**Doctoral Dissertation**

# A study of the genetic divergence of *Human mastadenovirus D* via comparative genomic analyses

# （比較ゲノム解析によるヒトアデノウイルス D 種の遺伝的多様性に関する研究）

**Graduate School of Information Science and Technology**
**Hokkaido University**

**GONZALEZ, Gabriel**

**May, 2015**

# Abstract

*Human mastadenovirus D* (HAdV-D) is an exceptionally type-rich human adenovirus species and causative agent of different diseases and neonatal fatalities, as well as an opportunistic pathogen in immune-compromised patients. A research group, including the author of this thesis, revealed that intertypic homologous recombination events between distant types have extensively diversified HAdV-D. This finding raised important questions, such as: (1) what mechanism has allowed frequent homologous recombination events between diverged types despite the fact that the homologous recombination rate is generally low between diverged sequences?; and (2) what has made it possible to produce functional recombinant genomes through homologous recombination events between distant genomes, even though it has been demonstrated that replacements of component proteins of a multiprotein system, such as the packaging system comprising several adenoviral proteins, with homologues from different types lead to malfunction? In order to address these questions concerning the mechanisms and processes that have generated functional recombinant HAdV-D forms via homologous recombination events, the author analyzed the evolutionary patterns of all available genome sequences of different HAdV-D types via three steps: i) identifying recombination events that have previously occurred in those genomes; ii) statistically analyzing the distribution of the recombination boundaries; and iii) detecting interregional coevolution conserved among different types. These analyses revealed that the recombination boundaries are concentrated in specific regions (referred to as hotspots), and these hotspots are located at genomic segments that are conserved over different genomes (referred to as universally conserved segments or UCSs), implying that these UCSs have participated in recombination initiation followed by the exchange of adjacent, highly diverged sections. In addition, the author developed a novel statistical method and showed the modularity of recombination, i.e., recombination events have exchanged specific blocks of the genome, thus allowing for the avoidance of the generation of nonfunctional chimeric proteins and chimeric combinations of physically interacting proteins. This evolutionary modularity was confirmed by the author's finding that the gene regions of physically interacting proteins have coevolved in different HAdV-D genomes in common, thereby suggesting strong purifying selection. It is therefore concluded that HAdV-D has diverged through frequent homologous recombination events initiated at UCSs and strong purifying selection against deleterious chimeric forms. The interregional coevolution analysis method that the author developed for this study provides a new means for predicting protein-protein and protein-DNA physical and functional interactions in other organisms.

# Acknowledgements

The author of this thesis would like to express the deepest gratitude to Professor Hidemi Watanabe, who has the attitude and the substance of a genius: he frequently and patiently exhorted his students to go deeper in the results and understand the reasons rather than just finding the results on their research. Without his guidance and persistent help this dissertation and related work would not have been possible.

Similarly, the same degree of gratitude is extended to Professor Kanako O. Koyanagi, whose invaluable time and advice always provided a constant source of insight and different perspectives to improve the work of her students. Without her involvement, not a single of the developed projects would have reached the exhibited clarity and maturity.

In addition, it is important to acknowledge Doctor Aoki Koki, who provided a bridge to adenovirus research and related topics by supplying the necessary knowledge from the clinical and epidemiological point of view regarding adenovirus and the different types of infections. His help was indispensable to connect all presented evolutionary analyses with the medical field.

In similar way, the author of this study wants to acknowledge the contributions and advices from Doctor Hisatoshi Kaneko, Doctor Nobuyoshi Kitaishi, Doctor Shigeaki Ohno and Doctor Tsuguto Fujimoto.

Additionally, a special acknowledgement and gratitude to the Japanese government for financing my studies and stay in Japan via the MEXT scholarship.

Last but not least, the results of the efforts reflected on this study are dedicated to my family and beloved ones, including her who provided me with inspiration to continue in many occasions. Particularly, this work is dedicated to my mother, brothers and sisters, who have never stopped providing support and encouragement.

# Publication Note

Chapter 3 is based on results presented at the 第 60 回日本ウイルス学会学術集会 (大阪, 日本 ) (2012) and 第 61 回日本ウイルス学会学術集会 (神戸 , 日本 ) (2013)

Chapter 4 is based on the paper (Gonzalez et al. 2014) published in *Gene*. It was also presented at the *11th International Adenovirus Meeting 2014*.

Chapter 5 is based on the paper (Gonzalez et al. 2015) published in *Journal of Virology*.

# Contents

# i. Abbreviations and Acronyms

| | |
|---|---|
| **Bp** | Base pair |
| **DBP** | DNA binding protein (section 1.3.2) |
| **DNA-pol** | DNA polymerase (section 1.3.2) |
| ***dN*** | Non-synonymous substitution rate (section 2.3) |
| ***dS*** | Synonymous substitution rate (section 2.3) |
| **E** | Early transcription region (section 1.3.3) |
| **EKC** | Epidemic keratoconjunctivitis (Chapter 1) |
| **GC%** | DNA ratio of guanine and cytosine (Chapter 1) |
| **GT** | Genotype (section 1.4) |
| **HAdV** | Human adenoviruses (section 1.4) |
| **HAdV-D** | Human mastadenovirus D (section 1.4) |
| **HLA** | Human leukocyte antigens (section 4.4.3) |
| **I** | Intermediate transcription region (section 1.3.3) |
| **INSDC** | International Nucleotide Sequence Database Collaboration (section 4.2.1) |
| **ITR** | Inverted terminal protein (section 1.3.2) |
| **Kbp** | Kilobase pairs |
| **L** | Late transcription region (section 1.3.3) |
| **MAANV** | Matrix of amino acids as numeric values (section 3.2.2) |
| **MGA** | Multiple genome alignment (section 4.2.1) |
| **MHC** | Major histocompability complex (section 4.4.3) |
| **MSVM** | Multiclass support vector machine (section 2.1) |
| **MWW** | Mann-Whitney-Wilcoxon rank test (section 3.2.5) |
| **NC** | Non-coding region (Table 3) |
| **NK** | Natural killer cell (section 1.3.3) |
| **Nt** | Nucleotide |
| **ORF** | Open reading frame (Chapter 3) |
| **OVO** | One-versus-one (section 2.1.2) |
| **PCR** | Polymerase chain reaction (Chapter 3) |
| **pI** | Isoelectric point (Chapter 3) |
| **pTP** | Pre-Terminal protein (section 1.3.2) |
| **RDP** | Recombination detection program (section 2.2) |
| **RFE** | Recursive feature elimination (section 3.2.3) |
| **RRE** | Reliable recombination event (section 4.2.2) |
| **SAdV** | Simian mastadenoviruses (Fig. 3.4) |
| **SVM** | Support vector machine (section 2.1) |
| **TN93** | Evolutionary model Tamura-Nei 1993 (section 1.4) |
| **TP** | Terminal protein (section 1.3.1) |
| **UCS** | Universally conserved segment (section 4.3.4) |
| **UTR** | Untranslated region (section 2.4) |
| **VT** | Vector of types (section 3.2.2) |

# ii. List of figures



# iii. List of Tables

# Chapter 1: Introduction

The technological progress in genetic material sequencing techniques combined with computational analyses for molecular evolution has revolutionized the understanding of viruses. Disciplines related to viral evolution, such as virology and epidemiology, have been enriched with deeper understanding of the interactions between viruses and hosts by demonstrating the mutual role played by both in modeling the evolution of each other. Furthermore, viruses impose selective pressures over infected host populations as host populations reciprocally impose immunological barriers against viruses. Such barriers represent selective pressures over viruses and result into viral divergence in order to successfully continue infecting and replicating in the host populations (Paterson et al. 2010).

The divergence of different kinds of viruses is due to a variety of evolutionary tendencies created by the distinct conditions and immune responses in the host populations. Such tendencies are identified as evolutionary mechanisms generating faster adaptation capabilities by increasing diversity such as small viral genomes with high mutational rates, e.g. single stranded RNA viruses (Duffy, Shackelton, Holmes 2008); while other viruses with longer genomes code proteins for modulating the immune response of the host and avoiding cellular apoptosis until the replication and assembly of the virions are complete, e.g. DNA viruses (O'Brien 1998). The characterization of the variety of viral adaptation mechanisms, which is evidenced in their evolution to continue infecting the respective host populations, provides understanding about the respective host. For instance, some types of cellular apoptosis have been characterized through the characterization of viral infectious cycles affecting cells (Chinnadurai 1998).

Several viruses use the human species as a host and represent cause of harms to the human health by causing diseases, which vary from minor discomforts to fatalities and in the duration of the symptoms from short outbreaks to latent infections with periodical outbursts. One family of such viruses is composed of human adenoviruses (HAdV), which represent a constant and worldwide cause of infections in a range of severity from subclinical cases in immune-competent patients to even lethal cases in immuno-compromised patients such as patients under radiotherapy, AIDS, malnutrition and newborns (Echavarria 2008). Besides,

HAdV are the main cause of epidemic infections regularly affecting populations, e.g. epidemic keratoconjunctivitis (EKC) in Japan (Ariga et al. 2005) and acute respiratory infection in military recruitments in the United States of America (Jacobs et al. 2004).

The advances in methods for sequencing genetic materials have allowed better analyses of isolated viral samples and have led to reports of increasing diversity in HAdV, particularly in *Human mastadenovirus D* (HAdV-D); in detail, it has been suggested that the recombination events between different lineages created new recombinant types (Aoki et al. 2008; Walsh et al. 2009; Kaneko et al. 2011a; Kaneko et al. 2011b; Robinson et al. 2011a; Singh et al. 2012; Zhou et al. 2012; Greber et al. 2013; Fujimoto et al. 2014). Moreover, such reports have brought controversy to the scientific community as the candidates to new types have been defined by computational-defined taxa (Seto et al. 2011), while traditional methods based on serological tests failed to provide an accurate distinction of these strains from previously characterized types. This controversy is the result of research suggesting these strains were simple variations of existing types (de Jong et al. 2008), while others considered these results as evidence of possible misclassifications by the methods based on serologic neutralization (Singh et al. 2012).

On the other hand, although recombination has been recognized as a characteristic on the adenoviral evolution implied as a possible adaptation mechanism to human immunological response (Lukashev et al. 2008), few reports have offered a proper explanation of how recombination events exchanged the sections that determine the divergence between distinguishable adenoviral lineages referred as types. Then, the present study aimed to answer the question as how these recombination events are related with the divergence in *Human mastadenovirus D*.

The report by Robinson et al. (2013), attempted to find a common pattern in frequently recombined sections, which are referred as recombination hotspots of the adenoviral genome, and pointed to the location of these hotspots correlated with a marked change in the nucleotide composition of the putative recombination breakpoints from high content of guanine and cytosine (GC%) to comparatively lower GC% and further suggesting this change could be the facilitator of the genomic exchanges (Robinson et al. 2013). Nevertheless, such a hypothesis failed to explain how these nucleotide patterns emerged in the adenoviral genome and evolved to eventually taking the function of facilitators for the recombination events in the particular location of these putative hotspots rather than in others genomic regions.

The evolution of the divergence of HAdV types could be the result of lateral transfer of genetic material that enabled combinations of characteristics and allowed a relatively faster divergence than the one such viruses could account by simple accumulation of point mutations;

thus representing a presumable advantage for adapting to different tissues and host populations. However, testing such a hypothesis first requires providing unambiguous criteria for type distinction. Then, comparisons of distinguishable adenoviral types could provide further explanation of the mechanisms enabling the recombination events between distant lineages while avoiding the disruption of functional protein pathways contained on the densely coded adenoviral genome.

To provide a proper thesis explaining the observations about the adenoviral type divergence in HAdV-D, this work presents innovative analyses based on the comparison of adenoviral complete genome sequences in the species looking for: (1) short informative genomic regions for type distinction (Chapter 3), (2) biases in the distribution of recombination events and (3) mechanisms enabling such biases (Chapter 4), (4) intergenomic nucleotide conservation along the genome (Chapter 4), and (5) evolutionary correlation between the different adenoviral genomic sections (Chapter 5).

## 1.1 Contributions of this thesis

The results discussed in this thesis provide insight into the differences between adenoviral types and the emergence of divergence in adenoviral species. They also offer an explanation to many of the main characteristics of the adenoviral genome structure and evolutionary trends in HAdV-D. Moreover, the author of this thesis presents evolutionary evidence supporting the existence of recombination boundary hotspots localized in the regions that determine the differences between lineages. These hotspots were enabled by the conservation of particular genomic sections through virtually all taxa levels. Additionally, they created a modular structure of exchange allowing particular modules to be exchanged with effects only over sections related with virus-host interactions, while sections coding adenoviral functions related with the viral DNA replication, packaging and structural integrity of the virions remained under a tight correlated evolution. Interestingly, the exchanged modules also showed evidence of an evolutionary relationship between them, which could be explained as a result of functional evolutionary constraints. Besides, the presented analyses can be easily extended to other genomes and used for assessing more characteristics beyond divergence. The following paragraphs show a brief description of the innovative computational analyses implemented in the course of this research and their main findings.

**Finding the most informative short genomic regions by SVM approach**

As introduced above, in the adenoviral research community there is an ongoing debate about the definition of adenoviral type and what kind of criteria should be used to distinguish between types. This thesis approaches the problem by using an algorithm based on support vector machines (SVM), which is a supervised machine learning algorithm, to look for the most informative regions in the adenoviral genome that simultaneously could classify the traditionally defined serotypes and the computationally-defined taxa called genotypes based solely in the coded protein sequences. This approach effectively considered the criteria for type classification proposed so far, and allowed simplifying the required data to a reduced number of short protein regions, enabling quick type identification while pointing to the common ground between the different sides of the discussion about type classification in the adenoviral research community (see section 1.4).

**Detecting recombination boundary hotspots and universally-conserved segments in HAdV-D**

HAdV-D has the highest number of types among HAdV species (43 out of 68). The reports of many of these types being the result of recombination events prompted the analysis of the distribution of recombination boundaries and led to the uncovering of recombination boundary hotspots. Further analyses demonstrated that exchanged regions corresponded to genomic sections diverged between different lineages. Nevertheless, the findings were counterintuitive as the homologous recombination rate in most genomic sequences is demonstrated to be inversely reduced by the divergence between sequences; therefore, to search for genomic characteristics that enabled the observed exchanges despite the divergence between sequences, the exposed research assessed the conservation along the genome with an innovative analysis based on synonymous-substitution rate ($d_S$) comparison. Consequently, a set of significantly conserved segments through all lineages denominated as universally-conserved segments (UCSs) was uncovered. The presence of the UCS near the recombination boundary hotspots provided regions where homologous recombination could initiate despite the high divergence on the rest of the exchanged region.

**Detecting evolutionary correlated modules**

Recombination events enabled by the recombination boundary hotspots succeeded despite the putatively deleterious effects due to disruption of functional pathways coded in the adenoviral genome. To explain these observations, this study analyzed the correlated evolution

along the different regions of the genome using an approach that assessed the correlation between their evolutionary distance matrices. Thus uncovered a set of windows regarded as the genomic basal sections of the HAdV-D. This set of windows contains the proteins in charge of viral DNA replication, protein maturation, assembly and packaging of virions, as also proteins of the virion capsid, which needs to be precisely assembled. Nevertheless, other regions were independent from these basal sections and evidenced a modularity that allowed them to be exchanged with putatively reduced deleterious effects on the adenoviral genomes. Such exchangeable regions showed correlation among them, and despite their variety of functional and structural roles, they had in common the characteristic of direct interaction with the host cell for evasion or modulation of the host defenses.

## 1.2 Organization of this thesis

To facilitate the interpretation of the results presented in this document, the rest of this introduction presents the biological background of adenoviruses with particular focus on HAdV-D. The following parts of the document are organized as follows. First, a brief description of other related works is presented in Chapter 2, which corresponds to the scientific framework used as the base to implement the analyses used along this thesis. Next, Chapter 3 presents an analysis of divergence between different lineages in HAdV that allows providing an unbiased distinction of adenoviral types, which is still a topic of discussion in the scientific community studying adenoviruses, as explained below in section 1.4. Then, Chapter 4 describes the recombination patterns in HAdV-D. Furthermore, the correlated evolution between genomic sections is contrasted with the recombination modules to evaluate the effects of the exchanged sections in the context of inter-genomic relationships in Chapter 5. Finally, Chapter 6 contrasts the implications of these results and presents an assessment of how well the thesis explains the reported observations and the future work to be developed.

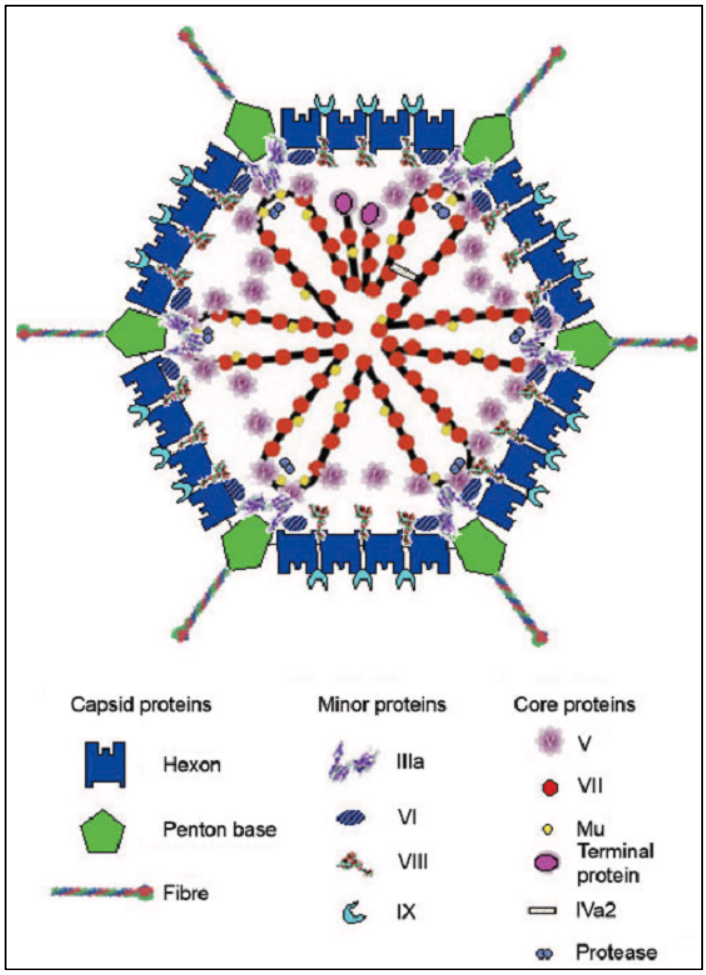


**Fig. 1.1 A schematic representation of adenoviral proteins forming a virion.** The figure shows the structure based on cryo-electron microscopy and crystallography. The location of the proteins inside the virion (minor proteins and distribution of DNA) remains as speculation. The capsid and minor proteins are well-defined and the diagram is not to scale. This figure was taken with permission from (Russell 2009). Copyright © Society of General Microbiology, *Journal of General Virology*, Vol. 90, no. 1, pp.1-20, 2009, 10.1099/vir.0.003087-0

## 1.3 Biological generalities about adenoviruses

The members of *Adenoviridae* family infecting humans have double stranded linear DNA genomes, which range between 30-37kbp, and are considered as medium size among DNA viruses (Davison, Benko, Harrach 2003; Harrach et al. 2011). The overall GC% content in the adenoviral DNA varies between 40.8% and 63.8% (Harrach et al. 2011). The paragraphs below introduce the main characteristics about different aspects of adenoviral structure, DNA replication and viral cycle.

### 1.3.1 Structure of the adenoviral virion

An adenoviral virion is composed of a linear DNA chain encapsidated in a non-enveloped icosahedral capsid with 70-90 nm in diameter. The capsid is composed by three kinds of major proteins: 240 non-vertex trimer proteins (hexons) and 12 vertex pentamer proteins (penton base) with an extension protruding from the virion surface composed of another trimer protein (fiber) (Fig. 1.1) (Russell 2009; Harrach et al. 2011).

Additionally, these major proteins are bound together by a set of minor proteins that work as the cement connecting and fixing them together. The structural minor proteins are: IIIa, VI, VIII and IX (Liu et al. 2010). The capsid structure is tightly packed as a shell protecting the contents of the virion composed of the adenoviral DNA, the protease protein and five other kinds of proteins fixing the viral core to capsid, namely, these proteins are: V, VII, X, IVa2 and terminal protein (TP) (Fig. 1.1) (Russell 2009).

The immune response of the host to the adenoviral infections involves developing antibodies mainly targeting the external structure of the virions, namely, the three major proteins of the capsid: hexon, penton base and fiber. Moreover, the antibodies target the exposed regions of these proteins, particularly the epitopes located in towers of the loops of hexon protein (Toogood, Crompton, Hay 1992), the RGD loop of penton base protein (Hong et al. 2003) and the knob of the fiber (Mei, Wadell 1996).

### 1.3.2 Adenoviral DNA replication

The adenoviral replication takes place in the host cell nucleus (Davison, Benko, Harrach 2003) and depending on the adenoviral species it requires recruitment of different host proteins to perform the process. In general, a protein-primed DNA replication strategy is used by all members of the *Adenoviridae* family (Davison, Benko, Harrach 2003). This strategy uses the viral coded DNA polymerase (DNA-pol), DNA binding protein (DBP) and the precursor terminal protein (pTP). Moreover, the DBP and DNA-pol bind in a complex that together with the genome-bound TP in the 5'-end of the inverted terminal region (ITR) of the linear DNA and other host proteins establish the pre-initiation complex for the replication (de Jong, Meijer, van der Vliet 2003; De Jong, Van Der Vliet, Brenkman 2003). Then, the pTP-DNA-pol complex jumps back from the binding position in the DNA strand, the pTP dissociates and the DNA-pol continues with the elongation of the complete strand mediated by the interaction with the DNA and the DBP (Parker et al. 1998; De Jong, Van Der Vliet, Brenkman 2003).

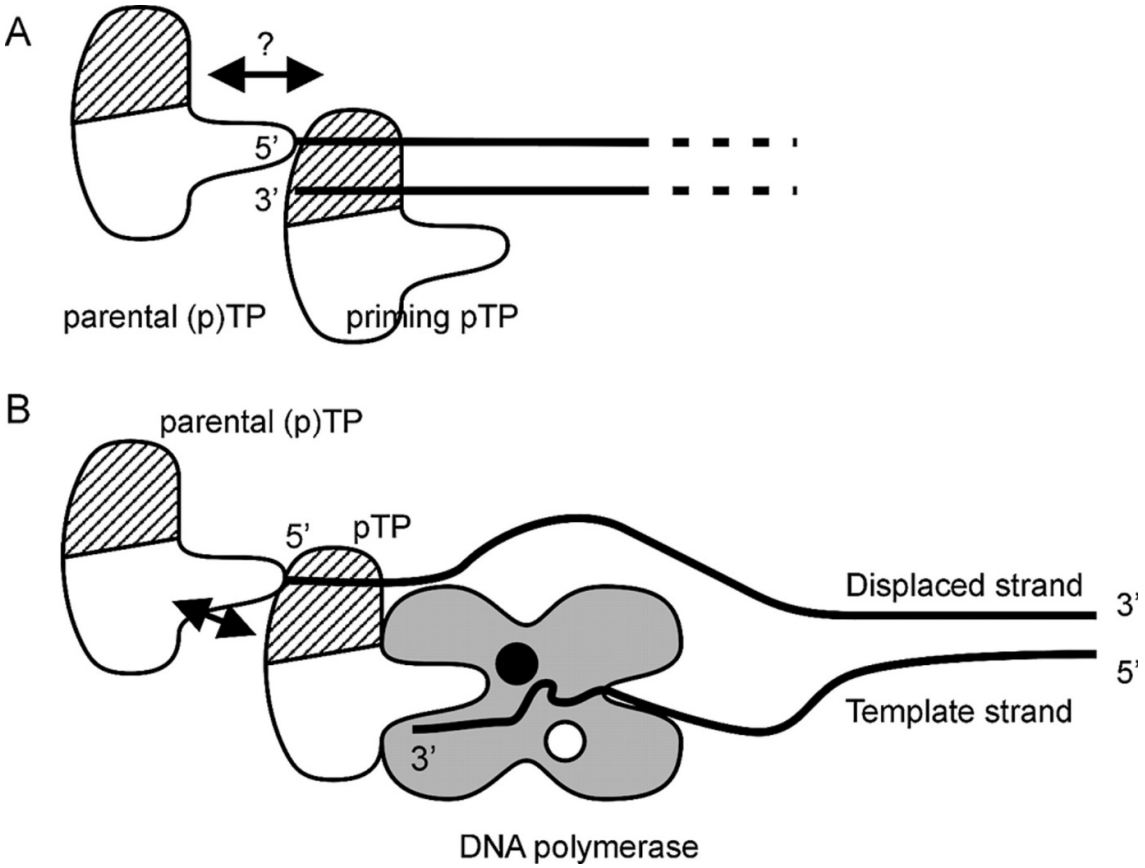


**Fig. 1.2 A schematic representation of adenoviral pre-initiation replication complex. (A)** A parental (p)TP molecule stays covalently bound to the 5' end of the adenoviral genome (lines) after having primed the preceding replication round. The observed cooperative binding to short DNA duplexes could mimic the interaction of an incoming priming pTP with the parental (p)TP (arrow). The ds/ssDNA binding region of pTP is indicated by hatching. **(B)** The ssDNA binding affinity of pTP could contribute to pTP/pol binding stability on an unwound origin structure, while the pTP region providing the priming activity should be close to the catalytic center (black circle) in the adenoviral DNA-pol. Potential stabilizing contacts can be made with the parental (p)TP (arrow) and two ssDNA contact points spaced by the dimensions of pTP: the displaced strand (bound in the DNA binding region) and the three 3' terminal nucleotides of the template strand (3')**.** 

In addition to the replication properties of the DNA-pol, this protein also presents a switch into exo-nuclease enzymatic activity to remove mismatched nucleotides without dissociating from the DNA strand that is being replicated allowing the DNA-pol to perform a proofread replication (King et al. 1997). This characteristic explains lower mutation rate in HAdV than in other shorter types of viruses, e.g. single stranded RNA viruses (Duffy, Shackelton, Holmes 2008).

### 1.3.3 Infectious cycle

The infectious cycle starts with the fiber attaching to a receptor in a susceptible cell that varies according to the type of the adenovirus (Zhang, Bergelson 2005). After attachment of

the fiber, interactions between motifs located in the penton base and cellular integrins activate the endocytosis of the virion and posterior delivery of the uncoated virion core into the host cell nucleus where is subject to mRNA transcription, virus DNA replication and assembly into new virions bound to infect other cells and repeat the cycle (Harrach et al. 2011).

The mRNA transcription of the products coded in the adenoviral genome follows an order transcribing different regions according to the stage of the infection. Moreover, many of the protein products are the result of alternative splicing of the mRNA (Sharp 1994). The transcription regions of the genome are named according to the transcription order as early regions (*E*), intermediate (*I*) and late (*L*), and refer to the early, intermediate and late stages of the infection, respectively. The early group codes proteins in charge of modulating the transcription processes in the host cell (*E1* and *E4*), assembling the viral DNA replication machinery described above (*E2*) and modulating and neutralizing host defense mechanisms (*E3*) such as inhibition of T-cell recognition and NK cell (Natural Killer) evasion (McSharry et al. 2008; Harrach et al. 2011). On the other hand, the intermediate (*IX* and *IVa2*) and late genes (*L1-L5*) involve the proteins related to the assembly and maturation of the virion (Davison, Benko, Harrach 2003).

## 1.4 Human adenovirus taxonomy

Because adenoviral species are usually confined to one or a few closely related host species, the host origin is one of the various criteria used to determine the species taxon of adenoviruses (Davison, Benko, Harrach 2003). The adenoviruses isolated from human tissues are named human adenovirus (HAdV) and are classified under the genus *Mastadenovirus*. The genus is further divided into species, which are named from the host and supplemented with letters of the alphabet. Particularly, until 2015, HAdV have been classified into seven separated species named with letters from *A* to *G*.

