In-silico assessment of implications of simple sequence repeats signature in 98 genomes of Polyomaviridae

Running Title: Microsatellites in Polyomaviridae genomes

Rezwanuzzaman Laskar¹, Md Gulam Jilani¹, Safdar Ali¹*

¹Clinical and Applied Genomics (CAG) Laboratory
Department of Biological Sciences, Aliah University, Kolkata, India

RL: rezwanuzzaman.laskar@gmail.com
MGJ: aman.jilani1@gmail.com

*Corresponding author:
Dr Safdar Ali
Assistant Professor,
Department of Biological Sciences,
Aliah University, IIA/27, Newtown
Kolkata 700160, India.
E-mail: safdar_mgl@live.in; ali@aliah.ac.in
Telephone No: 91-33-23416479
Fax: 91-33-29860252
Abstract

The simple sequence repeats (SSRs) are small 1-6bp tandem repeat elements present across diverse genomes and involved in gene regulation and evolution. Presently we analyzed SSRs in genomes of 98 species of family Polyomaviridae across four genera. The genome size ranged from 3962bp (BM87) to 7369bp (BM85) but maximum genomes were in the range of 5 to 5.5 kb. The GC% had an average of 42% and ranged between 34.69 (BM95) to 52.35 (BM81). A total of 3036 SSRs and 223 cSSRs were extracted using IMEx with incident frequency from 18 to 56 and 0 to 7 respectively. The most prevalent mono-nucleotide repeat motif was “T” (48.95%) followed by “A” (33.48%). “AT/TA” was the most prevalent dinucleotide motif closely followed by “CT/TC”. The distribution was expectedly more in coding region with 77.6% SSRs of which nearly half were in Large T Antigen (LTA) gene. Notably, most viruses with humans, apes and related species as host exhibited exclusivity of mono-nucleotide repeats in AT region, a proposed predictive marker for determination of humans as host in virus in course of its evolution. Each genome has a unique SSR signature which is pivotal for viral evolution particularly in terms of host divergence.

Keywords: Simple Sequence Repeats; Polyomaviridae; Prevalence, Distribution; Virus Host; Evolution
1. Introduction

Simple sequence repeats (SSRs) are ubiquitously present across the genomes of all organisms, albeit with different incidence, complexity and iterations. Ever since the identification of these repeats in multiple species, across coding and non-coding regions, their functional relevance has been explored at different levels (Chen et al., 2012; Gur-Arie et al., 2000; Kofler et al., 2008). Clinical relevance for humans is manifold. For instance, the expansion of these repeats through copy number alterations has been associated with enhancer amplification near oncogenes in cancer as well as in neuronal degradation in multiple neuropathies (Burguete et al., 2015; Hung et al., 2019).

Amongst various organisms, viruses are a unique platform to study SSRs owing to their small but rapidly evolving genomes. Further, the dependence of viruses on host cell for survival makes it an easy aspect to study in terms of genome features and evolution. SSRs have been reported to play a role in genome evolution (Bennetzen, 2000) and host range in viruses (Alam et al., 2019). Based on iterations and intervening sequences, tandemly repeated SSRs may be classified into interrupted, pure, compound, interrupted compound, complex or interrupted complex (Chambers and MacAvoy, 2000). At present, we have used Imperfect Microsatellite Extractor (IMEx) for extracting SSRs, wherein mono- to hexa-nucleotide repeat motifs are uncovered, imperfect microsatellites are allowed and compound microsatellites (cSSR: multiple SSRs separated by a distance of less than equal to dMAX) have a dMAX range of 10-50. So, the results need to be assessed within these parameters.

Present study focuses on extraction and analysis of microsatellites from genomes of 98 species of *Polyomaviridae*, which is a family of small, non-enveloped viruses that derives its name “Polyoma” from its ability to induce multiple tumors in its host. These viruses normally have
mammals, avians and fish as their hosts (Ahsan and Shah, 2006). The circular/linear genome generally encodes for two types of proteins. First, the early regulatory proteins which include large tumour antigen (LTAg), small tumour antigen (STAg), middle tumour antigen (MTAg), alternative tumour antigen (ATAg) and putative alternative large tumour antigen (PALTA)g. These are pivotal for replication, transcription and maturation of the virus during infection.

