# Interregional coevolution analysis revealing functional/structural interrelatedness between different genomic regions in *Human mastadenovirus D*

Gabriel Gonzalez,[a] Kanako O. Koyanagi,[a] Koki Aoki,[b] Hidemi Watanabe [a]#

Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Japan[a]; Department of Ophthalmology, Hokkaido University Graduate School of Medicine, Sapporo, Japan[b]

Running head: Interregional coevolution of human adenovirus genomes

#Address correspondence to Hidemi Watanabe, Graduate School of Information Science and Technology, Hokkaido University, Sapporo 060-0814, Japan.
Phone number: +81-11-706-6146
Fax number: +81-11-706-7340
E-mail: watanabe@ist.hokudai.ac.jp

Word count of Abstract: 247
Word count of Importance: 149
Word count of the text: 4264
Figures: 5
Tables: 1

## Abstract

*Human mastadenovirus D* (HAdV-D) is exceptionally rich in type among the seven human adenovirus species. This feature is attributed to frequent intertypic recombination events that have reshuffled orthologous genomic regions between different HAdV-D types. However, this trend appears to be paradoxical, as it has been demonstrated that the replacement of some of the interacting proteins for a specific function with other orthologues causes malfunction, indicating that intertypic recombination events may be deleterious. In order to understand why the paradoxical trend has been possible in HAdV-D evolution, we conducted an inter-regional coevolution analysis between different genomic regions of 45 different HAdV-D types and found that ca. 70% of the genome has coevolved, even though these are fragmented into several pieces via short intertypic recombination hotspot regions. Since it is statistically and biologically unlikely that all of the coevolving fragments have synchronously recombined between different genomes, it is probable that these regions have stayed in their original genomes during evolution as a platform for frequent intertypic recombination events in limited regions. It is also unlikely that the same genomic regions have remained almost untouched during frequent recombination events, independently, in all different types, by chance. In addition, the coevolving regions contain the coding regions of physically interacting proteins for important functions. Therefore, the coevolution of these regions should be attributed at least in part to natural selection due to common biological constraints operating on all types, including protein-protein interactions for essential functions. Our results predict additional unknown protein interactions.

## Importance

*Human mastadenovirus D*, an exceptionally type-rich human adenovirus species and causative agent of different diseases in a wide variety of tissues, including that of ocular region and digestive tract, as well as an opportunistic infection in immune-compromised patients, is known to have highly diverged through frequent intertypic recombination events; however, it has also been demonstrated that the replacement of a component protein of a multi-protein system with a homologous protein causes malfunction. The present study solved this apparent paradox by looking at which genomic parts have coevolved using a newly developed method. The results revealed that intertypic recombination events have occurred in limited genomic regions and been avoided in the genomic regions encoding proteins that physically interact for a given function. This approach detects purifying selection against recombination events causing the replacement of partial components of multi-protein systems and therefore predicts physical/functional interactions between different proteins and/or genomic elements.

## Introduction

Human adenoviruses (HAdVs) are members of the *Adenoviridae* family and non-enveloped viruses with an icosahedral nucleocapsid containing a linear double stranded DNA genome, the size of which ranges between 30 and 37 kbp (1). Each of the HAdV genomes contains 13 genes that encode approximately 40 different proteins, including those for the DNA replication machinery as well as modulation of the host immune response and the formation, assembly and packaging of virion structures (2, 3). HAdVs are classified into seven species, *Human mastadenovirus A* to *G* (HAdV-A to G), each of which consists of specific types and the number of constituent types greatly varies from species to species: four, eleven, five, 43, one, two and one for HAdV-A to G, respectively (4-9).

Among these human mastadenovirus species, HAdV-D is exceptional in that it is uniquely human-specific and extremely type-rich. The high diversity of HAdV-D is largely attributed to frequent intertypic recombination events within this species, including those between distantly related types (10). However, it appears to be paradoxical that new recombinant forms are generated via recombination between genomes of different types, especially distant ones, because random recombination events between highly diverged protein genes may result in chimeric proteins that are misfolded and malfunction. The potential negative effect of frequent recombination events has been discussed, and it has been implied that the chance of producing non-functional chimeric proteins via recombination between distant types may be avoided by biased modular exchanges of specific genomic segments via homologous recombination events that are initiated at universally conserved segments (UCSs) in the HAdV-D genomes (10). This implication seems to explain both the frequent occurrence of recombinant forms in HAdV-D and the existence of recombination boundary hot spots in HAdV-D genomes. However, it is still

probable that a molecular system that functions through physical protein-protein, protein-genome and/or other modes of interaction would malfunction when some of the interacting elements are replaced with different ones, including the corresponding homologues, from other genomes via intertypic recombination events. Indeed, it has been shown that the replacement of any of the adenoviral packaging proteins, *L1*:52/55K, *IVa2*:IVa2, *L4*:22K and *L1*:IIIa, with a homologue from a different serotype genome impairs the packaging function (11, 12).

In this study, we investigate evolutionary correlations between different regions of the HAdV-D genomes in order to determine whether different genomic regions, particularly, the regions encoding proteins that are functionally interrelated to one another, have recombined independently of one another and how such functional interrelatedness between different regions has shaped the evolutionary landscape of the HAdV-D genomes.

## Materials & Methods

**Genome sequences and multiple sequence alignment.** The available genome sequences of 43 HAdV-D types and two hybrid types (13) were obtained from the International Nucleotide Sequence Database Collaboration (INSDC) and aligned into a single multiple alignment using the iterative refinement method algorithm (FFT-NS-I) of MAFFT (14). A gap-free multiple genome alignment (MGA) was then obtained by removing gap-containing sites, and a molecular phylogenetic tree of the genomes (the genome tree) was constructed with the neighbor-joining (NJ) method (15) under the Tamura-Nei model (TN93) (16). The INSDC accession numbers of the considered types and two hybrids are: HAdV-8(AB448767), HAdV-9(AJ854486), HAdV-10(AB724351), HAdV-13(JN226747), HAdV-15(AB562586), HAdV-17(HQ910407),

HAdV-19(AB448771), HAdV-20(JN226749), HAdV-22(FJ619037), HAdV-23(JN226750), HAdV-24(JN226751), HAdV-25(JN226752), HAdV-26(EF153474), HAdV-27(JN226753), HAdV-28(FJ824826), HAdV-29(AB562587), HAdV-30(JN226755), HAdV-32(JN226756), HAdV-33(JN226758), HAdV-36(GQ384080), HAdV-37(AB448776), HAdV-38(JN226759), HAdV-39(JN226760), HAdV-42(JN226761), HAdV-43(JN226762), HAdV-44(JN226763), HAdV-45(JN226764), HAdV-46(AY875648), HAdV-47(JN226757), HAdV-48(EF153473), HAdV-49(DQ393829), HAdV-51(JN226765), HAdV-53(AB605243), HAdV-54(AB448770), HAdV-56(AB562588), HAdV-58(HQ883276), HAdV-59(JF799911), HAdV-60(HQ007053), HAdV-62(JN162671), HAdV-63(JN935766), HAdV-64(EF121005), HAdV-65(AP012285), HAdV-67(AP012302), HAdV-22/37(AB605240) and HAdV-22/37,8(AB605242).

**Simulation of genome evolution.** Forty-five artificial genomic sequences were generated by simulating sequence evolution using Mesquite version 2.75 (17). The following parameters for this sequence evolution were obtained from the real MGA: the tree topology, number of characters, ratio of invariant sites, alpha parameter of the gamma distribution of rate variance, nucleotide frequencies of A, T, G and C and transition/transversion ratio.

**Recombination event analysis.** Recombination events in the MGA were identified using the following seven algorithms available in the RDP 4.22β program (18): RDP (19), GENECONV (20), Chimaera (21), MaxChi (22), BootScan (23), SiScan (24) and 3Seq (25). We used the list of unique events that the RDP program produced by eliminating redundant events that were identified by different algorithms (18, 19). Among these recombination events, we used only those which were identified by >3 different algorithms with a Bonferroni-corrected *p*-value of < 0.001, which the RDP program calculated for each event (26, 27). We called these events reliable unique recombination events. Then, for each 200 nt sliding window, the number

of recombined regions that included the window region was counted. Close/distant types were defined in the same way as in reference (10).

