BBA - Molecular Basis of Disease xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

BBA - Molecular Basis of Disease

journal homepage: www.elsevier.com/locate/bbadis

Network-based method for mining novel HPV infection related genes using random walk with restart algorithm☆

Liucun Zhu[a,1], Fangchu Su[a,1], YaoChen Xu[b], Quan Zou[c,*]

[a] School of Life Sciences, Shanghai University, Shanghai 200444, China
[b] Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China
[c] School of Computer Science and Technology, TianJin University, Tianjin 300350, China

ARTICLE INFO

Keywords:
Human papillomavirus
Protein-protein interaction
Random walk with restart algorithm
GO terms
KEGG pathways

ABSTRACT

The human papillomavirus (HPV), a common virus that infects the reproductive tract, may lead to malignant changes within the infection area in certain cases and is directly associated with such cancers as cervical cancer, anal cancer, and vaginal cancer. Identification of novel HPV infection related genes can lead to a better understanding of the specific signal pathways and cellular processes related to HPV infection, providing information for the development of more efficient therapies. In this study, several novel HPV infection related genes were predicted by a computation method based on the known genes involved in HPV infection from HPVbase. This method applied the algorithm of random walk with restart (RWR) to a protein-protein interaction (PPI) network. The candidate genes were further filtered by the permutation and association tests. These steps eliminated genes occupying special positions in the PPI network and selected key genes with strong associations to known HPV infection related genes based on the interaction confidence and functional similarity obtained from published databases, such as STRING, gene ontology (GO) terms and KEGG pathways. Our study identified 104 novel HPV infection related genes, a number of which were confirmed to relate to the infection processes and complications of HPV infection, as reported in the literature. These results demonstrate the reliability of our method in identifying HPV infection related genes.

This article is part of a Special Issue entitled: Accelerating Precision Medicine through Genetic and Genomic Big Data Analysis edited by Yudong Cai & Tao Huang.

1. Introduction

Human papillomavirus (HPV) is a generic term that covers > 150 related viruses [1,2]. More than 40 HPV subtypes have been identified in humans, and most of them can only infect the genital areas of both males and females, constituting a significant pathogenic microorganism group in human bodies [3–5].

HPV has been regarded as one of the most common viral infection of reproductive tract all over the world [6]. Different subtypes of HPV may lead to various symptoms and complications. For HPV types 1 and 2, plantar warts have been confirmed to be a common symptom, while for HPV types 6 and 16, a specific anal dysplasia turns out to be a typical complication. These differences imply the interspecific pathogenic heterogeneity of this virus [7–9]. Furthermore, numerous infected patients may even have no noticeable symptoms [9,10]. This phenomenon have rendered less attention on HPV infection until its role in carcinogenesis was reported recently in various cancer types such as cervical cancer, anal cancer, oropharyngeal cancer, and vaginal cancer [11–15]. These studies revealed the strong associations of HPV infections with certain cancers, but the mechanisms of carcinogenesis of HPV infections remained unclear. Apart from the randomized DNA damage induced by HPV integration in the host genome, recent publications demonstrated that proteins E6 and E7 produced from the integrated DNA sequences as the main pathogenic factors [16,17]. Interestingly, these two proteins were confirmed to have interactions with the tumor suppressor protein p53 and pRb, respectively, suggesting their significant contributions to the oncogenic potentials of HPV virus [18,19]. These findings indicated the fundamental roles of protein-protein interactions (PPIs) between the virus and host proteins in tumorigenesis during HPV infection.

Recent efforts have been made to curate the human genes whose functions may be altered or disrupted by the integration process of HPV

☆ This article is part of a Special Issue entitled: Accelerating Precision Medicine through Genetic and Genomic Big Data Analysis edited by Yudong Cai & Tao Huang.

* Corresponding author.
E-mail addresses: zhuliucun@shu.edu.cn (L. Zhu), jjsfc@i.shu.edu.cn (F. Su), xuyaochen@sibcb.ac.cn (Y. Xu), zouquan@nclab.net (Q. Zou).

[1] These authors have contributed equally to this work.

https://doi.org/10.1016/j.bbadis.2017.11.021
Received 19 September 2017; Received in revised form 3 November 2017; Accepted 26 November 2017

infection and thus contribute to pathogenesis [20]. In this study, we took advantage of results coming out of this endeavor and identified a group of potential novel HPV infection related genes in a PPI network. The algorithm of random walk with restart (RWR) [21,22] was applied on the PPI network using known HPV infection related genes as seed nodes. After filtration of genes occupying special positions in the PPI network using the permutation test, genes with high interaction confidence and exhibiting functional similarity to known HPV infection related genes were selected through the association test by consulting the published databases such as STRING, gene ontology (GO) terms and KEGG pathway. We identified 104 key candidate genes that have strong associations to known HPV infection related genes and confirmed the involvement of some of them in HPV pathogenesis according to recent literatures.

2. Materials and methods

2.1. Dataset

Four hundred eighty-one HPV infection related genes were retrieved from HPVbase (http://crdd.osdd.net/servers/hpvbase/) [20], a comprehensive resource reporting three major efficacious cancer biomarkers: (1) integration and breakpoint events; (2) HPVs methylation patterns; (3) HPV mediated aberrant expression of distinct host microRNAs (miRNAs) [20]. We mapped these known HPV infection related genes onto a PPI network, in which proteins were represented by Ensembl IDs, leaving 464 Ensembl IDs. They are provided in Supplementary Material I. These 464 Ensembl IDs were selected as the seed nodes in the RWR algorithm to identify novel HPV infection related genes.