Second category of genes include those encoding for late structural proteins, which include the major capsid protein, VP1 and minor capsid proteins, VP2 and VP3. As the name suggests these are important for capsid formation (Meijden et al., 2015; Moens et al., 2011).

The extraction of SSRs in genomes of Polyomavirus and studying its incidence, distribution and complexity is an attempt to uncover its unique SSR signature which has the potential to not only predict the way these viruses are evolving but also which genome features and regions will be contributing for it. Understanding the viral genomics holds the key to combat unprecedented epidemiological outbreaks.

2. Materials and Methods

2.1. Genome sequences
Whole-genome sequence of 98 species of Alphapolyomavirus of family Polyomaviridae across 4 different genera which is listed in ICTV (https://talk.ictvonline.org/ictv-reports/ictv_online_report/dsDNA-viruses/w/polyomaviridae) was retrieved from NCBI (http://www.ncbi.nlm.nih.gov/). These include Alphapolyomavirus (43 species), Betapolyomavirus (33 species), Gammapolyomavirus (9 species) and Deltapolyomavirus (4 species). The study also included 9 species yet to be assigned Genera. The details of all the species included in the study (Genome type, Genera, Genome size, GC %, Host, Accession
number) have been summarized in Supplementary file 1. All the genomes were double stranded DNA, mostly circular except for 10 linear genomes. The information for all the known hosts for these viruses was assessed from Virus-Host Database(https://www.genome.jp/virushostdb/note.html).

2.2. Microsatellite extraction

Microsatellite extraction was carried out using the ‘Advance- Mode’ of IMEx with the parameters reported for HIV (Chen et al., 2012; Mudunuri and Nagarajaram, 2007) and as used for Mycobacteriophages (Alam et al., 2019). Briefly, the parameters included, Repeat, perfect; Repeat Size, all; Minimum Repeat Number 6 (mono-), 3 (di-), 3 (tri-), 3 (tetra-), 3 (penta-), 3 (hexa-); Maximum distance allowed between any two SSRs, dMAX: 10. Two SSRs separated by a distance of less than 10bp were thus treated as a single cSSR. In order to understand the SSR density across genomes, the dMAX was subsequently varied to 20, 30, 40, 50 and subsequent changes in cSSR incidence was recorded. It should be noted here that the maximum permissible dMAX value in IMEx is 50, because beyond that the fate of microsatellites is individualistic and hence clubbing it as cSSR becomes irrelevant. Other parameters were set to the defaults.

2.3. Statistical analysis

All statistical analyses performed on the spreadsheet using data Analysis ToolPak of MS Office Suite v2016. Linear regression was used to reveal the correlation between the relative abundance, relative density of microsatellites with genome size and GC%.

2.4. Dot Plot Analysis for Host Specificity
Dot plot analysis of two nucleic acid/protein sequences using Genome Pair Rapid Dotter (GEPARD) highlights the presence of SSRs within the genomes (Alam et al., 2019; Krumsiek et al., 2007) to ascertain their evolutionary relationships in context of repeats, reverse matches, and conserved domains. We used GEPARD v1.40 (Krumsiek et al., 2007) in the present study wherein, the viruses were selected for their host exclusivity and SSR sequence complexity; humans and other species, absolute mono-nucleotide repeat motif contribution by A/T and its reduced representation.

2.5. Evolutionary relationship

Alignment and phylogenetic reconstructions were performed using the function "build" of ETE3 v3.1.1 (Huerta-Cepas et al., 2016) as implemented on the GenomeNet (https://www.genome.jp/tools/ete/). The tree was constructed using FastTree v2.1.8 with default parameters (Price et al., 2009). Values at nodes are SH-like local support. The final tree for visualization was constructed utilizing the webtool iTOL (interactive Tree Of Life) (Letunic and Bork, 2019).

3. Results

3.1. Genome features

The genome size ranged from 3962bp (BM87) to 7369bp (BM85) but maximum genomes were in the range of 5 to 5.5 kb. However, the GC% with an average of 42% ranged between 34.69 (BM95) to 52.35 (BM81) but exhibits much more diversity as compared to genome size (Figure 1a, Supplementary file 1). In essence, the Polyomaviridae genomes are mostly of similar sizes, but its composition in terms of GC% is much more variable. If we hypothesize that SSR
incidence has an equal chance across whole genome, irrespective of the composition. then the
same should be reflected in the motifs of SSRs present. However, as discussed later, this is not
the case. There are several species which have mono-nucleotide motifs exclusively in the AT
region.

