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#{Address correspondence to Hidemi Watanabe, Graduate School of Information Science and Technology, Hokkaido University, Sapporo, 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, \( L_1:52/55K \), \( IVa2:IVa2 \), \( L_4:22K \) and \( L_1:IIIa \), 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, \(D^{w_x} = (d_{ij}^{w_x})\), were
calculated for \(i \text{ and } j = 1..N\) 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:

\[
r_{w_x w_y} = \frac{\sum_{i<j} (d_{ij}^{w_x} - \overline{D}_{w_x}) (d_{ij}^{w_y} - \overline{D}_{w_y})}{\sqrt{\sum_{i<j} (d_{ij}^{w_x} - \overline{D}_{w_x})^2 \sum_{i<j} (d_{ij}^{w_y} - \overline{D}_{w_y})^2}},
\]

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, \(D^{w_y'}\), 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 \(r_{w_x w_y'}\)'s
between \(D^{w_x}\) and all \(D^{w_y'}\) were calculated to obtain the percentage of cases showing \(r_{w_x w_y} \geq r_{w_x w_y'}\). 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 \( p_{ij} = \left( d_{\text{max}}^{\text{pol}} - d_{ij}^{\text{pol}} \right) / \sum_{k \neq i} N \left( d_{\text{max}}^{\text{pol}} - d_{ki}^{\text{pol}} \right) \), where pol represents the DNA polymerase region and \( d_{\text{max}}^{\text{pol}} \) is the largest value in \( D^{\text{pol}} \). 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 \( r_{wx,\text{pol}} \) 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, \( r_{wx,wy,\text{basal}} \), between windows \( w_x \) and \( w_y \) was calculated using the following formula:

\[
r_{wx,wy,\text{basal}} = \frac{\left( r_{wx,wy} - r_{wx,\text{basal}} r_{wy,\text{basal}} \right)}{\sqrt{(1 - r_{wx,\text{basal}}^2)(1 - r_{wy,\text{basal}}^2)}},
\]

where \( r_{wx,\text{basal}} \) and \( r_{wy,\text{basal}} \) are the correlation coefficients between \( D_{wx} \) and the distance matrix of the concatenated basal region, \( D_{\text{basal}} = (d_{ij}^{\text{basal}}) \), and between \( D_{wy} \) and \( D_{\text{basal}} \), respectively. The statistical significance of \( r_{wx,wy,\text{basal}} \) was estimated in the same way as for \( r_{wx,wy} \).

**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
substitutions, etc., independently of other lineages, and therefore their coevolution matrices should be full of high correlation coefficients. Indeed, this is demonstrated with the coevolution matrix of the artificial genomes that were computationally evolved via base substitutions (Fig. 3). However, as shown in Figure 1C, although a large part of the genome seems to have consistently coevolved, it is split into pieces by several uncorrelated regions. This coevolution pattern of the majority of the genome is highly consistent with Figures 1A and 1B. Therefore, we conclude that, while consistently coevolving major genomic regions have escaped most of the recombination events, the other regions have been reshuffled between different types.

In order to summarize the evolutionary heterogeneity in the coevolution matrix of the HAdV-D genomes, we calculated the ratio of the windows showing a significantly high correlation to each specific window, named the coevolution ratio in this study, along the genome (Fig. 1D) and made a histogram of the ratios (Fig. 4). The coevolution ratio histogram has two peaks at around 0.7 as the major site and around 0.1 as the minor site. The major peak corresponds to the plateaus of the correlation ratio plot (Fig. 1D), which match the positions of the major coevolving regions. Since these regions seem to be the rarest recombination regions and hence coevolving regions, we hereafter call them basal regions. On the other hand, the minor peak in the coevolution ratio histogram indicates that small portions of the genome have coevolved differently from the others. The coevolving regions for the minor peak are recognizable as small triangle-like shapes along the diagonal of the coevolution matrix (Fig. 1C). Note that a few windows do not show significant correlations with any or almost any of the other windows (grayed regions in Fig. 1) due to too the high level of sequence conservation in these windows (average distance and standard deviation are <0.014 and <0.008, respectively) to 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
elements, 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 EIA 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.
Evolutionary correlations between different genomic regions may be investigated by comparing their phylogenetic trees instead of employing distance matrices. Tree-based comparisons did not, however, produce so consistent results over the genome as the results of this study (data not shown). This is partly due to the loss of information that occurs when constructing phylogenetic trees from the distance matrices, which we directly used in the correlation analyses. The general trend that different tree construction algorithms and/or different parameter settings for the same tree construction algorithm can generate different trees, even from the same distance matrix, may be also a relevant source of the inconsistency observed in the tree-based analysis. Technical difficulty in measuring similarities between different trees is another issue in tree-based approaches. Although many algorithms for comparing tree topologies have been devised, e.g., the edit distance approach (57), the biological significance of the results of tree topology comparisons using these algorithms is not necessarily evident. Therefore, we decided to directly use distances to evaluate the evolutionary correlations between different genomic regions in this study.

