple heterogeneous experiments under the same perturbation factor(s)[16,36,37]. Thus, with consideration of asymmetric distribution of microarray data, we applied this scheme to generate PRL for individual virus, whereas equally-weighted majority-voting scheme was applied to generate cluster PRL. This avoids poor representation of datasets with small replication number for the heterogeneous experimental settings of collected datasets. And the contribution of each component virus’s expression profile to cluster signature is equally weighed. Meanwhile, identification virus clusters and their signatures advance the classification and characterization of viruses from the perspective of HTRs. This also offers a general approach to understand the molecular mechanisms underlying infections of the less-known from the well-known pathogens within clusters.

Moreover, our results are of significant interest to both biomedical researches exploring infection-related diseases and pharmaceutical industries developing broad-spectrum antimicrobials based on the host-targeting strategy. To this end, we provide a diagram of cluster infection-disease associations that accept support mostly from epidemiological studies disposing directionality potentials. With the increasing evidence from pathophysiological studies, the essential roles of specific signature genes and their therapeutic potential are being confirmed. Altogether, our analysis introduces new aspects in understanding the pathogenesis of infection-associated diseases, and emphasizes a shared genetic background in etiology interpretation.

The construction of our HTR network and identification of sub-network clusters largely relies on the availability of stand-format expression profiles of different cell types uninfected or upon infection of diverse pathogens in an endpoint-comparison mode, which limits their coverage. However, with the rapid growth of Library of Integrated Network-Based Cellular Signatures (LINCS) Program (http://www.lincsproject.org/), the power of expression data in building a network-based comprehensive view of disease states, drug actions and even infections is being well acknowledged. Thus, expression profile data of broader virus taxonomical diversity will become available and readily compatible to further improve the coverage of HTRN and sub-network clusters. Also, this information is important for further annotating the HTR similarities according to virus taxonomy. Moreover, incorporating more data of cell-type diversity will especially increase the characterization integrity of pathogen HTR individually or as a cluster component, and thereby uncover the common molecular mechanisms that different cell types coordinately initiate to maximize the likelihood of host detection of and confrontation to infection[3].

Another limiting factor lies in the infection-disease associations established on basis of a high enrichment of GWAS disease candidate genes in cluster signatures. Although combined with literature support the proposed associations could to some extent be validated epidemiologically, a proven causal relationship with confirmed pathogenesis role of specific dysregulated gene(s) still needs experimental validation. Nonetheless, these results highlight the determinant influence of host transcriptional changes on the pathogenesis and progress of both infectious diseases and multiple non-infectious human diseases, which also assist the identification of disease biomarkers[38–41] or prognostic indicators[42].

Although we have shown that the pathogen pairs in HTR network and pathogen clusters in sub-network of HTR network have significant HTR similarity in general, further comparison and analysis can be carried out by utilizing more *in vivo* expression profile data of tissue specificity, which better present the complex, multifactorial pathogen-host interactions[43,44] and explain the dynamic consequence to infectious microorganisms of different pathogenic nature[45]. Moreover, the differential expression patterns identified by comprehensive systematic analysis of cluster-specific signature hint at further experimental studies, the combination of which makes a powerful tool accelerating the understanding of exquisite host cellular molecular mechanisms against pathogenic infections[10,46,47]. Most importantly, in addition to a shared genetic background, many other factors might also contribute to the prognosis of complications in patients with infectious disease(s) or susceptibility to specific pathogen infection in patients with non-infectious disease(s). Thus, integrating relating factors in the follow-up studies of the infection-disease associations generated by our analysis will likely expand the knowledge of disease associations (especially for those with infectious etiologies) and finally show diagnostic and therapeutic values in the near future.