The correlation between genome size and GC content was ascertained with various SSR features.
SSR incidence was found to be significantly correlated (R^2 = 0.19, P < 0.05) with genome size
and GC content (R^2 = 0.08, P < 0.05). Though relative density and relative abundance were not
significantly correlated with genome size (R^2 = 0.01, P > 0.05; R^2 = 0.005, P > 0.05), significant
correlation was observed with GC content (R^2 = 0.20, P < 0.05; and R^2 = 0.23, P < 0.05)
respectively.

Further, cSSR incidence is significantly correlated with genome size (R^2 = 0.06, P < 0.05) but its
corresponding relative density (R^2 = 0.0038, P > 0.05) and relative abundance (R^2 = 0.004, P
> 0.05) shows no significant correlation therein. GC content is also significantly correlated for
cSSR incidence (R^2 = 0.06, P < 0.05), relative density (R^2 = 0.11, P < 0.05) and relative
abundance (R^2 = 0.08, P < 0.05).

3.2. Incidence of SSRs and cSSRs

A total of 3036 SSRs and 223 cSSRs were extracted from the 98 species of *Polyomaviridae*
(Supplementary files 2-4). The average distribution of SSRs and cSSRs per genome varied from
23 and 1.3 (*Gammapolyomavirus*) to 33 and 2.9 (*Betapolyomavirus*) respectively. Their
distribution across genera has been summarized in Table 1.

Maximum of 56 SSRs were present in BM85 whereas minimum of 18 was present in BM80 and
BM21. cSSR incidence ranged from 0 in seven species (BM99, BM82, BM76, BM59, BM24,
BM21, BM14) to 7 in two species (BM85 and BM84) (Figure 1a). Two interesting but contrasting observations can be made from this data. First, BM85 and BM84 with 7 cSSRs have 56 and 31 SSRs in a genome size of 7369 and 4697bp respectively (Supplementary file 2). What it essentially means is that though a longer genome should ideally account for more SSRs but the eventual clustering of SSRs reflected as cSSR incidence remains the same. Thus, the SSR rich regions of genome are independent of genome size. The second aspect is that the above observation is not the norm as is evident from the cSSR range of zero to seven. Multiple genomes of *Polyomaviridae* with varying number of SSRs have same number of cSSRs. This is highlighted by 29 species having 2 cSSRs (Figure 1a, Supplementary files 2-4) suggesting of a unique genome SSR signature.

In order to further highlight the regularity of this anomaly, we looked into cSSR%, which is percentage of SSRs present as cSSRs in a particular genome. Note, the variations in cSSR% are not only across different genera but even within, thereby negating the clustering of SSRs in a genera specific manner (Figure 2a). These are reflective of specific yet variable localizations and clustering of SSRs in a particular genome.

### 3.3. Relative abundance (RA) and relative density (RD) of SSRs and cSSRs

RA is the number of microsatellites present per kb of the genome whereas RD is the sequence space composed of SSRs of microsatellites per kb of the genome. So, these values are reflective of number of iterations of SSRs present. If the SSRs have a conserved tendency to be iterated, then higher incidence should correspond to elevated RD values. Also, a higher RA value should correspond to high RD value. As observed, BM65 has the highest RA and RD values of 9.32 and 80.4 respectively for SSRs which means, since more SSRs are present per kb of the genome,
more genome is comprised of SSRs. The corresponding lowest values for RA and RD was 3.39 (BM21) and 26.5 (BM80) respectively (Figure 1b, Supplementary files 2-4).

Similarly, the cSSR relative abundance (cRA) and relative density (cRD) was also studied. Since there were 7 species with no cSSR (Figure 1a), hence the minimum cRA and cRD values were zero for these species. The highest values for cRA and cRD were 1.490 (BM84) and 33.93 (BM95) respectively (Figure 1b, Supplementary files 2-4). This difference may be due to the differential composition of the cSSRs.