Each species consists of specific serotypes (no. 1-51, abbreviated as HAdV-1 to 51 in this document) and genotypes (no. 52-, or HAdV-GT52-) (Fig. 1.3). Serotypes have been defined serologically, i.e. specific immune sera inhibit or neutralize the reproduction of the virus (Takeuchi et al. 1999; Echavarria 2008), while genotypes have been proposed as computationally-defined taxonomic classes (Seto et al. 2011) but yet to be reviewed by the adenovirus research community to reach a consensus on the definition of genotype. For simplicity, these different types are collectively called types in this document. The number of

constitutive HAdV types varies widely from species to species: four, twelve, five, forty-three, one, two and one in *Human mastadenovirus A* to *G* (HAdV-A to G), respectively (Echavarria 2008; Matsushima et al. 2011; Walsh et al. 2011; Dehghan et al. 2012; Matsushima et al. 2013) (refer to Fig. 1.3 and Table 1 at the end of the chapter). Also, there are adenoviral samples isolated from primates that, because of multiple biological and genomic characteristics, have been classified under HAdV-B, -C, -E and -G (Harrach et al. 2011).

Certain types are predominantly associated with specific pathologies in specific tissues and their morbidity varies from type to type. However, it has been suggested that some adenoviral types could develop more than one class of pathology in different human populations according to the geographical location and the age of the patients or particular health conditions (Echavarria 2008). The divergence in HAdV has been attributed as one of the characteristics determining the variety of infections affecting tissues such as: respiratory, ocular surface, genitourinary, gastrointestinal, neurologic and cardiac (Robinson et al. 2011a). Additionally, studies reporting HAdV related to other health disorders such as obesity (Arnold et al. 2010) and opportunistic infections in patients receiving transplanted organs (Echavarria 2008).

The *Human mastadenovirus D* (HAdV-D) contains the most diverged number of types with 43 out of the reported 68 HAdV types (Fig. 1.3). Furthermore, HAdV-D is the only adenoviral species targeting exclusively human hosts contrasting with some types in HAdV-B, -C, -E and -G that have been isolated from other primates (Harrach et al. 2011). Other differences with the rest of species are the particularly high genome GC% (> 57%); eight predicted products in the *E3*; the clustering of its members in a single group, which by evolutionary distance under Tamura-Nei model (TN93) has an average distance between type members and to types of other species as 0.06 and as > 0.2, respectively; and relatively higher number of reported naturally occurring recombinant types.

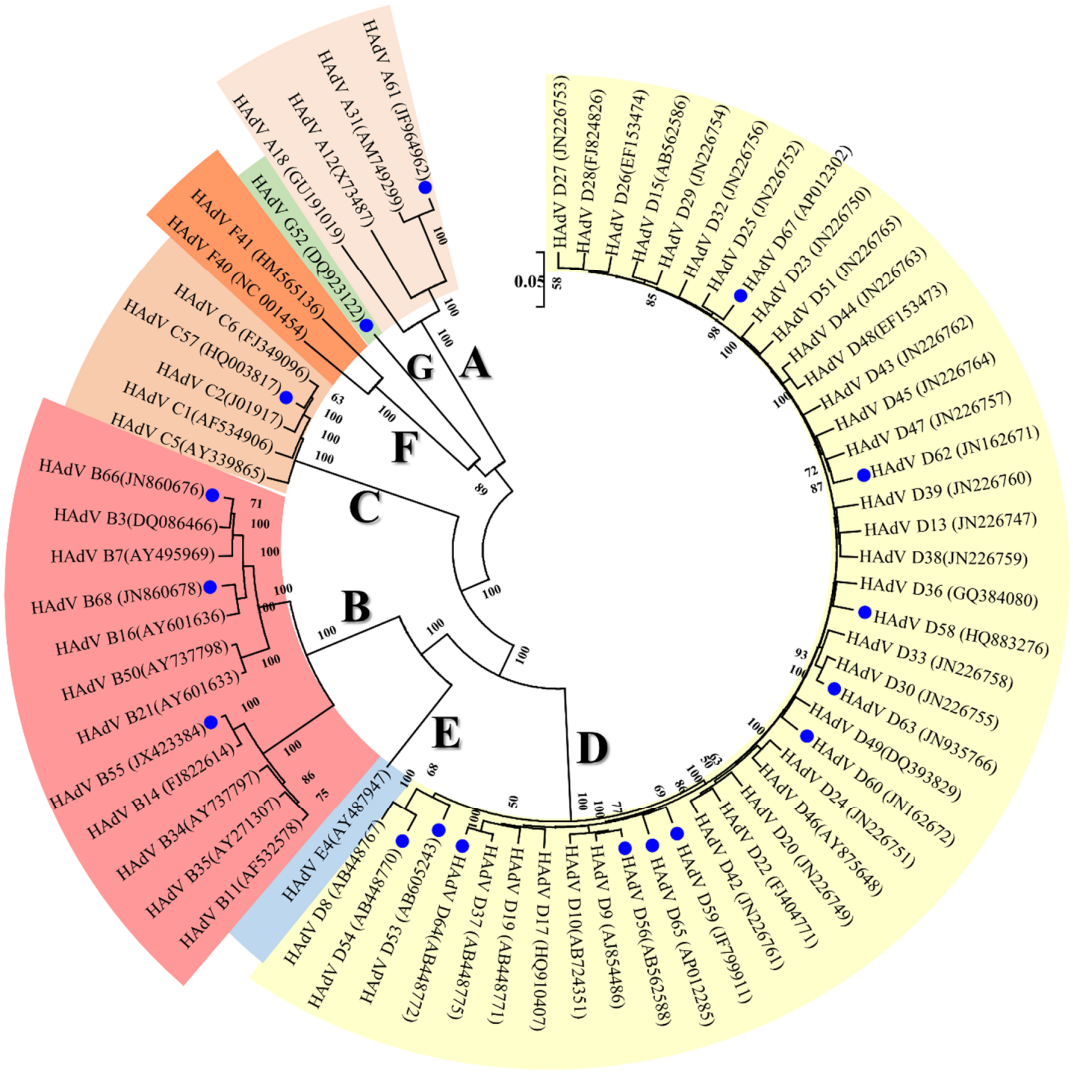


**Fig. 1.3 The phylogenetic tree of whole genome sequences for all accepted HAdV types.** The Neighbor-Joining tree was built with the whole genome sequences and removing gapped columns. The ungapped sites (27,515) were compared under Tamura-Nei evolutionary model. The percentage of bootstrap repetitions that supported the branching is shown next to the branches (<50 are omitted). Also, blue discs mark the genotypes. The human mastadenovirus species are shown at the root of the respective cluster.

**Table 1. Adenoviral sequences used along chapters in this thesis.**

| No. | Type[a] | Species | Accession No. |
|---|---|---|---|
| 1[b] | 12 | A | X73487 |
| 2[b] | 18 | A | GU191019 |
| 3[b] | 31 | A | AM749299 |
| 4 | 31 | A | KF268119 |
| 5[b] | GT61 | A | JF964962 |
| 6[b] | 3 | B | DQ086466 |
| 7 | 3 | B | AY599834 |
| 8 | 3 | B | AY599836 |
| 9 | 3 | B | DQ099432 |
| 10 | 3 | B | DQ105654 |
| 11 | 3 | B | JX423380 |
| 12 | 3 | B | JX423381 |
| 13 | 3 | B | JX423382 |
| 14 | 3 | B | KF268120 |
| 15 | 3 | B | KF268123 |
| 16 | 3 | B | KF268128 |
| 17 | 3 | B | KF268131 |
| 18 | 3 | B | KF268133 |
| 19 | 3 | B | KF268202 |
| 20[b] | 7 | B | KC857700 |
| 21 | 7 | B | AY495969 |
| 22 | 7 | B | AY594255 |
| 23 | 7 | B | AY594256 |
| 24 | 7 | B | AY601634 |
| 25 | 7 | B | GQ478341 |
| 26 | 7 | B | HQ659699 |
| 27 | 7 | B | JF800905 |
| 28 | 7 | B | JN860677 |
| 29 | 7 | B | JX423383 |
| 30 | 7 | B | JX423386 |
| 31 | 7 | B | JX423387 |
| 32 | 7 | B | JX423388 |
| 33 | 7 | B | JX625134 |
| 34 | 7 | B | KC440171 |
| 35 | 7 | B | KF268117 |
| 36 | 7 | B | KF268125 |
| 37 | 7 | B | KF268134 |
| 38 | 7 | B | KF268135 |
| 39 | 7 | B | KF268314 |
| 40 | 7 | B | KF268316 |
| 41[b] | 11 | B | AF532578 |
| 42 | 11 | B | AY163756 |
| 43 | 11 | B | AY598970 |
| 44 | 11 | B | FJ597732 |
| 45 | 11 | B | KF268121 |
| 46[b] | 14 | B | AY803294 |
| 47 | 14 | B | FJ822614 |
| 48 | 14 | B | JN032132 |
| 49 | 14 | B | JQ824845 |
| 50 | 14 | B | JX892927 |
| 51[b] | 16 | B | AY601636 |
| 52 | 16 | B | JN860680 |
| 53[b] | 21 | B | KJ364592 |
| 54 | 21 | B | AY601633 |
| 55 | 21 | B | KF528688 |
| 56 | 21 | B | KF577593 |
| 57 | 21 | B | KF577595 |
| 58 | 21 | B | KF577597 |
| 59 | 21 | B | KF577598 |
| 60 | 21 | B | KF802425 |

[a] GT stands for Genotype and S stands for Simian types

[b] Used for training in Chapter 3

[c] Used for recombination analysis in Chapter 4

[d] Used for co-evolutionary detection in Chapter 5

[e] Used for rooting HAdV-D in Chapter 4

**Table 1 (continued). Adenoviral sequences used along chapters in this thesis.**

| No. | Type[a] | Species | Accession No. |
|---|---|---|---|
| 61 | 21 | B | KF802426 |
| 62 | 21 | B | KF938575 |
| 63 | 21 | B | KJ364573 |
| 64 | 21 | B | KJ364574 |
| 65 | 21 | B | KJ364575 |
| 66 | 21 | B | KJ364576 |
| 67 | 21 | B | KJ364577 |
| 68 | 21 | B | KJ364578 |
| 69 | 21 | B | KJ364579 |
| 70 | 21 | B | KJ364580 |
| 71 | 21 | B | KJ364581 |
| 72 | 21 | B | KJ364582 |
| 73 | 21 | B | KJ364583 |
| 74 | 21 | B | KJ364584 |
| 75 | 21 | B | KJ364585 |
| 76 | 21 | B | KJ364586 |
| 77 | 21 | B | KJ364587 |
| 78 | 21 | B | KJ364588 |
| 79 | 21 | B | KJ364589 |
| 80 | 21 | B | KJ364590 |
| 81 | 21 | B | KJ364591 |
| 82[b] | 34 | B | AY737797 |
| 83 | 34 | B | KF268196 |
| 84[b] | 35 | B | KF268124 |
| 85 | 35 | B | AY128640 |
| 86 | 35 | B | AY271307 |
| 87[b] | 50 | B | AY737798 |
| 88 | 11+34 | B | KF906413 |
| 89 | 3+11 | B | EF564600 |
| 90 | 3+11 | B | EF564601 |
| 91 | 3+7 | B | JN860679 |
| 92[b] | GT55 | B | KF908851 |
| 93 | GT55 | B | FJ643676 |
| 94 | GT55 | B | JX123027 |
| 95 | GT55 | B | JX123028 |
| 96 | GT55 | B | JX123029 |
| 97 | GT55 | B | JX423384 |
| 98 | GT55 | B | JX423385 |
| 99 | GT55 | B | JX491639 |
| 100 | GT55 | B | KC857701 |
| 101 | GT55 | B | KJ883520 |
| 102 | GT55 | B | KJ883521 |
| 103 | GT55 | B | KJ883522 |
| 104b | GT66 | B | KF268126 |
| 105 | GT66 | B | JN860676 |
| 106[b] | GT68 | B | JN860678 |
| 107 | P16H3F16 | B | KF268315 |
| 108 | P35H34F7 | B | KF268328 |
| 109 | P3H3F7 | B | KF268311 |
| 110 | P3H7F3 | B | KF268132 |
| 111 | P7H3F3 | B | KF268210 |
| 112 | P7H3F3 | B | KF268212 |
| 113 | S27 | B | FJ025909 |
| 114 | S27 | B | FJ025928 |
| 115 | S28 | B | FJ025914 |
| 116 | S28 | B | FJ025915 |
| 117 | S29 | B | FJ025916 |
| 118 | S32 | B | FJ025911 |
| 119 | S33 | B | FJ025908 |
| 120 | S35 | B | FJ025910 |

[a] GT stands for Genotype and S stands for Simian types
[b] Used for training in Chapter 3
[c] Used for recombination analysis in Chapter 4
[d] Used for co-evolutionary detection in Chapter 5
[e] Used for rooting HAdV-D in Chapter 4

**Table 1 (continued). Adenoviral sequences used along chapters in this thesis.**

| No. | Type[a] | Species | Accession No. |
|---|---|---|---|
| 121 | S35 | B | FJ025912 |
| 122 | S41 | B | FJ025913 |
| 123 | S41 | B | FJ025927 |
| 124 | S46 | B | FJ025930 |
| 125 | S47 | B | FJ025929 |
| 126 | X | B | KF633445 |
| 127b | 1 | C | JX173086 |
| 128 | 1 | C | AF534906 |
| 129 | 1 | C | JX173078 |
| 130 | 1 | C | JX173080 |
| 131 | 1 | C | JX173082 |
| 132 | 1 | C | JX173083 |
| 133 | 1 | C | JX173085 |
| 134[b] | 2 | C | KF268310 |
| 135 | 2 | C | J01917 |
| 136 | 2 | C | JX173077 |
| 137 | 2 | C | JX173079 |
| 138 | 2 | C | JX173081 |
| 139 | 2 | C | JX173084 |
| 140 | 2 | C | KF268130 |
| 141[b] | 5 | C | AY601635 |
| 142 | 5 | C | AY339865 |
| 143 | 5 | C | KF268127 |
| 144 | 5 | C | KF268199 |
| 145 | 5 | C | KF429754 |
| 146 | 5 | C | M73260 |
| 147[b] | 6 | C | KF951595 |
| 148 | 6 | C | FJ349096 |
| 149 | 6 | C | HQ413315 |
| 150 | 6 | C | JX423389 |
| 151 | 6 | C | KF268129 |
| 152[b] | GT57 | C | HQ003817 |
| 153 | S31 | C | FJ025904 |
| 154 | S31 | C | FJ025906 |
| 155 | S34 | C | FJ025905 |
| 156 | S40 | C | FJ025907 |
| 157 | S40 | C | FJ025926 |
| 158 | S42 | C | FJ025902 |
| 159 | S42 | C | FJ025903 |
| 160 | S43 | C | FJ025900 |
| 161 | S44 | C | FJ025899 |
| 162 | S45 | C | FJ025901 |
| 163[b] | 8 | D | KF429751 |
| 164[c, d] | 8 | D | AB448767 |
| 165 | 8 | D | AB448768 |
| 166 | 8 | D | AB448769 |
| 167 | 8 | D | AB746853 |
| 168 | 8 | D | KF268118 |
| 169 | 8 | D | KF268198 |
| 170 | 8 | D | KF268205 |
| 171 | 8 | D | KF268321 |
| 172[b, c, d] | 9 | D | AJ854486 |
| 173b | 10 | D | AB695621 |
| 174 | 10 | D | JN226746 |
| 175[b, c, d] | 13 | D | JN226747 |
| 176[b] | 15 | D | KF268204 |
| 177[c, d] | 15 | D | AB562586 |
| 178[b] | 17 | D | KF268330 |
| 179 | 17 | D | AF108105 |
| 180[c, d] | 17 | D | HQ910407 |

[a] GT stands for Genotype and S stands for Simian types
[b] Used for training in Chapter 3
[c] Used for recombination analysis in Chapter 4
[d] Used for co-evolutionary detection in Chapter 5
[e] Used for rooting HAdV-D in Chapter 4

**Table 1 (continued). Adenoviral sequences used along chapters in this thesis.**

| No. | Type[a] | Species | Accession No. |
|---|---|---|---|
| 181[b] | 19 | D | JQ326209 |
| 182[c, d] | 19 | D | AB448771 |
| 183[b, c, d] | 20 | D | JN226749 |
| 184[b, c, d] | 22 | D | FJ619037 |
| 185 | 22 | D | FJ404771 |
| 186[b] | 23 | D | KF279629 |
| 187[c, d] | 23 | D | JN226750 |
| 188[b, c, d] | 24 | D | JN226751 |
| 189[b, c, d] | 25 | D | JN226752 |
| 190[b, c, d] | 26 | D | EF153474 |
| 191[b, c, d] | 27 | D | JN226753 |
| 192[b, c, d] | 28 | D | FJ824826 |
| 193[b] | 29 | D | JN226754 |
| 194[c, d] | 29 | D | AB562587 |
| 195[b, c, d] | 30 | D | JN226755 |
| 196[b, c, d] | 32 | D | JN226756 |
| 197[b, c, d] | 33 | D | JN226758 |
| 198[b, c, d] | 36 | D | GQ384080 |
| 199[b] | 37 | D | KF268203 |
| 200 | 37 | D | AB448775 |
| 201[c, d] | 37 | D | AB448776 |
| 202 | 37 | D | AB448777 |
| 203 | 37 | D | AB448778 |
| 204 | 37 | D | DQ900900 |
| 205 | 37 | D | KF268122 |
| 206[b, c, d] | 38 | D | JN226759 |
| 207[b, c, d] | 39 | D | JN226760 |
| 208[b] | 42 | D | KJ626292 |
| 209[c, d] | 42 | D | JN226761 |
| 210 | 42 | D | KJ626291 |
| 211[b] | 43 | D | KC529648 |
| 212[c, d] | 43 | D | JN226762 |
| 213[b, c, d] | 44 | D | JN226763 |
| 214[b, c, d] | 45 | D | JN226764 |
| 215[b, c, d] | 46 | D | AY875648 |
| 216[b, c, d] | 47 | D | JN226757 |
| 217[b, c, d] | 48 | D | EF153473 |
| 218[b, c, d] | 49 | D | DQ393829 |
| 219[b, c, d] | 51 | D | JN226765 |
| 220[c, d] | 22/37 | D | AB605240 |
| 221 | 22/37 | D | AB605241 |
| 222[c, d] | 22/37/8 | D | AB605242 |
| 223 | 65/48/NEW | D | AB765926 |
| 224[b] | GT53 | D | KF268197 |
| 225[c, d] | GT53 | D | AB605243 |
| 226 | GT53 | D | AB605244 |
| 227 | GT53 | D | AB605245 |
| 228 | GT53 | D | FJ169625 |
| 229[b] | GT54 | D | AB333801 |
| 230[b] | GT56 | D | KF268333 |
| 231[d] | GT56 | D | AB562588 |
| 232[c] | GT56 | D | HM770721 |
| 233 | GT56 | D | KF268209 |
| 234 | GT56 | D | KF268329 |
| 235[b] | GT58 | D | KF268319 |
| 236[c, d] | GT58 | D | HQ883276 |
| 237[b, c, d] | GT59 | D | JF799911 |
| 238[b, c, d] | GT60 | D | HQ007053 |
| 239 | GT60 | D | JN162672 |
| 240[b, c, d] | GT62 | D | JN162671 |

[a] GT stands for Genotype and S stands for Simian types
[b] Used for training in Chapter 3
[c] Used for recombination analysis in Chapter 4
[d] Used for co-evolutionary detection in Chapter 5
[e] Used for rooting HAdV-D in Chapter 4

**Table 1 (continued). Adenoviral sequences used along chapters in this thesis.**

| No. | Type[a] | Species | Accession No. |
|---|---|---|---|
| 241[b, c, d] | GT63 | D | JN935766 |
| 242[b] | GT64 | D | KF268213 |
| 243 | GT64 | D | AB448772 |
| 244 | GT64 | D | AB448773 |
| 245 | GT64 | D | AB448774 |
| 246[c, d] | GT64 | D | EF121005 |
| 247 | GT64 | D | JQ326207 |
| 248 | GT64 | D | JQ326208 |
| 249[b, c, d] | GT65 | D | AP012285 |
| 250[b, c, d] | GT67 | D | AP012302 |
| 251[b] | GT69 | D | JN226748 |
| 252 | P22H38F17 | D | KF268312 |
| 253 | P23H32F62 | D | KF268327 |
| 254 | P26H56F56 | D | KF268313 |
| 255 | P28H37F38 | D | KF268334 |
| 256 | P33H56F56 | D | KF268201 |
| 257 | P37H37F17 | D | KF268208 |
| 258 | P38H32F27 | D | KF268325 |
| 259 | P38H63F44 | D | KF268335 |
| 260 | P49H46F65 | D | KF268332 |
| 261 | P62H33F17 | D | KF268322 |
| 262 | P67H28F60 | D | KF268320 |
| 263 | P67H37F45 | D | KF268324 |
| 264 | P67H9F15 | D | KF268206 |
| 265 | P9H20F67 | D | KF268207 |
| 266 | P9H46F39 | D | KF268211 |
| 267[b] | 4 | E | AY599837 |
| 268 | 4 | E | AY487947 |
| 269 | 4 | E | AY594253 |
| 270 | 4 | E | AY594254 |
| 271 | 4 | E | AY599835 |
| 272 | 4 | E | EF371058 |
| 273 | 4 | E | KF006344 |
| 274 | S22 | E | AY530876 |
| 275 | S23 | E | AY530877 |
| 276 | S24 | E | AY530878 |
| 277 | S25 | E | AF394196 |
| 278 | S25 | E | FJ025918 |
| 279 | S26 | E | FJ025923 |
| 280 | S30 | E | FJ025920 |
| 281 | S36 | E | FJ025917 |
| 282 | S37 | E | FJ025919 |
| 283 | S37 | E | FJ025921 |
| 284 | S38 | E | FJ025922 |
| 285 | S39 | E | FJ025924 |
| 286 | Y25 | E | JN254802 |
| 287[b] | 40 | F | L19443 |
| 288[b] | 41 | F | KF303071 |
| 289 | 41 | F | AB728839 |
| 290 | 41 | F | DQ315364 |
| 291 | 41 | F | HM565136 |
| 292 | 41 | F | KF303069 |
| 293 | 41 | F | KF303070 |
| 294[b] | GT52 | G | DQ923122 |

[a] GT stands for Genotype and S stands for Simian types
[b] Used for training in Chapter 3
[c] Used for recombination analysis in Chapter 4
[d] Used for co-evolutionary detection in Chapter 5
[e] Used for rooting HAdV-D in Chapter 4

# Chapter 2: Framework of related works

## 2.1 Multiclass support vector machines (MSVM)

Multiclass support vector machines (MSVM) are supervised learning classification or supervised discrimination algorithms, which correspond to the process of categorizing samples based on past observations. The mathematical formulation for such a model is a function $y(w; x)$ that maps a single observation $x$ of $N$ features to a discrete class $c \in \{1, \ldots, C\}$, where $C$ is the number of different classes, by using the $N$ parameters of the model in the vector $w$, which is defined as the weight vector. Then, given $K$ observations along with their respective classes, these $K$ samples represent the past observations that are used for finding the parameter values in $w$ that fit the model for predicting the class of future observations (Bishop 2007; Psorakis, Damoulas, Girolami 2010). For better understanding of how $y$ accomplishes the multiclass classification, this section introduces the generalities for the simplest case of binary classification and how such approach is used to perform multiclass classification and recursive feature selection.

### 2.1.1 Binary support vector machines

Before deeper details in MSVM, this subsection presents how support vector machines (SVM) provide two-class classification, i.e. C=2. The two-class classification problem using linear models is represented in the basic form for linear classification as the function:

$$y(x) = w^T x + b \quad \text{(Eq. 1)}$$

Where $T$ means transposition, $x$ denotes a fixed feature-space of $N$ dimensions and $b$ stands for a bias parameter (Bishop 2007). The input is composed of $K$ observations with $K$ target values $t_1, \ldots, t_K$ where $t_i = -1$ or $t_i = 1$. Then, the calculation of the proper parameters for $y(x)$ is done by a dual approach, i.e. the data is considered as points in a multidimensional space and the problem is transformed into an algebraic problem.

To approach the problem of binary classification with SVM, the $K$ observations are used to find the parameters $w$ and $b$ to minimize the generalization error via maximizing the decision margin, which is defined as the shortest distance between the decision boundary and any of the samples used to train the model. The samples closer to the margin are named support

vectors and are the only observations conserved from the input for classifying future observations. The margin represents a hyper-plane crossing the *N*-dimensional space and dividing the *K* points into the two classes. The distance between the margin and the observation *x* is *y(x)* and the sign of such provides the classification. Then, the perpendicular distance *y* of a point *x* to the hyper-plane defined by *y(x)* = 0, *y* with the form of Eq. 1, is given by $|y(x)|/||w||$, with $|y(x)|$ and $||w||$ respectively representing the absolute value of *y(x)* and the norm of the vector *w*. Furthermore, the objective is only finding solutions to *w* and *b* for which all data points are correctly classified, i.e. $t_i y(x_i) > 0$ for all *i*. Then, the distance of a point $x_i$ to the decision surface can be written as:

$$\frac{t_i y(x_i)}{\|w\|} = \frac{t_i(w^T x_i + b)}{\|w\|} \qquad \text{Eq. 2}$$

Because the maximum margin is given by the distance to the closest point $x_i$ from the samples, the maximum solution to *w* and *b* is found by solving:

$$\arg max_{w,b} \left\{ \frac{1}{\|w\|} min_i [t_i(w^T x_i + b)] \right\} \qquad \text{F. 3}$$

However, the direct solution of this optimization problem is very complex. Thence, the problem is converted into an equivalent problem by first scaling *w* and *b*; then, the distance estimated by Eq. 2 remains the same but it is possible to set the equation for the closest point to the decision surface *k*:

$$t_k(w^T x_k + b) = 1 \qquad \text{Eq. 4}$$

And for the rest of *K-1* observations, the constraints are set in the form:

$$t_i(w^T x_i + b) \geq 1, \;\; i = 1, \dots, K \qquad \text{F. 5}$$

In order to provide some tolerance to misclassified data among the points used to train the data, the inequalities of F. 5 are modified to include slack variables $\xi_i \geq 0$, defined as $\xi_i = 0$ for those data points correctly classified and $\xi_i = |t_i - y(x_i)|$ for other points. This is known as soft margins, because they provide some tolerance to misclassified data. Then, F.5 is rewritten as:

$$t_i(w^T x_i + b) \geq 1 - \xi_i, \;\; i = 1, \dots, K \qquad \text{F. 6}$$

Now subject to the constraints of the form F. 6, the problem is converted into maximizing $||w||^{-1}$, which is equivalent to minimizing $||w||^2$ and so the optimization problem is re-written as:

$$\arg min_{w,b} \; C \sum_{i=1}^{K} \xi_i + \frac{1}{2}\|w\|^2 \qquad \text{F. 7}$$

Where, $C$ represents a trade-off parameter between slack variable penalty and the margin. To solve this problem, Lagrange multipliers $a_i \geq 0$ and $\mu_i \geq 0$ are added such as each $a_i$ and $\mu_i$ corresponds to one of the constraints in F. 6 and giving the Lagrangian function:

$$L(w, b, a) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{K}\xi_i - \sum_{i=1}^{K} a_i\{t_i(w^T x_i + b) - 1 + \xi_i\} - \sum_{i=1}^{K}\mu_i\xi_i \quad \text{F. 8}$$

with $a = (a_1, \ldots, a_K)^T$ and adding a minus sign in front of the Lagrange multiplier term to represent the minimizing goal for $w$ and $b$, and maximizing with respect to $a$. Then, setting the derivatives of $L(w, b, a)$ with respect to $w$, $b$ and $\xi_i$ equal to zero, the following equations are obtained:

$$w = \sum_{i=1}^{K} a_i t_i x_i \quad \text{F. 9}$$

$$0 = \sum_{i=1}^{K} a_i t_i \quad \text{F. 10}$$

$$a_i = C - \mu_i \quad \text{F. 11}$$

Then, F. 8 is transformed to the dual representation of the maximum margin problem by removing $w$ and $b$ using F. 9, F. 10 and F. 11 to substitute the values. Thence, the dual representation to be maximized is:

$$\tilde{L}(a) = \sum_{i=1}^{K} a_i - \frac{1}{2}\sum_{j=1}^{K}\sum_{i=1}^{K} a_i a_j t_i t_j x_i \cdot x_j \quad \text{F. 12}$$

Notice that $x_i \cdot x_j$ is a vector product that could be changed by a kernel function that could convert the vectorial space and extend the application of SVM to nonlinear spaces for classification (Bishop 2007). The maximization is subject to the constraints:

$$C \geq a_i \geq 0, i = 1, \ldots, K \quad \text{F. 13}$$

$$\sum_{i=1}^{K} a_i t_i = 0 \quad \text{F. 14}$$

This is a convex quadratic optimization problem, after solving it with quadratic programming, the optimum $a$ vector is an array where only the $a_i > 0$ represent the $x_i$ vectors that are part of the solution. Such vectors are the support vectors ($S$) and from the $K$ observations originally used to train the model, these are the ones used by the model to determine the margin and classify future observations. The vectors in $S$ are used to obtain $w$ by substituting $a$ in F. 9 and $b$ is obtained by averaging over all vectors in S:

$$b = \frac{1}{K_S}\sum_{s\in S}(t_s - \sum_{m\in S} a_m t_m x_m \cdot x_s) \quad \text{F. 15}$$

Then, the binary classification of future observation $x$ is performed by the sign of the function:

$$y(x) = (\sum_{s\in S} a_s t_s x_s \cdot x) + b \quad \text{F. 16}$$

### 2.1.2 Multiclass support vector machine

Different strategies for extending the SVM to classify samples into more than 2 classes, i.e. $C > 2$, have been proposed. For this document, the adopted approach to multiclass support vector machines (MSVM) is the one-versus-one (OVO) approach that construct $C(C-1)/2$ different binary SVM classifiers on all possible pairs of classes, and then it classifies the test points according to which class has the highest number of 'votes' (Bishop 2007; Zhou, Tuck 2007), where each 'vote' is a classification as that class among all the classifiers.

