Infection, Genetics and Evolution 24 (2014) 92–98

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

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

Frequency and distribution of simple and compound microsatellites in forty-eight Human papillomavirus (HPV) genomes

Avadhesh Kumar Singh [a,1], Chaudhary Mashhood Alam [b,1], Choudhary Sharfuddin [b], Safdar Ali [a,*]

[a] Department of Biomedical Sciences, SRCASW, University of Delhi, Vasundhara Enclave, New Delhi 110096, India
[b] Department of Botany, Patna University, Bihar 800005, India

ARTICLE INFO

Article history:
Received 18 January 2014
Received in revised form 2 March 2014
Accepted 12 March 2014
Available online 21 March 2014

Keywords:
Microsatellites
Human papillomavirus
Simple sequence repeats
Compound microsatellites

ABSTRACT

Simple sequence repeats (SSRs) are tandem-repeated sequences ubiquitously present but differentially distributed across genomes. Present study is a systematic analysis for incidence, composition and complexity of different microsatellites in 48 representative Human papillomavirus (HPV) genomes. The analysis revealed a total of 1868 SSRs and 120 cSSRs. However, four genomes (HPV-60, HPV-92, HPV-112 and HPV-136) lacked any cSSR content; while HPV-31 accounted for a maximum of 10 cSSRs. An overall increase in cSSR% with higher dMAX was observed. The SSRs and cSSRs were prevalent in coding regions. Poly(A/T) repeats were significantly more abundant than poly(G/C) repeats possibly due to high (A/T) content of the HPV genomes. Further, higher prevalence of di-nucleotide repeats over tri-nucleotide repeats may be attributed to instability of former because of higher slippage rate. An in-depth study of the satellite sequences would provide an insight into the imperfections and evolution of microsatellites.

1. Introduction

Human papillomaviruses (HPVs) are small non-enveloped viruses that contain circular, single molecule of double-stranded DNA genome of approximately 8 kb in size (Baker et al., 1991; Sapp et al., 1995). International Committee on the Taxonomy of Viruses (ICTV) has classified HPVs into distinct taxonomic family, the Papillomaviridae-distributed into five genera: Alphapapillomavirus, Betapapillomavirus, Gammapapillomavirus, Mupapillomavirus and Nupapillomavirus. Reportedly, HPVs are the long-sought, sexually-transmitted causative agents of cervical cancer (Zur Hausen, 2009), the most prevalent cancer in women, with an annual estimate of 530,000 new cases and over 270,000 deaths globally of which more than 85% of these deaths are in low- and middle-income countries (WHO, 2013). Of the different types of HPVs, 75% cause warts on skin called as cutaneous types of HPV while other 25% are mucosal types (affects mucous membrane)- at least 13 of mucosal HPVs are cancer-causing designated high-risk or oncogenic. HPV-16 and HPV-18 cause 70% of cervical cancers and pre-cancerous cervical lesions (WHO, 2013; Li et al., 2009).

To date, around 170 HPV types have been completely sequenced (Chouhy et al., 2013). All HPVs have the same general genome organization, which is functionally divided into three regions and typically contains seven or eight open reading frames (ORFs) (Baker et al., 1991; Sapp et al., 1995). The first, non-coding upstream regulatory region (URR) consists of core promoter along with enhancer and silencer sequences that regulate transcription of ORFs (Apt et al., 1996). The second is the early (E) region encoding for non-structural regulatory proteins E1, E2, E4, E5, E6 and E7, which are involved in viral genome replication, transcription, transformation and oncogenesis. The third is late (L) region, encoding structural proteins L1 (major capsid) and L2 (minor capsid) (Baker et al., 1991; Sapp et al., 1995). URR region is located between the early and the late regions (Longworth and Laimins, 2004), and contains the highest degree of genomic diversity (Apt et al., 1996). Different genotypes of HPVs are defined by variations in genomic sequence (>10%) in the E6, E7 and L1 ORFs (De Villiers, 2013; Bernard et al., 2013). The potency of high-risk HPV infections (HPV-16, HPV-18 and HPV-31) to progress to malignancy is attributed to the expression of the E6 and E7 oncogenes owing to their strong ability to degrade tumor suppressors p53 and retinoblastoma (RB) proteins in host, respectively, which is lacking in low-risk HPVs (HPV-6 and HPV-11) (Scheffner et al., 1990; McLaughlin-Drubin and Münger, 2009; Howie et al., 2009).