3.4. dMAX and cSSR

cSSR incidence is dependent on the allowed distance (dMAX) between two SSRs for it to be treated as one cSSR. Since cSSR is reflective of clustering of SSRs, and IMEx allows for dMAX values till 50, we analyzed cSSR incidence of Polyomaviridae genomes by varying the dMAX values from initial value of 10 to 20, 30, 40 and 50. Subsequently, % increase was calculated using the given formula.

\[
\% \text{ increase} = \left[ \frac{\text{cSSR incidence at dMAX}_n - \text{cSSR incidence at dMAX}_{(n-10)}}{\text{cSSR incidence at dMAX}_{(n-10)}} \right] \times 100
\]

This % increase was thereon plotted. Though maximum increase is observed for most species when dMAX increased from 10 to 20 as evident from the predominant black bar, it doesn’t conform to a pattern per se (Figure 2b). This means that even in species of the same family, SSRs chart their own path in terms of localizations in each genome.

3.5. SSR motif types and their prevalence
First, we wanted to ascertain the contribution of different repeat motif (mono- to hexa) to the overall SSRs incidence. The data was analysed separately for each of the genera. Also, the analysis was done in percentage and not absolute numbers to account for variable number of species across genera. Note that the data from species with unassigned genera wasn’t included herein. The contribution of mononucleotide repeats motifs ranged from 36% (Gammapolyomavirus) to 47% (Betapolyomavirus). Deltapolyomavirus had no incidence of penta- and hexa-nucleotide repeats whereas Gammapolyomavirus lacked hexanucleotide repeats. This can be attributed to fewer species in these genera. Gammapolyomavirus had the highest contribution from di-nucleotide repeats (39.42%) and the only genus to have more di-nucleotide repeats than mono-nucleotide repeats (Figure 3a, Supplementary files 2-3).

We thereon looked in to the motif composition of mono- and di-nucleotide repeats for their prevalence across the different genera of Polyomaviridae. For the mono-nucleotides, if we look at the overall data, the most prevalent repeat motif is “T” (48.95%) followed by “A” (33.48%). “T” also remains the most prevalent mono-nucleotide motif for Alpha-, Beta- and Delta-polyomavirus (47, 52 and 71 percent respectively). However, Gammapolyomavirus has a highest contribution from “C” (34.67%) followed by “T” (33.33%) (Figure 3b, Supplementary files 2-3). Interestingly, the same Gammapolyomavirus has the highest di-nucleotide repeat motif contribution from “AT/TA” (29.27%) motif while Alphapolyomavirus has its largest contribution from “CT/TC” (29.37). Overall, “AT/TA” was the most prevalent dinucleotide repeat motif closely followed by “CT/TC” (Figure 3c).
3.6. SSRs in coding regions

The assessment of SSRs distribution across genome revealed that non-coding region accounted for 679 SSRs (22.4%) whereas coding region comprised of 32 proteins/putative genes/ORFs housed 2357 (77.6%) of SSRs (Supplementary file 2). This aspect of prevalence in coding region conforms to earlier reports about viral SSRs (Alam et al., 2019, 2014).

Subsequently, we analyzed the SSR prevalence across different genes of the studied genomes. Six genes accounted for over 92% of SSRs. Overall, the LTag gene alone accounted for over 47% of total SSRs with VP1 gene a distant second at around 16% (Figure 3d). Thereafter, we dissected the data across different genera. Interestingly, though LTag gene takes the pole position in housing of SSRs across genera, its contribution varied. In *Betapolyomavirus*, it was accounting for one in every two SSR (49.54%) while in *Gammapolyomavirus*, approximately one in every three SSR was housed in LTag gene (35%). This difference permeates to all the genes, albeit to a lesser extent (Figure 3e, Supplementary files 2-3).

3.7. SSRs (Mono-nucleotide) specificity and host range exclusivity

The compilation of different SSRs contribution to overall incidence revealed an interesting observation. Eighteen species had one hundred percent mono-nucleotide SSRs comprising of A/T. Further, majority of these viruses had humans or members of the ape family as their hosts. To elucidate a possible pattern and significance of the same, we arranged all the studied species in decreasing order of their mono-nucleotide SSR contribution by A/T (Figure 4, Supplementary files 1-2). Notably, viruses with humans, apes and related species as hosts have a much higher A/T mono-nucleotide SSRs composition as compared to birds and fishes as hosts (Figure 4).
Using representative species (9 each) we thereon investigated whether the SSRs composition by A/T and the hosts reflect a pattern. Dotplot analysis was performed for nine species each with humans, apes and related species as hosts (Figure 5a) and nine species with birds, fishes and other species as hosts (Figure 5b). Interestingly, even though three species in figure 4 have 100% mono-nucleotide SSR contribution by A/T (same as Figure 5a), the overall number of dots (reflective of repeat sequences) is higher for all the genomes of figure 5a, representing humans and related species as hosts.