**Correlation analysis between different genomic regions.** At every window position $w_x$ in the MGA, a windowed sequence alignment was extracted from the MGA, and all pairwise evolutionary distances between the extracted windowed sequences, $D^{w_x} = \left(d_{ij}^{w_x}\right)$, were calculated for $i$ and $j$ = 1..$N$ under the TN93 model, where $w_x$ is given as the region of the MGA between 200 ($x$ – 1) + 1 and 200 ($x$ + 1) in bp for $x$>0 and $N$ is the number of the sequences (=45). Then, the correlation coefficient $r_{w_x w_y}$ between windows $w_x$ and $w_y$ was calculated for all possible combinations of $x$ and $y$ using the following formula:

$$r_{w_x w_y} = \sum_{i<j}^{N}\left(d_{ij}^{w_x} - \overline{D^{w_x}}\right)\left(d_{ij}^{w_y} - \overline{D^{w_y}}\right) \Big/ \left(\sqrt{\sum_{i<j}^{N}\left(d_{ij}^{w_x} - \overline{D^{w_x}}\right)^2}\sqrt{\sum_{i<j}^{N}\left(d_{ij}^{w_y} - \overline{D^{w_y}}\right)^2}\right),$$

where $\overline{D^{w_x}}$ and $\overline{D^{w_y}}$ represent the means of the non-diagonal elements of $D^{w_x}$ and $D^{w_y}$, respectively.

The statistical significance of $r_{w_x w_y}$ was estimated by a permutation test similar to the Mantel test (28). Specific numbers (1,000 in the present study) of null samples, $D^{w_y\prime}$, were generated by repeating specific times (=1,000) repositions of the rows and columns of $D^{w_y}$ symmetrically based on a reshuffled order of the sequences (see below). Then, the $r_{w_x w_y\prime}$'s between $D^{w_x}$ and all $D^{w_y\prime}$ were calculated to obtain the percentage of cases showing $r_{w_x w_y} \geq r_{w_x w_y\prime}$. If this percentage reached 99.75%, $w_x$ and $w_y$ were regarded as being significantly correlated at $p < 0.0025$.

In order to avoid overestimating the significance, the sequence order in the permutation test above was reshuffled by means of a phylogenetic permutation method (29), in which the current positions of sequences $i$ and $j$ ($i \neq j$), denoted by $O_i$ and $O_j$, in the sequence order were

exchanged at the probability $p_{ij} = \left(d_{\max}^{\mathrm{pol}} - d_{ij}^{\mathrm{pol}}\right) / \sum_{k \neq l}^{N} \left(d_{\max}^{\mathrm{pol}} - d_{kl}^{\mathrm{pol}}\right)$, where pol represents the DNA polymerase region and $d_{\max}^{\mathrm{pol}}$ is the largest value in $D^{\mathrm{pol}}$. Before starting the reshuffling, the sequence order was initialized as $O_i = i$ for $i$=1..$N$. The reshuffling step was repeated 1,000 times for a single test, and the DNA polymerase-coding region was chosen as the rarest region for recombination (10, 30) to obtain the fundamental phylogeny of the genomes.

**Identification of basal genomic regions.** Basal regions were defined as the regions that have evolved together with the DNA polymerase-coding region in the HAdV-D genomes, i.e., all windows showing significantly high $r_{w_x \mathrm{pol}}$ at $p < 0.0025$. The remaining genomic regions were called non-basal regions.

**Partial correlation analysis.** The partial correlation coefficient adjusted for the correlation to the basal regions, $r_{w_x w_y \_\mathrm{basal}}$, between windows $w_x$ and $w_y$ was calculated using the following formula:

$r_{w_x w_y \_\mathrm{basal}} = \left(r_{w_x w_y} - r_{w_x \mathrm{basal}} r_{w_y \mathrm{basal}}\right) / \sqrt{\left(1 - r_{w_x \mathrm{basal}}^2\right)\left(1 - r_{w_y \mathrm{basal}}^2\right)}$, where $r_{w_x \mathrm{basal}}$ and $r_{w_y \mathrm{basal}}$ are the correlation coefficients between $D^{w_x}$ and the distance matrix of the concatenated basal region, $D^{\mathrm{basal}} = \left(d_{ij}^{\mathrm{basal}}\right)$, and between $D^{w_y}$ and $D^{\mathrm{basal}}$, respectively. The statistical significance of $r_{w_x w_y \_\mathrm{basal}}$ was estimated in the same way as for $r_{w_x w_y}$.

**Prediction of domains in membrane proteins.** Transmembrane domains were predicted using the transmembrane hidden Markov model prediction TMHMM 2.0 program (31) with the predicted amino acid sequences of protein-coding regions in the *E3* region of the HAdV-8 genome (AB448767).

## Results

**Identification of intertypic recombination events.** We first aligned 45 different HAdV-D types, including two hybrids, and obtained a MGA of 33,645 gap-free sites. We then applied the RDP program (18) to the MGA and identified 195 reliable unique recombination events (see Supplementary Table 1 for the list of these recombination events), indicating that each genome has experienced ca. 4.3 intertypic recombination events on average in the past. Figures 1A and 1B show the positions of the 195 recombined regions and the number of the recombination events detected at every position along the genome, respectively. Figure 1B looks similar to the results of our previous study on recombination boundary hotspots (Figs. 2A and 2B in reference (10)). This is not a matter-of-course observation of recombination events, since the number of recombination events that involved a specific region is not necessarily related to the number of recombination events starting and/or ending in the region. The high correlation observed between these two values indicates frequent recombination events of short segments in the recombination boundary hotspots. This trend can be confirmed in the size distribution of the recombined regions (Fig. 2) in which the mean and median of the sizes of the recombined regions were 2.5 kbp and 1.2 kbp, respectively.

**Identification of coevolving genomic regions.** The biased distribution of short recombined regions to specific genomic positions, including recombination boundary hotspots (10), implies that remaining regions have stayed in the same genome, even during series of frequent recombination events in each lineage. In order to confirm this implication, we sought to identify significantly high evolutionary correlations between different genomic regions at $p < 0.0025$ (the upper left triangle of Fig. 1C; hereafter, we call this matrix the coevolution matrix). All genomic regions of haploid organisms usually evolve together during evolution via base

198 substitutions, etc., independently of other lineages, and therefore their coevolution matrices
199 should be full of high correlation coefficients. Indeed, this is demonstrated with the coevolution
200 matrix of the artificial genomes that were computationally evolved via base substitutions (Fig. 3).
201 However, as shown in Figure 1C, although a large part of the genome seems to have consistently
202 coevolved, it is split into pieces by several uncorrelated regions. This coevolution pattern of the
203 majority of the genome is highly consistent with Figures 1A and 1B. Therefore, we conclude that,
204 while consistently coevolving major genomic regions have escaped most of the recombination
205 events, the other regions have been reshuffled between different types.
206 In order to summarize the evolutionary heterogeneity in the coevolution matrix of the
207 HAdV-D genomes, we calculated the ratio of the windows showing a significantly high
208 correlation to each specific window, named the coevolution ratio in this study, along the genome
209 (Fig. 1D) and made a histogram of the ratios (Fig. 4). The coevolution ratio histogram has two
210 peaks at around 0.7 as the major site and around 0.1 as the minor site. The major peak
211 corresponds to the plateaus of the correlation ratio plot (Fig. 1D), which match the positions of
212 the major coevolving regions. Since these regions seem to be the rarest recombination regions
213 and hence coevolving regions, we hereafter call them basal regions. On the other hand, the minor
214 peak in the coevolution ratio histogram indicates that small portions of the genome have
215 coevolved differently from the others. The coevolving regions for the minor peak are
216 recognizable as small triangle-like shapes along the diagonal of the coevolution matrix (Fig. 1C).
217 Note that a few windows do not show significant correlations with any or almost any of the other
218 windows (grayed regions in Fig. 1) due to too the high level of sequence conservation in these
219 windows (average distance and standard deviation are <0.014 and <0.008, respectively) to
220 contain sufficient information on the divergence history of the HAdV-D genomes. We call these

invariant regions. As shown in Figure 4, the coevolution ratio distribution of the simulated genomes is strongly biased to 1.0, as expected.

**Defining basal and non-basal genomic regions.** In order to operationally identify the basal regions, we defined basal regions as regions having coevolved with the DNA polymerase-coding region. We chose the DNA polymerase-coding region as a reference simply because it has been recognized to be one of the rarest recombination regions in the HAdV-D genomes (10, 30). We confirmed that the choice of the reference region does not have a significant impact on the identification of the basal regions if it is chosen from the areas of the plateaus at around a 70% coevolution ratio, as indicated in Figure 1D (data not shown). The criterion for classifying a window as basal or non-basal was whether the window was significantly correlated with this reference region at $p < 0.0025$ (Fig. 1E). Based on this operational definition, 130 of 167 sliding windows (77.8%) were determined to be basal windows (Table 1). We confirmed that the thus identified basal regions corresponded to the distribution of the major peak of the coevolution ratio histogram (Fig. 4). The slight elevation of this ratio (77.8%) is due to the fact that a few windows that showed coevolution ratios of < 0.7 (Fig. 1D) had correlation coefficients to the DNA polymerase region with a *p*-value of < 0.0025 (Fig. 1E).

