Virus Research 240 (2017) 161–165

Contents lists available at ScienceDirect

Virus Research

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

Extent and evolution of gene duplication in DNA viruses

Yuxia Gao, Huayao Zhao, Yin Jin, Xiaoyu Xu, Guan-Zhu Han*

Jiangsu Key Laboratory for Microbes and Functional Genomics, Jiangsu Engineering and Technology Research Center for Microbiology, College of Life Sciences, Nanjing Normal University, Nanjing, Jiangsu 210046, China

ARTICLE INFO

Keywords:
DNA virus
Gene duplication
Genome evolution
Genome complexity

ABSTRACT

Gene duplication is the main source of genomic novelties and complexities for both eukaryotes and prokaryotes. In contrast, gene duplication appears to be infrequent in the RNA viruses. However, the extent and evolution of gene duplication in DNA viruses remains obscure. Here we perform a genome-wide analysis of gene duplication in the genomes of 250 DNA viruses that represent all known DNA viral genera. While no gene duplication event is identified in single stranded DNA (ssDNA) or reverse transcribing DNA viruses, gene duplication is frequent among double stranded DNA (dsDNA) viruses. For dsDNA viruses, the number of duplicate genes is significantly correlated with the genome complexity. We find that most of duplicate genes experienced purifying selection on average. Our results indicate that gene duplication play an important role in shaping the evolution of dsDNA viruses.

1. Introduction

DNA viruses include many important animal and human pathogens, such as poxviruses, herpesviruses, and hepatitis B virus. They are dramatically diverse in genome architectures, replication strategies, and host tropisms. Their genomes may consist of double-stranded DNA (dsDNA) or single-stranded DNA (ssDNA). They may replicate via RNA intermediates and require reverse transcription. Based on their RNA/DNA genomes and replication strategies, DNA viruses are typically classified into three classes, i.e. dsDNA, ssDNA, and dsDNA-RT viruses (Baltimore, 1971). In general, both ssDNA and dsDNA-RT viruses are small in size (in the order of several kilobases) and their genomes encode less than 10 genes. In contrast, dsDNA viruses exhibit highly variable genome sizes (~5000 to ~2.5×10^6 base pairs) and gene repertoires (5 to ~2600 genes) (Koonin et al., 2015a,b). For example, *Pandoravirus salinus*, the largest known virus, contains a 2.47-megabase genome and encodes ~2556 genes (Philippe et al., 2013).

Gene duplication is the primary source of new genes, playing a central role in shaping functional diversity and genome complexity (Zhang, 2003; He and Zhang 2005; Conant and Wolfe, 2008; Serres et al., 2009; Innan and Kondrashov, 2010). Gene duplication is a common occurrence for both eukaryotes and prokaryotes, accounting for large proportions of their gene repertoires (Zhang, 2003; He and Zhang 2005; Conant and Wolfe, 2008; Serres et al., 2009; Innan and Kondrashov, 2010). However, gene duplication is thought to be infrequent in RNA viruses (Simon-Loriere and Holmes, 2013). Among the rare gene duplication events documented in RNA viruses, the Vpx protein of primate lentiviruses originated from a duplication of the Vpr protein, both of which could antagonize the host restriction factor SAMHD1 and thus help lentiviruses evade hosts' immunity (Tristem et al., 1992; Lim et al., 2012). Gene duplication has been sporadically identified in DNA viruses, such as poxviruses and herpesviruses (McLysaght et al., 2003; Shackelton and Holmes, 2004; Davison, 2007; Elde et al., 2012; Filée, 2009; Filée, 2015). However, we still lack a comprehensive picture for the extent and evolution of gene duplication in DNA viruses.

The Red Queen hypothesis postulates that encounters of hosts and viruses result in a recurring series of adaptation and counter-adaptation (Van Valen, 1973; Dawkins and Krebs, 1979; Elde et al., 2012; Daugherty and Malik, 2012). The host-virus molecular 'arms race' is reflected by the fixation of excess non-synonymous substitutions in viral and host genes involved in the conflicts (Daugherty and Malik, 2012; Sironi et al., 2015). The role of gene duplication has been largely neglected in the host-virus arms races. Recently, the model poxvirus vaccinia was found to adapt rapidly against host antiviral defenses via duplication of anti-host factor K3L (Elde et al., 2012), which implies that gene duplication might play an important part in the evolution of Red Queen conflicts.

Here we performed a genome-wide analysis of the extent and evolution of gene duplication in the genomes of 250 DNA viruses (ssDNA, dsDNA-RT, and dsDNA viruses), which represent all currently known viral genera. We found gene duplication is frequent in the evolution of dsDNA viruses, but not in ssDNA or ds DNA-RT viruses. Moreover, we found a significant correlation between gene duplication and genome

* Corresponding author.
E-mail address: guanzhu@njnu.edu.cn (G.-Z. Han).

http://dx.doi.org/10.1016/j.virusres.2017.08.005
Received 4 June 2017; Received in revised form 10 August 2017; Accepted 10 August 2017
Available online 16 August 2017