Our method revealed coevolving genomic regions, which may be continuous or separate, and the identified coevolving regions contained the coding regions of proteins and/or DNA elements that physically interact with one another to function. This method is applicable not only to HAdV-D genomes, but also to any genome that has experienced recombination events and/or lateral gene transfers between different genomes, to detect interregional coevolution, which implies protein-protein and protein-DNA physical interactions. In addition, many protein function prediction methods have been devised thus far, e.g., homology-based methods, sequence motif-based methods, structure-based methods, genomic context-based 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.).


## Table 1: List of basal and non-basal regions with overlapping functional regions

<table>
<thead>
<tr>
<th>Basal/Non-Basal/Invariant</th>
<th>Basal</th>
<th>Non-basal</th>
<th>Basal</th>
<th>Non-basal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene/protein*</td>
<td>Start</td>
<td>End</td>
<td>Start</td>
<td>End</td>
</tr>
<tr>
<td>NC</td>
<td>1</td>
<td>600</td>
<td>1</td>
<td>625</td>
</tr>
<tr>
<td>E1A:12/13S</td>
<td>564</td>
<td>1113</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E1B:19K</td>
<td>1572</td>
<td>2120</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E1B:55K</td>
<td>1877</td>
<td>3364</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IX:IX</td>
<td>3449</td>
<td>3856</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IVa2:IVa2</td>
<td>5233</td>
<td>3900</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E2B:DNA pol</td>
<td>8278</td>
<td>5003</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E2B:pTP</td>
<td>10218</td>
<td>8323</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1:52/55K</td>
<td>10632</td>
<td>11750</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1:IIIa</td>
<td>11773</td>
<td>13470</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L2:Penent base</td>
<td>13524</td>
<td>15086</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L2:VII</td>
<td>15690</td>
<td>15680</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L2:V</td>
<td>16726</td>
<td>16950</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L3:VI</td>
<td>17006</td>
<td>17707</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L3:Hexon</td>
<td>17776</td>
<td>20604</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Annotation in HAdV-8 (AB448767)

<table>
<thead>
<tr>
<th>Functional Category</th>
<th>Characterized roles</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>NC</td>
<td>Starting point for replication</td>
<td>(47, 50)</td>
</tr>
<tr>
<td>E1A:12/13S</td>
<td>Promote p53-dependent apoptosis. Promotion of viral transcriptional activation</td>
<td>(53)</td>
</tr>
<tr>
<td>E1B:19K</td>
<td>Host modulation</td>
<td>(39)</td>
</tr>
<tr>
<td>E1B:55K</td>
<td>Host modulation</td>
<td>(39)</td>
</tr>
<tr>
<td>IX:IX</td>
<td>Structural</td>
<td>(2, 3, 33)</td>
</tr>
<tr>
<td>IVa2:IVa2</td>
<td>Core protein/Genome Packaging</td>
<td>(3)</td>
</tr>
<tr>
<td>E2B:DNA pol</td>
<td>Replication of the viral DNA</td>
<td>(47, 50)</td>
</tr>
<tr>
<td>E2B:pTP</td>
<td>Enables protein-priming start of the replication</td>
<td>(44)</td>
</tr>
<tr>
<td>L1:52/55K</td>
<td>Genome Packaging</td>
<td>(3)</td>
</tr>
<tr>
<td>L1:IIIa</td>
<td>Structural/Genome Packaging</td>
<td>(2, 3, 33)</td>
</tr>
<tr>
<td>L2:Penent base</td>
<td>Major capsid protein</td>
<td>(3)</td>
</tr>
<tr>
<td>L2:VII</td>
<td>Core protein/Maturation</td>
<td>(3)</td>
</tr>
<tr>
<td>L2:V</td>
<td>Core protein/Maturation</td>
<td>(3)</td>
</tr>
<tr>
<td>L3:VI</td>
<td>Structural/Maturation</td>
<td>(2, 3, 33)</td>
</tr>
<tr>
<td>L3:Hexon</td>
<td>Major capsid protein</td>
<td>(2, 3, 33)</td>
</tr>
</tbody>
</table>