In order to view the regional coevolution, rather than the basal one, we calculated partial correlation coefficients adjusted for the basal regions (see Materials and Methods). As shown in the lower right half of Figure 1C (named a non-basal coevolution matrix in this study), certain numbers of region pairs showed significantly high partial correlation coefficients (Supplementary Table 2), indicating that they have coevolved independently of the basal regions. Although some of the partial correlations, e.g., irregularly distributed small spots between basal

regions, are likely statistical errors, clear autocorrelations are seen as triangles along the diagonal of Figure 1C. Interestingly, correlations between separated non-basal regions were also observed (see below).

## Discussion

In the present study, we identified coevolving regions in the HAdV-D genomes and found that ca. 70% of the genome in total has coevolved as a whole even though it is split into several pieces by intervening genomic regions that have evolved differently. Since only a small number of recombination events were mapped in these major coevolving regions, and this seems to be the most probable explanation for the coevolution of such split regions, we regarded these regions as being evolutionarily basal regions, i.e., they have stayed in the same genomes during evolution as a platform/backbone of the recombination of non-basal regions. This observation is consistent with the finding in this study that most of the recombined regions are short (the median of the size is 1213 bp) and located in limited regions around the recombination boundary hotspots (10) and that such recombination hotspots intervene between basal regions. Our partial correlation analysis also showed that continuous non-basal regions are autocorrelated (Fig. 1C), evidencing the modularity of recombination, i.e., specific genomic segments have been recombined as modules (10). Note that this autocorrelation means that different parts of continuous regions have coevolved.

Since the homologous recombination mechanism, which has been speculated to be a mechanism responsible for the intertypic recombination events between HAdV-D genomes (10), does not generally specify the direction of recombination from the recombination initiation site

along the chromosome (32), it seems to be unlikely that only limited genomic regions have been frequently recombined in the HAdV-D genomes, even if homologous recombination events are initiated in highly conserved genomic regions, unless a specific site of a specific strand of the genome provides the 3' end of a single-stranded DNA for homologous recombination initiation. It also seems to be unlikely that under the conditions that HAdV-D genomes have experienced frequent recombination events (10), all different lineages of HAdV-D have escaped most of the recombination events in the basal regions by chance. Therefore, it would be more appropriate to reason that the coevolution between specific regions over different types is an outcome of purifying selection against intertypic recombination events within the basal regions and also in the non-basal evolving blocks due to biological constraints common to all types as well as the positive selection of specific forms recombined in the recombination hotspots.

As already mentioned above, it has been demonstrated using adenoviral packaging proteins, IVa2, *L1*:52/55K, IIIa and 22K, that replacing a component of a molecular system comprising several different protein components with a homologue protein from a different genome results in the functional impairment of the system (11, 12), indicating that such component replacements can be deleterious for the virus, and that recombinant forms bearing this type of change have been selectively eliminated. Interestingly enough, the coding regions of these proteins are all in basal or invariant region-containing basal regions. Including these examples, the basal regions were assigned in or over the coding regions of functionally interrelated (Table 1) and physically interacting protein genes: DNA polymerase, pTP and DNA-binding protein (DBP) for replication of the viral genome; IVa2, *L1*:52/55K, IIIa, 100K and 22K for packaging the viral DNA into an immature virion capsid (11, 12); IX, (IIIa,) VI and VIII as minor structural proteins for the interior and exterior of the virion nucleocapsid (2, 3, 33);

(IVa2,) VII, V and X for binding the viral DNA to the inside of the capsid and facilitating the transportation of the virion contents to the nucleus after viral entry (3); the protease for making IIIa, VI, VII, VIII, TP and X mature (34, 35); CR1α, RIDα, RIDβ and 14.7K for evading host cell apoptosis by blocking the host's TNF-R1, TRAIL-R1/-R2 and/or Fas (36-38); and 34K (*E4*:orf6), which interacts with *E1B*:55K to perform a variety of functions (39-46). Inverted terminal regions (ITRs) are known to contain replication origins where the DNA polymerase and TP bind together and initiate DNA replication (47, 48) and are also known to form the "panhandle structure" of the intra-molecular double helix for replication (49). DBP binds the viral DNA and enhances the replication initiation by DNA polymerase and pTP at the ITRs (47, 50). The multiple physical interactions of ITRs with DNA polymerase, pTP, DBP and themselves for replication are similar to the protein interactions mentioned above. These seem to indicate that multiple physical/functional protein-protein, protein-DNA, DNA-DNA interactions necessary for specific conserved functions may have prevented independent changes in these sequences during HAdV-D evolution. Although *E4*:orf1 and 12.5K also reside in basal regions, we have not succeeded in finding any reports of their specific functions, except for a report about the oncogenic activity of *E4*:orf1 of HAdV-9 in rats (51).

The remaining basal regions not yet mentioned above are mapped together with non-basal regions on the coding regions of 12/13S (we call it the *E1A* protein as well in accordance with previous works, e.g., reference (39)), *E1B*:19K, *E1B*:55K, penton base, hexon, gp19K and *E4*:orf2. Interestingly, the basal regions in the coding regions of two of the three major capsid proteins, penton base and hexon, correspond to the protein domains for the interaction with HAdV's IIIa, VI, VIII and IX, all of which are basal, as mentioned above and play a role together in cementing the bonds between hexons and penton bases to form the capsid

structure (2, 3, 11, 12, 33). While the macroscopic coevolution analysis showed that the whole coding region of fiber, the remaining major capsid protein, is a non-basal region (Fig. 1C), a finer scale analysis showed that a part of the N-terminal region is a short basal region located in the fiber tail domain (Fig. 5A) for anchoring the fiber in the penton base complex (2, 3, 33). These findings indicate that the protein domains/regions that physically interact with one another to form the viral capsid have also coevolved as basal regions. In the finer scale plot for the CR1α-CR1γ region in the *E3* gene, where only a single large non-basal region overlapping CR1β and CR1γ was detected by the macroscopic analysis, more complex basal-non-basal transitions were found (Fig. 5B). Interestingly, most of the basal windows, except for marginal ones, were mapped on or around transmembrane and cytoplasmic domains in CR1α and CR1β, implying that these regions may interact with other viral proteins. CR1α forms the CR1α-RIDαβ complex and co-immunoprecipitates with RIDβ, suggesting that CR1α interacts with RIDβ (36-38). The coding regions of RIDα and RIDβ are basal regions. These imply that the short basal region of CR1α may interact with RIDβ. We have no information about any interaction of CR1β with other adenoviral proteins.

Non-basal regions can be classified into two distinct types: invariant regions, which have been mentioned above, and variant regions. The invariant regions contain the major late promoter (MLP) and terminal regions of several coding/gene regions (Fig. 1C), implying the presence of conserved important functions in these regions. Eight variant non-basal regions (we call them non-basal regions below for simplicity) are found in or over the coding regions of the following proteins, most of which are categorized as either major capsid proteins, as mentioned above, or proteins for host modulation (Table 1): the penton base, hexon and fiber in the former category and 12/13S (the *E1A* protein), *E1B*:19K, *E1B*:55K, gp19K, CR1β, CR1γ, orf4, orf3 and

orf2 in the latter category (41, 52-54). The whole coding regions of CR1β, CR1γ, orf4 and orf3 reside within non-basal regions in the macroscopic analysis. The non-basal regions in three major capsid proteins contain epitope determinants: the RGD loop in penton base, loops L1 and L2 in the hexon and the shaft and knob in the fiber (55), all of which are exposed to the outside of the virions, without contact with other specific viral proteins, allowing these proteins to evolve independently of other adenoviral proteins. Similarly to these proteins, the predicted extracellular domains of the proteins shown in Figure 5B are largely non-basal regions. As seen in these non-basal regions, the co-evolvability is largely limited to themselves, i.e., autocorrelation, indicating modular recombination and functional relatedness. However, we found several cases of non-basal coevolution between separate genomic regions, even between distant regions, although it is difficult for us to tell which mechanisms and/or processes have produced such coevolution. Clear cases of such non-basal correlations are those between the coding regions of *E1A* (12S+13S), *E1B* (19K+55K) CR1β and fiber proteins (Fig. 1C), except between *E1A* proteins and the fiber (Supplementary Table 2). In addition, *E1A* and *E1B* show a correlation with different parts of IVa2. Although it is not known how *E1A* and *E1B* proteins are functionally related with IVa2, it has been demonstrated that 55K and IVa2 co-precipitate in immunoblot analyses (56), indicating their physical association. The partial correlations between *E1A, E1B* and IVa2 may be attributable to eight recombination events involving these regions between a limited set of types (Fig. 1A) (see Supplementary Table 1). Similarly, five recombination events were detected to exchange regions containing CR1β and the fiber. Although such minor co-recombination events seem to be a possible mechanism of the non-basal coevolution of closely located regions, how the distantly located regions, e.g., *E1A*-*E1B* and CR1β, have coevolved remains a challenging problem to address.