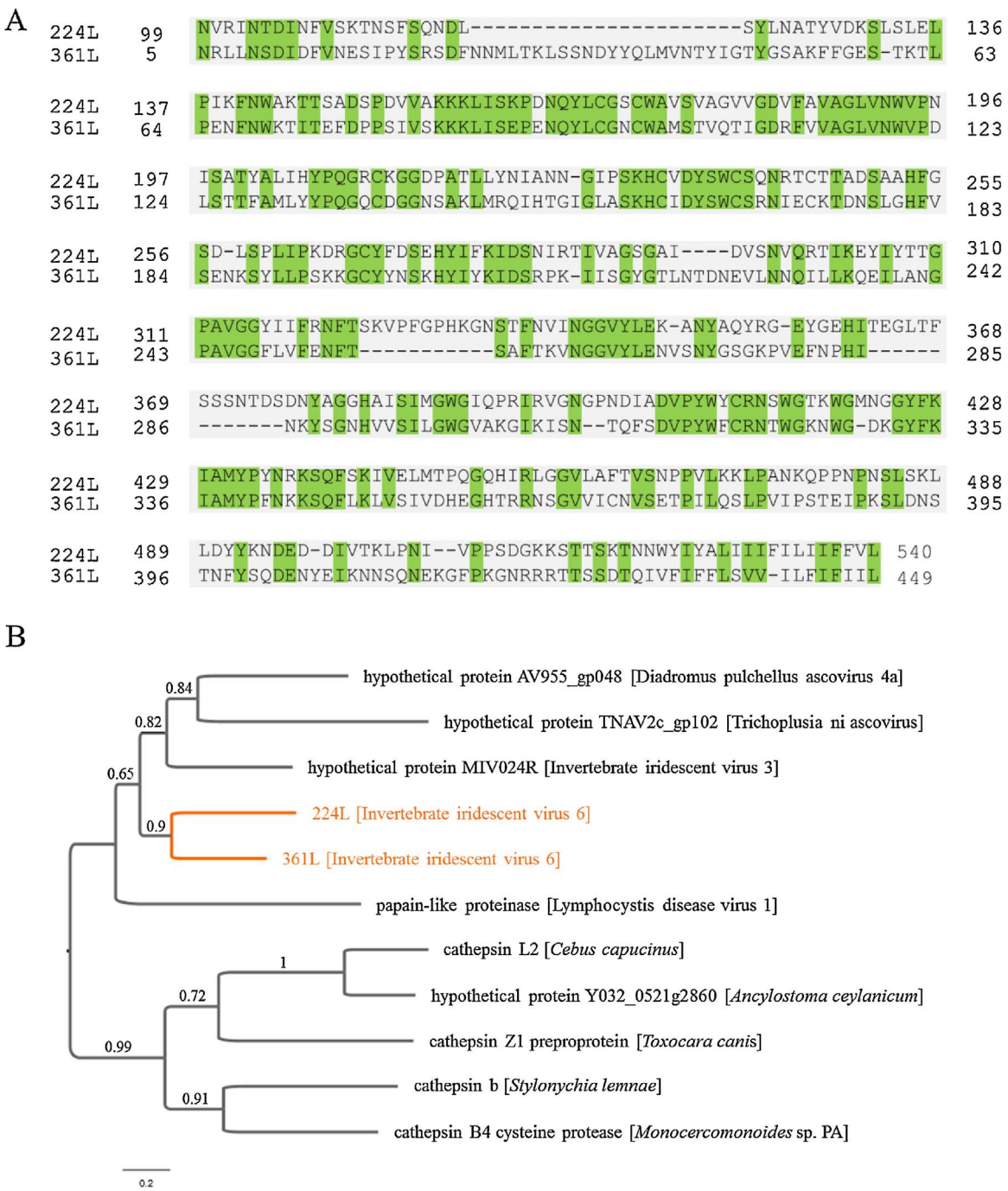

Fig. 1. An example of gene duplication in DNA virus, 224L and 361L proteins of invertebrate iridescent virus 6. (A) Alignment of 224L and 361L proteins of invertebrate iridescent virus 6. The identical amino acids are highlighted in green. (B) Phylogenetic tree of 224L, 361L, and their homologous proteins. The numbers on the node are SH-like values. Duplicate genes are labeled in orange.

complexity.

2. Results

2.1. Gene duplication is frequent in dsDNA viruses

We analyzed the genome-wide pattern of gene duplication within the genomes of 250 representative viruses that represent all known DNA virus genera. Our data sets include 40 ssDNA, 9 dsDNA-RT, and 201 dsDNA viruses (Table S1). Our genome-wide analyses identified a total of 612 genes that underwent gene duplication (thereafter each referred to as a gene family) in the genomes of 85 dsDNA viruses. For example, 224L protein (NP_149687) and 361L protein (NP_149824) of invertebrate iridescent virus 6 (*Iridovirus*, Iridoviridae) are derived from a gene duplication event (Fig. 1). 42.3% of the dsDNA viruses experienced certain degree of gene duplication, suggesting gene duplication is frequent in dsDNA viruses. These gene families contain a total of 1874 genes, with size varying from 2 to 61 genes. Among them, 68.8% (415/612) of the gene families contain two paralogs (Fig. 2). It should be noted that our estimates are conservative, because the gene duplication event occurring over long-time scale might be difficult to detect. It follows that gene duplication is more frequent in dsDNA viruses. However, we did not find any gene duplication event in ssDNA or dsDNA-RT viruses, even when a relaxed similarity search cut-off value