### 2.1.3 Multiple class support vector machine – Recursive feature elimination

The $w$ in Eq. 1 can be used to decide the relevance of each feature of the $N$ features in each $x$ observation (Guyon et al. 2002); the reason is the larger $|w_j|$ is, the more important role $j^{th}$ feature plays in the decision function Eq. 1, i.e. the feature has more relevance classifying the observation $x$. Then, by considering the squares of each element in F. 8 and making positive all entries, i.e. $w^2 \geq 0$, it is possible to rank all features according to relevance from the most to the less; the algorithm for recursive feature elimination (RFE) proceeds removing the less relevant feature and repeating the SVM optimization process over the $N$-1 remaining features. The RFE algorithm is used to find a subset of relevant features over some threshold of error according to a function that measures the accuracy of the classifications or it could be executed until all features are ranked by the order of elimination from the training data (Guyon et al. 2002).

To extend the RFE algorithm to multiclass models, the $w^2$ vectors from the $C(C\text{-}1)/2$ binary classifiers in a given MSVM-OVO model are added and then the features are ranked according to the total of the addition. Thus, the resulting total $w$ vector is summarizing the role of each feature along all the binary classifiers. The algorithm is depicted as follows (Zhou, Tuck 2007):

```
Algorithm: MSVM-RFE
Input: n classified observations <<x_i, c_i>>
       m number of features to be selected
Output: the m indexes of selected most relevant features
Initialize F to all indexes in the full feature set; //F is the set of selected features
p = number of indexes of features in F; //starts with the N
while p > m do
       Train a multiclass SVM with features in F; //weight vectors w_r = [w_r1, …, w_rp]^T
       Calculate the ranking criteria for F: criteria_i = Σ_{r=1}^{C(C-1)/2} w_ri^2;
       Order criteria from smallest to largest;
       Remove the index of the feature with smallest c_i from F;
       p = number of indexes of features remaining in F;
end
return F
```

## 2.2 Algorithms for detection of recombination events between sequences

On a recombination event, a copy of one sequence is inserted into a second sequence of genetic material resulting in a new sequence, which contains pieces of both original recombinant parents. Moreover, the homologous recombination is a type of genetic recombination where the exchange represents a copy of the minor parental donor into the homologous section of the major parental molecule. The exchange is enabled by the similarity between the involved sequences.

To identify the results of such recombination events in the evolutionary history of any genetic material, there are several algorithms that compare a group of sequences and assign probabilities to possible recombinant events between them. These algorithms also provide the most likely recombination boundaries of the event, i.e. the points in the major parental sequence where the copy of the minor parent started and ended. The following subsections provide a description of the recombination detection algorithms used in Chapter 4 and is available as part of the suite of algorithms in the recombination detection program RDP version 4.22β (Heath et al. 2006).

**RDP method**

The RDP method (Martin, Rybicki 2000) screens for recombination events analyzing every possible unique set of three sequences from a multiple sequence alignment using the following three-step procedure:

1. In each unique set of three sequences sampled from an alignment, which is hereafter called triplet, all phylogenetically uninformative sites are discarded. Uninformative sites are defined as follows: given an unweighted pair group method with arithmetic mean (UPGMA) dendrogram built from the alignment, in any particular triplet with sequences A, B and C, assume A and B are more closely related to one another than to C; then, the uninformative sites are those that are identical in all three sequences, different in all three sequences, or are not present in any member of a group of reference sequences.

2. A sliding window is moved one nucleotide at a time along the informative sites. The average identity percentage is calculated at each position for each of the three possible pairs of sequences: A-B, A-C and B-C. Potential recombinant regions are identified as regions where the identity percentage of A-B is lower than that of A-C or B-C.
3. In a potential recombinant region the probability that a particular group of nucleotide identities occurred by chance is estimated using a binomial distribution. A p-value is calculated from this probability by multiplying it by the number of unique windows examined. A Bonferroni corrected p-value, which is corrected p-value accounting for multiple comparisons, is calculated from this p-value by multiplying it by the total number of triplets examined within the alignment.

**GENECONV**

The GENECONV algorithm (Padidam, Sawyer, Fauquet 1999) searches for regions within a sequence alignment in which pairs of sequences are similar enough to suspect that one of them may be the result of a recombination event. The basic set of steps followed by this algorithm is as follows:

1. Monomorphic sites, i.e. sites with identical value through all sequences in the alignment, are excluded as a control for constant or highly selected sites. The result is an alignment of polymorphic sites.
2. For every possible pair of sequences in the alignment, regions are found that are either, (a) identical and unusually long for that pair of sequences, or (b) with an unusually high similarity degree. The similarity is scored based on a scheme where (a) matches (or concordant sites) count as +1 and (b) there is a penalty for mismatches (or discordant sites). The mismatch penalty depends on the density of polymorphic sites between the two sequences and on a user-specified mismatch intensity parameter or G-scale.
3. P-values are assigned to high scoring regions, also called fragments with high scoring aligned pairs (HSAPs). The p-values assigned to these regions are derived through (a) slow but accurate process of permutations and/or (b) a BLAST derived Karlin and Altschul (KA) method (Karlin, Altschul 1990). Although approximate, Bonferroni corrected KA p-values are generally far more conservative than p-values from permutations process. The Bonferroni correction simply involves multiplication of pair-wise KA p-values by the number of pair-wise comparisons made during an analysis.

**MaxChi**

This algorithm is named after maximum $\chi^2$ because it uses a $\chi^2$ distribution to determine the positions of breakpoints between analyzed sequences (Smith 1992). Given an alignment the algorithm examines pairs of sequences and searches for recombination breakpoints by looking for significant differences in the proportions of variable and non-variable polymorphic alignment positions in adjacent regions of sequence. The algorithm follows these steps:

1. All monomorphic sites in an alignment are discarded, leaving an alignment of polymorphic sites.
2. For every possible sequence pair in the alignment, a sliding window of set length with a partition at its center is moved along the sequences one nucleotide at a time.
3. At each window position a 2 x 2 $\chi^2$ value is calculated as an expression of the difference in the number of matched/mismatched sites between the pair of sequences on either side of the central partition.
4. In the plot of $\chi^2$ values along the length of the alignment, peaks indicate potential recombination breakpoints.

**Chimaera**

The Chimaera algorithm (Posada, Crandall 2001) is a modification of the MaxChi algorithm with a difference in the way the polymorphic sites are selected. Also, Chimaera can only be used to assess triplets of sequences. Each sequence in a triplet is in turn examined to determine if it could be a recombinant product of the remaining two sequences in the triplet following these steps:

1. The algorithm discards all monomorphic sites and sites at which neither of the two "parental" sequences matches the sequence being tested as "recombinant". The three sequences are compressed into a linear string of 1's and 0's with 1 representing a match of the recombinant with one parent and 0 representing a match with the other.
2. A sliding window of set length with a partition at its center is moved along the string of 1's and 0's one position at a time.
3. At each window position a 2 x 2 $\chi^2$ value is calculated as an expression of the difference in the proportion of 1's and 0's on either side of the central partition. The plot of $\chi^2$ values along the length of the alignment indicates potential recombination breakpoints in the peaks.

**BootScan**

The sliding window algorithm BootScan (Salminen et al. 1995) looks for recombination events in a sequence by comparing the changes in the bootstrap support of the phylogenetic trees and changes in the recombined sequence association with the putative recombinant parents. The version used by RDP program uses an automated search for recombination signals that analyzes the alignment without prior identification of recombinant and non-recombinant sequences. This algorithm follows the next steps:

1. A window of set size is moved along the alignment a specified number of nucleotides at a time.
2. Bootstrap replicates of each window are constructed and pair-wise distances are calculated so that they can be used as either themselves for a pair-wise distance BOOTSCAN or in a tree in either UPGMA or neighbor joining.
3. At each window position the relative grouping (based either on pair-wise distances or tree positions from step 2) of every possible sequence triplet in the alignment is determined over all bootstrap replicates.
4. After completing the last window in the sliding window scan, stored bootstrap data on pair-wise sequence relationships in every possible sequence triplet over all windows is scanned for alterations in relative bootstrap support for sequence pairs. High degrees of bootstrap support alternating between two different sequence pairs are indicative of potential recombination events. The recombination breakpoints are assumed to be near the midpoint between the transitions from high bootstrap values grouping the recombinant sequence with one sequence to high values grouping the recombinant sequence with a third sequence.
5. For the statistical support either a binomial distribution or $\chi^2$ p-values can be calculated for identified regions.

**SiScan**

The sister scanning (SiScan) (Gibbs, Armstrong, Gibbs 2000) examines every possible triplet in an alignment for evidence of recombination events following the next steps:

1. A fourth sequence is either constructed by "horizontal" randomization of one of the sequences in a triplet or drawn from the alignment, e.g. the most diverged sequence in the alignment.

2. A window of set length is moved along the alignment of four sequences a set number of nucleotides at a time. If a randomized sequence is being used, a new randomized sequence is made for every window in a horizontal randomization.
3. Each column of the alignment is sorted into one of fifteen different categories.
4. The nucleotides in each column of the alignment are then randomized in a "vertical" randomization to produce a predetermined number of permuted alignments. The number of columns falling into the fifteen different categories is determined for each of the permuted alignments.
5. At every window position a Z-test is used to determine whether the number of columns in that window corresponding to any of the 15 site categories differed significantly from those determined for the vertically randomized alignments.

**3Seq**

The algorithm 3Seq (Boni, Posada, Feldman 2007) also examines triplets by focusing in the polymorphic sites similarly to Chimaera. The algorithm proceeds as follows:

1. 3Seq discards all monomorphic sites and sites at which neither of the two "parental" sequences matches the sequence analyzed as "recombinant". The three sequences are compressed into a linear string of +1's and -1's, with +1 and -1 representing a match with either one parent or the other, respectively.
2. Starting at each end of the -1 & +1 sequence a running total of the -1's and +1's is recorded at each new position.
3. The maximum difference in the running total across any two sites in the sequence is then noted together with the distance between the sites.
4. Whereas the sites bounding the maximum change in the running total indicate the most probable positions of potential recombination breakpoints, the difference between the running totals recorded at the sites and the number of nucleotides separating them can be used to either calculate a p-value, or read a p-value from a pre-computed p-value table.

## 2.3 Synonymous substitution rate

Those nucleotide mutations in genomic regions that are translated into proteins by the transcription and translation processes, i.e. coding regions, without changes in the underlying

amino acid sequence are named synonymous substitutions because these are silent changes in reference to the coded protein sequence; for instance a mutation in the first position of the codon TTA to CTA has no impact in the translated amino acid sequence because both codons code for leucine. On the other hand, nucleotide mutations that affect the amino acid sequence are named non-synonymous substitutions; for instance a mutation in the second position of the codon CAA to CTA changes the coded amino acid from glutamine to leucine with effects on the coded protein (Nei, Kumar 2000). Since synonymous substitutions are apparently free from natural selection, the rate of synonymous substitutions ($d_S$) is often assumed to be the rate of neutral nucleotide substitution (Miyata, Yasunaga, Nishida 1980). Additionally, the synonymous substitution rate is usually similar for many genes along the genome and is manifold higher than the non-synonymous substitution rate ($d_N$), which varies from gene to gene and it is assumed to be due to purifying selection (Kimura 1984).

Although there are several methods for estimating the $d_S$ and correcting for multiple mutations in the same codon, the following explanation is focused only in the Pamilo-Bianchi-Li method (Li 1993; Pamilo, Bianchi 1993), which calculates the $d_S$ accounting for the different likelihood of transitional and transversional mutations. Due to the purpose of this document only this method is covered; however, for a broader description of other methods refer to Nei, Kumar (2000) and Yang (2006).

To calculate the $d_S$ between two sequences, let each nucleotide site be classified according to the position in the respective codon as non-degenerate, two-fold degenerate and four-fold degenerate according to how many of the three possible mutations are synonymous. Furthermore, the classification of non-degenerate position is for positions in which any of the three mutations result in a different coded amino acid; while in a two-fold degenerate site, only one change is synonymous, and any mutation in a four-fold degenerate site results in the same amino acid. Additionally, the third position of ATT, which codes for isoleucine, is three-fold degenerate but is often grouped as the two-fold degenerate site. Then, let $L_0$, $L_2$ and $L_4$ be the total number of sites in the three degeneracy classes averaged over the two sequences, respectively, and by using K80 evolutionary distance model (Kimura 1980) to estimate the number of transitions and transversions per site within each degeneracy class, represented as $A_i$ and $B_i$, respectively, with $i = 0, 2, 4$. Then, $d_S$ and $d_N$ are estimated as follows:

$$d_S = \frac{L_2 A_2 + L_4 A_4}{L_2 + L_4} + B_4 \qquad 2.3.1$$

$$d_N = A_0 + \frac{L_0 B_0 + L_2 B_2}{L_0 + L_2} \qquad 2.3.2$$

## 2.4 Correlated evolution

The neutral theory of molecular evolution states that selectively neutral molecular changes are fixed by genetic drift (Kimura 1984); therefore, it suggests constant fixation rates of mutations throughout the evolutionary time and homogeneous distribution of substitution rates through generations (Codoner, Fares 2008). However, a nucleotide position in the DNA molecule depends on the other surrounding nucleotides for taking a meaning, e.g. coding an amino acid in coding regions or forming a binding motif in an untranslated region; therefore, determining if a mutation is selectively neutral is in function of its effects in more complex interactions (Robinson et al. 2003). Additionally, other selective forces enter in action when the mutations are due to recombination processes and the chance of disrupting protein interactions is high.

For instance, consider a protein that is part of a functional pathway involving another protein. Then, the amino acid motifs where the first protein binds to the motifs in the second protein share evolutionary constraints. These constraints are reflected as any mutation altering the tertiary shape of the first protein could enhance or eliminate the interfacing characteristics with the second protein (Goh et al. 2000). Another example at nucleotide level happens when a change in a nucleotide motif in the untranslated region (UTR) enhances or diminishes the expression of the protein, e.g. deletions in 5'-UTR sequence affect protein expression in yeast (Dvir et al. 2013). Also, recombination can result in deleterious combinations, for example when adenoviral recombinant types where proteins in charge of the viral DNA packaging are replaced by orthologous genes from different types, the packaging machinery becomes inoperative and the virions are not assembled.

The constraints shared by different molecules are named correlated evolution. There are different kinds of constraints that lead to correlated evolution: phylogenetic inertia, such as the correlation shared between neighboring genes exposed to similar evolutionary forces; structural requirements, such as proteins that bind together to perform a function; functional constraints, such as proteins that evolve similarly due to the similar function; and stochastic constraints, such as those signals of correlation due to the effect of chance but not related to any other of the constraints. Moreover, two entities are considered to coevolve when selective pressures in one specific entity drive the evolution in the other entity and when adaptation happens, it affects both. The entities under co-evolution go from nucleotides, to amino-acids, to proteins, to cells and even organisms (Codoner, Fares 2008).

Several different methods have been devised to assess for correlation between different entities. Furthermore, to identify co-evolution between nucleotide sequences or correlated changes in proteins, different methods rely on multiple diverged sequences to explore how the mutations evidence any consistent parallel evolution through all sequences. One of such approaches is based on distance matrices; moreover, the rationale behind this approach is that the correlation between the phylogenetic patterns of one entity should reflect the other entity throughout multiple sequences, i.e. reflecting the accumulation of correlated mutations in both entities (Waddell, Kishino, Ota 2007). For other methods based on mutual information see (Dunn, Wahl, Gloor 2008) and (Martin et al. 2005b).

The distance matrix-based method takes the two pairwise distance matrices corresponding to two entities. The assessed evolutionary distance varies according to the goal and the characteristics of the entities, and some examples of used distances are: evolutionary nucleotide distance, evolutionary amino acid distance and cophenetic distance based on the phylogenetic tree. Then, both distance matrices are tested for correlation by different mathematical methods, one of which uses the Pearson's correlation test; also, to estimate the statistical support and remove the possibility of stochastic correlation, different methods are available. Among such methods, this document presents the permutation methods based on Mantel correlation test and partial Mantel correlation test (Legendre 2000) despite the fact that the original Mantel test performs statistically poorly because it largely overlooks the phylogenetic correlation between the sequences when performing the permutation test. Therefore, to correct flaws in statistically analysis of Mantel test, it is possible to add a modification by permuting sequences of the distance matrices giving a higher probability to permute between closely related sequences and accounting the underlying phylogenetic relationships (Lapointe, Garland 2001).

# Chapter 3: Support vector machine approach to identify informative genomic regions for cost-efficient type classification of human adenoviruses

Infection surveillance programs monitoring yearly human adenoviral diseases in the general population provide clinical and epidemiological data about thousands of infectious samples in different countries. The proper classification of these samples allows them to be associated with accumulated information about each adenoviral type. One solution for an effective type classification is based on complete genome sequencing of each sample. However, this is a costly and unfeasible task given the thousands of cases throughout the year. In addition, it is known that the sequence variability of adenoviral genomes varies greatly by genomic region. This study aimed to search for specific genomic regions that are short enough for standard PCR amplification and Sanger sequencing but contain sufficient information about type divergence and type specificity for cost-efficient, feasible and accurate monitoring of human adenoviruses. By applying a support vector machine approach and a statistical test to the translated sequences of the ORFs in the genomes of 69 human adenoviral types in the seven human adenoviral species, we identified dense regions of informative sites that are short in length (33-150 nt), and combinations of two or more of these regions contain sufficient information for accurate type classification of 294 different sequences. Therefore, accurate and cost-efficient type identification of large numbers of samples can be achieved by sequencing these short but informative regions in the adenoviral genomes.

## 3.1 Introduction

Human adenoviruses (HAdVs) are a worldwide cause of infectious disease that affect a variety of tissues (e.g., respiratory, ocular, gastrointestinal, and urinary) (Benkő 2008) and are linked to obesity (Arnold et al. 2010). HAdVs belong to the genus *Mastadenovirus* of the family *Adenoviridae* and are classified into seven species (HAdV-A to -G) (Jones et al. 2007). Each species consists of specific serologically defined serotypes (Echavarria 2008) (1–51, abbreviated in this chapter as HAdV followed by the species and the serotype number) and computationally defined taxonomic classes called genotypes (GT) (Seto et al. 2011) (52 and above, or HAdV-G GT52); for clarity, these different types are collectively referred as types in this chapter. The number of constitutive types varies from 1 to more than 44 among species (Echavarria 2008; Matsushima et al. 2011; Walsh et al. 2011; Dehghan et al. 2012; Matsushima et al. 2013). Additionally, some adenoviral strains isolated from other primates have been classified as HAdV-B, -C, -E and -G owing to similar biological and genetic characteristics (Harrach et al. 2011).

Surveillance programs monitoring adenoviral infections in different human populations constantly provide data about infection patterns, emergence, and disappearance of adenoviral agents. Their function is particularly important considering adenoviral infections are estimated to cause 8% of clinically relevant viral diseases worldwide (Adhikary, Hanaoka, Fujimoto 2014). The proper classification of HAdV strains enables association with available clinical data about infections and improves understanding of their epidemiological and evolutionary characteristics. A suggested method for classification is based on whole genome sequencing (Liu et al. 2012). Although such an approach provides proper classification of any strain, it involves sequencing the genomic regions required for classification along with regions highly conserved among strains; therefore, whole genome sequencing is not the most effective or feasible solution for screening and classifying thousands of samples (Adhikary, Hanaoka, Fujimoto 2014). On the other hand, the accuracy of type classification through partial sequencing of adenoviral strains remains to be determined and subsequently standardized, to be consistent throughout different studies and disciplines, particularly for its application in infection surveillance and epidemiological studies.

Although recent reports have suggested the use of genomic regions for type classification based on their high evolutionary divergence between types (Kaneko et al. 2009; Seto et al. 2011; Matsushima et al. 2014), an analytical approach is still required to determine

which adenoviral genomic regions are the most informative for accurate type classification. Therefore, the optimal regions were searched for on the adenoviral genome to determine the type by analyzing the amino acid sequences encoded in 294 HAdV strains from the 51 accepted serotypes, 18 proposed genotypes, 26 intermediate types, and 28 types isolated from other primates. A support vector machine (SVM) approach was used to determine regions that maximize type classification information while minimizing the amount of genomic data required for an unambiguous type classification system by using consistently informative regions throughout the seven HAdV species.

## 3.2. Materials & Methods

### 3.2.1. Adenoviral sequences and amino acid multiple sequence alignment

The amino acid sequences of proteins encoded by HAdV species were extracted from all available complete genome sequences (n = 294) of multiple strains from the International Nucleotide Sequence Database Collaboration (Table 1) and translated into the respective amino acid sequences following the annotations of HAdV-A12 (X73487), HAdV-B11(AF532578), HAdV-C2 (J01917), HAdV-D8 (AB448767), HAdV-E4 (AY458656), HAdV-F41 (HM565136), and HAdV-G GT52 (DQ923122). Extracted amino acid sequences were aligned into multiple sequence alignments for each predicted protein using the Iterative Refinement Method algorithm (FFT-NS-I) of MAFFT (Katoh et al. 2002). Sequences filled with gaps were included in the multiple sequence alignments representing the particular strains from species where those proteins are absent. The absent proteins by species were *E3*:(CR1α, gp19K, CR1β and CR1γ) in HAdV-A; *E3*:CR1γ in HAdV-C and -G; and *E3*:(12.2K, CR1α, and CR1γ) and *E4*:ORF1 in HAdV-F. Although *E3*:gp19K is suggested to be absent in HAdV-F and HAdV-G (Robinson et al. 2011a), BLAST alignment (Altschul et al. 1997) allowed the identification of homologous regions, which are presumably not expressed. Additionally, the *L5*:short-fibers of HAdV-F and -G were ignored for further analyses because they are characteristic of only 3 of the 69 considered types; analogously, *E3*:CR1δ belongs only to HAdV-E. Finally, all protein multiple sequence alignments were concatenated, resulting in a single amino acid multiple sequence alignment with 12,939 sites.

### 3.2.2 Transformation of sequences into numeric vectors

The amino acid multiple sequence alignment and the taxonomical classification of the sequences (types) were respectively transformed into a matrix of amino acids as numeric values (MAANV) and a vector of types (VT) (class labels), such that the *i*-th amino acid sequence in the alignment and its respective type ($1 \leq i \leq 294$) were represented by the *i*-th row of MAANV and the *i*-th entry of VT, respectively, while the *j*-th column of MAANV represented the *j*-th position of the amino acid multiple sequence alignment ($1 \leq j \leq 12{,}939$). The value of each entry in MAANV was the isoelectric point (pI) assigned to the respective amino acid (Liu et al. 2004) or 0 for gaps.

### 3.2.3 Informative sites ranked by multiclass support vector machines

To identify the most relevant amino acid sites for type classification, this study implemented a modified version of the SVM for RFE (Guyon et al. 2002) adapted to allow MSVM-RFE in a one-versus-one approach (Zhou, Tuck 2007) (see section 2.1). The input for MSVM-RFE was a subset of MAANV containing exactly one sequence for each of the 69 HAdV types (Table 1) and the respective subset of entries with the assigned types in VT, and the ranked features corresponded to the columns of MAANV. The implemented modified version of the MSVM-RFE algorithm first removed columns with less than five different values for all considered rows in MAANV because those sites provide little distinguishing information among types and to reduce the execution time due to the MSVM-RFE is an algorithm of time complexity $O(nm^3)$, with n as the number of types (=69) and m as the number of considered sites. Then, the MSVM-RFE ranked the remaining columns.

### 3.2.4 Assessing significance of informative windows

The MSVM-RFE produced the most relevant individually ranked columns (Guyon et al. 2002); however, these did not necessarily represent the most informative regions for adenoviral type classification, i.e. continuous genomic regions with significantly more characteristic sites for distinguishing among types, since the relevance of sites in the region must be collectively assessed. To identify collectively informative regions for type classification, an *N*-site ($N = 51$) sliding window analysis was performed over the MAANV columns with *N*/2-site steps and considered only windows with all sites on the same protein. The collective ranking significance of each window was assessed with a bootstrap test that assigned significance as the proportion of 100,000 randomly assembled sets of *N*-sites from all MAANV that had more highly ranked sites than the tested window.

### 3.2.5 Cross-validation test

A cross-validation test was used to measure and compare the classification hit ratio of different sets of sites, i.e. regions or combination of regions, independently from the effects of data. For each repetition, the test randomly separated the 294 rows of MAANV, representing the sequences of 294 strains (Table 1) into five disjoint subsets. For each subset the type classification of its elements was predicted by an MSVM classification model trained with the combined four remaining subsets. Each prediction was counted as either: (1) correct or (2) incorrect. A prediction for a strain whose type was not present in any other of the four subsets was ignored. The hit ratio of the subset was then calculated as the number of correct predictions divided by the total number of correct and incorrect predictions. The hit ratio for the test was estimated by averaging among the hit ratios of the five subsets. To rule out any bias in the hit ratio measurement produced by random combinations of data splits, 100 repetitions were performed for any tested set of sites. Different sets of sites were compared by the respective samples of 100 hit ratios with the Mann-Whitney-Wilcoxon test (MWW) with matched samples.

## 3.3 Results & Discussion

### 3.3.1 Informative windows spread throughout the genome

The SVM approach used the amino acid sequences encoded in each strain to combine the evolutionary information in each sequence and the physicochemical properties of the expressed proteins, ultimately leading to differences between the types, such as host immune reaction, tissue affinity, and viral-host protein interactions. Although this study represents the results for each amino acid by its respective pI, to reflect unique charge properties, other amino acid properties produced congruent results, as long as the specific property analyzed allowed different amino acids to be compared on a scale (e.g. volume and hydrophobicity; data not shown).

The MSVM-RFE algorithm was trained with sites having at least five different amino acids in the alignment of 69 adenoviral types isolated from humans (Table 1). The presented implementation ranked 1,456 of the 12,939 sites represented in MAANV (Fig. 3.1A), while the remaining sites were ranked as less informative because they exhibited low variability among the seven species (data not shown).