359 Evolutionary correlations between different genomic regions may be investigated by
360 comparing their phylogenetic trees instead of employing distance matrices. Tree-based
361 comparisons did not, however, produce so consistent results over the genome as the results of
362 this study (data not shown). This is partly due to the loss of information that occurs when
363 constructing phylogenetic trees from the distance matrices, which we directly used in the
364 correlation analyses. The general trend that different tree construction algorithms and/or different
365 parameter settings for the same tree construction algorithm can generate different trees, even
366 from the same distance matrix, may be also a relevant source of the inconsistency observed in
367 the tree-based analysis. Technical difficulty in measuring similarities between different trees is
368 another issue in tree-based approaches. Although many algorithms for comparing tree topologies
369 have been devised, e.g, the edit distance approach (57), the biological significance of the results
370 of tree topology comparisons using these algorithms is not necessarily evident. Therefore, we
371 decided to directly use distances to evaluate the evolutionary correlations between different
372 genomic regions in this study.
373 Our method revealed coevolving genomic regions, which may be continuous or
374 separate, and the identified coevolving regions contained the coding regions of proteins and/or
375 DNA elements that physically interact with one another to function. This method is applicable
376 not only to HAdV-D genomes, but also to any genome that has experienced recombination
377 events and/or lateral gene transfers between different genomes, to detect interregional
378 coevolution, which implies protein-protein and protein-DNA physical interactions. In addition,
379 many protein function prediction methods have been devised thus far, e.g., homology-based
380 methods, sequence motif-based methods, structure-based methods, genomic context-based
381 methods, including those using information about gene fusion in the Rosetta stone approach and

the co-location/co-expression, and network-based methods (58). However, our present method does not belong to any of these categories. Therefore, this study provides an additional new means of predicting the functions of proteins/DNA regions and protein/DNA interactions.

## Acknowledgements

This work was partly supported by MEXT KAKENHI Grant Number 22125009 (to H. W.).

## References


1. **Harrach B, Benkö M, Both GW, Brown M, Davison AJ, Echavarría M, Hess M, S. JM, Kajon A, Lehmkuhl HD, Mautner V, Mittal SK, Wadell G.** 2011. Family Adenoviridae, p 125-141. *In* Viruses ICoTo, King AMQ, Adams MJ, Lefkowitz EJ, Carstens EB (ed), Virus Taxonomy: Classification and Nomenclature of Viruses: Ninth Report of the International Committee on Taxonomy of Viruses. Academic Press.
2. **Davison AJ, Benko M, Harrach B.** 2003. Genetic content and evolution of adenoviruses. J. Gen. Virol. **84:**2895-2908.
3. **Russell WC.** 2009. Adenoviruses: update on structure and function. J. Gen. Virol. **90:**1-20.
4. **Matsushima Y, Shimizu H, Phan TG, Ushijima H.** 2011. Genomic characterization of a novel human adenovirus type 31 recombinant in the hexon gene. J. Gen. Virol. **92:**2770-2775.
5. **Dehghan S, Liu EB, Seto J, Torres SF, Hudson NR, Kajon AE, Metzgar D, Dyer DW, Chodosh J, Jones MS, Seto D.** 2012. Five genome sequences of subspecies B1 human adenoviruses associated with acute respiratory disease. J. Virol. **86:**635-636.
6. **Walsh MP, Seto J, Liu EB, Dehghan S, Hudson NR, Lukashev AN, Ivanova O, Chodosh J, Dyer DW, Jones MS, Seto D.** 2011. Computational analysis of two species C human adenoviruses provides evidence of a novel virus. J. Clin. Microbiol. **49:**3482-3490.
7. **Matsushima Y, Shimizu H, Kano A, Nakajima E, Ishimaru Y, Dey SK, Watanabe Y, Adachi F, Mitani K, Fujimoto T, Phan TG, Ushijima H.** 2013. Genome sequence of a novel virus of the species Human adenovirus D associated with acute gastroenteritis. Genome Announcements **1:**e00068-00012.
8. **Robinson CM, Seto D, Jones MS, Dyer DW, Chodosh J.** 2011. Molecular evolution of human species D adenoviruses. Infect Genet Evol **11:**1208-1217.
9. **Aoki K, Benko M, Davison AJ, Echavarria M, Erdman DD, Harrach B, Kajon AE, Schnurr D, Wadell G, Members of the Adenovirus Research C.** 2011. Toward an integrated human adenovirus designation system that utilizes molecular and serological data and serves both clinical and fundamental virology. J Virol **85:**5703-5704.
10. **Gonzalez G, Koyanagi KO, Aoki K, Kitaichi N, Ohno S, Kaneko H, Ishida S, Watanabe H.** 2014. Intertypic modular exchanges of genomic segments by homologous recombination at universally conserved segments in human adenovirus species D. Gene **547:**10-17.
11. **Ma HC, Hearing P.** 2011. Adenovirus structural protein IIIa is involved in the serotype specificity of viral DNA packaging. J. Virol. **85:**7849-7855.
12. **Wohl BP, Hearing P.** 2008. Role for the L1-52/55K protein in the serotype specificity of adenovirus DNA packaging. J. Virol. **82:**5089-5092.
13. **Kaneko H, Aoki K, Ishida S, Ohno S, Kitaichi N, Ishiko H, Fujimoto T, Ikeda Y, Nakamura M, Gonzalez G, Koyanagi KO, Watanabe H, Suzutani T.** 2011. Recombination analysis of intermediate human adenovirus type 53 in Japan by complete genome sequence. J. Gen. Virol. **92:**1251-1259.
14. **Katoh K, Misawa K, Kuma K, Miyata T.** 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. **30:**3059-3066.

15. **Saitou N, Nei M.** 1987. The Neighbor-Joining method - a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:**406-425.
16. **Tamura K, Nei M.** 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial-DNA in humans and chimpanzees. Mol. Biol. Evol. **10:**512-526.
17. **Shull VL, Vogler AP, Baker MD, Maddison DR, Hammond PM.** 2001. Sequence alignment of 18S ribosomal RNA and the basal relationships of Adephagan beetles: Evidence for monophyly of aquatic families and the placement of Trachypachidae. Syst Biol **50:**945-969.
18. **Heath L, van der Walt E, Varsani A, Martin DP.** 2006. Recombination patterns in aphthoviruses mirror those found in other picornaviruses. J. Virol. **80:**11827-11832.
19. **Martin D, Rybicki E.** 2000. RDP: detection of recombination amongst aligned sequences. Bioinformatics **16:**562-563.
20. **Padidam M, Sawyer S, Fauquet CM.** 1999. Possible emergence of new geminiviruses by frequent recombination. Virology **265:**218-225.
21. **Posada D, Crandall KA.** 2001. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. Proceedings of the National Academy of Sciences of the United States of America **98:**13757-13762.
22. **Smith JM.** 1992. Analyzing the mosaic structure of genes. J. Mol. Evol. **34:**126-129.
23. **Martin DP, Posada D, Crandall KA, Williamson C.** 2005. A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. AIDS Res. Hum. Retroviruses **21:**98-102.
24. **Gibbs MJ, Armstrong JS, Gibbs AJ.** 2000. Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. Bioinformatics **16:**573-582.
25. **Boni MF, Posada D, Feldman MW.** 2007. An exact nonparametric method for inferring mosaic structure in sequence triplets. Genetics **176:**1035-1047.
26. **Lefeuvre P, Lett JM, Reynaud B, Martin DP.** 2007. Avoidance of protein fold disruption in natural virus recombinants. PLoS Pathogens **3:**e181.
27. **Lefeuvre P, Lett JM, Varsani A, Martin DP.** 2009. Widely conserved recombination patterns among single-stranded DNA viruses. J. Virol. **83:**2697-2707.
28. **Mantel N.** 1967. The detection of disease clustering and a generalized regression approach. Cancer research **27:**209-220.
29. **Lapointe FJ, Garland T.** 2001. A generalized permutation model for the analysis of cross-species data. J Classif **18:**109-127.
30. **Liu H, Naismith JH, Hay RT.** 2000. Identification of conserved residues contributing to the activities of adenovirus DNA polymerase. J Virol **74:**11681-11689.
31. **Sonnhammer EL, von Heijne G, Krogh A.** 1998. A hidden Markov model for predicting transmembrane helices in protein sequences. Proc Int Conf Intell Syst Mol Biol **6:**175-182.
32. **Krejci L, Altmannova V, Spirek M, Zhao XL.** 2012. Homologous recombination and its regulation. Nucleic acids research **40:**5795-5818.
33. **Liu H, Jin L, Koh SB, Atanasov I, Schein S, Wu L, Zhou ZH.** 2010. Atomic structure of human adenovirus by cryo-EM reveals interactions among protein networks. Science **329:**1038-1043.
34. **Matthews DA, Russell WC.** 1995. Adenovirus protein-protein interactions: molecular-parameters governing the binding of protein VI to hexon and the activation of the adenovirus 23k protease. J. Gen. Virol. **76:**1959-1969.