3.8. Phylogenetic tree of Polyomaviridae

Subsequently, we constructed the phylogenetic tree of the 98 Polyomaviridae genomes and observed that all the viruses aren’t evolved together as per their hosts. However, hosts do reflect in the tree. Multiple places of clustering of virus with same or related hosts can be observed (Figure 6). The fact that all viruses with human or same hosts don’t follow the pattern is only indicative of other players in genome evolution besides hosts.

We thereon superimposed the data for percentage mono-nucleotide SSR contribution by AT region, the phylogenetic analysis and the known hosts. For the sake of clarity, hosts of only those species with >90% mono-nucleotide SSR contribution from AT region are shown as illustrations here, though the complete information is provided in Figure 4. We hypothesize that the presence of mono-repeats in the AT region is somehow providing for viral host flexibility and interchangeability. The same is supported by transmission of HIV which had origins in monkey and Coronavirus which had originally bats as host (19). Both the species are hosts for Polyomavirus genomes having exclusive or near exclusive contribution of mono-SSRs from AT region.
4. Discussion

Owing to the variable nature of the A/T and G/C regions of the DNA, often these sequences exhibit specific attributes. The significance of AT repeats in strand slippage and copy number polymorphism is well documented (Katti et al., 2001). Though this implies GC content to be an important aspect for SSR studies but it is not necessarily the case primarily because of two reasons. First, because of uneven distribution of SSRs across any genome as observed herein and reported elsewhere as well (Alam et al., 2019, 2013; Chen et al., 2012). This means that certain regions of the genome will have more SSRs than others but this is not driven by GC content. For instance, there are 18 species herein where the complete mono-nucleotide SSRs are localized to the A/T region. The fact that these genomes have maximum GC content of 52%, proves the point. We believe and propose that this unevenness in distribution is not random but with a purpose; most probably host range, as discussed later. Secondly, the prevalence of repeats is dependent on size of repeat motifs, as in what’s applicable to mono-nucleotides, isn’t true for di-nucleotides and it also varies from one genus to another. However, two exceptions both in *Gammapolyomavirus* deserve mention. First, it’s the only genera to have maximum mono-nucleotide SSRs as “C”. It’s a deviation from AT region being hub for shorter repeat motifs. Contrastingly, it returns to expected lines with “AT/TA” being the most represented di-nucleotide repeat motif. Secondly, we should bear in mind that this genus has lesser number of species (nine) but that may be looked with multiple perspectives. Either we consider the fewer species as the reason for the aberrant observation or we can assume this uniqueness is the reason for fewer species in *Gammapolyomavirus*. We believe in the latter.
The study of cSSRs has always been relevant with SSRs owing to their involvement in functional aspects like regulation of gene expression (Chen et al., 2011; Kashi and King, 2006). Essentially, cSSR is a reflection of accumulation of SSRs in the genome. Higher cSSR incidence refers to SSRs present in close proximity to each other and with these being sources of variations and genome evolution (Kim et al., 2008; Madsen et al., 2008), we further looked at cSSRs in terms of cSSR% and by varying dMAX. An increase in cSSR incidence with increasing dMAX is expected and observed as well (Figure 2b). However, the increase not conforming to any pattern as visible by the different lengths of differently coloured lines is indicative of each genomes’ uniqueness. The few instances wherein negative percentage is observed is owing to merging of two independent cSSRs into one with increasing dMAX, thus leading to a decrease in cSSR incidence. Also, the cSSR% varies not only across the genera of Polyomaviridae but also within the species of same genera (Figure 2a). In spite of these variations, of all the reported cSSRs, only 17 are composed of three SSRs and 3 of four SSRs. Rest all are of two SSRs only. There is only one species BM97 which has two cSSRs of more than 3 SSRs each. Other genomes have a single representation only. All the above figures are for dMAX of 10 (Supplementary file 4).