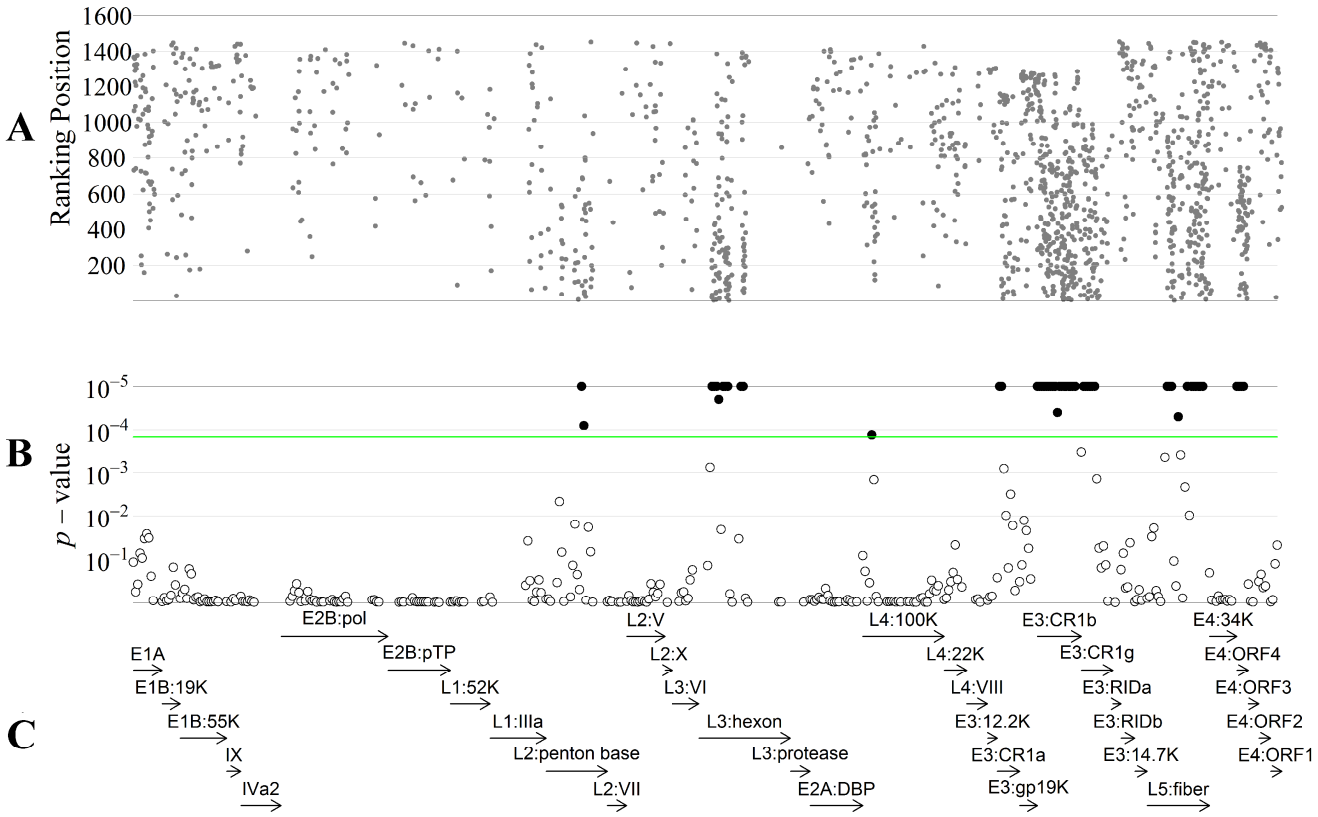


**Fig. 3.1 Informative windows for type classification.** The abscissa represents the amino acid sites of the concatenated proteins reflected as columns in the matrix of amino acid numeric values (MAANV). **A**. The ranking position is shown in the ordinate for sites considered in the MSVM-RFE as the number of sites remaining to be ranked. **B**. The ordinate shows the ratio of 100,000 bootstrapped random samples with more highly ranked positions in the MSVM-RFE than the window. The green line indicates the significance threshold, $p < 1.5 \times 10\text{-}4$ ($p < 0.05$ with Bonferroni correction for 341 windows). **C**. Positions of considered proteins corresponding to MAANV sites.

The MSVM-RFE approach simultaneously explored a multidimensional space and searched for the optimal combination of features, such as amino acid sites, for modeling a classification function that reflects current taxonomical systems based on serotypes and genotypes. Additionally, training the model required as little as a single observation for each type, which was convenient because some types have been isolated and sequenced only once (Table 1).

Top-ranked, individual sites are not necessarily in close proximity; therefore, an *N*-site sliding window analysis ($N = 51$) was implemented to assess the collective ranking significance within a window, allowing to maximize the number of consecutive informative sites in short regions. This analysis could lead to specific targets for identification methods such as PCR sequencing of characteristic genomic segments. For this purpose, the window and step size were set in the range to search for epitope clusters, which could comprise 9–25 amino acids (Koren et al. 2007).

The sliding window analysis revealed significance for 53 of the 341 windows ($p < 1.5 \times 10^{-4}$, bootstrap test with Bonferroni correction; $p < 0.05$) that represented candidates for adenoviral type classifier regions (Fig. 3.1B), hereafter referred to as candidates. Furthermore, 22 of these candidates coincided with the predicted locations of epitope determinants in different species such as the RGD loop in *L2*:penton base (Stewart et al. 1997), the variable loops of *L3*:hexon (Liu et al. 2010), and the *L5*:fiber, which also determines tissue affinity (Chiu et al. 2001).

The remaining 31 candidates were located in regions characterized as highly divergent proteins between types (Benkő 2008). In particular, 26 candidates were located in three *E3* proteins, CR1α, CR1β, and CR1γ, which are variable in function between types (Windheim et al. 2013). The final five candidates were from *E4*:ORF4 and *L4*:100K.

### 3.3.2 Hit ratio comparisons

There was a risk of overfitting the data; therefore, a cross-validation test was implemented to compare and judge the classification hit ratio for each of the candidates independently of the results of the MSVM-RFE. Each candidate is henceforth referred to by its protein of origin followed by the first site of the 51-amino acid window in the protein, e.g., *L3*:hexon-301 refers to the window starting at site 301 of the *L3*:hexon protein and ending at site 351. Additionally, to establish a baseline for the classification hit ratio, the set of sites corresponding to each protein and an additional set including all sites in MAANV was tested, referred as the ALL set, which represented whole genome-based classification. Thus, the

results allowed comparisons of classification hit ratios from the whole genome, single proteins, and candidates uncovered by the MSVM-RFE approach.

The cross-validation test was conducted to measure the hit ratio of the model, independent of the trained data and the effects of different data combinations on the sensitivity of the method to distinguish between closely related types based on different weight distributions among the considered sites. Additionally, although different repetitions for the same set of sites had different random data splits, different sets of sites were tested under the same splits of data; therefore, the hit ratio distribution was comparable among candidates, proteins, the ALL set, and combinations of these sets without any bias due to effects of random data splits.

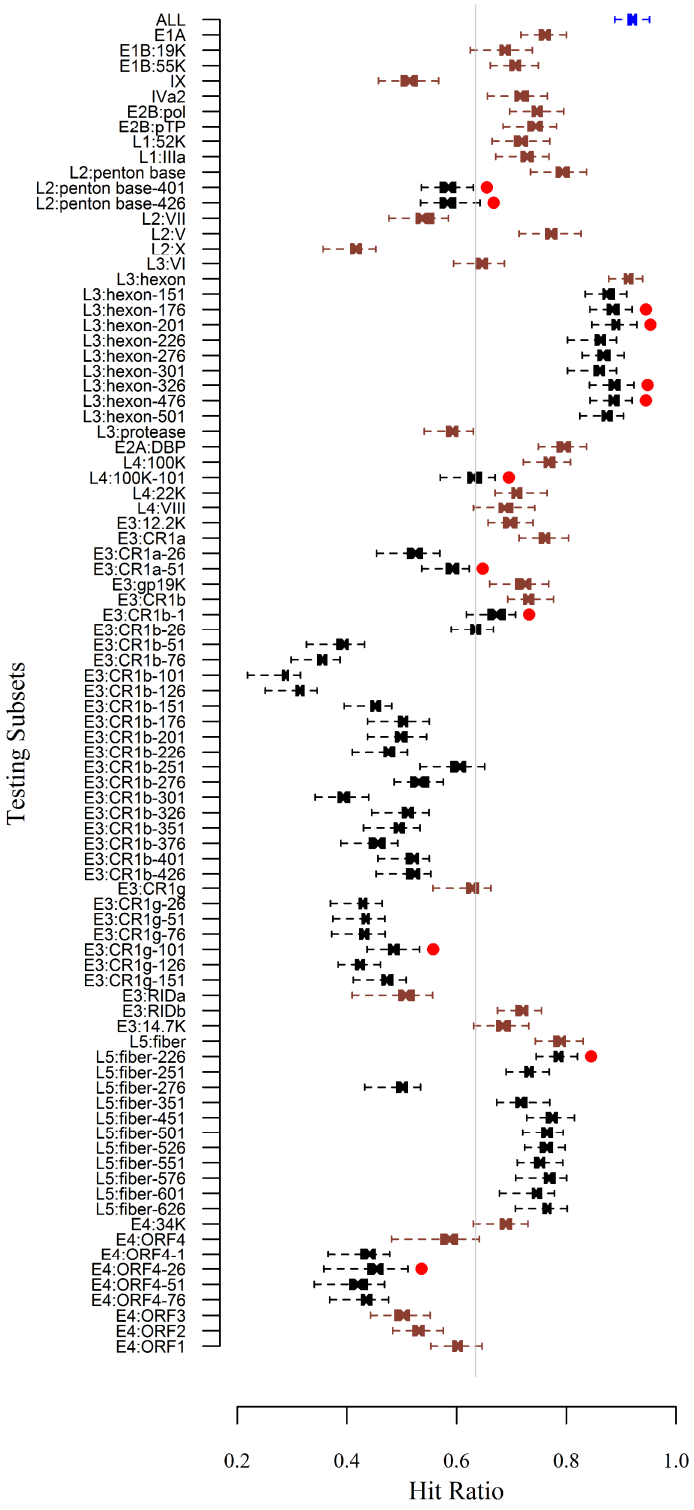


**Fig. 3.2 Boxplot of the hit ratio by sets of sites.** The ordinate represents the sets of sites considered on each tested subset. The abscissa shows the hit ratio. The boxes represent the distribution of the hit ratio through the cross-validation tests. The blue, brown, and black boxes represent the ALL set, protein sets, and candidate sets, respectively. The gray line indicates the average hit ratio of all considered sets of sites. Red dots next to the box plots of candidate sets indicate candidate sets that performed significantly better than other for each respective protein ($p < 0.005$, MWW).

The comparison of the hit ratio for each tested set of columns demonstrated that major virion structural proteins, i.e. *L3*:hexon, *L2*:penton base, and *L5*:fiber, performed better than other proteins (Fig. 3.2), which was consistent with the reported locations of epitopes through different adenoviral species (Stewart et al. 1997; Chiu et al. 2001; Liu et al. 2010) and reflected the serological results used to define the serotypes. However, 31 of the 53 candidate windows had lower average hit ratios than the average of sets of considered sites (<0.65) (Fig. 3.2) despite the significant number of positions ranked as relevant by MSVM-RFE (Fig. 3.1A). This could arise from variability in the number of types among species; therefore, regions relevant for distinguishing between types for species with a high number of types such as HAdV-D (44 types) and -B (12 types) were reflected in the ranking, e.g., *E3*:(CR1β and CR1γ). This result indicated the importance of the hit ratio test for validating the generality and extensibility of any particular region for all seven species.

### 3.3.3 Classification performance of pairs of regions

As expected, any classification based on an independent protein was outperformed by the ALL set; however, it is still possible that a combination of best-performing candidates spread throughout the genome achieves a hit ratio similar to the ALL set while minimizing the number of required sites. To search for such a combination of candidates, different sets of these candidate sites were combined.

My approach aimed to compile a combined set of informative regions that could produce a hit ratio comparable to the one exhibited by the ALL set; to do this, different combinations of candidate regions were assessed, firstly by combining pairs, then triads, and so on. The first attempt at a combinatorial analysis only considered pairs of candidates from the eight different proteins identified with significantly informative candidates (Fig. 3.1). The shown approach chose only the candidates with the highest significant hit ratio (at $p < 0.005$, MWW) among the windows in the respective proteins: *L2*:penton base (two candidates were considered), *L3*:hexon (four candidates were considered), *L4*:100K, *E3*:(CR1α, CR1β, and CR1λ), *L5*:fiber, and *E4*:ORF4 (Fig. 3.2). These candidates represent slightly different amino acid positions in each species with different coding regions in their genomic sequences (Table 2); it is noted that genomes of human adenoviruses are syntenic between different types (Davison, Benko, Harrach 2003), which results in very stable multiple sequence alignments. The ALL set was included as the baseline for the hit ratio test compared to the 66 combinations of candidates. Consequently, tested pairs compared the hit ratio of 102 sites to the baseline of 12,939 sites.

**Table 2. Amino acid and nucleotide positions of candidates for adenoviral type classifiers by species**[a]

| Candidates | | HAdV-A12 (X73487) | HAdV-B11 (AF532578) | HAdV-C2 (J01917) | HAdV-D8 (AB448767) | HAdV-E4 (AY458656) | HAdV-F41 (HM565136) | HAdV-GGT52 (DQ923122) |
|---|---|---|---|---|---|---|---|---|
| ***L2*:penton base [401–451]** | Protein[b] | 287–301 | 329–358 | 328–359 | 303–324 | 315–335 | 297–311 | 296–310 |
| | Genome[c] | 14,255–14,296 | 14,670–14,756 | 15,135–15,182 | 14,451–14,495 | 14,766–14,786 | 14,121–14,162 | 14,018–14,059 |
| ***L2*:penton base [426–476]** | Protein | 294–301 | 342–362 | 341–371 | 311–324 | 321–337 | 304–311 | 303–310 |
| | Genome | 14,273–14,296 | 14,706–14,768 | 15,168–15,261 | 14,454–14,495 | 14,775–14,825 | 14,139–14,162 | 14,036–14,059 |
| ***L3*:hexon [176–226]** | *Protein* | 138–170 | 154–188 | 164–202 | 146–181 | 139–174 | 138–168 | 138–169 |
| | *Genome* | 18,154–18,249 | 18,717–18,818 | 19,327–19,443 | 18,214–18,318 | 18,665–18,769 | 18,001–18,090 | 17,758–17,850 |
| ***L3*:hexon [201–251]** | *Protein* | 152–194 | 170–213 | 184–225 | 162–204 | 156–199 | 152–193 | 152–194 |
| | *Genome* | 18,196–18,321 | 18,762–18,893 | 19,390–19,512 | 18,262–18,387 | 18,713–18,844 | 18,043–18,165 | 17,800–17,925 |
| ***L3*:hexon [326–376]** | *Protein* | 242–290 | 273–319 | 288–334 | 265–311 | 261–305 | 252–298 | 244–290 |
| | *Genome* | 18,466–18,609 | 19,074–19,211 | 19,702–19,839 | 18,571–18,708 | 19,031–19,162 | 18,343–18,480 | 18,076–18,213 |
| ***L3*:hexon [476–526]** | *Protein* | 390–430 | 419–459 | 434–479 | 411–453 | 405–447 | 398–436 | 390–430 |
| | *Genome* | 18,907–19,029 | 19,509–19,631 | 20,137–20,274 | 19,006–19,134 | 19,460–19,588 | 18,778–18,894 | 18511–18,633 |
| ***L4*:100K [101–151]** | Protein | 52–73 | 78–112 | 59–90 | 37–53 | 63–92 | 50–63 | 36–58 |
| | Genome | 22,848–22,913 | 23,664–23,768 | 24,282–24,377 | 22,880–22,930 | 23,527–23,616 | 22,650–22,691 | 22,297–22,365 |
| ***E3*:CR1α [51–101]** | Protein | NA | 32–65 | 42–91 | 42–89 | 44–89 | NA | 51–92 |
| | Genome | NA | 27,552–27,650 | 28,293–28,442 | 26,525–26,668 | 27,408–27,545 | NA | 26,181–26,306 |
| ***E3*:CR1-β [1–51]** | Protein | NA | 1–39 | 1–13 | 1–40 | 1–42 | 1–50 | 1–41 |
| | Genome | NA | 28,356–28,472 | 29,546–29,584 | 27,494–27,571 | 28,449–28,574 | 26,772–26,888 | 26,552–26,674 |
| ***E3*:CR1-γ [101–151]** | Protein | NA | 69–92 | NA | 69–116 | NA | NA | NA |
| | Genome | NA | 29,126–29,194 | NA | 28,900–29,040 | NA | NA | NA |
| ***L5*:fiber [226–276]** | Protein | 223–266 | 78–104 | 186–231 | 91–129 | 91–129 | 169–213 | 168–212 |
| | Genome | 30,034–30,165 | 31,046–31,123 | 31,585–31,683 | 31,055–31,171 | 31,931–32,038 | 30,439–30,573 | 30,159–30,293 |
| ***E4*:ORF4 [26–76]** | Protein | 25–75 | 25–75 | 25–75 | 24–74 | 25–75 | 26–76 | 25–75 |
| | Genome | 32,382–32,534 | 33,018–33,170 | 34,148–34,270 | 33,098–33,250 | 34,246–34,368 | 32,841–32,993 | 32,567–32,680 |

[a] NA, absent in the species

[b] Amino acid position in the protein without gaps

[c] Nucleotide position in the genome without gaps

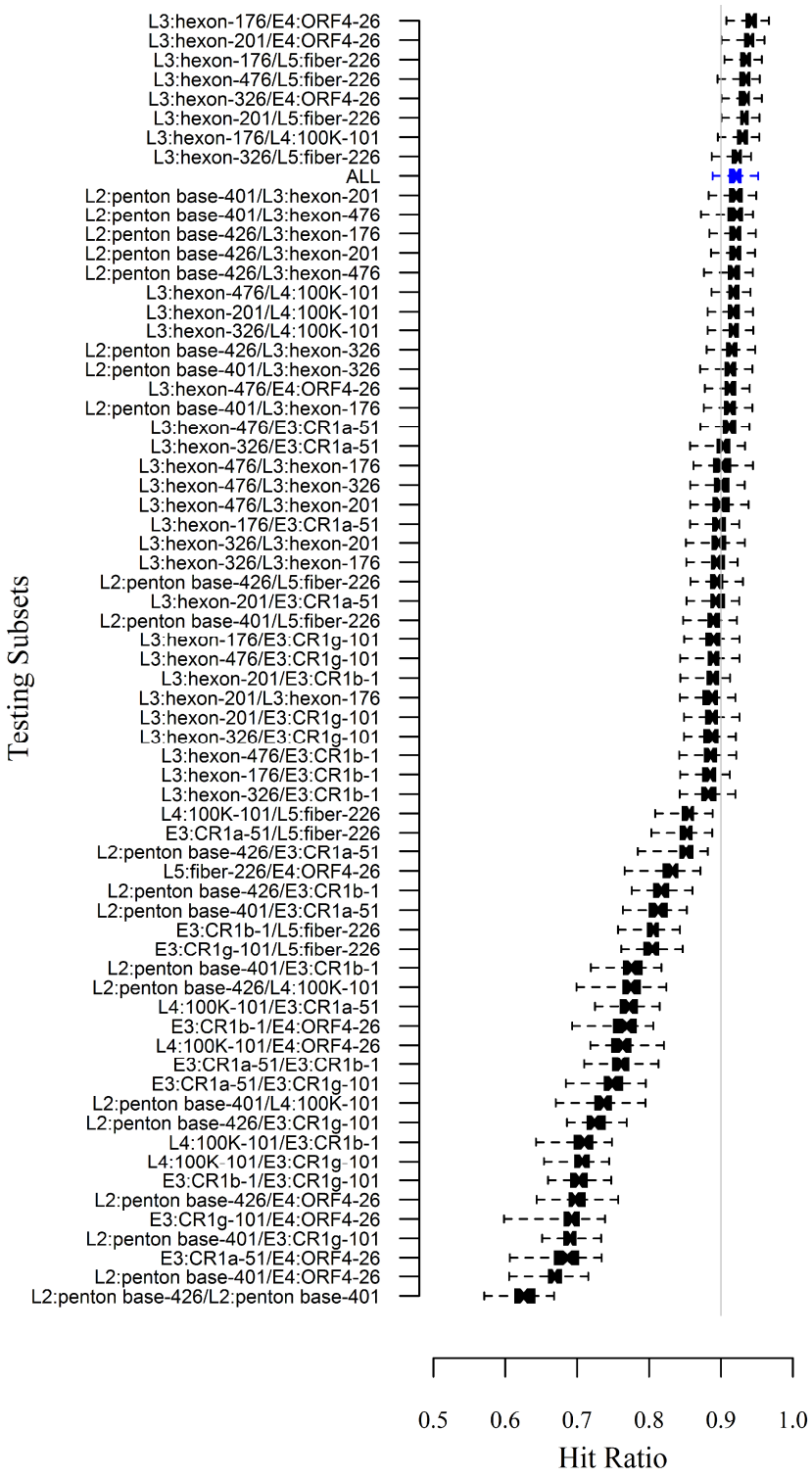


**Fig. 3.3 Boxplot of the hit ratios of pairs of candidates.** The ordinate presents the sets of sites. The abscissa shows the hit ratio from the cross-validation test. The boxes represent the distribution of the hit ratio through repetitions of the cross-validation test. The blue and black boxes represent the ALL set and the pairs of candidate sets, respectively. The gray line indicates the 0.9 hit ratio.

The hit ratio distributions of the paired-sets were sorted in descending order according to their average (Fig. 3.3). Interestingly, the ALL set was significantly outperformed by seven combinations of candidates ($p < 0.005$, MWW): L3:hexon-176/E4:ORF4-26, L3:hexon-201/E4:ORF4-26, L3:hexon-176/L5:fiber-226, L3:hexon-476/L5:fiber-226, L3:hexon-326/E4:ORF4-26, L3:hexon-201/L5:fiber-226, and L3:hexon-176/L4:100K-101. Moreover, combinations of hexon candidates and candidates from other proteins were found in all of the top sets with an average hit ratio above 0.9. Although this result confirmed the importance of the hexon protein for type classification, as indicated by serology-based methods, it also suggested that candidate sets from other proteins play a complementary role in completing the information required for accurate type classification.

Although other pairs were outperformed by the ALL set, they showed improved hit ratios than single proteins. For instance, combination of the *L2*:penton base-401 and *L5*:fiber-226 significantly outperformed the hit ratio of the whole penton base and fiber proteins ($p < 1 \times 10^{-24}$, MWW).

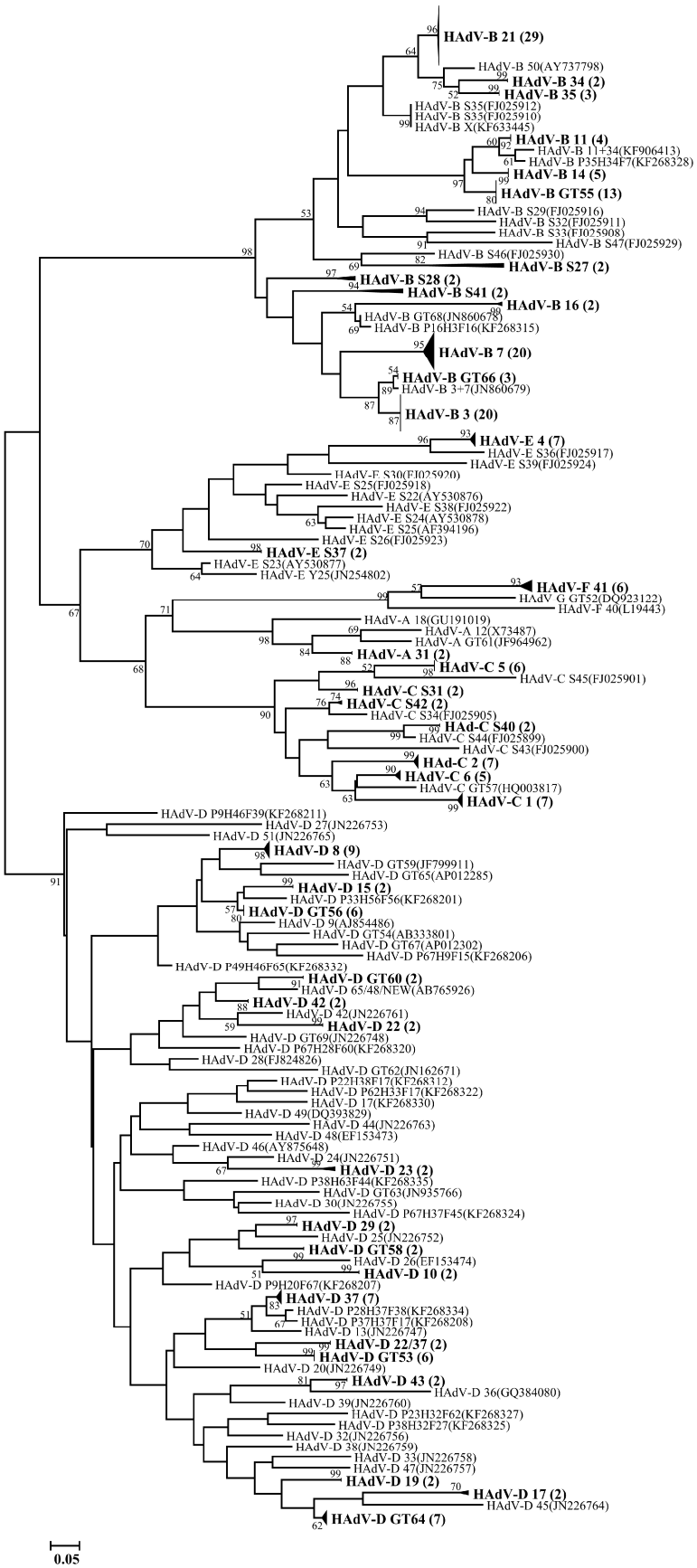
HAdV-B 21 (29)
HAdV-B 50(AY737798)
HAdV-B 34 (2)
HAdV-B 35 (3)
HAdV-B S35(FJ025912)
HAdV-B S33(FJ025910)
HAdV-B X(KF633445)
HAdV-B 11 (4)
HAdV-B 11+34(KF906413)
HAdV-B P35H34F7(KF268328)
HAdV-B 14 (5)
HAdV-B GT55 (13)
HAdV-B S29(FJ025916)
HAdV-B S32(FJ025911)
HAdV-B S33(FJ025908)
HAdV-B S47(FJ025929)
HAdV-B S46(FJ025930)
HAdV-B S27 (2)
HAdV-B S28 (2)
HAdV-B S41 (2)
HAdV-B 16 (2)
HAdV-B GT68(JN860678)
HAdV-B P16H3F16(KF268315)
HAdV-B 7 (20)
HAdV-B GT66 (3)
HAdV-B 3+7(JN860679)
HAdV-B 3 (20)
HAdV-E 4 (7)
HAdV-E S36(FJ025917)
HAdV-E S39(FJ025924)
HAdV-E S30(FJ025920)
HAdV-E S25(FJ025918)
HAdV-E S22(AY530876)
HAdV-E S38(FJ025922)
HAdV-E S24(AY530878)
HAdV-E S25(AF394196)
HAdV-E S26(FJ025923)
HAdV-E S37 (2)
HAdV-E S23(AY530877)
HAdV-E Y25(JN254802)
HAdV-F 41 (6)
HAdV-G GT52(DQ923122)
HAdV-F 40(L19443)
HAdV-A 18(GU191019)
HAdV-A 12(X73487)
HAdV-A GT61(JF964962)
HAdV-A 31 (2)
HAdV-C 5 (6)
HAdV-C S45(FJ025901)
HAdV-C S31 (2)
HAdV-C S42 (2)
HAdV-C S34(FJ025905)
HAd-C S40 (2)
HAdV-C S44(FJ025899)
HAdV-C S43(FJ025900)
HAd-C 2 (7)
HAdV-C 6 (5)
HAdV-C GT57(HQ003817)
HAdV-C 1 (7)
HAdV-D P9H46F39(KF268211)
HAdV-D 27(JN226753)
HAdV-D 51(JN226765)
HAdV-D 8 (9)
HAdV-D GT59(JF799911)
HAdV-D GT65(AP012285)
HAdV-D 15 (2)
HAdV-D P33H56F56(KF268201)
HAdV-D GT56 (6)
HAdV-D 9(AJ854486)
HAdV-D GT54(AB333801)
HAdV-D GT67(AP012302)
HAdV-D P67H9F15(KF268206)
HAdV-D P49H46F65(KF268332)
HAdV-D GT60 (2)
HAdV-D 65/48/NEW(AB765926)
HAdV-D 42 (2)
HAdV-D 42(JN226761)
HAdV-D 22 (2)
HAdV-D GT69(JN226748)
HAdV-D P67H28F60(KF268320)
HAdV-D 28(FJ824826)
HAdV-D GT62(JN162671)
HAdV-D P22H38F17(KF268312)
HAdV-D P62H33F17(KF268322)
HAdV-D 17(KF268330)
HAdV-D 49(DQ393829)
HAdV-D 44(JN226763)
HAdV-D 48(EF153473)
HAdV-D 46(AY875648)
HAdV-D 24(JN226751)
HAdV-D 23 (2)
HAdV-D P38H63F44(KF268335)
HAdV-D GT63(JN935766)
HAdV-D 30(JN226755)
HAdV-D P67H37F45(KF268324)
HAdV-D 29 (2)
HAdV-D 25(JN226752)
HAdV-D GT58 (2)
HAdV-D 26(EF153474)
HAdV-D 10 (2)
HAdV-D P9H20F67(KF268207)
HAdV-D 37 (7)
HAdV-D P28H37F38(KF268334)
HAdV-D P37H37F17(KF268208)
HAdV-D 13(JN226747)
HAdV-D 22/37 (2)
HAdV-D GT53 (6)
HAdV-D 20(JN226749)
HAdV-D 43 (2)
HAdV-D 36(GQ384080)
HAdV-D 39(JN226760)
HAdV-D P23H32F62(KF268327)
HAdV-D P38H32F27(KF268325)
HAdV-D 32(JN226756)
HAdV-D 38(JN226759)
HAdV-D 33(JN226758)
HAdV-D 47(JN226757)
HAdV-D 19 (2)
HAdV-D 17 (2)
HAdV-D 45(JN226764)
HAdV-D GT64 (7)
0.05

**Fig. 3.4 Phylogenetic tree of amino acid sequences using the concatenation of best candidates.** An unrooted neighbor-joining tree of the amino acid concatenation of three of the best performing candidates and positions with less than 95% site coverage were eliminated leaving 63 sites. Bootstrapped confidence values calculated with 1,000 replicates are shown beside branches as percentages (>50). Numbers in parentheses indicate the number of collapsed strains from the same type. The tree is drawn to scale: branch lengths represent the number of amino acid substitutions per site computed using the JTT matrix-based method. HAdV: human mastadenovirus; SAdV: simian mastadenovirus.