Abbreviations: HPV, Human papillomaviruses; ICTV, International Committee on the Taxonomy of Viruses; ORF, open reading frames; URR, upstream regulatory region; SSR, simple sequence repeats; NCBI, National Center for Biotechnology Information; IMEx, imperfect microsatellite extraction; RA, relative abundance; RD, relative density.

* Corresponding author. Tel.: +91 11 22623503; fax: +91 11 22623504.
E-mail addresses: safdar_mgl@live.in, alisafd@gmail.com (S. Ali).
[1] Both authors contributed equally.

http://dx.doi.org/10.1016/j.meegid.2014.03.010
1567-1348/

Simple sequence repeats (SSRs), also called as mini- or microsatellites, are DNA/RNA stretches of 1–6 (or more) bp unit of tandem-repeated sequences in a genome (Chen et al., 2009; Chen et al., 2010; Alam et al., 2013, 2014). These sequences are highly divergent and ubiquitously distributed in viral, prokaryotic and eukaryotic genomes and can occupy both the coding and non-coding sequences (i.e. 3′-UTR, 5′-UTR, exons and introns) (Alam et al., 2013, 2014; Mrázek et al., 2007; Tóth et al., 2000). SSRs are the product of either de novo genesis or adoptive genesis (Kim et al., 2008), and the generation and instability of SSRs are primarily due to errors of DNA replication and/or repair machinery (Tóth et al., 2000; Katti et al., 2001). Owing to their abundance, ubiquity, simplicity, variation, multi-allelic nature among genomes and potential of abundant polymorphisms, SSRs are highly regarded as valuable source of genetic markers and genome diversity, and have been broadly applied in various areas, including determination of evolutionary relationships, comparative genome analyses and establishment of genetic maps (Pearson et al., 2005; Kashi and King, 2006; Deback et al., 2009). Variable length of microsatellites affects local DNA structure or the encoded proteins (Mrázek et al., 2007) thereby having implication in gene regulation, transcription and protein function (Kashi and King, 2006; Usdin, 2008). Also, not universally, genome features such as size and GC content influence the occurrence of microsatellites (Dieringer and Schlötterer, 2003; Coenye and Vandamme, 2005) and the polymorphism therein (Kelkar et al., 2008). Presumably, because SSRs affect the regulation of gene activity, chromatin organization, DNA replication, recombination, cell cycle and mismatch repair, their genomic distribution is non-random (Li et al., 2004).

Presence of interruptions between two or more microsatellites has revealed their different types, such as interrupted, pure, compound, interrupted compound, complex and interrupted complex (Chambers and MacAvoy, 2000). This study primarily focuses on pure and compound microsatellites (cSSRs, two or more microsatellites adjacent to each other). Interestingly, they are more abundant in coding regions than those in non-coding regions in eukaryotes (Tóth et al., 2000; Metzgar et al., 2000) and in some prokaryotes (Li et al., 2004), possibly due to increased selection in coding regions (Ellegren, 2004; Karaoglu et al., 2005), and in viruses due to high coding density (Chen et al., 2009; Alam et al., 2014). The cSSRs comprised 4–25% of genomes of *Homo sapiens*, *Macaca mulatta*, *Mus musculus* and *Rattus norvegicus*, and included some highly polymorphic compound repeats such as (dCdA) *n*(dG–dT)*n* (Weber, 1990; Bull et al., 1999; Kofler et al., 2008). Furthermore, *Escherichia coli* genomes had a frequency of 1.75–2.85% while those from HIV type-1 genomes had up to 24.24% cSSRs, suggesting the variations across genomes (Chen et al., 2012). An in-depth study of the diversifications in satellite sequences would provide insight into the imperfections and evolution of microsatellites.