Methods

GEO microarray data filtering

To make screening of the NCBI Gene Expression Omnibus (GEO) datasets more feasible, we used the GEOmetadb: GEO Microarray Search Tool[48]. The SQLite database of GEO was downloaded from GEOmetadb website on April 27 2013, and the database was renewed to April 25 2013. It contains 37, 781 projects including 958 projects containing expression profiles generated on HG-U133A cartridge arrays (Platform GPL96) and 2, 979 projects containing expression profiles generated on HG-U133 Plus 2.0 (Platform GPL570). In order to make full use of the data, projects using HG-U133A cartridge arrays (GPL96) or HG-U133 Plus 2.0 (GPL570) were used in our research, because these two platforms were widely used and the probes of Platform GPL96 are almost covered by Platform GPL570. We manually checked the descriptions of these datasets to find out projects containing expression profiles of host cellular responses to pathogenic infections. The projects selected should comply with the following principles: (I) The project should contain at least one sample of untreated specific pathogen infection; (II) The project should contain at least one sample of control (It could be uninfected, mock-infected, health or other blank controls defined by submitters); (III) The project would be discarded if more than ten percent of probes miss data values in its series matrix file. Finally 82 datasets representing host transcriptional responses (HTRs) to 50 pathogens were picked out for the following approach.

Pairing samples

We paired infect samples and control samples in accordance with the following principles:

(I) Samples of pathogen infections were paired with samples uninfected with other same experiment conditions such as cell types and culture time.

(II) Infection samples measured at the very beginning of infection (usually marked as time course 0) were not taken as infection samples.

(III) Samples measured before infection or measured at infection time course 0 would be taken as controls to infection if there were no control samples measured after time course 0.

(IV) If the number of control samples exceeded corresponded infect samples, the excess control samples will be deserted.

(V) If the number of infect samples exceeded corresponded control samples, the control samples will be recycled to pair excess infect samples which means some control samples will be paired to more than one infect samples.

Generating probe rank lists (PRLs) for each pair of samples

Each pair of samples contains one sample of infected and one of control. Probes with the deficiency of values in some datasets were excluded in our study. We generated the intersection of probes of each dataset to get the final probes shared by all datasets. And the 22,160 probes that have values in all datasets remained, covering more than 99% of HG-U133A cartridge arrays. The values of sample data values in series matrix file were transferred to count if they had been log transferred. These probes were ranked according to the expression change got by comparing the corresponded infect sample and control sample. The most common ranking method for pairwise gene expression profiles is to simply rank the genes according the fold change as compared to vehicles. While using fold change may introduce some false differential genes with low expression levels. First, values of each pair of samples less than a threshold value were set to that value. The lower quartile of each pair of samples was selected as the corresponding threshold value. In this way, the probe sets whose infection/vehicle ratio equaled one are placed in the middle of the rank list. Then we sub-sort them using a smaller threshold (the former threshold divided by ten). We repeat this process to retain probes with fold change of one until all probes are properly sorted. The sorted lists of probes named

Probe Rank Lists (PRLs) represented the regulation level considering both the expression change fold and expression values, and the probes representing most up (or down) regulated genes would get top (or bottom) ranks in PRL.

Merging probe rank lists of same virus type

The PRL of the same pathogen (representing HTRs to infection on different cell types, infection time and individuals) were combined to a single PRL to represent the HTRs to the pathogen with the R package GeneExpressionSignature[49] according a hierarchical majority-voting scheme previously described[20,50]. First, Spearman's Foot rule[36] were used to measure the distances between each two PRLs, and then we built a minimum spanning tree with the Kruskal Algorithm strategy[51] based on the distances calculated. And finally we merged the PRLs representing infections of same pathogens according to the minimum spanning tree using the Borda Merging Method[37]. After merging, HTRs to each kind of pathogen were represented by corresponded merged PRLs. Pathogens within the same species were considered as same pathogens. To be noted, several kinds of oncogenic Human papillomavirus including HPV-16, HPV-18, HPV-31, HPV-33, HPV-35, HPV-58, HPV-66 were represented by HPV, because their expression profiles from centralized projects were similar to each other.