### 3.3.4 Phylogenetic analysis of the best-performing candidates

A phylogenetic tree was prepared from the concatenated amino acid sequences of the three best-performing candidates in paired tests: *L2*: penton base-401, *L3*: hexon-176, and *L5*: fiber-226, to corroborate the findings that such regions provide a clear separation of the HAdV types (Fig. 3.4). The phylogenetic tree for the concatenation of all proteins was prepared in a similar way using MEGA6 (Tamura et al. 2013) (Fig. 3.5).

The high bootstrap values (>70%) supporting the clusters of strains from the same type evidenced the classification power of the 153-amino acid concatenation. Specifically, the average difference in amino acid substitutions per site in the alignment between any two taxa from two different type clusters was 1.05, whereas the average difference between any two taxa in the same type cluster was 0.03. Comparatively, the respective values calculated over the ALL set were 0.27 and 0.009. Such a contrast supports the utility of the concatenated regions for type classification by reflecting low differences with high statistical support between samples of the same type and high differences between different types. Thus, a proper type classification was achieved with ~1.2% of the total amino acid sites (153 of 12,939). The difference between taxa representing the type clusters in the phylogenetic tree corroborated the type predictive accuracy evidenced during the hit ratio tests, which was extended to other types of *HAdV* isolated exclusively from primates, and it demonstrated the power of the candidates for classifying unknown types.

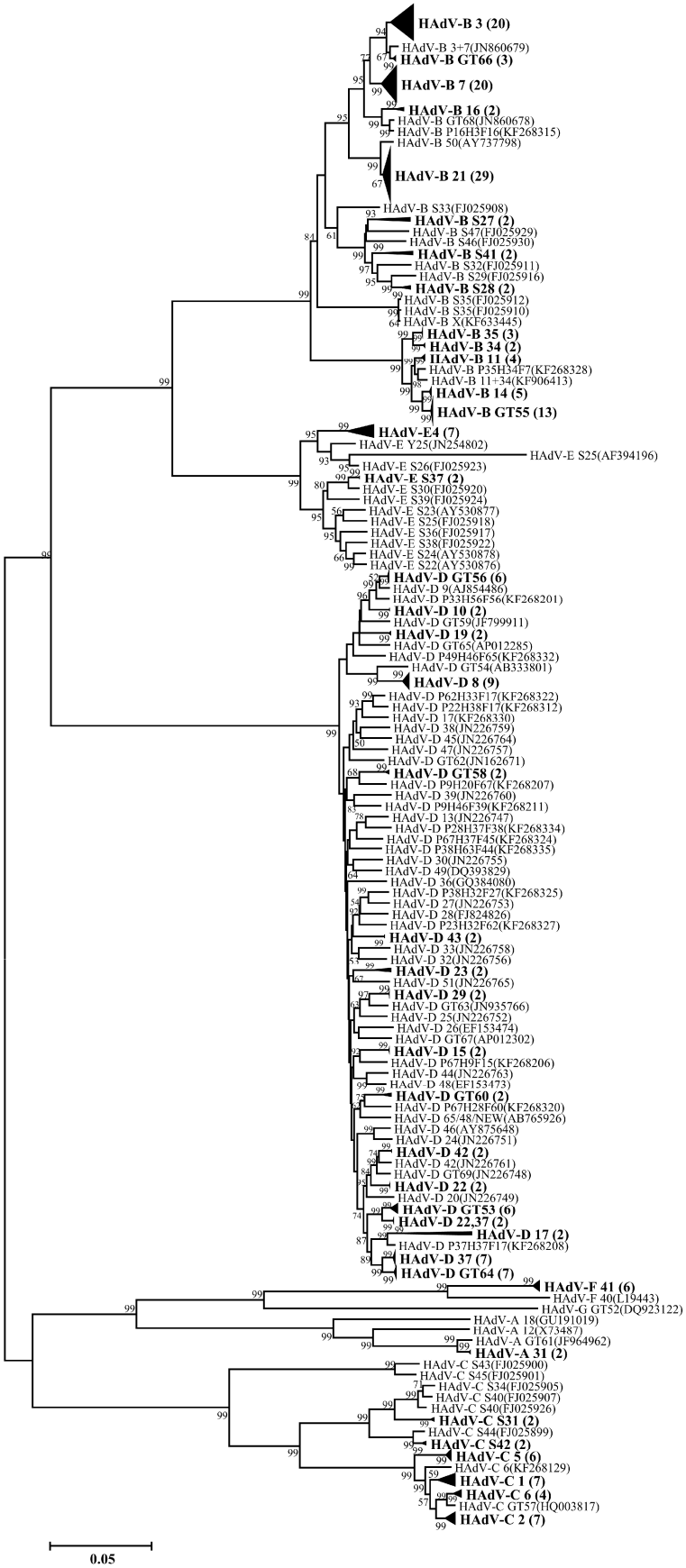
HAdV-B 3 (20)
HAdV-B 3+7(JN860679)
HAdV-B GT66 (3)
HAdV-B 7 (20)
HAdV-B 16 (2)
HAdV-B GT68(JN860678)
HAdV-B P16H3F16(KF268315)
HAdV-B 50(AY737798)
HAdV-B 21 (29)
HAdV-B S33(FJ025908)
HAdV-B S27 (2)
HAdV-B S47(FJ025929)
HAdV-B S46(FJ025930)
HAdV-B S41 (2)
HAdV-B S32(FJ025911)
HAdV-B S29(FJ025916)
HAdV-B S28 (2)
HAdV-B S35(FJ025912)
HAdV-B S35(FJ025910)
HAdV-B X(KF633445)
HAdV-B 35 (3)
HAdV-B 34 (2)
HAdV-B 11 (4)
HAdV-B P35H34F7(KF268328)
HAdV-B 11+34(KF906413)
HAdV-B 14 (5)
HAdV-B GT55 (13)
HAdV-E4 (7)
HAdV-E Y25(JN254802)
HAdV-E S25(AF394196)
HAdV-E S26(FJ025923)
HAdV-E S37 (2)
HAdV-E S30(FJ025920)
HAdV-E S39(FJ025924)
HAdV-E S23(AY530877)
HAdV-E S25(FJ025918)
HAdV-E S36(FJ025917)
HAdV-E S38(FJ025922)
HAdV-E S24(AY530878)
HAdV-E S22(AY530876)
HAdV-D GT56 (6)
HAdV-D 9(AJ854486)
HAdV-D P33H56F56(KF268201)
HAdV-D 10 (2)
HAdV-D GT59(JF799911)
HAdV-D 19 (2)
HAdV-D GT65(AP012285)
HAdV-D P49H46F65(KF268332)
HAdV-D GT54(AB333801)
HAdV-D 8 (9)
HAdV-D P62H33F17(KF268322)
HAdV-D P22H38F17(KF268312)
HAdV-D 17(KF268330)
HAdV-D 38(JN226759)
HAdV-D 45(JN226764)
HAdV-D 47(JN226757)
HAdV-D GT62(JN162671)
HAdV-D GT58 (2)
HAdV-D P9H20F67(KF268207)
HAdV-D 39(JN226760)
HAdV-D P9H46F39(KF268211)
HAdV-D 13(JN226747)
HAdV-D P28H37F38(KF268334)
HAdV-D P67H37F45(KF268324)
HAdV-D P38H63F44(KF268335)
HAdV-D 30(JN226755)
HAdV-D 49(DQ393829)
HAdV-D 36(GQ384080)
HAdV-D P38H32F27(KF268325)
HAdV-D 27(JN226753)
HAdV-D 28(FJ824826)
HAdV-D P23H32F62(KF268327)
HAdV-D 43 (2)
HAdV-D 33(JN226758)
HAdV-D 32(JN226756)
HAdV-D 23 (2)
HAdV-D 51(JN226765)
HAdV-D 29 (2)
HAdV-D GT63(JN935766)
HAdV-D 25(JN226752)
HAdV-D 26(EF153474)
HAdV-D GT67(AP012302)
HAdV-D 15 (2)
HAdV-D P67H9F15(KF268206)
HAdV-D 44(JN226763)
HAdV-D 48(EF153473)
HAdV-D GT60 (2)
HAdV-D P67H28F60(KF268320)
HAdV-D 65/48/NEW(AB765926)
HAdV-D 46(AY875648)
HAdV-D 24(JN226751)
HAdV-D 42 (2)
HAdV-D 42(JN226761)
HAdV-D GT69(JN226748)
HAdV-D 22 (2)
HAdV-D 20(JN226749)
HAdV-D GT53 (6)
HAdV-D 22,37 (2)
HAdV-D 17 (2)
HAdV-D P37H37F17(KF268208)
HAdV-D 37 (7)
HAdV-D GT64 (7)
HAdV-F 41 (6)
HAdV-F 40(L19443)
HAdV-G GT52(DQ923122)
HAdV-A 18(GU191019)
HAdV-A 12(X73487)
HAdV-A GT61(JF964962)
HAdV-A 31 (2)
HAdV-C S43(FJ025900)
HAdV-C S45(FJ025901)
HAdV-C S34(FJ025905)
HAdV-C S40(FJ025907)
HAdV-C S40(FJ025926)
HAdV-C S31 (2)
HAdV-C S44(FJ025899)
HAdV-C S42 (2)
HAdV-C 5 (6)
HAdV-C 6(KF268129)
HAdV-C 1 (7)
HAdV-C 6 (4)
HAdV-C GT57(HQ003817)
HAdV-C 2 (7)
0.05

**Fig. 3.5 Phylogenetic tree of amino acid sequences using the concatenation of ALL proteins.** An unrooted neighbor-joining tree of the amino acid concatenation of all proteins and positions with less than 95% site coverage were eliminated leaving 9497 sites. Bootstrapped confidence values calculated with 1,000 replicates are shown beside branches as percentages (>50). Numbers in parentheses indicate the number of collapsed strains from the same type. The tree is drawn to scale: branch lengths represent the number of amino acid substitutions per site computed using the JTT matrix-based method. HAdV: human mastadenovirus; SAdV: simian mastadenovirus.

To assess the applicability of the characterized best performing candidates in type classification, a particular strain was analyzed: HAdV-D P65H48F60 (AB765926), which was considered to be a variant of HAdV-D48 by a partial hexon analysis and characterized as the result of a complex evolutionary process by a genome analysis (Fujimoto et al. 2014). Such a strain is easily detectable as a strain different to HAdV-D48 with the three of the best performing candidates concatenated above (Fig. 3.4) without requiring more than three partial sequences targeting the corresponding coding areas (Table 2) and performing a search for the closest type with BLAST (Zhang et al. 2000). The results are HAdV-D GT65 (e-value < $10^{-15}$, 100% query coverage), HAdV-D48 (e-value < $3 \times 10^{-48}$, 100% query coverage) and HAdV-D GT60 (e-value < $10^{-46}$, 100% query coverage), for penton base-401, hexon-176, and fiber-226, respectively. These results were consistent with using complete protein coding regions despite the fact that these candidates represented < 5% of these regions. This suggests that short informative regions could reduce the effort required to classify the sample types, which is important when working with the large quantities of samples faced by surveillance programs.

## 3.4 Conclusion

The SVM approach reported here determined adenoviral genomic regions sufficient for type classification. This approach provided an appropriate and replicable test framework to assess and compare different proposed sets of regions, such as whole genome or single protein, for detection and classification of HAdV types. The results of this analytical approach have practical applications for type identification and rapid classification in infection surveillance programs. PCR sequencing methods targeting the characterized regions that encode the best candidates maximize type classification information while minimizing the required genomic region to be sampled. This way, more strains can be accurately classified at a lower cost, within a shorter timeframe, and with increased sensitivity to spot novel types. Also, since this approach employed general methods for sequence analyses and machine learning; thereby, it

is applicable to any other taxa to be classified if the genome sequences with annotation data and classification information are available.

# Chapter 4: Intertypic modular exchanges of genomic segments by homologous recombination at universally conserved segments in *Human mastadenovirus D*

*Human mastadenovirus D*, which is composed of clinically and epidemiologically important pathogens worldwide, contains more diverged type than any other species of the genus Mastadenovirus, although the mechanisms accounting for the high level of diversity remain to be disclosed. Recent studies of known and new types of HAdV-D have indicated that intertypic recombination between distant types contributes to the increasing diversity of the species. However, such findings raise the question as to how homologous recombination events occur between diversified types since homologous recombination is suppressed as nucleotide sequences diverge. In order to address this question, the distribution of the recombination boundaries was investigated in comparison with the landscape of intergenomic sequence conservation assessed according to the synonymous substitution rate ($d_S$). The results revealed that specific genomic segments are conserved between even the most distantly related genomes; these segments were called "universally conserved segments" (UCSs). These findings suggest that UCSs facilitate homologous recombination, resulting in intergenomic segmental exchanges of UCS-flanking genomic regions as recombination modules. With the aid of such a mechanism, the haploid genomes of HAdV-Ds may have been reshuffled, resulting in chimeric genomes out of diversified repertoires in the HAdV-D population analogous to the MHC region reshuffled via crossing over in vertebrates. In addition, some HAdVs with chimeric genomes may have had the opportunity to avoid host immune responses thereby causing epidemics.

## 4.1 Introduction

As presented in Chapter 1, HAdV-D is exceptionally type-rich among the HAdV species. Tissue tropism varies even between types of the same species; in particular, the types of HAdV-D range widely from the cornea/conjunctiva to the respiratory tract and intestines, causing epidemic keratoconjunctivitis outbreaks (Fujimoto et al. 2012; Centers for Disease, Prevention 2013), respiratory diseases (Robinson et al. 2011b), acute gastroenteritis (Matsushima et al. 2013) and obesity (Arnold et al. 2010). Moreover, HAdV-D is only human-specific and has recently been reported to have an increasing number of new types (Ishiko et al. 2008; Robinson et al. 2009; Kaneko et al. 2011b). Due to these clinical characteristics, *HAdV-D* is recognized to be one of the most virologically important adenoviral species.

The HAdV-D genotypes have been recognized as new recombinant forms of known types, e.g., the HAdV-GT53 (Aoki et al. 2008; Walsh et al. 2009; Kaneko et al. 2011a), HAdV-GT56 (Kaneko et al. 2011b; Robinson et al. 2011b), HAdV-GT63 (Singh et al. 2012) and the HAdV-GT64 (Zhou et al. 2012), except for HAdV-GT54, which has been recognized to be a new type based on genome typing methods (Ishiko et al. 2008; Jin et al. 2011) and its genomic sequence identity (Kaneko et al. 2011c). Additionally, it has been reported that adenoviral intertypic recombination events have frequently taken place (Lukashev et al. 2008), even between distantly related types of the same species, as found in the genome of HAdV-GT53 derived from HAdV-22, HAdV-37 and HAdV-8 (Aoki et al. 2008; Walsh et al. 2009; Kaneko et al. 2011a). However, it has not yet been well studied how such intertypic recombination events have been involved in the evolution of HAdVs and what mechanisms have allowed for recombination even between distantly related viruses. This chapter addressed these points using 45 available genome sequences of HAdV-D (Table 1), the most abundant HAdV species with respect to genomic data, in order to identify possible molecular mechanisms behind the associated genome evolution.

## 4.2 Materials & Methods

### 4.2.1 The data

The complete genome sequences of 45 types of HAdV-D available from the International Nucleotide Sequence Database Collaboration (INSDC) database (Table 1) were aligned into a multiple genome alignment (MGA) using the Iterative Refinement Method

algorithm (FFT-NS-I) of MAFFT (Katoh et al. 2002). The length of the thus produced MGA was 35,807 sites, including gaps. After removing all gaps, the regions of the DNA polymerase (3,273 sites) were extracted for further analyses. A rooted neighbor-joining tree of the translated sequence of the DNA polymerase was constructed by adding outgroups (HAdV-C2 and HAdV-C5) under the Tamura-Nei model (TN93) (Tamura, Nei 1993) and confirming the results with a bootstrap test of 1,000 replicates.

### 4.2.2 Counting recombination boundaries

The recombination events in the genome sequences were identified using the algorithms introduced in section 2.2. The seven algorithms are available in the RDP 4.22β program (Heath et al. 2006): RDP (Martin, Rybicki 2000), GENECONV (Padidam, Sawyer, Fauquet 1999), Chimaera (Posada, Crandall 2001), MaxChi (Smith 1992), BootScan (Martin et al. 2005a), SiScan (Gibbs, Armstrong, Gibbs 2000) and 3Seq (Boni, Posada, Feldman 2007). Among the thus identified recombination event signals, only those which were identified by >3 different algorithms with a Bonferroni-corrected p-value of < 0.001 in each of the algorithms were used for the further analyses described below in order to reduce the false-positive rate (Lefeuvre et al. 2007; Lefeuvre et al. 2009). Then, the list of unique events was referred as reliable recombination events (RREs) identified by the RDP programs produced by eliminating redundantly events identified using different algorithms (Martin, Rybicki 2000; Heath et al. 2006). From this list, the number of boundaries of RREs was counted along the MGA with a sliding window of a 200-nt width with a 100-nt step. Then, a binomial test was applied to each window in order to assess the significance of the number of recombination boundaries observed in the window.

### 4.2.3 Detection of coevolving regions

Using the MGA data, the evolutionary collinearity of the genomic segments was analyzed using the same approach as the Recombination Region Count Matrix (Lefeuvre et al. 2007). In this analysis, the regions from positions $x$ to $y$ in the MGA for all possible combinations of $x$ and $y$ ranging from the left to the right termini of the MGA were considered, and the number of cases in which either $x$ or $y$ but not both were located in the region of an RRE was counted for all possible pairs of genomes. This analysis identified which regions have tended to evolve together.

### 4.2.4 Identification of variable and universally conserved segments

For the statistical evaluations of sequence conservation, the average pairwise evolutionary distance between 45 genomes was calculated for each sliding window with a 200-nt width and a 100-nt step after all gapped alignment sites were removed from the MGA to create a gap-free MGA. For each window position, $\bar{d}(w) = \left(\frac{N(N-1)}{2}\right)^{-1} \sum_{i<j}^{N} d(i,j,w)$ and the standard deviation of $d(i,j,w)$ were calculated, where $N$ represents the number of the genomes compared, i.e., 45 in the present chapter, and $d(i,j,w)$ represents the pairwise evolutionary distance between sequences $i$ and $j$ at window $w$ of the gap-free MGA under the TN93 model (Tamura, Nei 1993).

For the synonymous substitution rate ($d_S$) analyses (see section 2.3), firstly the amino acid sequence alignment of each ORF in the genomes were assessed and then back-translated to the original codon alignment in order to avoid inappropriate gap insertions causing wrong $d_S$ values due to unnecessary frame shifts. Gapped sites were eliminated from the codon alignment prior to the $d_S$ calculations. Each codon alignment was scanned with a 204-nt window, which corresponds to 68 codons, and a 102-nt step in order to calculate $d_S(i, j, k, c)$, which represents the $d_S$ value between sequences $i$ and $j$ calculated at codon position $c$ of the window in gene $k$ ($1 \leq k \leq 13$, gene $U$ was excluded from this analysis due to its short length). All $d_S(i, j, k, c)$ values were estimated using R library *seqinr* (Charif et al. 2005). To test whether the set of the $d_S$'s at specific window position C in the specific gene K were significantly lower than the $d_S$'s located at the other positions in the same gene (gene K), a Mann–Whitney Wilcoxon rank test was performed for $D_S^{K,C} < d_S^{K,C}$, where $D_S^{K,C}$ is the set of all $d_S$ values at window position C in gene K, and $d_S^{K,C}$ is the set of all $d_S$ values in gene K but $D_S^{K,C}$. Similarly, the non-coding regions were analyzed, including the intergenic, intron and UTRs, using $d(i, j, w)$ instead of $d_S$.

## 4.3 Results

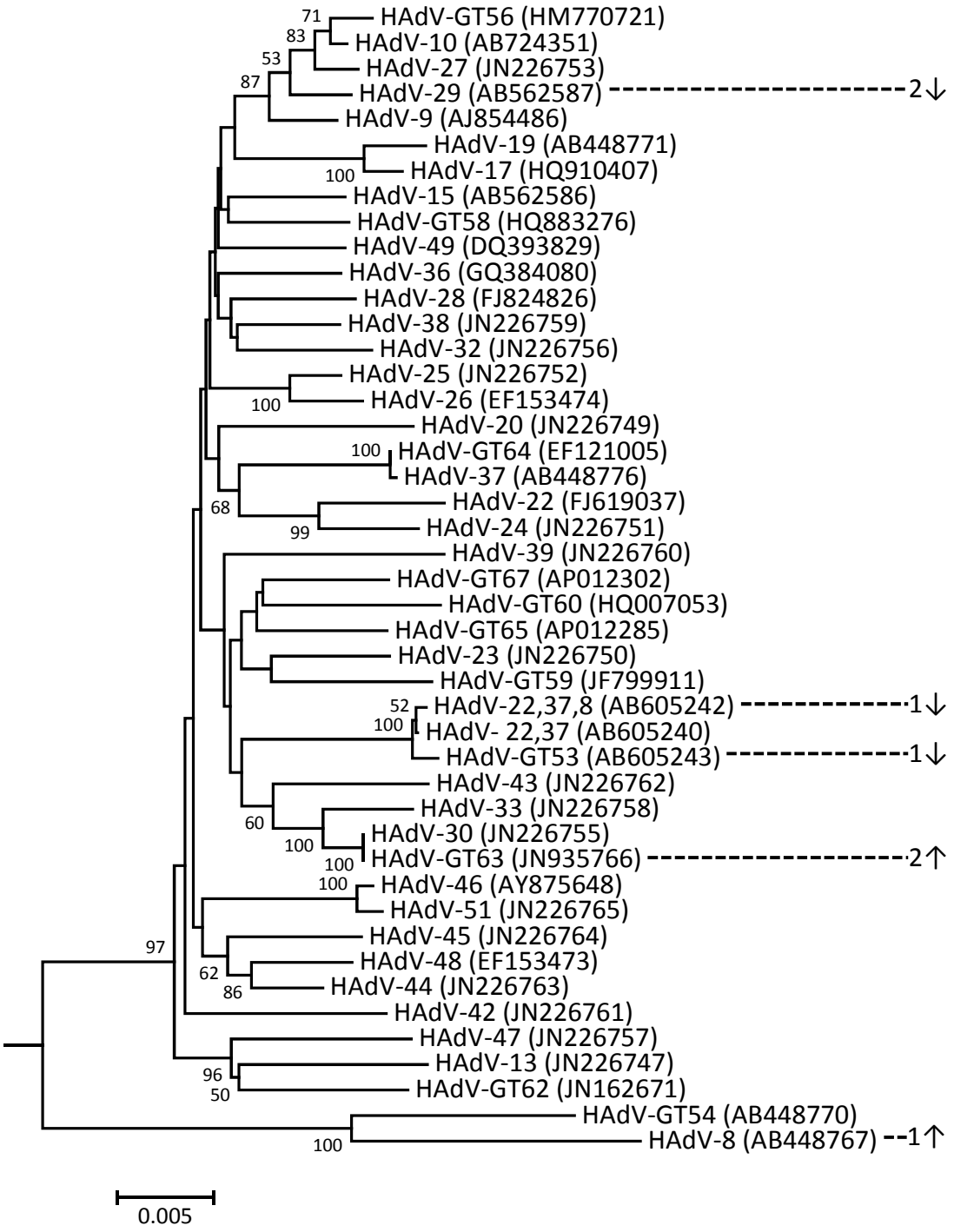


**Fig. 4.1 Neighbor-joining tree of the *E2B* DNA polymerase gene nucleotide sequences.** This tree was rooted with two HAdV-C sequences as outgroups (omitted from the figure). Bootstrap confidence values calculated with 1,000 replicates are shown beside branches in percentages (>50). Two examples of reliable recombination events that were detected between the types with RDP are indicated by the same number ("1" or "2") and arrows. 

### 4.3.1. Detection of recombination signals

The MGA of 45 HAdV-D was built with two outgroup genomes (Table 1). The average evolutionary distance at the nucleotide sequence level between these HAdV-D genomes was 0.06, with the largest distance being 0.10 between HAdV-8 and HAdV-39. Then a molecular phylogenetic tree of the HAdV-D types was constructed using the nucleotide sequences of their DNA polymerase genes instead of the whole genome sequences in order to obtain a firm reference phylogeny between the HAdV-D types (Fig. 4.1). Recent reports have shown that certain HAdV-Ds have experienced intertypic genomic recombination events, as exemplified above (Aoki et al. 2008; Walsh et al. 2009; Kaneko et al. 2011a; Kaneko et al. 2011c; Singh et al. 2012; Zhou et al. 2012); however, the polymerase gene is one of the most conserved and almost recombination-free regions in the HAdV-D genome (Liu, Naismith, Hay 2000). Based on whether the evolutionary distance between the polymerase genes of each possible pair of 45 HAdV types was less or greater than the average (0.02), the pairs of 45 HAdV types were classified into two classes: close and distant, respectively.

Using the RDP program (Heath et al. 2006), recombination signals were searched in the MGA, each of which consisted of a recombined region and two boundaries. Initially, 384 signals were detected using this program, 188 of which were recognized to be unique and reliable (RREs), showing that each genome contains ca. 4.2 recombination signals on average. Among these signals, 124 were identified between distant types. Two examples between HAdV-GT53 and HAdV-8 (Aoki et al. 2008; Walsh et al. 2009; Kaneko et al. 2011a) and between HAdV-GT63 and HAdV-29 (Singh et al. 2012) are depicted in Figure 4.1. These 188 RREs were further analyzed in the following sections.