35. **Perez-Berna AJ, Ortega-Esteban A, Menendez-Conejero R, Winkler DC, Menendez M, Steven AC, Flint SJ, de Pablo PJ, San Martin C.** 2012. The role of capsid maturation on adenovirus priming for sequential uncoating. J. Biol. Chem. **287:**31582-31595.
36. **Benedict CA, Norris PS, Prigozy TI, Bodmer JL, Mahr JA, Garnett CT, Martinon F, Tschopp J, Gooding LR, Ware CF.** 2001. Three adenovirus E3 proteins cooperate to evade apoptosis by tumor necrosis factor-related apoptosis-inducing ligand receptor-1 and-2. J Biol Chem **276:**3270-3278.
37. **Tollefson AE, Toth K, Doronin K, Kuppuswamy M, Doronina OA, Lichtenstein DL, Hermiston TW, Smith CA, Wold WSM.** 2001. Inhibition of TRAIL-induced apoptosis and forced internalization of TRAIL receptor 1 by adenovirus proteins. J Virol **75:**8875-8887.
38. **Schneider-Brachert W, Tchikov V, Merkel O, Jakob M, Hallas C, Kruse ML, Groitl P, Lehn A, Hildt E, Held-Feindt J, Dobner T, Kabelitz D, Kronke M, Schutze S.** 2006. Inhibition of TNF receptor 1 internalization by adenovirus 14.7K as a novel immune escape mechanism. J Clin Invest **116:**2901-2913.
39. **Berk AJ.** 2005. Recent lessons in gene expression, cell cycle control, and cell biology from adenovirus. Oncogene **24:**7673-7685.
40. **Leppard KN.** 1997. E4 gene function in adenovirus, adenovirus vector and adeno-associated virus infections. J Gen Virol **78:**2131-2138.
41. **Chinnadurai G.** 1998. Control of apoptosis by human adenovirus genes. Semin Virol **8:**399-408.
42. **Boyer JL, Ketner G.** 2000. Genetic analysis of a potential zinc-binding domain of the adenovirus E4 34k protein. J Biol Chem **275:**14969-14978.
43. **Rubenwolf S, Schutt H, Nevels M, Wolf H, Dobner T.** 1997. Structural analysis of the adenovirus type 5 E1B 55-kilodalton-E4orf6 protein complex. J Virol **71:**1115-1123.
44. **Querido E, Morrison MR, Chu-Pham-Dang H, Thirlwell SWL, Boivin D, Branton PE.** 2001. Identification of three functions of the adenovirus E4orf6 protein that mediate p53 degradation by the E4orf6-E1B55K complex (vol 75, pg 699, 2001). J Virol **75:**2508-2508.
45. **Blackford AN, Grand RJA.** 2009. Adenovirus E1B 55-Kilodalton protein: multiple roles in viral infection and cell transformation. J Virol **83:**4000-4012.
46. **Kato SEM, Huang WY, Flint SJ.** 2011. Role of the RNA recognition motif of the E1B 55 kDa protein in the adenovirus type 5 infectious cycle. Virology **417:**9-17.
47. **de Jong RN, van der Vliet PC, Brenkman AB.** 2003. Adenovirus DNA replication: protein priming, jumping back and the role of the DNA binding protein DBP. Curr. Top. Microbiol. Immunol. **272:**187-211.
48. **Webster A, Leith IR, Nicholson J, Hounsell J, Hay RT.** 1997. Role of preterminal protein processing in adenovirus replication. J Virol **71:**6381-6389.
49. **Lippe R, Graham FL.** 1989. Adenoviruses with nonidentical terminal sequences are viable. J Virol **63:**5133-5141.
50. **Brenkman AB, Breure EC, van der Vliet PC.** 2002. Molecular architecture of adenovirus DNA polymerase and location of the protein primer. J. Virol. **76:**8200-8207.
51. **Javier RT.** 1994. Adenovirus type-9 E4 open reading frame-1 encodes a transforming protein required for the production of mammary-tumors in rats. J Virol **68:**3917-3924.
52. **Miron MJ, Gallouzi IE, Lavoie JN, Branton PE.** 2005. Nuclear localization of the adenovirus E4orf4 protein is mediated through an arginine-rich motif and correlates with cell death (vol 23, pg 7458, 2004). Oncogene **24:**4162-4162.

53. **Zheng ZM.** 2010. Viral oncogenes, noncoding RNAs, and RNA splicing in human tumor viruses. Int J Biol Sci **6:**730-755.
54. **McSharry BP, Burgert HG, Owen DP, Stanton RJ, Prod'homme V, Sester M, Koebernick K, Groh V, Spies T, Cox S, Little AM, Wang ECY, Tomasec P, Wilkinson GWG.** 2008. Adenovirus E3/19K promotes evasion of NK cell recognition by intracellular sequestration of the NKG2D ligands, major histocompatibility complex class I chain-related proteins A and B. J Virol **82:**4585-4594.
55. **Gahery-Segard H, Farace F, Godfrin D, Gaston J, Lengagne R, Tursz T, Boulanger P, Guillet JG.** 1998. Immune response to recombinant capsid proteins of adenovirus in humans: antifiber and anti-penton base antibodies have a synergistic effect on neutralizing activity. J. Virol. **72:**2388-2397.
56. **Harada JN, Shevchenko A, Shevchenko A, Pallas DC, Berk AJ.** 2002. Analysis of the adenovirus E1B-55K-anchored proteome reveals its link to ubiquitination machinery. J Virol **76:**9194-9206.
57. **Bille P.** 2005. A survey on tree edit distance and related problems. Theor Comput Sci **337:**217-239.
58. **Lee D, Redfern O, Orengo C.** 2007. Predicting protein function from sequence and structure. Nat Rev Mol Cell Bio **8:**995-1005.

551 **Table 1** List of basal and non-basal regions with overlapping functional regions