2.2. PPI network

Proteins in human body, both intracellular and intercellular, rarely play their functions alone and are highly regulated. Proteins that form a PPI always participate in the same metabolic pathways, biological processes, or assemble into protein complexes. Thus, the PPI information can be used to infer novel HPV infection related genes based on known HPV infection related genes. Up to now, several studies have been presented to investigate the properties of proteins and genes using PPI information [23–34].

In this study, we used the PPI information of human that was obtained from STRING [35] (http://string-db.org/, Version 10.0), a well-known public database collecting PPIs derived from 9,643,763 proteins and 2031 organisms. These PPIs were collected from five main sources: (1) genomic context predictions, (2) high throughput lab experiments, (3) (conserved) co-expression, (4) automated textmining, and (5) previous knowledge in the database. Apparently, these PPIs reflected both the direct (physical) and indirect (functional) associations between proteins. To retrieve the PPIs in human, a file named '9606.protein.links.v10.txt.gz' was downloaded, which contained 4,274,001 human PPIs covering 19,247 proteins. Each PPI consisted of two Ensembl IDs and an interaction score that ranged from 150 to 999. A larger score meant that the corresponding PPI was more likely to occur. For formulation, the score of a PPI with proteins p_a and p_b was denoted by $S_I(p_a,p_b)$. The PPI network, denoted by G, defined 19,247 proteins as nodes, and each edge represented one PPI, i.e., two nodes were adjacent if and only if their corresponding proteins constituted a PPI. Furthermore, the score of each PPI was assigned to the corresponding edge as its weight.

2.3. RWR algorithm

To tackle different biological problems successfully from the perspective of network method [36], we designed a network that was able to execute on the PPI network mentioned in Section 2.2. The RWR algorithm [21,22] is a classic ranking algorithm, which starts from some seed nodes to search possible novel nodes and ranks them from high to low probabilities. It has been adopted to search novel disease genes or other related problems [21,37–39]. Hence, this study also used the RWR algorithm to search possible novel HPV infection related genes based on known HPV infection related genes in the PPI network. Before the algorithm was executed on the PPI network, an initialization vector P_0 was constructed, which contained 19,247 components, each representing the probability for one node (gene) in the PPI network to be a novel HPV infection related gene. Because there were 464 known HPV infection related genes in the PPI network, their corresponding components in P_0 were set to 1/464, and others, zero.

The RWR algorithm simulated a random walker that moved on the PPI network starting from the 464 genes. For formulation, let P_i denote a vector representing the probability of each node after the i-th moving procedure stopped. After each moving procedure, P_i was updated, which was calculated as follows:

$$P_{i+1} = \boldsymbol{A}^T P_i + cP_0 \quad (1)$$

where $\boldsymbol{A}$ was the column-wise normalized adjacency matrix of the PPI network G, and c was the restart probability indicating the importance of the seed nodes ($c = 0.8$ in this study). When the $L1$-norm of the difference of two successive vectors was less than 1E-06, i.e., $||P_{i+1} - P_i||_{L1} < 10^{-6}$, the algorithm was considered converged, and vector P_{i+1} was the output. According to P_{i+1}, each node except the 464 genes received a probability that denoted the possibility for its corresponding gene to be a novel HPV infection related gene. Genes with larger probabilities were more likely to be the novel HPV infection related genes. 10^{-5} was set as the threshold of the probability to select candidate genes as suggested by several previous studies [40–43]. Thus, we also used it to pick up significant candidate genes, i.e., genes with probabilities larger than 10^{-5} were selected. For convenience, the selected genes were called RWR genes.

2.4. Permutation test

Among the RWR genes selected by the RWR algorithm, a number shared little or even no associations with HPV infection. The reason that they were still selected was that they occupied special positions in the PPI network such that the structure of the PPI network might have influenced results of the RWR algorithm. Some significance tests could be applied to filter this type of genes [44]. In this study, we adopted the permutation test [27,45,46], which was performed in three steps:

(1) 1000 gene sets, denoted as S_1, S_2, …, S_{1000}, were randomly constructed, each containing 464 genes;
(2) For each set, the RWR algorithm was executed on the PPI network G using genes in the set as seed nodes. The probability for each RWR gene to be novel HPV infection related gene was computed.
(3) After the RWR algorithm had been applied on all randomly produced sets, the p-value was calculated for each RWR gene, which was defined by

$$\text{p-value}(g) = \Theta/1000 \quad (2)$$

where Θ was the number of randomly produced sets on which the probability for the RWR gene g to be novel HPV infection related gene was larger than that on the 464 known HPV infection related genes. It is necessary to point out that the outcome of Eq. (2) is an approximation of the true p-value because it is impossible to test all randomly produced sets. For the p-value 0, we denoted it as " < 0.001" to indicate that its true value was between 0 and 0.001, while we directly used the approximation to denote the true value for other cases. A RWR gene receiving a high p-value was not specific to HPV infection because it was easily identified by several randomly produced sets. Thus, only the RWR genes with low p-values should be selected. A significance level of 0.05 was adopted in this study to refine the RWR genes. The remaining