### Reference

(47, 50, 53, 39, 39, 2, 3, 33, 3, 3)
| 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 | 24019 | Genome Packaging | Viral genome packaging. | (11, 12) |
| L4:32K | 24732 | 25115 | Genome Packaging | Viral genome packaging by interacting with IVa2. | (3) |
| L4:VIII | 25444 | 26127 | Structural/Maturation | Minor structural protein. | (2, 3, 33) |
| E3:100K | 22772 | 24919 | Genome Packaging | Viral genome packaging by interacting with IVa2. | (3) |
| E3:22K | 24732 | 25115 | Genome Packaging | 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:CR1β | 27452 | 28666 | Host modulation | Not fully characterized. | |
| E3:CR1γ | 28693 | 29472 | Host modulation | Not fully characterized. | |
| E3:14.7K | 30139 | 30531 | Host modulation | Inactivate p53 and p53-dependent apoptosis. Viral late gene expression. | (38, 41) |
| E3:12.5K | 26128 | 26448 | Host modulation | Major structural protein, binding to cellular receptor, tissue affinity. | (33) |
| L5:Fiber | 30785 | 31873 | Major capsid protein | Major structural protein, binding to cellular receptor, tissue affinity. | (33) |
| E3:orf4 | 33319 | 32957 | Host modulation | Lysis of infected cell. | (41, 52) |
| E3:orf3 | 33675 | 33322 | Host modulation | Promotion of viral gene expression and replication. | (40) |
| E3:orf2 | 34064 | 33672 | Host modulation | Not fully characterized. | |
| E3: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
evolution using Mesquite version 2.75 under the following conditions: the tree topology = the
genome tree; the number of characters = 33,645; the ratio of invariant sites = 0.65; the alpha
parameter of the gamma distribution of rate variance = 0.477; nucleotide frequencies of A, T,
G, and C = 0.22, 0.21, 0.29, and 0.28, respectively; transition/transversion ratio = 1.57. The
abscissa and ordinate (the x and y axes) of this matrix represent the physical positions in the
simulated MGA, and each point (x, y) of the matrix shows the Mantel’s correlation
coefficient between windows x and y. The correlation coefficient ranges from near zero
(yellow) to near one (red). Independent and diagonal windows are colored black and gray,
respectively.

**FIG 4 Histogram of significant correlation ratios.** The abscissa is the ratio of the number of
windows that show a significant correlation coefficient to a specific window against the total
number of windows (=167). The left and right ordinates are for the absolute frequencies (bars)
and relative cumulative frequencies (lines), respectively. The gray bars and line are for the
simulated data. The absolute frequencies in basal and non-basal regions of the real data are
shown in black and mesh bars, respectively, together with cumulative frequencies in the black
line.

**FIG 5 Finer-scale coevolution analysis.** The results of the finer scale analyses are depicted
for two highlighted regions, (A) fiber and (B) E3 region. The abscissae show the position in
the HAdV-8 genome (AB448767). The left ordinates represent the correlation coefficients
between each 100 bp-window and the entire DNA polymerase-coding region. Significantly
correlated windows ($p<0.0025$) are shown with disks, equivalent to a basal region, and the
others are presented with open circles. Protein-coding regions are shown with arrows below
each plot. The predicted extracellular (O), transmembrane (M) and cytoplasmic (I) regions are shown in the coding regions of CR1α, gp19K, CR1β, and CR1γ.