| Basal/Non-Basal/Invariant | MGA | | HAdV-8 (AB448767) | | Annotation in HAdV-8 (AB448767) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Start** | **End** | **Start** | **End** | ***Gene***:protein[a] | **Start** | **End** | **Functional Category** | **Characterized roles** | **Reference** |
| Basal | 1 | 600 | 1 | 625 | **NC** | | | DNA replication | Starting point for replication | (47, 50) |
| Non-basal | 601 | 800 | 626 | 831 | ***E1A*:12/13S** | 564 | 1113 | Host modulation | Promote p53-dependent apoptosis. Promotion of viral transcriptional activation | (53) |
| Basal | 801 | 1600 | 832 | 1648 | | 1209 | 1420 | | | |
| | | | | | ***E1B*:19K** | 1572 | 2120 | Host modulation | Suppresses both p53-dependent and -independent apoptosis induced during adenovirus infection | (39) |
| Non-basal | 1601 | 1800 | 1649 | 1878 | | | | | | |
| | | | | | ***E1B*:55K** | 1877 | 3364 | Host modulation | Inactivate p53 and p53-dependent apoptosis | (39) |
| Basal | 1801 | 2600 | 1879 | 2648 | | | | | | |
| Non-basal | 2601 | 3200 | 2649 | 3248 | | | | | | |
| Invariant | 3201 | 3400 | 3249 | 3451 | | | | | | |
| Basal | 3401 | 4200 | 3452 | 4267 | ***IX*:IX** | 3449 | 3856 | Structural | Minor structural protein, stabilizer of the capsid | (2, 3, 33) |
| | | | | | ***IVa2*:IVa2** | 5233 | 3900 | Core protein/Genome Packaging | Genome packaging process | (3) |
| Invariant | 4201 | 4400 | 4268 | 4467 | | | | | | |
| Basal | 4401 | 5400 | 4468 | 5467 | | | | | | |
| | | | | | ***E2B*:DNA pol** | 8278 | 5003 | DNA replication | Replication of the viral DNA | (47, 50) |
| Invariant | 5401 | 5600 | 5468 | 5667 | | | | | | |
| Basal | 5601 | 8200 | 5668 | 8267 | | | | | | |
| Invariant | 8201 | 8400 | 8268 | 8473 | ***E2B*:pTP** | 10218 | 8323 | DNA replication | Enables protein-priming start of the replication | (47) |
| Basal | 8401 | 11600 | 8474 | 11797 | | | | | | |
| | | | | | ***L1*:52/55K** | 10632 | 11750 | Genome Packaging | Virus assembly by interacting with pVII | (3) |
| | | | | | ***L1*:IIIa** | 11773 | 13470 | Structural/Genome Packaging | Minor structural protein | (2, 3) |
| Invariant | 11601 | 11800 | 11798 | 11997 | | | | | | |
| Basal | 11801 | 14000 | 11998 | 14233 | | | | | | |
| | | | | | ***L2*:Penton base** | 13524 | 15086 | Major capsid protein | Major structural protein, binds the fiber to the capsid, endocytosis of the virion | (3) |
| Non-basal | 14001 | 14200 | 14234 | 14493 | | | | | | |
| Basal | 14201 | 17200 | 14497 | 17642 | | | | | | |
| | | | | | ***L2*:VII** | 15090 | 15680 | Core protein/Maturation | Viral DNA import to the nucleus of the infected cell | (3) |
| | | | | | ***L2*:V** | 15713 | 16696 | Core protein/Maturation | Suggested as mediator of viral DNA transport to the nucleus of the infected cell | (3) |
| | | | | | ***L2*:X** | 16726 | 16950 | Core protein/Maturation | Suggested as mediator of viral prechromatin condensation and facilitate packaging of the core complex | (3) |
| | | | | | ***L3*:VI** | 17006 | 17707 | Structural/Maturation | Minor structural protein, activation of protease | (2, 3, 33) |
| Invariant | 17201 | 17400 | 17643 | 17852 | | | | | | |
| | | | | | ***L3*:Hexon** | 17776 | 20604 | Major capsid protein | Major structural protein | (2, 3, 33) |
| Basal | 17401 | 17600 | 17853 | 18052 | | | | | | |
| Non-basal | 17601 | 18600 | 18053 | 19164 | | | | | | |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Basal | 18601 | 26600 | 19165 | 27259 | ***L3*:protease** | 20607 | 21230 | Genome Packaging/Maturation | Cleavages proteins into maturity the adenoviral proteins: IIIa, VI, VII, VIII, pTP and X | (34, 35) |
| | | | | | ***E2A*:DBP** | 22755 | 21283 | DNA replication | Binds single-stranded DNA displaced during genome replication and is required for initiation and elongation of replication | (47) |
| | | | | | ***L4*:100K** | 22772 | 24919 | Genome Packaging | Viral genome packaging | (11, 12) |
| | | | | | ***L4*:22K** | 24732 | 25115 | Genome Packaging | Viral genome packaging by interacting with IVa2 | (3) |
| | | | | | ***L4*:VIII** | 25444 | 26127 | Structural/Maturation | Minor structural protein | (2, 3, 33) |
| | | | | | ***E3*:12.5K** | 26128 | 26448 | Host modulation | | |
| | | | | | ***E3*:CR1α** | 26402 | 26956 | Host modulation | Down modulation of TRAIL receptors | (36) |
| | | | | | ***E3*:gp19K** | 26953 | 27426 | Host modulation | Inhibition T-cell recognition and NK cell activation | (54) |
| Non-basal | 26601 | 28400 | 27260 | 29467 | ***E3*:CR1β** | 27452 | 28666 | Host modulation | *Not fully characterized* | |
| | | | | | ***E3*:CR1γ**[b] | 28693 | 29472 | Host modulation | *Not fully characterized* | |
| Basal | 28401 | 29600 | 29468 | 30728 | ***E3*:RIDα** | 29479 | 29754 | Host modulation | Evasion of TNF-apoptosis | (36) |
| | | | | | ***E3*:RIDβ** | 29757 | 30146 | Host modulation | Evasion of TNF-apoptosis | (36) |
| | | | | | ***E3*:14.7K** | 30139 | 30531 | Host modulation | Evasion of TNF-apoptosis | (38, 41) |
| Non-basal | 29601 | 30800 | 30729 | 32020 | ***L5*:Fiber** | 30785 | 31873 | Major capsid protein | Major structural protein, binding to cellular receptor, tissue affinity | (33) |
| Basal | 30801 | 31800 | 32021 | 33020 | ***E4*:34K**[b] | 33027 | 32149 | Host modulation | Inactivate p53 and p53-dependent apoptosis. Viral late gene expression | (41, 42) |
| Non-basal | 31801 | 32800 | 33021 | 34024 | ***E4*:orf4** | 33319 | 32957 | Host modulation | Lysis of infected cell | (41, 52) |
| | | | | | ***E4*:orf3** | 33675 | 33322 | Host modulation | Promotion of viral gene expression and replication | (40) |
| | | | | | ***E4*:orf2** | 34064 | 33672 | Host modulation | *Not fully characterized* | |
| Basal | 32801 | 33401 | 34025 | 34633 | ***E4*:orf1** | 34302 | 34105 | Host modulation | Lytic infection and oncogenesis | (40) |
| | | | | | **NC** | | | DNA replication | Starting point for replication | (47, 50) |

[a] NC stands for non-coding region

[b] A small section of the coding region (<10 nt) falls in a different region

## Figure Legends.

**FIG 1 Regional recombination and coevolution.** The abscissae of all panels represent the positions adjusted to HAdV-8 (AB448767) as reference. (A) **Positions of the 195 identified recombined regions.** Each black line represents a recombined region. (B) **Number of recombination events in each 200 bp window.** Green and magenta dots mean basal and non-basal regions, respectively. The upper dots in a lighter color and the lower dots in a darker color represent the counts of all recombined regions and those between distant types only, respectively. (C) **Merged coevolution (upper left) and partial coevolution (lower right) matrices.** The values of significant correlation/partial correlation coefficients ($p <$ 0.0025) are shown using color gradient ranging from near 0 (yellow) to 1.0 (red). The diagonal is shown in gray. The basal (green), non-basal (magenta) and invariant (gray) regions are indicated at the bottom of the matrix (details are in Table 1). Genes (thick arrows) and protein-coding regions (black arrows) are shown around the matrix. (D) **Ratio of the windows showing significant correlations to the window at each position.** (E) **Correlation coefficient of each window against the entire DNA polymerase-coding region.** Significant ($p < 0.0025$) correlation coefficients are shown with green discs, corresponding to the basal regions, and the others are presented with magenta circles.

**FIG 2 Distribution of recombined segments lengths**. The abscissa shows the different lengths while the ordinate shows the frequency by each size category. Lengths of the recombined segments were adjusted to match Fig. 1A.

**FIG 3 Evolutionary correlation on simulated sequences**. The simulated multiple genome alignment of 45 artificial genomic sequences were generated by simulating sequence