### 4.3.2 Distribution of the detected recombined boundaries

The distribution of the boundaries of the RREs was investigated as whether were randomly distributed over the genome by using a binomial test of the frequency of the boundaries falling in each sliding window. This test calculated how likely were these recombination boundaries to be detected by chance in the observed number of each window under the expected number of recombination boundaries per nucleotide (0.0105=the total number of boundaries (376) divided by the number of sites of genome alignment (35,807)).

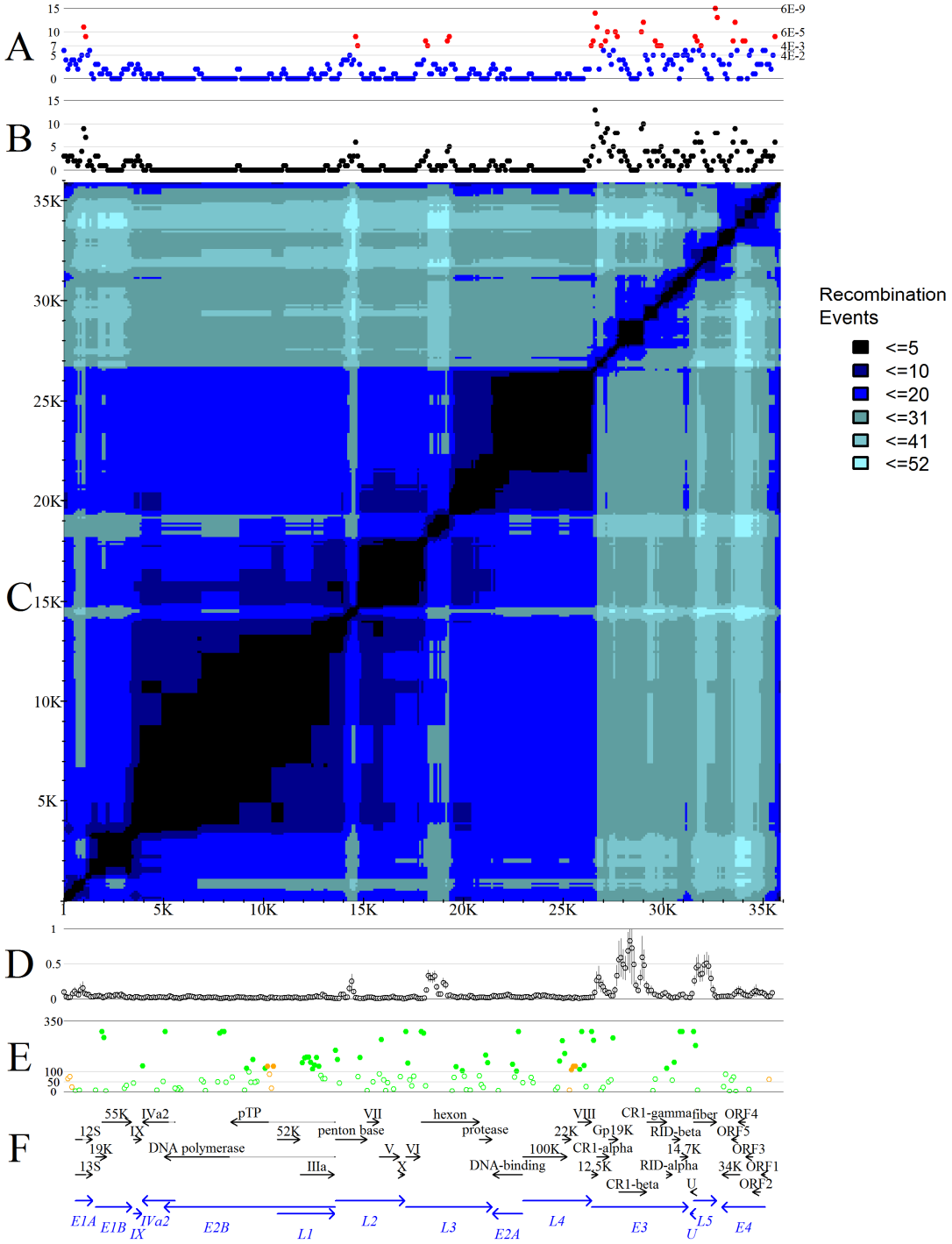


**Fig. 4.2 Sliding window analyses of the MGA.** The abscissae represent the position in the MGA. (A) The left ordinate shows the number of recombination boundaries at each window position, and the right ordinate shows the binomial probability of the number of recombination boundaries at each window position. The points $p<0.01$ are shown in red. (B) The left ordinate shows the number of recombination boundaries between distant genomes at each window position. (C) Evolutionary collinearity interruption matrix. The abscissa and ordinate (the $x$ and $y$ axes) of this symmetric matrix represent the physical positions in the MGA, and each point $(x, y)$ in this matrix shows the number of recombined regions of RREs including $x$ or $y$ but not both according to the color code presented on the right side of the matrix. (D) The average evolutionary distance between 45 genomes at each window position. The standard deviations are shown with error bars. (E) The positions of the UCSs. The ordinate

shows the probability of a Mann–Whitney Wilcoxon rank test to a $-\log_{10}$ scale. The green and orange filled dots indicate the UCSs ($p<10^{-100}$), and the empty green and orange dots show significantly low $d_S$ and $d$ with $p<0.01$, respectively. (F) The positions of the protein-coding exons and *genes* (blue) in the HAdV-D genome. The arrowheads indicate the 3' end of the coding region, and the gray portion shows the introns between coding exons.


Thirty-three among 357 windows showed significantly high numbers of recombination boundaries ($p<0.01$, red dots in Fig. 4.2A), accounting for 150 of the 376 boundaries, i.e., 40% of the identified recombination boundaries sit on less than 9% of the genome, with these "recombination boundary hotspots" being mostly concentrated in the following genes (the protein product names are in parentheses): *E1A* (13S/12S), *L2* (penton base), *L3* (hexon), *L4* (VIII), *E3* (12.5kDa, CR1α, gp19k, CR1β and CR1γ), *L5* (fiber) and *E4* (ORF4, ORF5 and 34kDa). These results clearly demonstrate that the recombination boundaries are localized in specific genomic regions and not randomly distributed.

In addition, 248 of 376 boundaries (66%) were found between distant genomes, and their distribution (Fig. 4.2B) was highly correlated with that observed between the close genomes, at $r$=0.94 (Pearson correlation test, $p$ < 2E-16). These findings appear to be paradoxical because homologous recombination occurs less frequently as the difference between DNA sequences to be recombined increases (Bailis, Rothstein 1990; Nassif, Engels 1993; Selva et al. 1995; Vulic, Lenski, Radman 1999; Eppley et al. 2007). The recombination events identified in this study were assumed to occur via homologous recombination between equivalent/orthologous loci of different genomes because none of the genomes analyzed contained intragenomic rearrangements causing changes in gene order and/or gene orientation, that is, all genomes are perfectly syntenic to each other, and the recombination events observed in HAdV-D genomes are like crossing over between homologous chromosomes in eukaryotes.

### 4.3.3. Recombined regions and collinearity analysis

In addition to the distribution of the recombination boundaries, the regional preference of the recombined regions was evaluated. Instead of simply counting how many RREs in each region of the genome are involved, how frequently the coevolution of each genomic region has been disrupted by 188 RREs (Fig. 4.2C) was investigated. Each dark rectangular region along the diagonal in Figure 4.2C indicates that the region corresponding to the diagonal of the rectangular region has tended to evolve together. These regions were called recombination modules. The regions consist of small modules that have been reshuffled between different lineages at many different positions, e.g., the region ranging from 26.6 k to 35.7 k, while those of large modules have remained stable, e.g., the region ranging from 3.2 k to 14 k. The modules

were mostly flanked by recombination hotspots (red dots in Fig. 4.2A), indicating that the RREs involved exchanges of modules. It should also be noted that the region from 3.2 k to 21.5 k appears to be a large module interrupted by two modules, since dark colors extend over this range, meaning that, for example, the region around 4 k evolved with the region around 20 k together. Therefore, the hotspots around 14.5 k and 18-19.2 k may not have served as one end of the modules 3.2 k-14 k or 15 k-18 k or 19.2 k-26.6 k for the exchanges, but rather have served as the ends of short modules at 14-14.5 k and 18 k-19.2 k for exchanges, respectively. These regions correspond exactly to the RGD loop region in the penton base and the two loop regions in the hexon. Such modules flanked by recombination hotspots were called frequently recombined modules. Other frequently recombined modules include those between 14.2 k-14.8 k, 18.3 k-19.3 k, 26.8 k-30.7 k, 31.6 k-32.8 k and 33.1 k-35.1 k, corresponding to the following protein products and protein domains, the variable loop and the RGD loop in the penton base, both loops in the hexon, the *E3* region, *L5* (fiber) and *E4*, respectively. The regions between 0.8 k - 1.1 k and 1.8 k - 3.2 k, corresponding to *E1A*:13S/12S and *E1B*:55K, respectively, are less frequently exchanged than the frequently recombined modules but still showed some modularity.

### 4.3.4. Identification of universally conserved segments (UCSs)

In order to investigate the cause of the biased distribution of recombination boundaries and modules and the possible mechanisms that allow for homologous recombination between distant genomes, the sequence conservation level along the genome was assessed, which is known to be directly related to homologous recombination efficiency (Bailis, Rothstein 1990; Nassif, Engels 1993; Selva et al. 1995; Vulic, Lenski, Radman 1999; Eppley et al. 2007). As shown in Figures 4.2C and 4.2D, a sliding window analysis of the average evolutionary distance between 45 genomes revealed that many frequently recombined modules overlap with the variable segments of the genome, defined as those with a significantly higher divergence level with $p<0.05$, including those between 18.2 k-18.8 k, 26.7 k-26.9 k, 27.7 k-28.7 k, 28.8 k-29.3 k and 31.6 k-32.6 k (Fig. 4.2D). This result continues to appear paradoxical, as previously mentioned.

Such a puzzle was approached by assessing the conservation in the flanking regions of the recombination boundary hotspots. Homologous recombination events are initiated at a point at which the nucleotide sequences are almost perfectly matched between the recombined DNAs (Selva et al. 1995; Opperman, Emmanuel, Levy 2004). However, the RDP program only detects recombined regions by searching for changes in the patterns of nucleotide variation

between the genomes compared; hence, the recombination boundaries themselves may not be included in the recombined regions detected due to the high level of conservation, meaning that the true boundaries may be just outside of the detected regions. In order to determine whether this is the case, significantly conserved segments in the genomes were sought by looking at the synonymous substitution rate, $d_S$, in the protein-coding regions (~93% of the HAdV genomes) to eliminate the effects of selective conservation at the amino acid sequence level and simple nucleotide divergence in the intergenic regions.

Among the regions of a significantly low $d_S$ at $p<0.01$, those of $p<10^{-100}$ are called universally conserved segments or UCSs, which are present in virtually all genomes (Fig. 4.2E). Compared with the other plots in Figure 4.2, UCSs were found around highly variable regions, and, more interestingly, such UCSs on either or both sides of the highly variable regions coexisted with the recombination boundary hotspots in their proximity. Furthermore, 17 of the 33 recombination boundary hotspots (red dots in Fig. 4.2A) were found within a 300-nt range from one or more UCSs (Fig. 4.2E). The recombination boundary hotspots in or around the penton base, hexon and fiber genes were found to be just beside the UCSs of $p=10^{-206}$, $10^{-298}$ and $10^{-298}$, respectively.

## 4.4 Discussion

### 4.4.1 Recombination boundary hotspots in HAdV-D

This study demonstrated the presence of recombination boundary hotspots in specific regions in the HAdV-D genome. The biased distribution of recombined regions may be a result of the natural selection of specific recombination products out of those randomly occurring over the genome or may be due to the fact that recombination events occur in limited genomic regions. If the former is the case, it is strongly suggested that the recombined regions are highly important for the fitness of viruses. Indeed, it has been reported that recombination hotspots are found in the coding regions of the major capsid proteins, penton base, hexon and fiber (Robinson et al. 2013), which contain epitopes and serotype determinants (Eiz, Adrian, Pring-Akerblom 1997; Hong et al. 2003), and 11 of the 33 hotspots were confirmed to be found in the coding regions of these major capsid proteins as well as one in the coding region of a minor capsid protein (VIII) (see section 4.3.2). Another 10 hotspots reside in the coding regions of other specific functions, two in each of 13S/12S of gene *E1* (*E1A*:13S/12S), *E4*:34kDa and *E4*:ORF4, which are all involved in apoptosis of the virus-infected host cells (Chinnadurai

1998; Pechkovsky, Salzberg, Kleinberger 2013), and two in each of *E3*:gp19k and *E3*:CR1-β for modulating host's immune reaction (Wold, Tollefson 1998; McSharry et al. 2008; Windheim et al. 2013). Interestingly, these functions are all related to virus-host interactions. Although another 10 remaining hotspots are found in three function-unknown proteins, *E3*:12.5kDa, *E3*:CR1-α and *E3*:CR1-γ, these proteins are encoded by the same *E3* gene as gp19k and CR1-β, implying they may function together (Horwitz 2004). The last hotspot of 33 is located in an ITR.

The functional bias of the hotspot-containing coding regions to virus-host interactions contrasts markedly with recombination boundary-rare regions or "deserts" (Fig. 4.2A), which contain the coding regions of 14 proteins for viral DNA replication (DNA polymerase, pTP and DNA-binding protein), DNA encapsidation (52K and IVa2) and stabilization of the viral core and the major capsid proteins in the structure of the virion (IX, IIIa, VII, V, X, VI, protease, 100K and 22K) (Davison, Benko, Harrach 2003; Russell 2009), apparently all unrelated to virus-host interactions. This correlation between recombination boundary density and functional relatedness of proteins to virus-host interaction implies natural selection of recombined products involved in such interactions.

### 4.4.2 UCSs involved in exchanges of regions between distant genomes

Although it is not yet clear what exact molecular mechanisms are involved, it is obvious that different HAdV-D genomes have been recombined only at orthologous sites, strongly suggesting that homologous recombination mechanisms play a role in recombination events producing RREs. The nucleotide divergence levels of the frequently recombined modules (see section 4.3.3) were much higher than those of the other regions (Fig. 4.2D). This finding is paradoxical since homologous recombination is suppressed as the divergence level between two nucleotide sequences increases, especially at the sites of recombination initiation (Selva et al. 1995; Opperman, Emmanuel, Levy 2004). In order to identify the biological mechanisms underlying these apparently perplexing observations, sequence conservation was investigated between different genomes and found that UCSs reside in the vicinity of highly variable regions (Fig. 4.2E). The average distance between 45 genomes was 0.03 in these UCSs and 0.10 in the other regions, indicating that UCSs have more chances to become the subject of homologous recombination than other regions (Bailis, Rothstein 1990; Nassif, Engels 1993; Selva et al. 1995; Vulic, Lenski, Radman 1999; Eppley et al. 2007).

As mentioned above, the RDP program detects recombined regions based on significant changes in several sequence features of genomic segments, including the molecular

phylogenetic relationships, meaning that the real recombination boundaries, especially at homologous recombination initiation sites, may not be detected as part of the RDP-detected recombined regions due to the high level conservation required for homologous recombination initiation. Therefore, the UCSs next to the recombination boundary hotspots would have been involved in the initiation of homologous recombination events, even those observed between distantly related genomes, although they were not included in the RREs by the RDP program.

### 4.4.3 Modular exchange via homologous recombination

Homologous recombination events can take place or be initiated at virtually every location over the genomes as stochastic events if the genomes are closely related to each other, whereas such events can only take place at highly conserved sites if the genomes are distantly related. Since it is very unlikely that the highly conserved sites between different independently evolving genomes have remained unchanged by chance, the conserved sites, including the UCSs in the present study, may have undergone negative selection due to some functional constraints. Interestingly enough, many UCSs were found over the HAdV-D genome; however, the most significant UCSs reside in the vicinity of recombination boundary hotspots (Fig. 4.2). Therefore, it is probable that one of the functional constraints on these UCSs is homologous recombination. In this case, negative natural selection has acted on the conserved sites around the identified recombination boundary hotspots via positive effects of homologous recombination on the fitness of the recombinant viruses.

One of such possible positive effects is the drastic changes of viral antigenicity and/or pathogenicity by acquiring largely different sequences from distant viruses. This evolutionary process allows the viruses to undergo more rapid changes via the parallel accumulation of mutations and lateral transmission that combine mutations in different viruses instead of relying on the sequential accumulation of mutations via vertical transmission. Indeed, highly variable segments, e.g., the fiber and E3 gene regions, seem to have been frequently recombined as modules (Fig. 4.2). Through such positive effects of homologous recombination events, negative selection may have acted on the edges of the recombined regions. Furthermore, the positive selection on high variability in specific regions may have carved the recombination probability landscape while maintaining UCSs at a high level of recombination probability, inducing negative selection on their sequences. The genomes thus shaped should possess the property of exchanging specific genomic modules between distant genomes via homologous recombination events at UCSs. Such modularity of exchanged segments is indeed observed in Figure 4.2C. The modules seem to correspond to specific functional units, e.g., genes and

functional domains, in the genome; for example, the hexon loop 1 and loop 2 regions (Ebner, Pinsker, Lion 2005), the variable loop and RGD loop domain of the penton base, the protein products in the *E3* region and the fiber gene. This modular exchange property may have increased the success rate of generating functional recombinant virus genomes, even in cases in which highly diverged genomic segments are exchanged. In general, random recombination events between diverged genes may cause protein fold disruption (Lefeuvre et al. 2007). Similar observations have been made for distantly related lambdoid bacteriophages (Clark et al. 2001). It is also interesting that similar features have been found in the human MHC region, which is characterized by highly polymorphic loci and clusters of meiotic recombination hotspots, in contrast to the only limited levels of recombination observed in between the hotspots, thus resulting in contrasting effects on the linkage disequilibrium patterns (Jeffreys, Kauppi, Neumann 2001; Cullen et al. 2002; Lam et al. 2013) and exchanges of regions as haplotypes, e.g., human leukocyte antigens (HLAs). In relation to adenovirus, it is noteworthy that the E3 product known as E3:gp19K is in charge of (1) inhibiting the transport of HLA class I (HLA-I) to the cell surface, impeding the antigen presentation by retaining the major histocompatibility class I complex (MHC-I) in the endoplasmic reticulum, and (2) the evasion of natural killer cell (NK) recognition (Wold, Tollefson 1998; McSharry et al. 2008).

The UCSs may have served to produce different chimeric proteins, e.g., hexon proteins, some of which helped the recombinant virus to escape from the host defense systems, increasing the viral fitness. This hypothesis is consistent with the results showing that switching the knob of the fiber on oncolytic adenoviruses helps to avoid neutralization (Raki et al. 2011), and experiments demonstrating effective adenovirus neutralization have been carried out based on the synergistic effects of antibodies against the hexon, penton base and fiber, while changing at least one of these proteins in a re-infection partially delays the effective neutralization of the virus by the host immune system (Gahery-Segard et al. 1998).

## 4.5 Conclusion

This chapter presented the hypothesis that homologous recombination events have likely been initiated at UCSs, enabling the exchanges of specific less conserved genomic regions between even distant types of HAdV-D. The effects of positive selection on specific genomic regions may have shaped the modular structures as homologous recombination units, and homologous recombination of highly variable modules between distant types may have played an important evolutionary role in rapidly producing new types of HAdV-D, thereby causing frequent widespread epidemics of different types of diseases worldwide.

# Chapter 5: Interregional coevolution analysis revealing functional/structural interrelatedness between different genomic regions in *Human mastadenovirus D*

*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 (see chapter 4). 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 orthologs 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, an inter-regional coevolution analysis was conducted 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. The presented results predict additional unknown protein interactions.

## 5.1 Introduction

As mentioned earlier in Chapter 1, 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 (Harrach et al. 2011). 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 (Davison, Benko, Harrach 2003; Russell 2009). 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, forty-three, one, two and one for HAdV-A to G, respectively (Aoki et al. 2011; Matsushima et al. 2011; Robinson et al. 2011a; Walsh et al. 2011; Dehghan et al. 2012; Matsushima et al. 2013).

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 (Gonzalez et al. 2014) , as demonstrated in Chapter 4. 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 (Gonzalez et al. 2014). 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 (Wohl, Hearing 2008; Ma, Hearing 2011).

In this chapter, evolutionary correlations between different regions of the HAdV-D genomes were investigated 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.

## 5.2 Materials & Methods

### 5.2.1 Data

The available genome sequences of 43 HAdV-D types and two hybrid types (Kaneko et al. 2011a) were obtained from the International Nucleotide Sequence Database Collaboration (INSDC) (Table 1) and aligned into a single multiple alignment using the iterative refinement method algorithm (FFT-NS-I) of MAFFT (Katoh et al. 2002). 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 (Saitou, Nei 1987) under the Tamura-Nei model (TN93) (Tamura, Nei 1993).

### 5.2.2 Simulation of genome evolution

Forty-five artificial genomic sequences were generated by simulating sequence evolution using Mesquite version 2.75 (Shull et al. 2001). 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.

### 5.2.3 Recombination event analysis

Recombination events in the MGA were identified using following seven algorithms described in section 2.2 and available in the RDP 4.22b program (Heath et al. 2006): RDP (Martin, Rybicki 2000), GENECONV (Padidam, Sawyer, Fauquet 1999), Chimaera (Posada, Crandall 2001), MaxChi (Smith 1992), BootScan (Martin et al. 2005a), SiScan (Gibbs, Armstrong, Gibbs 2000) and 3Seq (Boni, Posada, Feldman 2007). The list of unique events that the RDP program produced was used by eliminating redundant events that were identified

by different algorithms (Martin, Rybicki 2000; Heath et al. 2006). Among these recombination events, this study 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 (Lefeuvre et al. 2007; Lefeuvre et al. 2009). These events were called 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 chapter 4.

### 5.2.4 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 (Mantel 1967). 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 (Lapointe, Garland 2001), 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 $p_{ij} = \left(d_{\max}^{\mathrm{pol}} - d_{ij}^{\mathrm{pol}}\right) \Big/ \sum_{k\neq l}^{N}\left(d_{\max}^{\mathrm{pol}} - d_{kl}^{\mathrm{pol}}\right)$, 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 (Liu, Naismith, Hay 2000; Gonzalez et al. 2014) to obtain the fundamental phylogeny of the genomes.

### 5.2.5 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.

### 5.2.6 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 $r_{w_x w_y}$.

### 5.2.7 Prediction of domains in membrane proteins

Transmembrane domains were predicted using the transmembrane hidden Markov model prediction TMHMM 2.0 program (Sonnhammer, von Heijne, Krogh 1998) with the predicted amino acid sequences of protein-coding regions in the *E3* region of the HAdV-8 genome (AB448767).

## 5.3. Results

### 5.3.1 Identification of intertypic recombination events

The study firstly aligned 45 different HAdV-D types, including two hybrids, and obtained a MGA of 33,645 gap-free sites. Then, the RDP program was applied (Heath et al. 2006) to the MGA and identified 195 reliable unique recombination events, indicating that each genome has experienced ca. 4.3 intertypic recombination events on average in the past. Figures 5.1A and 5.1B show the positions of the 195 recombined regions and the number of the recombination events detected at every position along the genome, respectively. Figure 5.1B looks similar to the results of the chapter 4 on recombination boundary hotspots (Figs. 4.2A and 4.2B). 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. 5.2) in which the mean and median of the sizes of the recombined regions were 2.5 kbp and 1.2 kbp, respectively.

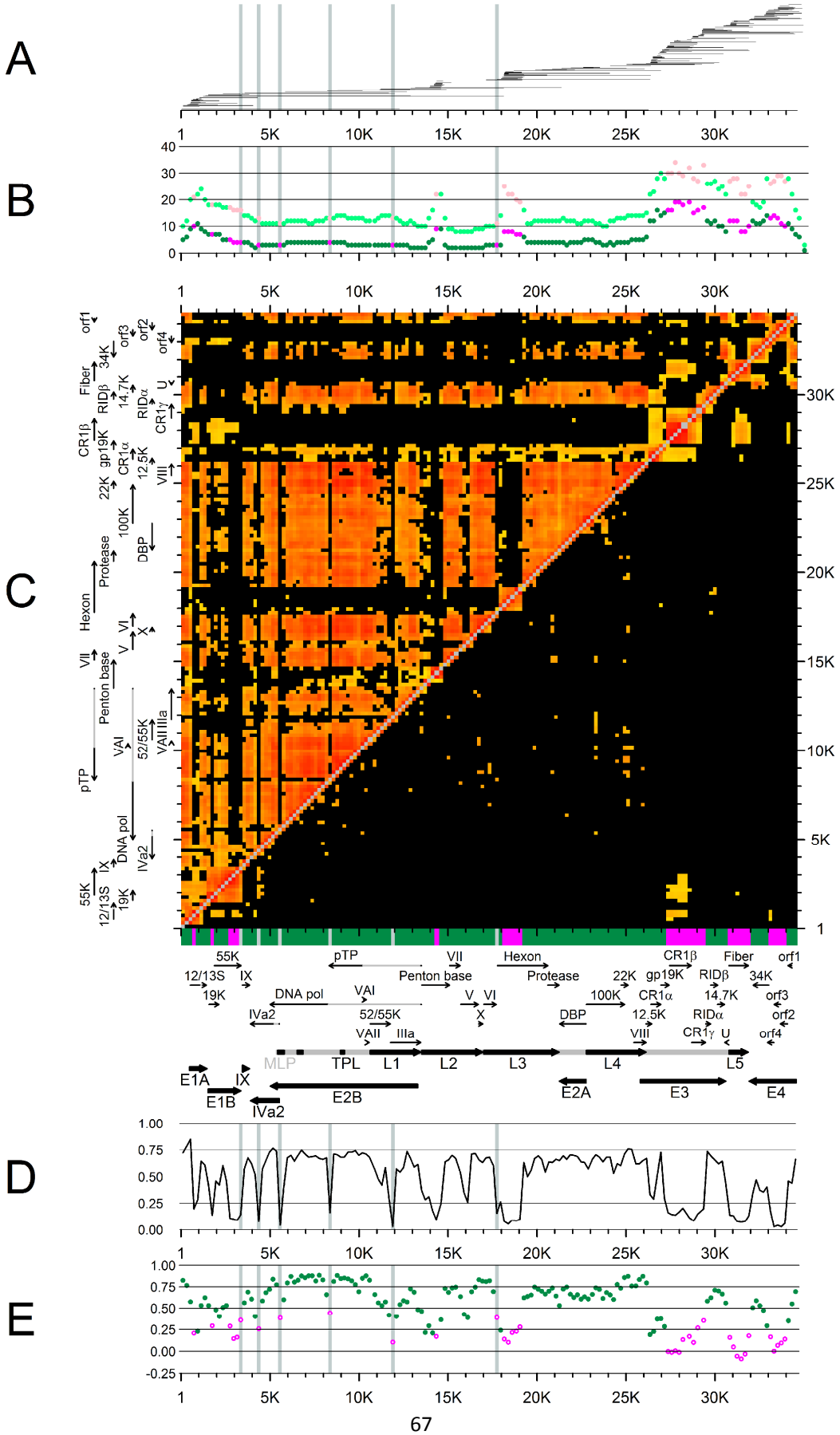
A
B
C
D
E
1
5K
10K
15K
20K
25K
30K
40
30
20
10
0
1.00
0.75
0.50
0.25
0.00
-0.25
orf1
orf3
orf2
orf4
34K
Fiber
14.7K
RIDβ
U
RIDα
CR1γ
CR1β
19K
gp19K
CR1α
12.5K
22K
VIII
100K
DBP
Protease
Hexon
VI
X
V
VII
Penton base
IIIa
52/55K
VAII
VAI
pTP
DNA pol
IVa2
IX
55K
12/13S
19K
MLP
TPL
L1
L2
L3
L4
L5
E1A
E1B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

**Fig. 5.1 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 3). 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.  *Journal of Virology*, 2015, doi:10.1128/JVI.00515-15.