Calculating distances to measure HTR similarities between virus pairs

First, we extracted signatures for each virus. A signature is a group of genes that may serve as a synthetic descriptor of a particular biological action which may be key genes to some diseases, cellular response to a kind of drug or other bioprocesses. And in our study, the signatures are subsets of the most significantly regulated genes in the general cellular responses to virus infections, that is to say genes in signatures were those seemed to be consistently up- (or down-) regulated in host responses to the infection of corresponded viruses. We selected the top- and bottom-ranked 250 genes of each PRL as signatures of each virus. The size of signatures was determined according to the parameter as previously reported[16,49]. To evaluate how similar the HTRs of two different viruses are, we used Gene Set Enrichment Analysis (GSEA)[23] to quantify whether genes in the signature of virus A were also tend to be placed at the top (or bottom) in the PRL of virus B. Use $\{up_A, down_A\}$ to represent the signature of virus A, and the enrichment score of up_A(or $down_A$) in the PRL of virus B presented by $ES_B^{up_A}$ (or $ES_B^{down_A}$) would be high if the corresponded genes tended to be placed at the top in the PRL of B. And the similarity between HTRs to virus A and B were finally expressed by distances between them drawing from the enrichment scores of their signatures in the opponent's PRL. We defined the distance between HTRs to virus A and B in accordance with the Average Enrichment-Score Distance as previously defined[16].

Significance Test of virus-virus HTR Similarities

To validate the significance of the distances between viruses, we used the same algorithm to calculate distances between random arranged PRLs of the same size as those used in our study (i.e. 22,160) to obtain a control. And we repeated this experiment for one million times. Then, we computed a P value for each virus-virus distance by comparing the actual distance value to the distribution of distances obtained on random data. The P value determines the probability that each distance between two viruses reach values lower than or equal to its actual value is due to chance. Finally, the original P values were then converted into FDR[52].

Identification of Virus Clusters based on HTR Similarity Clustering and Homology Analysis of Component Viruses

The affinity propagation algorithm[24] was used in the identification of virus clusters with significantly similar internal HTRs, i.e., clusters. And the distances between each virus were taken to generate clusters of viruses by this algorithm. And for each cluster, a virus was elected as exemplar considered as best representative of the features of all cluster members. Specifically, viral pathogens were partitioned to a quantity of clusters automatically calculated by the affinity propagation algorithm, and for a virus was designated as exemplar. Then, nodes corresponding to viruses were colored differently according to their clusters, and nodes with a < 1 distance (i.e., a positive correlation between them) were connected with edges. And we clustered the exemplars again to obtain second-level clusters, and add edges with a distance below one between the exemplars within same second-level clusters. This process was repeated over new level exemplars until convergence, and the edge width was negatively correlated to their correlated distances.

Extracting Signature Genes for Non-viral and Viral Clusters and Statistical Analysis

To explore the common characteristics of HTRs to viruses in a same cluster, we identified a set of host cellular genes consistently up- (or down-) regulated upon the infection of cluster viruses, named cluster signature (CS). To this end, we merged the PRLs of pathogens in the same cluster using the Borda Merging Method to get a cluster PRL in which probes consistently up- (or down-) regulated were top (or bottom) ranked. Probes with the highest or lowest 250 rank in cluster PRL were picked out to represent the most up- or down-regulated probes for each cluster and genes corresponding to these probes were identified as signatures of each cluster (Table S3). Then we statistically analyzed the CS gene counts with respect to the shared number of clusters (Table S4).

Gene Ontology (GO) Biological Processes (BP) and Kyoto Encyclopaedia of Genes and Genomes (KEGG) Annotations of Cluster Signatures

We submitted each cluster signature to the DAVID Functional Annotation online service (http://david.abcc.ncifcrf.gov/)[53,54] to identify the enriched bioprocesses and pathways in cluster signatures. In our study, the GOTERM_BP_FAT item and KEGG_PATHWAY item were used to annotate cluster signatures with default parameters. The analysis results downloaded from DAVID website can be found in Table S5 and S6. And the most enriched GO BP terms (MEGTs, $p < 0.001$) and KEGG pathway terms (MEKTs, $p < 0.1$) clustered from up- and down-regulated signature genes were shown in a heat map representation. And the p value threshold 0.1 for KEGG pathway terms was specifically chosen to ensure reliability of identified MEKTs, considering the fact that genes recorded in KEGG pathways are much less than those recorded in GO BPs