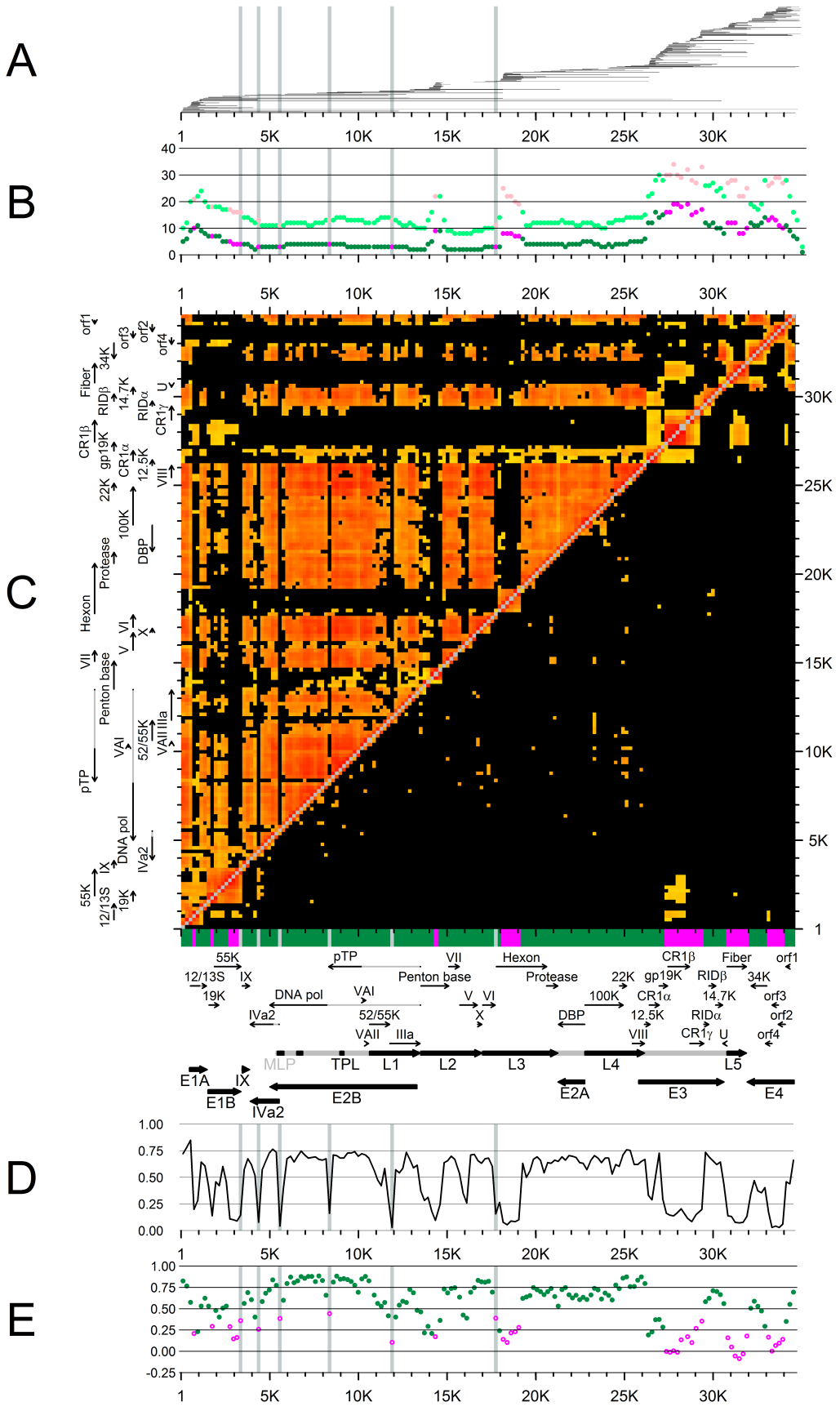
578 evolution using Mesquite version 2.75 under the following conditions: the tree topology = the
579 genome tree; the number of characters = 33,645; the ratio of invariant sites = 0.65; the alpha
580 parameter of the gamma distribution of rate variance = 0.477; nucleotide frequencies of A, T,
581 G, and C = 0.22, 0.21, 0.29, and 0.28, respectively; transition/transversion ratio = 1.57. The
582 abscissa and ordinate (the x and y axes) of this matrix represent the physical positions in the
583 simulated MGA, and each point (x, y) of the matrix shows the Mantel's correlation
584 coefficient between windows x and y. The correlation coefficient ranges from near zero
585 (yellow) to near one (red). Independent and diagonal windows are colored black and gray,
586 respectively.
587
588 **FIG 4 Histogram of significant correlation ratios.** The abscissa is the ratio of the number of
589 windows that show a significant correlation coefficient to a specific window against the total
590 number of windows (=167). The left and right ordinates are for the absolute frequencies (bars)
591 and relative cumulative frequencies (lines), respectively. The gray bars and line are for the
592 simulated data. The absolute frequencies in basal and non-basal regions of the real data are
593 shown in black and mesh bars, respectively, together with cumulative frequencies in the black
594 line.
595
596 **FIG 5 Finer-scale coevolution analysis.** The results of the finer scale analyses are depicted
597 for two highlighted regions, (A) fiber and (B) *E3* region. The abscissae show the position in
598 the HAdV-8 genome (AB448767). The left ordinates represent the correlation coefficients
599 between each 100 bp-window and the entire DNA polymerase-coding region. Significantly
600 correlated windows ($p$<0.0025) are shown with disks, equivalent to a basal region, and the
601 others are presented with open circles. Protein-coding regions are shown with arrows below

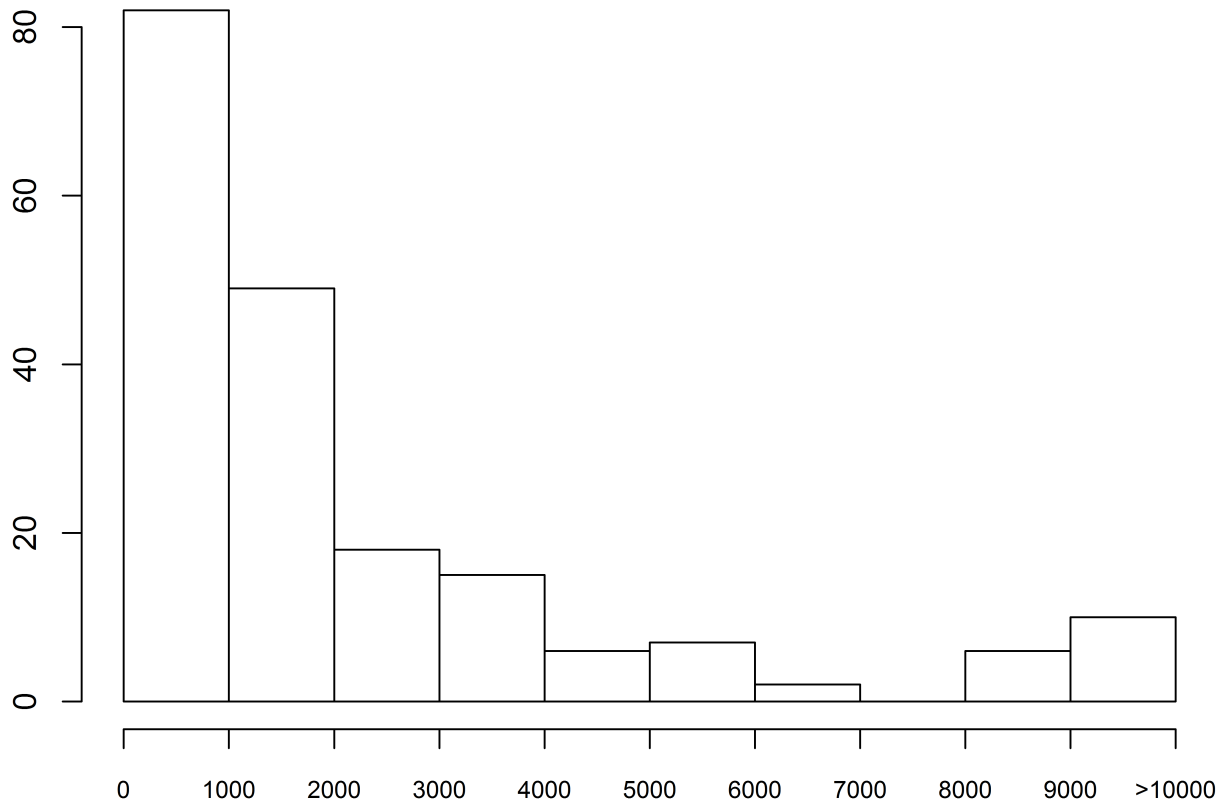
602 each plot. The predicted extracellular (O), transmembrane (M) and cytoplasmic (I) regions