The distribution of around 78% SSRs across coding regions is in accordance with other viral genomes though the gene specific data (Figure 3d-e) exhibits uniqueness to Polyomaviridae genomes. The overlap of genes is reflected by LTA/STA or VP2/VP3 representation. Presence of SSRs in these overlapping regions can be influential in the scenario that an alteration there would have an impact on two genes simultaneously. The cSSRs constitution ranged from two to four SSRs, albeit with divergent motifs as mentioned above. The distribution of SSRs failed to conform to a pattern. Thus, we can affirm that the genome specific clustering of SSRs is not only...
unique but regulated as well. This may be an attempt of the genome to shield itself from changes as clustering of SSRs will lead to developing hotspots for mutations.

Though the overall evolution of viruses is guided by multiple factors like host range and genome features, the number and composition of mono-nucleotide SSRs showed a correlation with the hosts and we believe the data has the foundation of predicting the future hosts for any viral species. Our hypothesis stems from the fact that there were eighteen genomes which exhibited mono-nucleotide repeats being exclusively restricted to the AT region. A closer analysis (Figure 4) revealed a pattern suggesting humans or related hosts in those genomes. On widening our analysis, we can say with confidence that the contribution of mono-nucleotide SSRs from AT region is pivotal for host range determination. Earlier studies on evolution of *Polyomavirus* have suggested gene duplications and inversions as sources for variations in genome size and also predicted their prior existence in invertebrate hosts indicating an evolving virus family in terms of host (Buck et al., 2016). This becomes all the more relevant when we look at the suggested organisms on the basis of this study to share a common/interchangeable host range for viruses. This includes monkeys (HIV) and Bats (*Coronavirus*) (Parrish et al., 2008). We accept that the correlation between mono-repeat from AT region and host is not universal suggesting other influencing factors but it’s presence in species across genera demands further authentication of the idea.

The fact that mutational and evolving nature of viral genes are responsible for host susceptibility and genome evolution makes it important and interesting to explore the role of microsatellites present as unique SSR signature in each genome to be able to predict and be prepared for yet to happen evolutionary pathogenesis of viruses.
Acknowledgements

The authors thank Department of Biological Sciences, Aliah University, Kolkata, India for all the financial and infrastructural support provided.

Authors Disclosure Statement

No competing financial interests exist

Authors' contributions

RL performed all the analysis of extracted SSRs and prepared all the figures and tables. MGJ carried out the extraction of microsatellites from IMEx. SA supervised the whole study and prepared the manuscript.

References


The Ancient Evolutionary History of Polyomaviruses. PLOS Pathogens 12, e1005574. https://doi.org/10.1371/journal.ppat.1005574


Figure 1: A) Genome features and SSR/cSSR incidence of Polyomaviridae genomes. Though genome size is predominantly around 5-5.5kb as evident by fairly constant level of red bars whereas the corresponding GC variations (superimposed black bars) have a much broader range. Also, note the diversity in SSRs incidence in genomes of similar length. Furthermore, higher SSR incidence doesn’t necessarily translate to more cSSRs. B) Relative abundance (RA) and relative density (RD) of SSRs and cSSRs. RA is the number of microsatellites present per kb of the genome whereas RD is the sequence space composed of SSRs of microsatellites per kb of the genome. The varying peaks signify the presence of a unique SSR signature for each genome.
Figure 2: A) cSSR% in the studied *Polyomaviridae* genomes. Percentage of individual SSRs as part of cSSRs is cSSR%. The data for all the genera are differentially coloured. Not only there is diversity across the genera but also within the genomes of the same genera as well. Interestingly, BM84 which has the highest cSSR% is yet to be classified into any genera. B) Percentage increase in cSSR incidence with increasing dMAX (10 to 50). Note the non-linearity in increase. Negative bars represent decrease in cSSR incidence when two cSSRs merge into one with increasing dMAX.
Figure 3: A) SSR incidence and motif length. An increase in repeat motif resulted in lesser incidence, inverse proportionality, which is expected. However, two observations should be noted. Firstly, *Gammapolyomavirus* is the only genera where the highest incidence is of di-nucleotide repeat motifs. All others have mono-nucleotide motif as most represented along expected lines. Second, the fall in incidence from mono- to di-nucleotide motif SSRs is the least in *Deltapolyomavirus*.  
B) Mono-nucleotide motif composition. In spite of varying GC percentage (Figure 1), the mono-nucleotide motif composition is very much biased towards A/T across all genera. Total represents overall data.  
C) Di-nucleotide motif composition. Though AT/TA is the most represented di-nucleotide repeat motif overall, it doesn’t enjoy the same status across all genera, with *Alphapolyomavirus* being the exception. Here, CT/TC has the highest incidence closely followed by AT/TA.  
D) Distribution of SSRs (%) across different proteins. Overall, LTA accounted for over 47% of all SSRs in the coding region with VP1 coming a distant second at around 16%. Only the 6 proteins which accounted for highest SSRs were included, the rest have been collectively taken as “Others”.  
E) SSRs contribution (%) by proteins across different genera. Herein, subtle variations are visible. Though LTA gene accounts for maximum SSRs in the coding genome across all the genera but the contributing percentage varies from 35% in *Gammapolyomavirus* to almost 50% in *Betapolyomavirus*.
Figure 4: Genomes with decreasing % of A/T mono-nucleotides repeat motif. Though, not perfect, the similar values for humans and related species suggests host range dependency on SSR distribution across AT genome regions. Higher the contribution of mono-nucleotide repeat motifs from AT region, greater are the chances that it will have humans or related species as its host.
Figure 5: Dotplot analysis of *Polyomaviridae* genomes with A) Human, apes or related species as hosts with mono-nucleotide repeat motif contribution of 100% from the AT region and B) Divergent hosts with varying mono-nucleotide repeats in the AT region.
Figure 6: Phylogenetic and Host range Analysis. The phylogenetic tree is based on whole genome sequence alignment with few important observations. First, the unassigned species are sharing nodes with different genera and hence their cumulative data need to be assessed with care. Secondly, the circles representing mono-nucleotide SSR contribution indicate that those genomes with exclusive mono-nucleotide SSR in the AT region are distributed across all genera, albeit with varying frequency. Thirdly, the selective representation of host for genomes has been done in two categories, those with exclusive mono-SSRs in AT region (100% indicated by complete orange circle) and those with (90≤ mono-SSRs in AT region <100). It suggests their host range potential which is supported by recent Coronavirus transmission from bats.
Table 1: SSR and cSSR incidence across the different genera of *Polyomaviridae*