Functional Annotation Clustering and Network Construction

To build each cluster signature a functional annotation clustering network, the enriched annotation terms including GO terms (p < 0.001) and KEGG pathways (p < 0.1) of each cluster were clustered due to numbers of their shared genes by affinity propagation algorithm[24] (Table S7). And for each cluster, an annotation term was elected by this algorithm as the exemplar. Each cluster was made up of several annotation terms with similar descriptions and sharing genes, therefore, the exemplar annotation term was able to represent corresponding cluster. We manually classified the 36 exemplars into five first-level and 14 second-level categories, which were correspondingly numbered. And we colored each cluster based on their exemplar first-level classification and used the exemplar number to represent each cluster in annotation clustering map which showed the genes contained by each cluster. And gene count in each annotation cluster was divided by total enriched gene count of corresponding cluster to get the percentage of clustered gene count.

Mapping of Disease/Trait genes to Cluster Signatures

A Catalog of Published Genome-Wide Association Studies[55] (GWAS Catalog) available at: www.genome.gov/gwastudies, accessed in July 15, 2013 was used as a reference of disease/trait related genes. Genes reported adjacent to SNPs annotated to be associated with diseases/traits were mapped to cluster signatures. And we listed all the mapped records of each cluster signatures (Table S8).

We used the plain guilt by association strategy to build infection-disease relations. To explore the possible associations between viral infections and human diseases, we built a viral infection-disease network, in which virus cluster was connected by an edge to a well-defined human disease if the mapped genes of 3 different disease-associated SNPs were shared in viral cluster signature. A uniform threshold value 3 was set to ensure the reliability of identified viral infection-disease associations, because using hypergeometric distribution test, the probability of > 3 gene overlap between 500 signature genes and GWAS genes associated with individual disease is 0.0129, if 20 GWAS genes associated with single disease are covered by the 20,000 genome genes in DNA microarray.

Validation of Virus-Disease Associations

Generally, the keywords for searching literatures related specific viral infection-disease associations are in the form of “xxx virus AND xxx (i.e., GWAS disease name). We have virologist perform the expert interpretation of Abstracts and full articles when necessary to identify the association types and designate publications that support the identified association.

Conclusions

Through analysis of host expression profiles to various virus infections, we identified different virus infection modes and characterized them with signature genes. The functional annotation of signature genes revealed different virus-host interaction mechanisms. Moreover, association analysis of signature genes and GWAS genes predicted infection-disease relations at high confirmation rates for well-studied viruses, indicating their power in associating infections with diseases.

Competing financial interests

The authors declare no competing financial interests.

Author contributions

L.H., H.B., X.Z., and X.B. designed the research; L.H., H.B., H.H., and X.Q. selected and collected the expression profile datasets, L.H. and Y.L. performed the computational and statistical analysis, L.H., H.B., and X.B. wrote the paper. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by National Nature Science Foundation of China (No.U1435222, No.81273488), National Key Technologies R&D Program for New Drugs of China (2012ZX09301-003), the Program of International S&T Cooperation (2014DFB30020) and the General Financial Grant from the China Postdoctoral Science Foundation (2013M532211).

Notes and references

1 Z. Spirer and A. Barzilai, *Harefuah*, 2012, **151**, 483–487, 496.
2 N. Higa, C. Toma, T. Nohara, N. Nakasone, G. Takaesu and T. Suzuki, *Trends Microbiol.*, 2013, **21**, 342–349.
3 C. T. Ng, L. M. Snell, D. G. Brooks and M. B. A. Oldstone, *Cell Host Microbe*, 2013, **13**, 652–664.
4 L. Eckmann, *Curr. Opin. Gastroenterol.*, 2006, **22**, 95–101.
5 J. R. Jonsson, D. M. Purdie, A. D. Clouston and E. E. Powell, *Mol. Diagn. Ther.*, 2008, **12**, 209–218.
6 Y. H. Kidane, C. Lawrence and T. M. Murali, *PloS One*, 2013, **8**, e58553.
7 A. K. Zaas, T. Burke, M. Chen, M. McClain, B. Nicholson, T. Veldman, E. L. Tsalik, V. Fowler, E. P. Rivers, R. Otero, S. F. Kingsmore, D. Voora, J. Lucas, A. O. Hero, L. Carin, C. W. Woods and G. S. Ginsburg, *Sci. Transl. Med.*, 2013, **5**, 203ra126.
8 B. Joseph, M. Frosch, C. Schoen and A. Schubert-Unkmeir, *Methods Mol. Biol. Clifton NJ*, 2012, **799**, 267–293.
9 K. McGuire and E. J. Glass, *Vet. Immunol. Immunopathol.*, 2005, **105**, 259–275.
10 N. D. Maynard, D. N. Macklin, K. Kirkegaard and M. W. Covert, *Mol. Syst. Biol.*, 2012, **8**, 567.
11 S. Nusser-Stein, A. Beyer, I. Rimann, M. Adamczyk, N. Piterman, A. Hajnal and J. Fisher, *Mol. Syst. Biol.*, 2012, **8**, 618.
12 D. Walsh and I. Mohr, *Nat. Rev. Microbiol.*, 2011, **9**, 860–875.