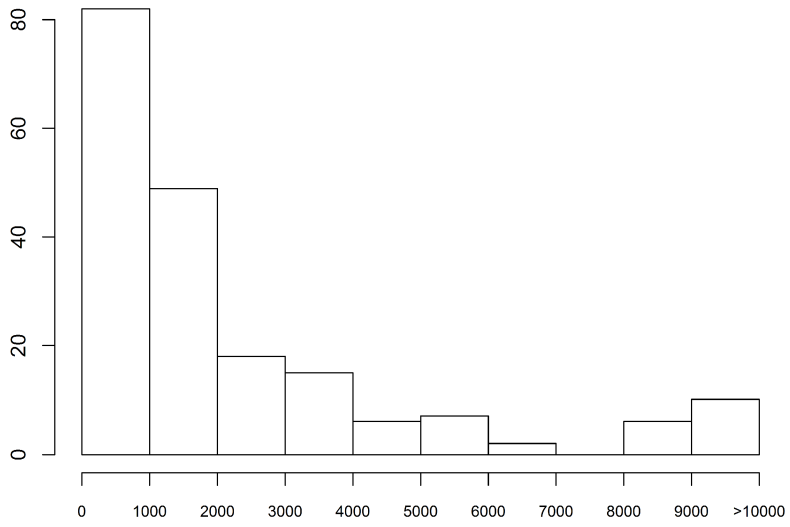


**Fig. 5.2 Distribution of recombined segments lengths.** The abscissa shows the different lengths while the ordinate shows the frequency by each size category. Lengths of the recombined segments were adjusted to match Fig. 5.1A.  *Journal of Virology*, 2015, doi:10.1128/JVI.00515-15.

### 5.3.2 Identification of coevolving genomic regions

The biased distribution of short recombined regions to specific genomic positions, including recombination boundary hotspots (refer to Chapter 4), 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, this analysis sought to identify significantly high evolutionary correlations between different genomic regions at $p < 0.0025$ (the upper left triangle of Fig. 5.1C; hereafter, this matrix is called 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. 5.3). However, as shown in Figure 5.1C, 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 5.1A and 5.1B. Therefore, it was concluded that, while consistently coevolving major genomic regions have escaped most of the recombination events, the other regions have been reshuffled between different types.

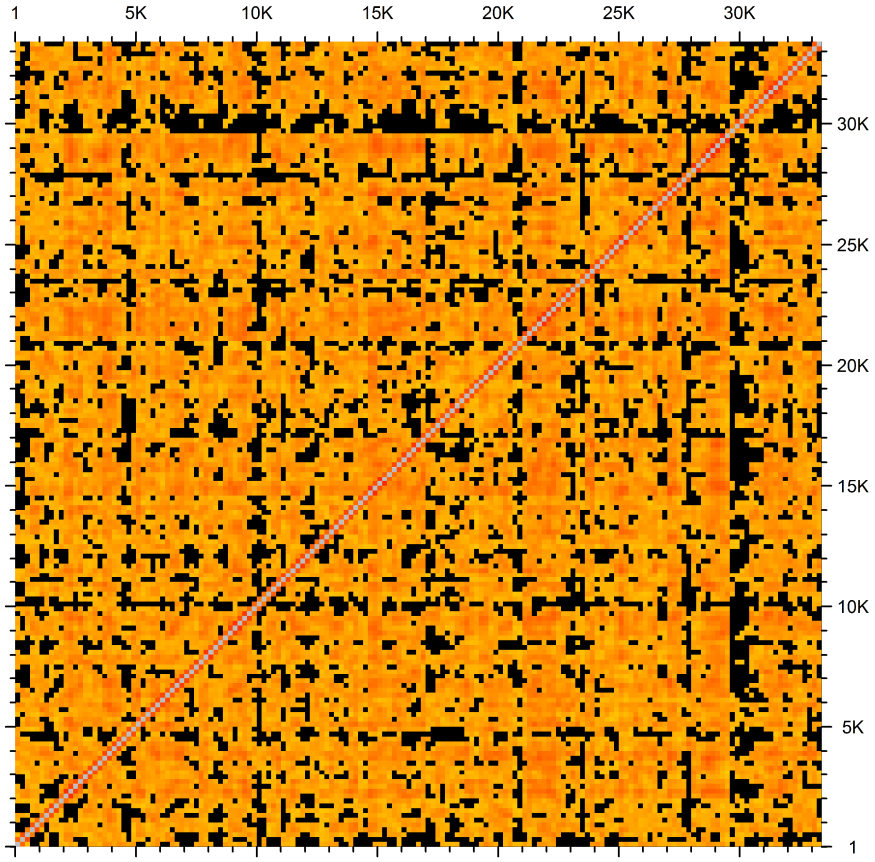

**Fig. 5.3 Evolutionary correlation on simulated sequence.** The simulated multiple genome alignment of 45 artificial genomic sequences were generated by simulating sequence evolution using Mesquite version 2.75 under the following conditions: the tree topology = the genome tree; the number of characters = 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.  *Journal of Virology*, 2015, doi:10.1128/JVI.00515-15.

In order to summarize the evolutionary heterogeneity in the coevolution matrix of the HAdV-D genomes, the ratio of the windows showing a significantly high correlation to each specific window was calculated, named the coevolution ratio in this study, along the genome (Fig. 5.1D) and made a histogram of the ratios (Fig. 5.4). 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. 5.1D), which match the positions of the major coevolving regions. Since these regions seem to be the rarest recombination regions and hence coevolving regions, hereafter they are called 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. 5.1C). Note that a few windows do not show significant correlations with any or almost any of the other windows (grayed regions in Fig. 5.1) due to 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. These were called invariant regions. As shown in Figure 5.4, the coevolution ratio distribution of the simulated genomes is strongly biased to 1.0, as expected.

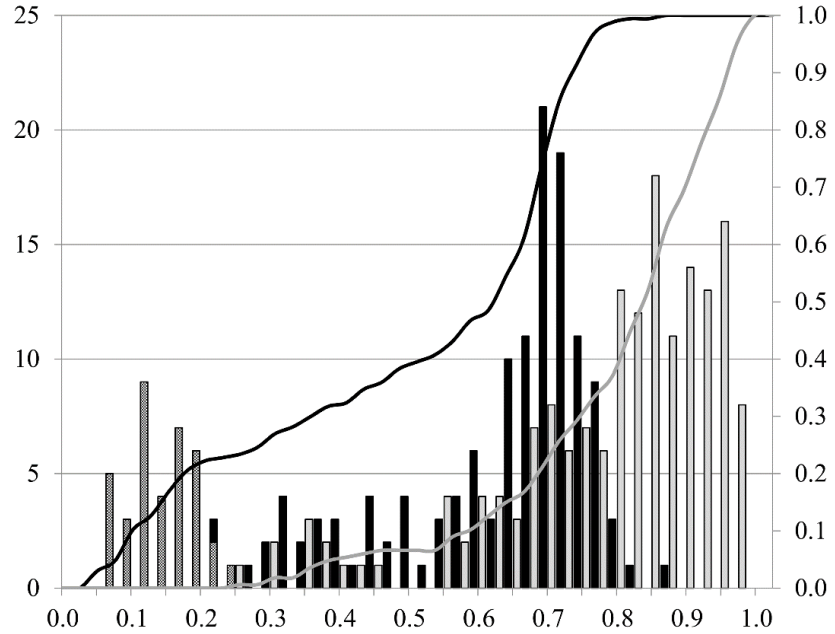


**Fig. 5.4 Histogram of significant correlation ratios.** The abscissa is the ratio of the number of windows that show a significant correlation coefficient to a specific window against the total number of windows (=167). The left and right ordinates are for the absolute frequencies (bars) and relative 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. Copyright © American Society for Microbiology, *Journal of Virology*, 2015, doi:10.1128/JVI.00515-15.

### 5.3.3 Defining basal and non-basal genomic regions

In order to operationally identify the basal regions, basal regions were defined as regions having coevolved with the DNA polymerase-coding region. The DNA polymerase-coding region was chosen as a reference simply because it has been recognized to be one of the rarest recombination regions in the HAdV-D genomes (Liu, Naismith, Hay 2000; Gonzalez et al. 2014). It was 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 5.1D (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. 5.1E). Based on this operational definition, 130 of 167 sliding windows (77.8%) were determined to be basal windows (Table 3). It was confirmed that the thus identified basal regions corresponded to the distribution of the major peak of the coevolution ratio histogram (Fig. 5.4). 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. 5.1D) had correlation coefficients to the DNA polymerase region with a *p*-value of < 0.0025 (Fig. 5.1E).

## Table 3[+]. 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 | (Brenkman, Breure, van der Vliet 2002; De Jong, Van Der Vliet, Brenkman 2003) |
| | | | | | ***E1A*:12/13S** | 564 | 1113 | Host modulation | Promote p53-dependent apoptosis. Promotion of viral transcriptional activation | (Zheng 2010) |
| Non-basal | 601 | 800 | 626 | 831 | | | | | | |
| Basal | 801 | 1600 | 832 | 1648 | | 1209 | 1420 | | | |
| | | | | | ***E1B*:19K** | 1572 | 2120 | Host modulation | Suppresses both p53-dependent and -independent apoptosis induced during adenovirus infection | (Berk 2005) |
| Non-basal | 1601 | 1800 | 1649 | 1878 | | | | | | |
| | | | | | ***E1B*:55K** | 1877 | 3364 | Host modulation | Inactivate p53 and p53-dependent apoptosis | (Berk 2005) |
| 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 | (Davison, Benko, Harrach 2003; Russell 2009; Liu et al. 2010) |
| | | | | | ***IVa2*:IVa2** | 5233 | 3900 | Core protein/Genome Packaging | Genome packaging process | (Russell 2009) |
| Invariant | 4201 | 4400 | 4268 | 4467 | | | | | | |
| Basal | 4401 | 5400 | 4468 | 5467 | | | | | | |
| | | | | | ***E2B*:DNA pol** | 8278 | 5003 | DNA replication | Replication of the viral DNA | (Brenkman, Breure, van der Vliet 2002; De Jong, Van Der Vliet, Brenkman 2003) |
| 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 | (De Jong, Van Der Vliet, Brenkman 2003) |
| Basal | 8401 | 11600 | 8474 | 11797 | | | | | | |
| | | | | | ***L1*:52/55K** | 10632 | 11750 | Genome Packaging | Virus assembly by interacting with pVII | (Russell 2009) |
| | | | | | ***L1*:IIIa** | 11773 | 13470 | Structural/Genome Packaging | Minor structural protein | (Davison, Benko, Harrach 2003; Russell 2009) |
| 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 | (Russell 2009) |
| 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 | (Russell 2009) |
| | | | | | ***L2*:V** | 15713 | 16696 | Core protein/Maturation | Suggested as mediator of viral DNA transport to the nucleus of the infected cell | (Russell 2009) |
| | | | | | ***L2*:X** | 16726 | 16950 | Core protein/Maturation | Suggested as mediator of viral prechromatin condensation and facilitate packaging of the core complex | (Russell 2009) |
| | | | | | ***L3*:VI** | 17006 | 17707 | Structural/Maturation | Minor structural protein, activation of protease | (Davison, Benko, Harrach 2003; Russell 2009; Liu et al. 2010) |
| Invariant | 17201 | 17400 | 17643 | 17852 | | | | | | |
| | | | | | ***L3*:Hexon** | 17776 | 20604 | Major capsid protein | Major structural protein | (Davison, Benko, Harrach 2003; Russell 2009; Liu et al. 2010) |
| Basal | 17401 | 17600 | 17853 | 18052 | | | | | | |



[a] NC stands for non-coding region

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

## Table 3. List of basal and non-basal regions with overlapping functional regions (continued)

| 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 |
| 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 | (Matthews, Russell 1995; Perez-Berna et al. 2012) |
| | | | | | ***E2A*:DBP** | 22755 | 21283 | DNA replication | Binds single-stranded DNA displaced during genome replication and is required for initiation and elongation of replication | (De Jong, Van Der Vliet, Brenkman 2003) |
| | | | | | ***L4*:100K** | 22772 | 24919 | Genome Packaging | Viral genome packaging | (Wohl, Hearing 2008; Ma, Hearing 2011) |
| | | | | | ***L4*:22K** | 24732 | 25115 | Genome Packaging | Viral genome packaging by interacting with IVa2 | (Russell 2009) |
| | | | | | ***L4*:VIII** | 25444 | 26127 | Structural/Maturation | Minor structural protein | (Davison, Benko, Harrach 2003; Russell 2009; Liu et al. 2010) |
| | | | | | ***E3*:12.5K** | 26128 | 26448 | Host modulation | | |
| | | | | | ***E3*:CR1α** | 26402 | 26956 | Host modulation | Down modulation of TRAIL receptors | (Benedict et al. 2001) |
| | | | | | ***E3*:gp19K** | 26953 | 27426 | Host modulation | Inhibition T-cell recognition and NK cell activation | (McSharry et al. 2008) |
| 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 | (Benedict et al. 2001) |
| | | | | | ***E3*:RIDβ** | 29757 | 30146 | Host modulation | Evasion of TNF-apoptosis | (Benedict et al. 2001) |
| | | | | | ***E3*:14.7K** | 30139 | 30531 | Host modulation | Evasion of TNF-apoptosis | (Chinnadurai 1998; Schneider-Brachert et al. 2006) |
| Non-basal | 29601 | 30800 | 30729 | 32020 | ***L5*:Fiber** | 30785 | 31873 | Major capsid protein | Major structural protein, binding to cellular receptor, tissue affinity | (Liu et al. 2010) |
| Basal | 30801 | 31800 | 32021 | 33020 | ***E4*:34K[b]** | 33027 | 32149 | Host modulation | Inactivate p53 and p53-dependent apoptosis. Viral late gene expression | (Chinnadurai 1998; Boyer, Ketner 2000) |
| Non-basal | 31801 | 32800 | 33021 | 34024 | ***E4*:orf4** | 33319 | 32957 | Host modulation | Lysis of infected cell | (Chinnadurai 1998; Miron et al. 2005) |
| | | | | | ***E4*:orf3** | 33675 | 33322 | Host modulation | Promotion of viral gene expression and replication | (Leppard 1997) |
| | | | | | ***E4*:orf2** | 34064 | 33672 | Host modulation | *Not fully characterized* | |
| Basal | 32801 | 33401 | 34025 | 34633 | | | | | | |
| | | | | | ***E4*:orf1** | 34302 | 34105 | Host modulation | Lytic infection and oncogenesis | (Leppard 1997) |
| | | | | | **NC** | | | DNA replication | Starting point for replication | (Brenkman, Breure, van der Vliet 2002; De Jong, Van Der Vliet, Brenkman 2003) |

[a] NC stands for non-coding region

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

In order to view the regional coevolution, rather than the basal one, partial correlation coefficients were calculated adjusted for the basal regions (see Materials and Methods). As shown in the lower right half of Figure 5.1C (named a non-basal coevolution matrix in this study), certain numbers of region pairs showed significantly high partial correlation coefficients (Table 4 at the end of this chapter), 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 5.1C. Interestingly, correlations between separated non-basal regions were also observed (see below).

## 5.4. Discussion

In the present chapter, coevolving regions were identified 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, these regions were regarded 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 (section 4.4.1) and that such recombination hotspots intervene between basal regions. The presented partial correlation analysis also showed that continuous non-basal regions are autocorrelated (Fig. 5.1C), evidencing the modularity of recombination, i.e., specific genomic segments have been recombined as modules, as suggested in section 4.4.3. 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 (section 4.4.2), does not generally specify the direction of recombination from the recombination initiation site along the chromosome (Krejci et al. 2012), 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 (section 4.4.2), 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 strains recombined in 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 (Wohl, Hearing 2008; Ma, Hearing 2011), 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 3) and physically interacting protein genes: DNA polymerase, pTP and 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 (Wohl, Hearing 2008; Ma, Hearing 2011); IX, (IIIa,) VI and VIII as minor structural proteins for the interior and exterior of the virion nucleocapsid (Davison, Benko, Harrach 2003; Russell 2009; Liu et al. 2010); (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 (Russell 2009); the protease for making IIIa, VI, VII, VIII, TP and X mature (Matthews, Russell 1995; Perez-Berna et al. 2012); CR1α, RIDα, RIDβ and 14.7K for evading host cell apoptosis by blocking the host's TNF-R1, TRAIL-R1/-R2 and/or Fas (Benedict et al. 2001; Tollefson et al. 2001; Schneider-Brachert et al. 2006); and 34K (*E4*:orf6), which interacts with *E1B*:55K to perform a variety of functions (Leppard 1997; Rubenwolf et al. 1997; Chinnadurai 1998; Boyer, Ketner 2000; Querido et al. 2001; Berk 2005; Blackford, Grand 2009; Kato, Huang, Flint 2011). Inverted terminal regions (ITRs) are known to contain replication origins where the DNA polymerase and TP bind together and initiate DNA replication (Webster et al. 1997; De Jong, Van Der Vliet, Brenkman 2003) and are also known to form the "panhandle structure" of the intra-molecular double helix for replication (Lippe, Graham 1989). DBP binds the viral DNA

and enhances the replication initiation by DNA polymerase and pTP at the ITRs (Brenkman, Breure, van der Vliet 2002; De Jong, Van Der Vliet, Brenkman 2003) (Fig. 1.2). 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, on the time of this study it did not succeed in finding any reports of their specific functions, except for a report about the oncogenic activity of *E4*:orf1 of HAdV-9 in rats (Javier 1994).

The remaining basal regions not yet mentioned above are mapped together with non-basal regions on the coding regions of 12/13S (it was called the *E1A* protein as well in accordance with previous works, e.g., reference (Berk 2005)), *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 (Davison, Benko, Harrach 2003; Wohl, Hearing 2008; Russell 2009; Liu et al. 2010; Ma, Hearing 2011). While the macroscopic coevolution analysis showed that the whole coding region of fiber, the remaining major capsid protein, is a non-basal region (Fig. 5.1C), 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. 5.5A) for anchoring the fiber in the penton base complex (Davison, Benko, Harrach 2003; Russell 2009; Liu et al. 2010). 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. 5.5B). 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 (Benedict et al. 2001; Tollefson et al. 2001; Schneider-Brachert et al. 2006). The coding regions of RIDα and RIDβ are basal regions. These imply that the short basal region of CR1α may interact with RIDβ. There is no information about any interaction of CR1β with other adenoviral proteins.

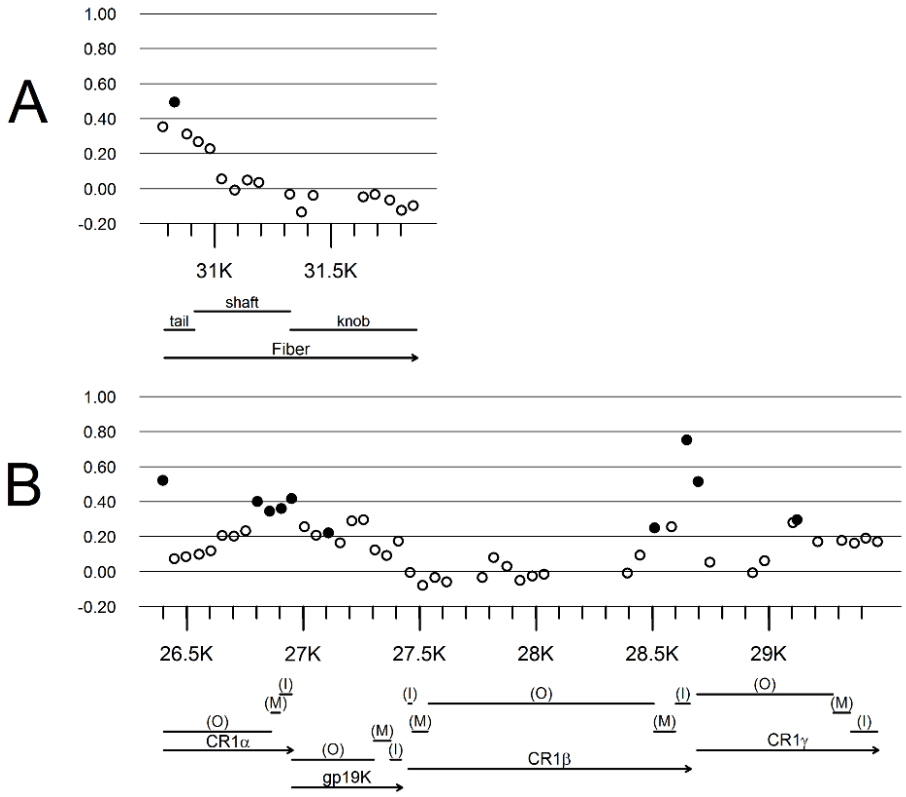


**Fig. 5.5 Finer-scale coevolution analysis.** The results of the finer scale analyses are depicted for two highlighted regions, (A) fiber and (B) E3 region. The abscissae show the position in the HAdV-8 genome (AB448767). The left ordinates represent the correlation coefficients between each 100 bp-window 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. Copyright © American Society for Microbiology, *Journal of Virology*, 2015, doi:10.1128/JVI.00515-15.

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. 5.1C), implying the presence of conserved important functions in these regions. Eight variant non-basal regions (they are called 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 3): 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 (Chinnadurai 1998; Miron et al. 2005; McSharry et al. 2008; Zheng 2010). 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 (Gahery-Segard et al. 1998), 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 5.5B 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 unit/relatedness. However, several cases of non-basal coevolution were found between separate genomic regions, even between distant regions, although it is difficult 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. 5.1C), except between *E1A* proteins and the fiber (Table 4 at the end of this chapter). 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 (Harada et al. 2002), 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. 5.1A). 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, e.g. comparing the clusters or the branches lengths, did not, however, produce so consistent results over the genome as the results of this chapter (data not shown). This is partly due to the loss of information that occurs when constructing phylogenetic trees from the distance matrices, which were 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 (Bille 2005), the biological significance of the results of tree topology comparisons using these algorithms is not necessarily evident. Therefore, it was

decided to directly use distances to evaluate the evolutionary correlations between different genomic regions in this study.

This 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 (Lee, Redfern, Orengo 2007). However, the 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.

**Table 4[+]. Pairs of significantly partially correlated coding regions[a]**

| Protein 1 | Windows | Protein 2 | Windows |
|---|---|---|---|
| **12/13S** | 626-1032 | **19K** | 1649-2249 |
| **12/13S** | 626-832 | **55K** | 2249-3452 |
| **12/13S** | 832-1032 | **55K** | 2249-3249 |
| **12/13S** | 1243-1443 | **IX** | 3655-3861 |
| **12/13S** | 626-1032 | **IVa2** | 4068-4268 |
| **12/13S** | 626-1032 | **gp19K** | 27260-27464 |
| **12/13S** | 626-1032 | **CR1ß** | 27464-28640 |
| **12/13S** | 626-1032 | **34K** | 33021-33221 |
| **19K** | 1649-1849 | **55K** | 2249-3452 |
| **19K** | 1849-2049 | **55K** | 2449-3452 |
| **19K** | 2049-2249 | **55K** | 2649-3452 |
| **19K** | 1649-1849 | **IVa2** | 4268-4468 |
| **19K** | 1849-2049 | **IVa2** | 4268-4668 |
| **19K** | 2049-2249 | **IVa2** | 4268-4468 |
| **19K** | 1649-2049 | **gp19K** | 27260-27464 |
| **19K** | 1649-2049 | **CR1ß** | 27464-28460 |
| **19K** | 2049-2249 | **CR1ß** | 27464-28698 |
| **19K** | 1849-2049 | **Fiber** | 31141-31374 |
| **55K** | 2249-2449 | **IVa2** | 4268-4668 |
| **55K** | 2649-3049 | **IVa2** | 4268-4468 |
| **55K** | 3049-3452 | **IVa2** | 4268-4668 |
| **55K** | 2249-2449 | **gp19K** | 27260-27464 |
| **55K** | 2249-2449 | **CR1ß** | 27464-28460 |
| **55K** | 2449-2649 | **CR1ß** | 27882-28460 |
| **55K** | 2649-3049 | **CR1ß** | 27882-28460 |
| **IVa2** | 4068-4268 | **DNA pol** | 8068-8268 |
| **IVa2** | 4668-4868 | **DNA pol** | 7868-8068 |
| **DNA pol** | 5268-5468 | **pTP** | 10086-10287 |
| **DNA pol** | 5668-5868 | **pTP** | 9074-9283 |
| **DNA pol** | 6868-7068 | **pTP** | 8474-8874 |
| **DNA pol** | 7068-7268 | **pTP** | 8674-8874 |
| **DNA pol** | 7668-8068 | **pTP** | 8674-8874 |
| **DNA pol** | 7668-7868 | **IIIa** | 12798-12998 |
| **DNA pol** | 8068-8268 | **VII** | 15597-15804 |
| **DNA pol** | 7868-8068 | **X** | 16838-17038 |
| **DNA pol** | 6868-7068 | **VI** | 17238-17443 |
| **DNA pol** | 8268-8474 | **VI** | 17643-17853 |
| **DNA pol** | 6268-6468 | **DBP** | 21570-21770 |
| **DNA pol** | 5268-5468 | **100K** | 23206-23406 |
| **DNA pol** | 6868-7268 | **22K** | 25021-25224 |
| **DNA pol** | 6868-7068 | **VIII** | 25848-26048 |
| **pTP** | 9074-9483 | **52/55K** | 10783-10983 |
| **pTP** | 9483-9686 | **Penton base** | 14927-15128 |
| **pTP** | 8674-8874 | **V** | 16638-16838 |
| **pTP** | 8674-8874 | **X** | 16838-17038 |
| **pTP** | 8674-9283 | **VI** | 17238-17443 |
| **pTP** | 10086-10287 | **Hexon** | 19765-19965 |
| **pTP** | 8674-8874 | **Protease** | 20765-20965 |
| **pTP** | 8674-8874 | **100K** | 24818-25021 |
| **pTP** | 9283-9483 | **100K** | 24406-24606 |
| **pTP** | 9283-9483 | **100K** | 24818-25021 |
| **pTP** | 9483-9686 | **100K** | 23206-23406 |
| **pTP** | 9483-9686 | **100K** | 24818-25021 |
| **pTP** | 10086-10287 | **100K** | 23206-23406 |
| **pTP** | 10086-10287 | **100K** | 24818-25021 |
| **pTP** | 8474-8874 | **22K** | 25021-25224 |
| **pTP** | 9283-9686 | **22K** | 25021-25224 |
| **pTP** | 10086-10287 | **22K** | 25021-25224 |
| **pTP** | 8874-9074 | **VIII** | 25848-26048 |
| **VAI** | 10287-10489 | **V** | 16638-16838 |
| **VAII** | 10489-10783 | **V** | 16638-16838 |
| **52/55K** | 10983-11383 | **IIIa** | 13410-13619 |
| **52/55K** | 10783-10983 | **100K** | 24406-24606 |
| **52/55K** | 10983-11383 | **100K** | 23206-23406 |
| **52/55K** | 10983-11383 | **100K** | 24818-25021 |
| **52/55K** | 11383-11798 | **100K** | 23206-23406 |
| **52/55K** | 10983-11183 | **22K** | 25021-25224 |



[a] Positions relative to HAdV-8 (AB448767)