Although there are accumulating evidences confirming the role of microsatellites in generating genomic diversity, evolutionary relationships and phenotypic changes, such studies are scarce in case of HPVs. Here, we systematically analyzed the incidence, composition and complexity of different microsatellites in HPV genomes that may help understand the functional aspects and adaptation the hosts.

2. Materials and methods

2.1. HPV genome sequences

The whole-genome sequence of 48 randomly-selected HPVs was accessed from National Center for Biotechnology Information (NCBI) GenBank database (http://www.ncbi.nlm.nih.gov/), and exhaustively analyzed for simple and compound microsatellites. These represented all the five genera as follows; Alphapapillomavirus ($N = 14$), Betapapillomavirus ($N = 5$), Gammapapillomavirus ($N = 19$), Mupapillomavirus ($N = 2$) and Nupapillomavirus ($N = 1$), and also unclassified HPVs (7). Genome sizes of analyzed HPVs ranged from 7100 nucleotides (HPV-48; accession no. U31789) to 8033 nucleotides (HPV-90; accession no. AY057438). Relevant features of these genomes have been summarized in Table 1.

2.2. Retrieving microsatellites and their analyses

A whole-genome search for the distribution of the simple and the compound microsatellites was performed using the IMEx software (Mudunuri and Nagarajaram, 2007). Previous reports on eukaryotes and *E. coli* have elucidated microsatellites with lengths of 12 nucleotides or more (Tóth et al., 2000), but HPV genomes did not yield any results following those parameters, possibly due to their smaller genome size. Subsequently, IMEx software was exploited using the 'Advanced Mode' as previously reported for HIV (Chen et al., 2012), tobamovirus (Alam et al., 2013) and carlavirus (Alam et al., 2014) genomes. Briefly parameters were set using Type of Repeat: perfect; Repeat Size: all; Minimum Repeat Number: 6, 3, 3, 3, 3, 3; Maximum distance allowed between any two SSRs (dMAX): 10 bp (10–50 bp is used for seven randomly selected HPV genomes). The other parameters were set as default. cSSRs were not standardized in order to determine real composition.

2.3. Statistical analysis

All the simple mathematical calculations were performed using Microsoft Office Excel 2010. However, the Pearson correlation coefficient (r) was calculated using GraphPad Prism Software, version 5 (La Jolla, CA, USA) to evaluate the influence of genome size and GC content, if any, on SSRs and cSSRs. A P-value <0.05 was considered to be significant.

3. Results

3.1. Occurrence of SSRs

The analysis revealed a total of 1868 SSRs unevenly distributed across all HPV types included in this study (Table 1, Supplementary Table 1, Fig. 1). Numbers of SSRs per genome ranged from 26 in HPV-131 (accession no. GU117631) to 66 in HPV-31 (accession no. J04353) (Table 1, Fig. 1A). A highly variant relative abundance (RA) of SSRs was observed that ranged from 3.62 bp/kb (HPV-131) to 8.34 bp/kb (HPV-31) (Table 1, Fig. 1B). Likewise, the relative density (RD) varied from 23.95 bp/kb (HPV-131 to 59.15 bp/kb (HPV-31) (Table 1, Fig. 1C).

3.2. Occurrence of cSSRs

The investigation of HPV genomes resulted in an observation of a total of 120 cSSRs. Despite high incidence of SSRs, four genomes (HPV-60, HPV-92, HPV-112 and HPV-136) lacked any cSSR content; however, HPV-31 accounted for maximum 10 cSSRs in its genome (Table 1, Supplementary Table 2, Fig. 1A). cSSRs in genome of HPV-31 exhibited maximum RA (1.26 kb/bp) and RD (27.3 bp/kb) whereas four genomes (HPV-60, HPV-92, HPV-112 and HPV-136) lacked any RA or RD (Table 1, Supplementary Table 2, Fig. 1B–C). The percentage of individual microsatellite (SSR) being the part of a cSSR (i.e. cSSR%) was zero in HPV-60, HPV-92, HPV-112 and HPV-136 (29, 29, 30 and 37 SSRs, respectively) while it was the highest (15.15%) in HPV-31 (66 SSRs) (Table 1, Supplementary Table 2, Fig. 1D).