<table>
<thead>
<tr>
<th>S No</th>
<th>Genera</th>
<th>No of Species</th>
<th>No of Species</th>
<th>Average SSR per Species</th>
<th>cSSR incidence</th>
<th>Average cSSR per Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><em>Alphapolyomavirus</em></td>
<td>43</td>
<td>1315</td>
<td>30.58</td>
<td>80</td>
<td>1.86</td>
</tr>
<tr>
<td>2</td>
<td><em>Betapolyomavirus</em></td>
<td>33</td>
<td>1090</td>
<td>33.03</td>
<td>96</td>
<td>2.9</td>
</tr>
<tr>
<td>3</td>
<td><em>Deltapolyomavirus</em></td>
<td>04</td>
<td>108</td>
<td>27</td>
<td>6</td>
<td>1.5</td>
</tr>
<tr>
<td>4</td>
<td><em>Gammapolyomavirus</em></td>
<td>09</td>
<td>208</td>
<td>23.11</td>
<td>12</td>
<td>1.33</td>
</tr>
<tr>
<td>5</td>
<td>Unassigned Species</td>
<td>09</td>
<td>315</td>
<td>35</td>
<td>29</td>
<td>3.22</td>
</tr>
<tr>
<td></td>
<td><strong>Total</strong></td>
<td><strong>98</strong></td>
<td><strong>3036</strong></td>
<td><strong>223</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Supporting information

Supplementary file S1: Details of *Polyomaviridae* genomes used in the study. Genome features (NCBI Accession number, size, shape and GC%); virus details (Classification, name, abbreviation) and Host details are summarized.

Supplementary file S2: SSRs and cSSRs extracted form studied *Polyomaviridae* genomes. RA is the number of microsatellites present per kb of the genome whereas RD is the sequence space composed of SSRs of microsatellites per kb of the genome. The corresponding value for cSSR has been represented by cRA and cRD. Overall SSR incidence, tract size as well as for individual repeat motifs along with their localization and other features are mentioned.

Supplementary file S3: Raw data for distribution of SSRs found in the *Polyomaviridae* genomes. This includes consensus sequence, number of iterations, tract size, their start and end positions along with localization (Coding/Non coding).

Supplementary file S4: Raw data for distribution of cSSRs found in the *Polyomaviridae* genomes. This includes composition of the cSSRs, their iterations and their positioning on the genome (start/end).