13 I. Ioannidis, B. McNally, M. Willette, M. E. Peeples, D. Chaussabel, J. E. Durbin, O. Ramilo, A. Mejias and E. Flaño, *J. Virol.*, 2012, **86**, 5422–5436.

14 J. B. Vos, N. A. Datson, A. H. van Kampen, A. C. Luyf, R. M. Verhoosel, P. L. Zeeuwen, D. Olthuis, K. F. Rabe, J. Schalkwijk and P. S. Hiemstra, *BMC Genomics*, 2006, **7**, 9, 1-10.

15 R. G. Jenner and R. A. Young, *Nat. Rev. Microbiol.*, 2005, **3**, 281–294.

16 F. Iorio, R. Bosotti, E. Scacheri, V. Belcastro, P. Mithbaokar, R. Ferriero, L. Murino, R. Tagliaferri, N. Brunetti-Pierri, A. Isacchi and others, *Proc. Natl. Acad. Sci.*, 2010, **107**, 14621–14626.

17 J. T. Dudley, M. Sirota, M. Shenoy, R. K. Pai, S. Roedder, A. P. Chiang, A. A. Morgan, M. M. Sarwal, P. J. Pasricha and A. J. Butte, *Sci. Transl. Med.*, 2011, **3**, 96ra76.

18 M. Sirota, J. T. Dudley, J. Kim, A. P. Chiang, A. A. Morgan, A. Sweet-Cordero, J. Sage and A. J. Butte, *Sci. Transl. Med.*, 2011, **3**, 96ra77.

19 J. Kim, V. T. Vasu, R. Mishra, K. R. Singleton, M. Yoo, S. M. Leach, E. Farias-Hesson, R. J. Mason, J. Kang, P. Ramamoorthy, J. A. Kern, L. E. Heasley, J. H. Finigan and A. C. Tan, *Bioinformatics*, 2014, 2393-2398.

20 R. Gao, B. Cao, Y. Hu, Z. Feng, D. Wang, W. Hu, J. Chen, Z. Jie, H. Qiu, K. Xu, X. Xu, H. Lu, W. Zhu, Z. Gao, N. Xiang, Y. Shen, Z. He, Y. Gu, Z. Zhang, Y. Yang, X. Zhao, L. Zhou, X. Li, S. Zou, Y. Zhang, X. Li, L. Yang, J. Guo, J. Dong, Q. Li, L. Dong, Y. Zhu, T. Bai, S. Wang, P. Hao, W. Yang, Y. Zhang, J. Han, H. Yu, D. Li, G. F. Gao, G. Wu, Y. Wang, Z. Yuan and Y. Shu, *N. Engl. J. Med.*, 2013, **368**, 1888–1897.

21 J. Lamb, *Nat. Rev. Cancer*, 2007, **7**, 54–60.

22 J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J.-P. Brunet, A. Subramanian, K. N. Ross, M. Reich, H. Hieronymus, G. Wei, S. A. Armstrong, S. J. Haggarty, P. A. Clemons, R. Wei, S. A. Carr, E. S. Lander and T. R. Golub, *Science*, 2006, **313**, 1929–1935.

23 A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub and E. S. Lander, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 15545–15550.