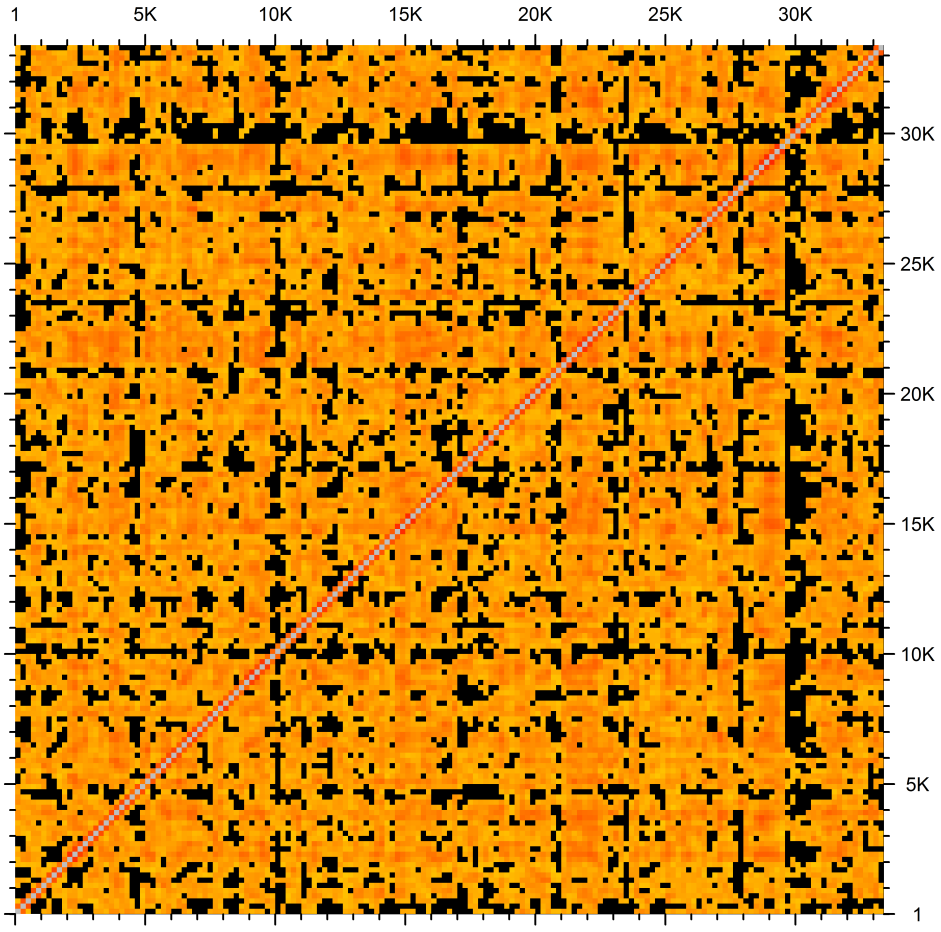
603 are shown in the coding regions of CR1α, gp19K, CR1β, and CR1γ.

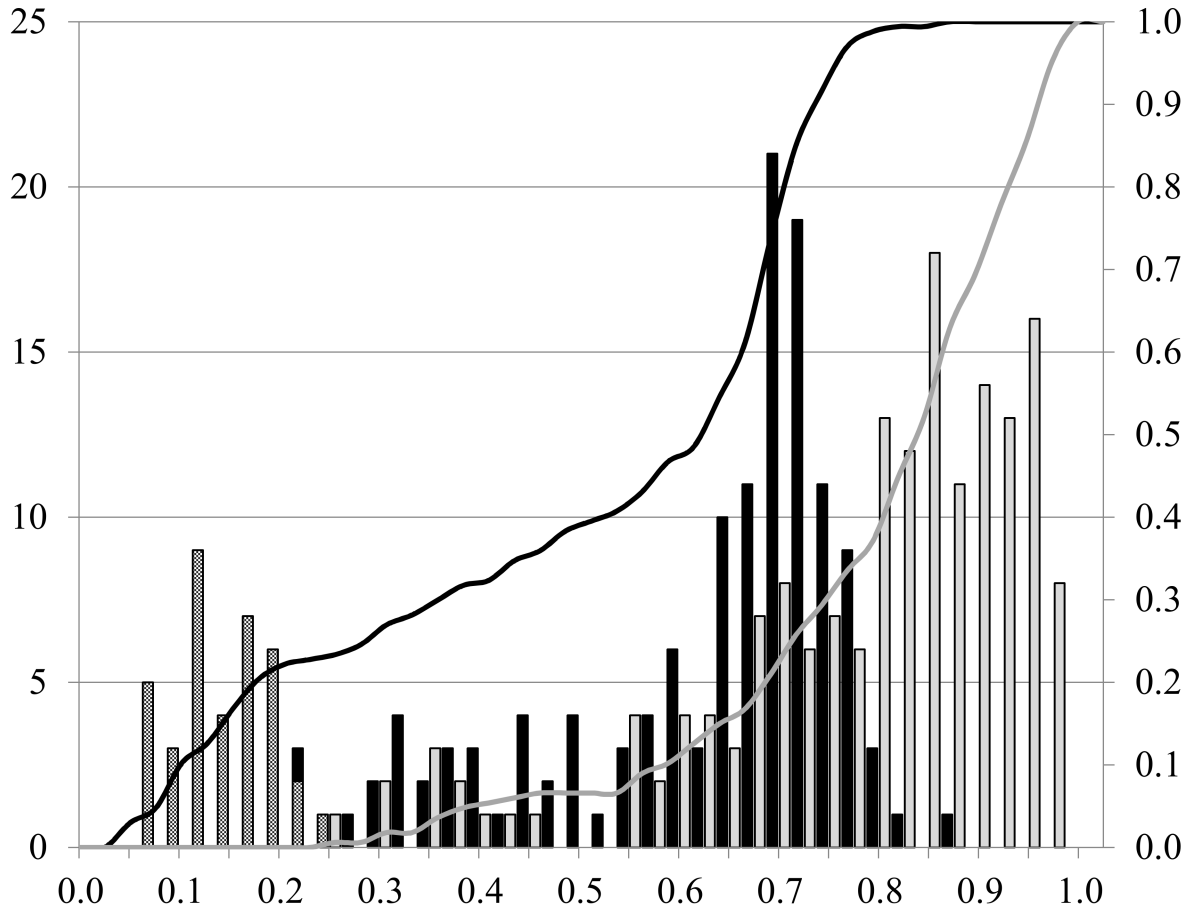
604

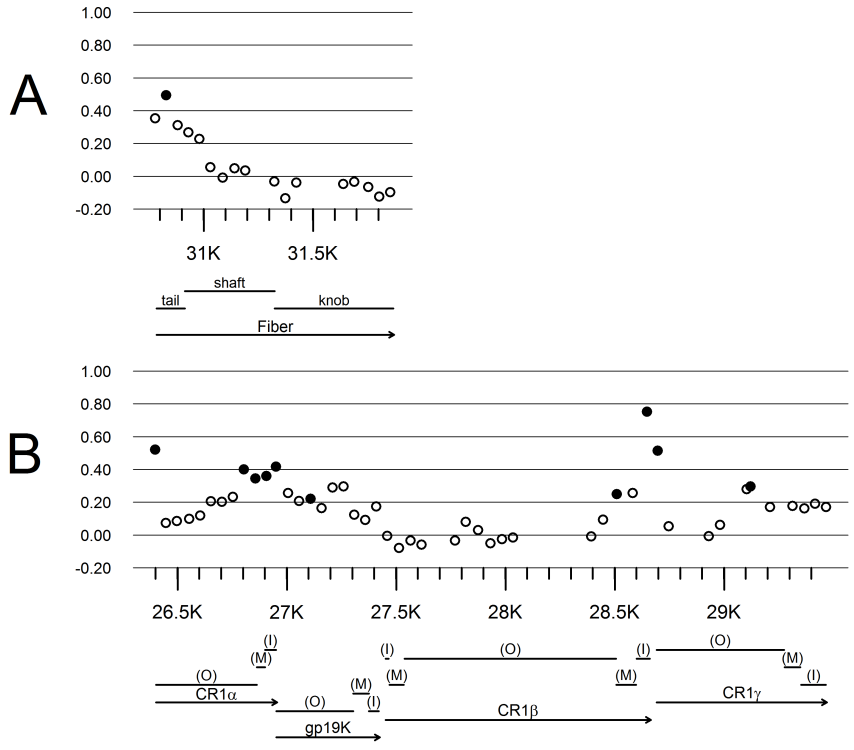
605


A
B
C
D
E
1
5K
10K
15K
20K
25K
30K
40
30
20
10
0
orf1
orf3
orf2
orf4
Fiber
34K
RIDβ
14.7K
RIDα
U
CR1γ
CR1β
gp19K
CR1α
12.5K
VIII
22K
100K
DBP
Protease
Hexon
VI
X
V
VII
Penton base
VAI
52/55K
VAII
IIIa
pTP
DNA pol
IVa2
IX
55K
12/13S
19K
MLP
TPL
L1
L2
L3
L4
L5
E1A
E1B
E2B
E2A
E3
E4
1.00
0.75
0.50
0.25
0.00
-0.25


0
20
40
60
80
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
>10000


1
5K
10K
15K
20K
25K
30K
30K
25K
20K
15K
10K
5K
1


25
20
15
10
5
0
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0


A
1.00
0.80
0.60
0.40
0.20
0.00
-0.20
31K
31.5K
shaft
tail
knob
Fiber
B
1.00
0.80
0.60
0.40
0.20
0.00
-0.20
26.5K
27K
27.5K
28K
28.5K
29K
(I)
(M)
(O)
CR1α
(O)
(M)
(I)
gp19K
(I)
(O)
(M)
(M)
CR1β
(I)
(O)
(M)
(I)
CR1γ