24 B. J. Frey and D. Dueck, *Science*, 2007, **315**, 972 –976.

25 J. Peiris, S. Lai, L. Poon, Y. Guan, L. Yam, W. Lim, J. Nicholls, W. Yee, W. Yan, M. Cheung, V. Cheng, K. Chan, D. Tsang, R. Yung, T. Ng and K. Yuen, *The Lancet*, 2003, **361**, 1319–1325.

26 S. Vandini, P. Bottau, G. Faldella and M. Lanari, *BioMed Res. Int.*, **2015**, 875723.

27 B. Am, L. Ev, S. Iv, D. Ma and M. ’ianova Li, *Vopr. Virusol.*, 1986, **32**, 724–729.

28 K. Honjo, R. van Reekum and N. P. L. G. Verhoeff, *Alzheimers Dement.*, 2009, **5**, 348–360.

29 N. T. Funderburg, D. A. Zidar, C. Shive, A. Lioi, J. Mudd, L. W. Musselwhite, D. I. Simon, M. A. Costa, B. Rodriguez, S. F. Sieg and M. M. Lederman, *Blood*, 2012, **120**, 4599–4608.

30 J. J. Hwang, J. Wei, S. Abbara, S. K. Grinspoon and J. Lo, *J. Acquir. Immune Defic. Syndr. 1999*, 2012, **61**, 359–363.

31 M. Rotger, T. R. Glass, T. Junier, J. Lundgren, J. D. Neaton, E. S. Poloni, A. B. van ’t Wout, R. Lubomirov, S. Colombo, R. Martinez, A. Rauch, H. F. Günthard, J. Neuhaus, D. Wentworth, D. van Manen, L. A. Gras, H. Schuitemaker, L. Albini, C. Torti, L. P. Jacobson, X. Li, L. A. Kingsley, F. Carli, G. Guaraldi, E. S. Ford, I. Sereti, C. Hadigan, E. Martinez, M. Arnedo, L. Egaña-Gorroño, J. M. Gatell, M. Law, C. Bendall, K. Petoumenos, J. Rockstroh, J.-C. Wasmuth, K. Kabamba, M. Delforge, S. De Wit, F. Berger, S. Mauss, M. de Paz Sierra, M. Losso, W. H. Belloso, M. Leyes, A. Campins, A. Mondi, A. De Luca, I. Bernardino, M. Barriuso-Iglesias, A. Torrecilla-Rodriguez, J. Gonzalez-Garcia, J. R. Arribas, I. Fanti, S. Gel, J. Puig, E. Negredo, M. Gutierrez, P. Domingo, J. Fischer, G. Fätkenheuer, C. Alonso-Villaverde, A. Macken, J. Woo, T. McGinty, P. Mallon, A. Mangili, S. Skinner, C. A. Wanke, P. Reiss, R. Weber, H. C. Bucher, J. Fellay, A. Telenti, P. E. Tarr, MAGNIFICENT Consortium, INSIGHT and Swiss HIV Cohort Study, *Clin. Infect. Dis. Off. Publ. Infect. Dis. Soc. Am.*, 2013, **57**, 112–121.

32 R. Apps, Y. Qi, J. M. Carlson, H. Chen, X. Gao, R. Thomas, Y. Yuki, G. Q. Del Prete, P. Goulder, Z. L. Brumme, C. J. Brumme, M. John, S. Mallal, G. Nelson, R. Bosch, D. Heckerman, J. L. Stein, K. A. Soderberg, M. A. Moody, T. N. Denny, X. Zeng, J. Fang, A. Moffett, J. D. Lifson, J. J. Goedert, S. Buchbinder, G. D. Kirk, J. Fellay, P. McLaren, S. G. Deeks, F. Pereyra, B. Walker, N. L. Michael, A. Weintrob, S. Wolinsky, W. Liao and M. Carrington, *Science*, 2013, **340**, 87–91.

33 R. Mechelli, R. Umeton, C. Policano, V. Annibali, G. Coarelli, V. A. G. Ricigliano, D. Vittori, A. Fornasiero, M. C. Buscarinu, S. Romano, M. Salvetti and G. Ristori, *PLoS ONE*, 2013, **8**.

34 G. Hu and P. Agarwal, *PLoS ONE*, 2009, **4**, e6536.

35 M. Iskar, M. Campillos, M. Kuhn, L. J. Jensen, V. van Noort and P. Bork, *PLoS Comput Biol*, 2010, **6**, e1000925.

36 P. Diaconis and R. L. Graham, *J. R. Stat. Soc. Ser. B Methodol.*, 1977, **39**, 262–268.

37 S. Lin and others, *Stat. Appl. Genet. Mol. Biol.*, 2010, **9**, Article20.

38 J. Correale, M. Farez and G. Razzitte, *Ann. Neurol.*, 2008, **64**, 187–199.

39 I. Hirsch, C. Caux, U. Hasan, N. Bendriss-Vermare and D. Olive, *Trends Immunol.*, 2010, **31**, 391–397.

40 L. McCoy, I. Tsunoda and R. S. Fujinami, *Autoimmunity*, 2006, **39**, 9–19.

41 F. Plattner and D. Soldati-Favre, *Annu. Rev. Microbiol.*, 2008, **62**, 471–487.

42 A. K. Zaas, M. Chen, J. Varkey, T. Veldman, A. O. 3rd Hero, J. Lucas, Y. Huang, R. Turner, A. Gilbert, R. Lambkin-Williams, N. C. Øien, B. Nicholson, S. Kingsmore, L. Carin, C. W. Woods and G. S. Ginsburg, *Cell Host Microbe*, 2009, **6**, 207–217.

43 Y. Sui, R. Potula, D. Pinson, I. Adany, Z. Li, J. Day, E. Buch, J. Segebrecht, F. Villinger, Z. Liu, M. Huang, O. Narayan and S. Buch, *J. Med. Primatol.*, 2003, **32**, 229–239.

44 L. A. Taylor, C. M. Carthy, D. Yang, K. Saad, D. Wong, G. Schreiner, L. W. Stanton and B. M. McManus, *Circ. Res.*, 2000, **87**, 328–334.

45 J. B. Domachowske, C. A. Bonville, A. J. Easton and H. F. Rosenberg, *J. Infect. Dis.*, 2002, **186**, 8–14.

46 M. Brandes, F. Klauschen, S. Kuchen and R. N. Germain, *Cell*, 2013, **154**, 197–212.

47 I.-M. Wang, B. Zhang, X. Yang, J. Zhu, S. Stepaniants, C. Zhang, Q. Meng, M. Peters, Y. He, C. Ni, D. Slipetz, M. A. Crackower, H. Houshyar, C. M. Tan, E. Asante-Appiah, G. O’Neill, M. Jane Luo, R. Thieringer, J. Yuan, C.-S. Chiu, P. Yee Lum, J. Lamb, Y. Boie, H. A. Wilkinson, E. E. Schadt, H. Dai and C. Roberts, *Mol. Syst. Biol.*, 2012, **8**, 594.

48 Y. Zhu, S. Davis, R. Stephens, P. S. Meltzer and Y. Chen, *Bioinformatics*, 2008, **24**, 2798–2800.

49 F. Li, Y. Cao, L. Han, X. Cui, D. Xie, S. Wang and X. Bo, *OMICS J. Integr. Biol.*, 2013, **17**, 116–118.

50 F. Iorio, R. Tagliaferri and D. di Bernardo, *J. Comput. Biol. J. Comput. Mol. Cell Biol.*, 2009, **16**, 241–251.

51 T. H. Cormen, C. E. Leiserson, R. L. Rivest and C. Stein, *Introd. Algorithms*, 1990, 561–579.

52 Y. Benjamini and Y. Hochberg, *J. R. Stat. Soc. Ser. B Methodol.*, 1995, **57**, 289–300.

53 B. T. S. Da Wei Huang and R. A. Lempicki, *Nat. Protoc.*, 2008, **4**, 44–57.

54 B. T. Sherman and R. A. Lempicki, *Nucleic Acids Res.*, 2009, **37**, 1–13.

55 L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins and T. A. Manolio, *Proc. Natl. Acad. Sci.*, 2009, **106**, 9362–9367.