**Table 4. Pairs of significantly partially correlated coding regions**[a] (continued)

| Protein 1 | Windows | Protein 2 | Windows |
|---|---|---|---|
| **IIIa** | 11798-11998 | **Penton base** | 14927-15128 |
| **IIIa** | 12998-13410 | **V** | 16638-16838 |
| **IIIa** | 11798-11998 | **Protease** | 20965-21165 |
| **IIIa** | 11798-12198 | **100K** | 23206-23406 |
| **Penton base** | 14927-15128 | **Protease** | 20765-21165 |
| **Penton base** | 14927-15128 | **100K** | 23206-23406 |
| **Penton base** | 14927-15128 | **100K** | 24818-25021 |
| **VII** | 15128-15334 | **Protease** | 20765-21165 |
| **VII** | 15128-15334 | **100K** | 23206-23406 |
| **V** | 16638-16838 | **VI** | 17238-17443 |
| **V** | 16638-16838 | **Protease** | 20765-20965 |
| **V** | 16638-16838 | **22K** | 25021-25224 |
| **X** | 16838-17038 | **Protease** | 20765-20965 |
| **X** | 16838-17038 | **22K** | 25021-25224 |
| **VI** | 17038-17238 | **Hexon** | 19565-19765 |
| **VI** | 17238-17643 | **Protease** | 20765-20965 |
| **VI** | 17643-17853 | **RIDα** | 29668-29888 |
| **Hexon** | 19565-19765 | **100K** | 24818-25021 |
| **Hexon** | 19565-19965 | **22K** | 25021-25224 |
| **Hexon** | 18053-18284 | **CR1α** | 26857-27060 |
| **Hexon** | 18739-18939 | **RIDα** | 29668-29888 |
| **Hexon** | 18739-19165 | **RIDß** | 29888-30101 |
| **Protease** | 20765-20965 | **100K** | 24606-24818 |
| **Protease** | 20765-20965 | **22K** | 25021-25224 |
| **DBP** | 21570-21770 | **100K** | 23806-24206 |
| **DBP** | 21770-21970 | **100K** | 23806-24006 |
| **DBP** | 22170-22370 | **100K** | 22784-23206 |
| **DBP** | 22170-22784 | **100K** | 23606-24006 |
| **12.5K** | 26248-26448 | **CR1α** | 26857-27060 |
| **12.5K** | 26248-26654 | **gp19K** | 27060-27464 |
| **12.5K** | 26248-26448 | **CR1ß** | 27464-28460 |
| **12.5K** | 26448-26654 | **CR1ß** | 27464-28698 |
| **12.5K** | 26248-26654 | **CR1γ** | 28698-28933 |
| **12.5K** | 26248-26654 | **Fiber** | 31374-31589 |
| **CR1α** | 26654-26857 | **gp19K** | 27260-28698 |
| **CR1α** | 26857-27060 | **CR1ß** | 27464-28698 |
| **CR1α** | 26654-27060 | **CR1γ** | 28698-28933 |
| **gp19K** | 27260-27464 | **CR1ß** | 27882-28698 |
| **gp19K** | 27260-27464 | **CR1γ** | 28698-29265 |
| **gp19K** | 27060-27260 | **Fiber** | 30931-31141 |
| **gp19K** | 27260-27464 | **Fiber** | 30931-32021 |
| **gp19K** | 27260-27464 | **orf2** | 34025-34225 |
| **CR1ß** | 27464-28094 | **CR1γ** | 28698-29265 |
| **CR1ß** | 28094-28460 | **CR1γ** | 28933-29265 |
| **CR1ß** | 27464-28460 | **Fiber** | 30931-32021 |
| **CR1ß** | 28460-28698 | **Fiber** | 31141-31804 |
| **CR1ß** | 27464-28460 | **orf2** | 34025-34225 |
| **CR1γ** | 28933-29265 | **RIDα** | 29668-29888 |
| **CR1γ** | 28698-28933 | **Fiber** | 31141-31804 |
| **CR1γ** | 28933-29265 | **Fiber** | 31141-31374 |
| **14.7K** | 30304-30504 | **34K** | 32221-32421 |
| **14.7K** | 30304-30504 | **34K** | 32621-32821 |
| **14.7K** | 30504-30729 | **34K** | 32221-33021 |
| **14.7K** | 30504-30729 | **orf2** | 34025-34225 |
| **14.7K** | 30304-30729 | **orf1** | 34225-34427 |
| **Fiber** | 30931-31141 | **34K** | 32221-32821 |
| **Fiber** | 31804-32021 | **34K** | 32421-33021 |
| **Fiber** | 31141-31374 | **orf3** | 33622-33825 |
| **Fiber** | 30931-31141 | **orf2** | 34025-34225 |
| **Fiber** | 31141-31589 | **orf2** | 33825-34225 |
| **Fiber** | 31589-32021 | **orf2** | 34025-34225 |
| **Fiber** | 30931-31141 | **orf1** | 34225-34427 |
| **Fiber** | 31804-32021 | **orf1** | 34225-34427 |
| **34K** | 33021-33221 | **orf3** | 33622-33825 |
| **34K** | 32221-33021 | **orf2** | 34025-34225 |
| **34K** | 33021-33221 | **orf2** | 33825-34225 |
| **34K** | 32221-33221 | **orf1** | 34225-34427 |
| **orf4** | 33221-33422 | **orf2** | 33825-34225 |
| **orf3** | 33422-33622 | **orf2** | 34025-34225 |

[a] Positions relative to HAdV-8 (AB448767)

# Chapter 6: Conclusion

This thesis brings evidence to the scientific knowledge supporting an explanation to the increasing type divergence of HAdV-D by a recombination mechanism that has facilitated the exchange of evolutionary modules among the diversified types resulting in novel infectious recombinant products. Moreover, these modules were demonstrated to be independent from the rest of the genome despite conserving a strong correlation inside them and in some cases between them. Despite their different functional roles, these independent modules shared the characteristic of viral interactions with the host such as immune reaction modulation or the interaction with the cellular receptors at the target tissues for initiating the infection. The variability characterized in these regions and their direct involvement determining the host immune reaction make them the main candidates for detection of the different varieties of types in the different species. On the other hand, other genomic regions were considered as a set of basal genes that coded the functions for the support of the replication cycle of the adenoviral genome and assembly of the virion capsids; these regions showed lowered variability between types and lower frequency of recombination events affecting them.

The exposed results account for a proper evolutionary mechanism model that balances the divergence and conservation between adenoviral types. Such a model correctly explains the location of the recombination boundaries in hotspots near highly diverged regions. Moreover, a great part of the characterized divergence between types occurred particularly in regions with functions related to direct virus-host interactions such as epitopes, e.g. hexon and fiber, which were the main target of the early classification system that relied on the reaction of the host immune system for the type determination, i.e. serotyping. Therefore, the recombination events that shuffled these regions, combined with their independent evolution from the rest of the genome allowed the observed divergence in the candidates to new adenoviral types. On the other hand, the conservation observed in other regions is presumably the result of higher selective pressure against changes affecting the interactions between proteins coded in basal regions, i.e. mutations and recombinations in these regions have higher chance of reducing the viral fitness. These regions are presumably affected deleteriously by recombination events that disrupt important functional pathways necessary in specific stages of the virus replication or maturation.

The exposed mechanisms supporting the frequent recombination are different from other mechanisms hypothesized for frequent recombination of adenoviruses introduced in Chapter 1 because the presented mechanism of UCSs and modular exchange can be assumed to be the result of progressive development over time and the evidence can be assessed as shown in Chapter 4 and 5, respectively. More specifically, although the results allow for recombination events happening in other regions outside the recombination boundary hotspots, the exposed scientific model assigns to such events a lower probability of success due to putatively deleterious effects over the basal sections of the adenoviral genome. On the other hand, other available hypotheses require the evolution of particular nucleotide composition to take the function of recombination hotspot boundaries. Also, such hypotheses barely explained the reason for developing those nucleotide patterns exclusively in the recombination hotspots.

The model proposed in the previous chapters provides answers to the questions as to how the evolutionary mechanisms of UCSs can explain the particular location of the recombination hotspots that enabled (1) the frequent exchange of diverged genetic material (2) without disrupting the adenoviral evolutionary functional pathways of basal regions, which were the two main objectives of this study. Additionally, this model creates a framework that allows further questioning about how to explain the localization of apparent hotspots of divergence matching the recombined modules.

The localized divergence could be just partly or completely attributed to the effects introduced by the putatively frequent recombination processes. This hypothesis is dismissed when considering the fact that the accepted model of homologous recombination events suggests the mutational effects change the recombination boundaries rather than the recombined region between them. One plausible alternative involves the independence exhibited by these modules from the rest of the genome. Then, such independence could be the results of a relatively relaxation in the selective pressures over these modules reflected in a relatively faster accumulation of divergence on them than the one observed in other genomic regions such as the basal regions (Chapter 5). This observation explains also the variability detected by the machine learning approach and used to predict the types based on differences on these non-basal regions. However, this alternative hypothesis still requires rigorous testing under the neutral theory of molecular evolution for its confirmation and it constitutes one of the future works to be developed as continuation of the presented work.

# iv. References


Adhikary, AK, N Hanaoka, T Fujimoto. 2014. Simple and cost-effective restriction endonuclease analysis of Human Adenoviruses. Biomed Research International 2014:Article ID 363790.

Altschul, SF, TL Madden, AA Schaffer, JH Zhang, Z Zhang, W Miller, DJ Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389-3402.

Aoki, K, M Benko, AJ Davison, M Echavarria, DD Erdman, B Harrach, AE Kajon, D Schnurr, G Wadell, C Members of the Adenovirus Research. 2011. Toward an integrated human adenovirus designation system that utilizes molecular and serological data and serves both clinical and fundamental virology. Journal of Virology 85:5703-5704.

Aoki, K, H Ishiko, T Konno, Y Shimada, A Hayashi, H Kaneko, T Ohguchi, Y Tagawa, S Ohno, S Yamazaki. 2008. Epidemic keratoconjunctivitis due to the novel hexon-chimeric-intermediate 22,37/H8 human adenovirus. Journal of Clinical Microbiology 46:3259-3269.

Ariga, T, Y Shimada, K Shiratori, et al. 2005. Five new genome types of adenovirus type 37 caused epidemic keratoconjunctivitis in Sapporo, Japan, for more than 10 years. Journal of Clinical Microbiology 43:726-732.

Arnold, J, M Janoska, AE Kajon, D Metzgar, NR Hudson, S Torres, B Harrach, D Seto, J Chodosh, MS Jones. 2010. Genomic characterization of human adenovirus 36, a putative obesity agent. Virus Research 149:152-161.

Bailis, AM, R Rothstein. 1990. A defect in mismatch repair in Saccharomyces cerevisiae stimulates ectopic recombination between homeologous genes by an excision repair dependent process. Genetics 126:535-547.

Benedict, CA, PS Norris, TI Prigozy, JL Bodmer, JA Mahr, CT Garnett, F Martinon, J Tschopp, LR Gooding, CF Ware. 2001. Three adenovirus E3 proteins cooperate to evade apoptosis by tumor necrosis factor-related apoptosis-inducing ligand receptor-1 and-2. Journal of Biological Chemistry 276:3270-3278.

Benkő, M. 2008. Adenoviruses: Pathogenesis. In: BWJM Editors-in-Chief: , MHVv Regenmortel, editors. Encyclopedia of Virology (Third Edition). Oxford: Academic Press. p. 24-29.

Berk, AJ. 2005. Recent lessons in gene expression, cell cycle control, and cell biology from adenovirus. Oncogene 24:7673-7685.

Bille, P. 2005. A survey on tree edit distance and related problems. Theoretical Computer Science 337:217-239.

Bishop, CM. 2007. Pattern Recognition and Machine Learning (Information Science and Statistics): Springer, New York. p. 326-349.

Blackford, AN, RJA Grand. 2009. Adenovirus E1B 55-Kilodalton protein: multiple roles in viral infection and cell transformation. Journal of Virology 83:4000-4012.

Boni, MF, D Posada, MW Feldman. 2007. An exact nonparametric method for inferring mosaic structure in sequence triplets. Genetics 176:1035-1047.

Boyer, JL, G Ketner. 2000. Genetic analysis of a potential zinc-binding domain of the adenovirus E4 34k protein. Journal of Biological Chemistry 275:14969-14978.

Brenkman, AB, EC Breure, PC van der Vliet. 2002. Molecular architecture of adenovirus DNA polymerase and location of the protein primer. Journal of Virology 76:8200-8207.

Centers for Disease, C, Prevention. 2013. Adenovirus-associated epidemic keratoconjunctivitis outbreaks - four States, 2008-2010. MMWR Morb Mortal Wkly Rep 62:637-641.

Charif, D, J Thioulouse, JR Lobry, G Perriere. 2005. Online synonymous codon usage analyses with the ade4 and seqinR packages. Bioinformatics 21:545-547.

Chinnadurai, G. 1998. Control of apoptosis by human adenovirus genes. Seminars in Virology 8:399-408.

Chiu, CY, E Wu, SL Brown, DJ Von Seggern, GR Nemerow, PL Stewart. 2001. Structural analysis of a fiber-pseudotyped adenovirus with ocular tropism suggests differential modes of cell receptor interactions. Journal of Virology 75:5375-5380.
Clark, AJ, W Inwood, T Cloutier, TS Dhillon. 2001. Nucleotide sequence of coliphage HK620 and the evolution of lambdoid phages. Journal of Molecular Biology 311:657-679.
Codoner, FM, MA Fares. 2008. Why should we care about molecular coevolution? Evol Bioinform Online 4:29-38.
Cullen, M, SP Perfetto, W Klitz, G Nelson, M Carrington. 2002. High-resolution patterns of meiotic recombination across the human major histocompatibility complex. American Journal of Human Genetics 71:759-776.
Davison, AJ, M Benko, B Harrach. 2003. Genetic content and evolution of adenoviruses. J Gen Virol 84:2895-2908.
de Jong, JC, AD Osterhaus, MS Jones, B Harrach. 2008. Human adenovirus type 52: a type 41 in disguise? Journal of Virology 82:3809; author reply 3809-3810.
de Jong, RN, LAT Meijer, PC van der Vliet. 2003. DNA binding properties of the adenovirus DNA replication priming protein pTP. Nucleic Acids Res 31:3274-3286.
De Jong, RN, PC Van Der Vliet, AB Brenkman. 2003. Adenovirus DNA replication: Protein priming, jumping back and the role of the DNA binding protein DBP. Adenoviruses: Model and Vectors in Virus-Host Interactions 272:187-211.
Dehghan, S, EB Liu, J Seto, et al. 2012. Five genome sequences of subspecies B1 human adenoviruses associated with acute respiratory disease. Journal of Virology 86:635-636.
Duffy, S, LA Shackelton, EC Holmes. 2008. Rates of evolutionary change in viruses: patterns and determinants. Nature Reviews: Genetics 9:267-276.
Dunn, SD, LM Wahl, GB Gloor. 2008. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics 24:333-340.
Dvir, S, L Velten, E Sharon, D Zeevi, LB Carey, A Weinberger, E Segal. 2013. Deciphering the rules by which 5 '-UTR sequences affect protein expression in yeast. Proc Natl Acad Sci U S A 110:E2792-E2801.
Ebner, K, W Pinsker, T Lion. 2005. Comparative sequence analysis of the hexon gene in the entire spectrum of human adenovirus serotypes: phylogenetic, taxonomic, and clinical implications. Journal of Virology 79:12635-12642.
Echavarria, M. 2008. Adenoviruses in immunocompromised hosts. Clinical Microbiology Reviews 21:704-715.
Eiz, B, T Adrian, P Pring-Akerblom. 1997. Recombinant fibre proteins of human adenoviruses Ad9, Ad15 and Ad19: localization of the haemagglutination properties and the type-specific determinant. Res Virol 148:5-10.
Eppley, JM, GW Tyson, WM Getz, JF Banfield. 2007. Genetic exchange across a species boundary in the archaeal genus ferroplasma. Genetics 177:407-416.
Fujimoto, T, Y Matsushima, H Shimizu, Y Ishimaru, A Kano, E Nakajima, AK Adhikary, N Hanaoka, N Okabe. 2012. A molecular epidemiologic study of human adenovirus type 8 isolates causing epidemic keratoconjunctivitis in Kawasaki City, Japan in 2011. Japanese Journal of Infectious Diseases 65:260-263.
Fujimoto, T, S Yamane, T Ogawa, et al. 2014. A novel complex recombinant form of type 48-related human adenovirus species D isolated in Japan. Japanese Journal of Infectious Diseases 67:416-416.
Gahery-Segard, H, F Farace, D Godfrin, J Gaston, R Lengagne, T Tursz, P Boulanger, JG Guillet. 1998. Immune response to recombinant capsid proteins of adenovirus in humans: antifiber and anti-penton base antibodies have a synergistic effect on neutralizing activity. Journal of Virology 72:2388-2397.
Gibbs, MJ, JS Armstrong, AJ Gibbs. 2000. Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. Bioinformatics 16:573-582.

Goh, CS, AA Bogan, M Joachimiak, D Walther, FE Cohen. 2000. Co-evolution of proteins with their interaction partners. Journal of Molecular Biology 299:283-293.
Gonzalez, G, KO Koyanagi, K Aoki, N Kitaichi, S Ohno, H Kaneko, S Ishida, H Watanabe. 2014. Intertypic modular exchanges of genomic segments by homologous recombination at universally conserved segments in human adenovirus species D. Gene 547:10-17.
Gonzalez, G, KO Koyanagi, K Aoki, H Watanabe. 2015. Interregional Coevolution Analysis Revealing Functional and Structural Interrelatedness between Different Genomic Regions in Human Mastadenovirus D. Journal of Virology 89:6209-6217.
Greber, UF, N Arnberg, G Wadell, M Benko, EJ Kremer. 2013. Adenoviruses - from pathogens to therapeutics: a report on the 10th International Adenovirus Meeting. Cellular Microbiology 15:16-23.
Guyon, I, J Weston, S Barnhill, V Vapnik. 2002. Gene selection for cancer classification using support vector machines. Machine Learning 46:389-422.
Harada, JN, A Shevchenko, A Shevchenko, DC Pallas, AJ Berk. 2002. Analysis of the adenovirus E1B-55K-anchored proteome reveals its link to ubiquitination machinery. Journal of Virology 76:9194-9206.
Harrach, B, M Benkö, GW Both, et al. 2011. Family Adenoviridae. In: ICoTo Viruses, AMQ King, MJ Adams, EJ Lefkowitz, EB Carstens, editors. Virus Taxonomy: Classification and Nomenclature of Viruses: Ninth Report of the International Committee on Taxonomy of Viruses: Academic Press. p. 125-141.
Heath, L, E van der Walt, A Varsani, DP Martin. 2006. Recombination patterns in aphthoviruses mirror those found in other picornaviruses. Journal of Virology 80:11827-11832.
Hong, SS, NA Habib, L Franqueville, S Jensen, PA Boulanger. 2003. Identification of adenovirus (ad) penton base neutralizing epitopes by use of sera from patients who had received conditionally replicative ad (addl1520) for treatment of liver tumors. Journal of Virology 77:10366-10375.
Horwitz, MS. 2004. Function of adenovirus E3 proteins and their interactions with immunoregulatory cell proteins. Journal of Gene Medicine 6 Suppl 1:S172-183.
Ishiko, H, Y Shimada, T Konno, A Hayashi, T Ohguchi, Y Tagawa, K Aoki, S Ohno, S Yamazaki. 2008. Novel human adenovirus causing nosocomial epidemic keratoconjunctivitis. Journal of Clinical Microbiology 46:2002-2008.
Jacobs, SC, AJ Davison, S Carr, AM Bennett, R Phillpotts, GW Wilkinson. 2004. Characterization and manipulation of the human adenovirus 4 genome. Journal of General Virology 85:3361-3366.
Javier, RT. 1994. Adenovirus type-9 E4 open reading frame-1 encodes a transforming protein required for the production of mammary-tumors in rats. Journal of Virology 68:3917-3924.
Jeffreys, AJ, L Kauppi, R Neumann. 2001. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nature Genetics 29:217-222.
Jin, XH, K Aoki, N Kitaichi, T Ariga, S Ishida, S Ohno. 2011. Genome variability of human adenovirus type 8 causing epidemic keratoconjunctivitis during 1986-2003 in Japan. Mol Vis 17:3121-3127.
Jones, MS, 2nd, B Harrach, RD Ganac, MM Gozum, WP Dela Cruz, B Riedel, C Pan, EL Delwart, DP Schnurr. 2007. New adenovirus species found in a patient presenting with gastroenteritis. Journal of Virology 81:5978-5984.
Kaneko, H, K Aoki, S Ishida, et al. 2011a. Recombination analysis of intermediate human adenovirus type 53 in Japan by complete genome sequence. Journal of General Virology 92:1251-1259.
Kaneko, H, K Aoki, S Ohno, et al. 2011b. Complete Genome Analysis of a Novel Intertypic Recombinant Human Adenovirus Causing Epidemic Keratoconjunctivitis in Japan. Journal of Clinical Microbiology 49:484-490.

Kaneko, H, T Iida, H Ishiko, T Ohguchi, T Ariga, Y Tagawa, K Aoki, S Ohno, T Suzutani. 2009. Analysis of the complete genome sequence of epidemic keratoconjunctivitis-related human adenovirus type 8, 19, 37 and a novel serotype. Journal of General Virology 90:1471-1476.
Kaneko, H, T Suzutani, K Aoki, et al. 2011c. Epidemiological and virological features of epidemic keratoconjunctivitis due to new human adenovirus type 54 in Japan. Br J Ophthalmol 95:32-36.
Karlin, S, SF Altschul. 1990. Methods for Assessing the Statistical Significance of Molecular Sequence Features by Using General Scoring Schemes. Proc Natl Acad Sci U S A 87:2264-2268.
Kato, SEM, WY Huang, SJ Flint. 2011. Role of the RNA recognition motif of the E1B 55 kDa protein in the adenovirus type 5 infectious cycle. Virology 417:9-17.
Katoh, K, K Misawa, K Kuma, T Miyata. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30:3059-3066.
Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution 16:111-120.
Kimura, M. 1984. The Neutral Theory of Molecular Evolution: Cambridge University Press.
King, AJ, WR Teertstra, L Blanco, M Salas, PC van der Vliet. 1997. Processive proofreading by the adenovirus DNA polymerase. Association with the priming protein reduces exonucleolytic degradation. Nucleic Acids Res 25:1745-1752.
Koren, E, AS De Groot, V Jawa, et al. 2007. Clinical validation of the "in silico" prediction of immunogenicity of a human recombinant therapeutic protein. Clinical Immunology 124:26-32.
Krejci, L, V Altmannova, M Spirek, XL Zhao. 2012. Homologous recombination and its regulation. Nucleic Acids Res 40:5795-5818.
Lam, TH, M Shen, JM Chia, SH Chan, EC Ren. 2013. Population-specific recombination sites within the human MHC region. Heredity (Edinb) 111:131-138.
Lapointe, FJ, T Garland. 2001. A generalized permutation model for the analysis of cross-species data. Journal of Classification 18:109-127.
Lee, D, O Redfern, C Orengo. 2007. Predicting protein function from sequence and structure. Nat Rev Mol Cell Biol 8:995-1005.
Lefeuvre, P, JM Lett, B Reynaud, DP Martin. 2007. Avoidance of protein fold disruption in natural virus recombinants. PLoS Pathog 3:e181.
Lefeuvre, P, JM Lett, A Varsani, DP Martin. 2009. Widely conserved recombination patterns among single-stranded DNA viruses. Journal of Virology 83:2697-2707.
Legendre, P. 2000. Comparison of permutation methods for the partial correlation and partial Mantel tests. Journal of Statistical Computation and Simulation 67:37-73.
Leppard, KN. 1997. E4 gene function in adenovirus, adenovirus vector and adeno-associated virus infections. J Gen Virol 78:2131-2138.
Li, WH. 1993. Unbiased Estimation of the Rates of Synonymous and Nonsynonymous Substitution. Journal of Molecular Evolution 36:96-99.
Lippe, R, FL Graham. 1989. Adenoviruses with nonidentical terminal sequences are viable. Journal of Virology 63:5133-5141.
Liu, EB, DA Wadford, J Seto, et al. 2012. Computational and serologic analysis of novel and known viruses in species human adenovirus D in which serology and genomics do not correlate. PLoS One 7:e33212.
Liu, H, L Jin, SB Koh, I Atanasov, S Schein, L Wu, ZH Zhou. 2010. Atomic structure of human adenovirus by cryo-EM reveals interactions among protein networks. Science 329:1038-1043.
Liu, H, JH Naismith, RT Hay. 2000. Identification of conserved residues contributing to the activities of adenovirus DNA polymerase. Journal of Virology 74:11681-11689.
Liu, HX, RS Zhang, XJ Yao, MC Liu, ZD Hu, BT Fan. 2004. Prediction of the isoelectric point of an amino acid based on GA-PLS and SVMs. J Chem Inf Comput Sci 44:161-167.

Lukashev, AN, OE Ivanova, TP Eremeeva, RD Iggo. 2008. Evidence of frequent recombination among human adenoviruses. Journal of General Virology 89:380-388.
Ma, HC, P Hearing. 2011. Adenovirus structural protein IIIa is involved in the serotype specificity of viral DNA packaging. Journal of Virology 85:7849-7855.
Mantel, N. 1967. The detection of disease clustering and a generalized regression approach. Cancer Res 27:209-220.
Martin, D, E Rybicki. 2000. RDP: detection of recombination amongst aligned sequences. Bioinformatics 16:562-563.
Martin, DP, D Posada, KA Crandall, C Williamson. 2005a. A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. AIDS Res Hum Retroviruses 21:98-102.
Martin, LC, GB Gloor, SD Dunn, LM Wahl. 2005b. Using information theory to search for co-evolving residues in proteins. Bioinformatics 21:4116-4124.
Matsushima, Y, E Nakajima, M Ishikawa, A Kano, A Komane, T Fujimoto, N Hanaoka, N Okabe, H Shimizu. 2014. Construction of new primer sets for corresponding to genetic evolution of human adenoviruses in major capsid genes through frequent recombination. Japanese Journal of Infectious Diseases 67:495-502.
Matsushima, Y, H Shimizu, A Kano, et al. 2013. Genome sequence of a novel virus of the species Human adenovirus D associated with acute gastroenteritis. Genome Announc 1:e00068-00012.
Matsushima, Y, H Shimizu, TG Phan, H Ushijima. 2011. Genomic characterization of a novel human adenovirus type 31 recombinant in the hexon gene. Journal of General Virology 92:2770-2775.
Matthews, DA, WC Russell. 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.
McSharry, BP, HG Burgert, DP Owen, et al. 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. Journal of Virology 82:4585-4594.
Mei, YF, G Wadell. 1996. Epitopes and hemagglutination binding domain on subgenus B:2 adenovirus fibers. Journal of Virology 70:3688-3697.
Miron, MJ, IE Gallouzi, JN Lavoie, PE Branton. 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.
Miyata, T, T Yasunaga, T Nishida. 1980. Nucleotide sequence divergence and functional constraint in mRNA evolution. Proc Natl Acad Sci U S A 77:7328-7332.
Nassif, N, W Engels. 1993. DNA homology requirements for mitotic gap repair in Drosophila. Proc Natl Acad Sci U S A 90:1262-1266.
Nei, M, S Kumar. 2000. Molecular Evolution and Phylogenetics: Oxford University Press.
O'Brien, V. 1998. Viruses and apoptosis. Journal of General Virology 79 ( Pt 8):1833-1845.
Opperman, R, E Emmanuel, AA Levy. 2004. The effect of sequence divergence on recombination between direct repeats in Arabidopsis. Genetics 168:2207-2215.
Padidam, M, S Sawyer, CM Fauquet. 1999. Possible emergence of new geminiviruses by frequent recombination. Virology 265:218-225.
Pamilo, P, NO Bianchi. 1993. Evolution of the Zfx and Zfy Genes - Rates and Interdependence between the Genes. Molecular Biology and Evolution 10:271-281.
Parker, EJ, CH Botting, A Webster, RT Hay. 1998. Adenovirus DNA polymerase: domain organisation and interaction with preterminal protein. Nucleic Acids Res 26:1240-1247.
Paterson, S, T Vogwill, A Buckling, et al. 2010. Antagonistic coevolution accelerates molecular evolution. Nature 464:275-278.

Pechkovsky, A, A Salzberg, T Kleinberger. 2013. The adenovirus E4orf4 protein induces a unique mode of cell death while inhibiting classical apoptosis. Cell Cycle 12:2343-2344.
Perez-Berna, AJ, A Ortega-Esteban, R Menendez-Conejero, DC Winkler, M Menendez, AC Steven, SJ Flint, PJ de Pablo, C San Martin. 2012. The role of capsid maturation on adenovirus priming for sequential uncoating. Journal of Biological Chemistry 287:31582-31595.
Posada, D, KA Crandall. 2001. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. Proc Natl Acad Sci U S A 98:13757-13762.
Psorakis, I, T Damoulas, MA Girolami. 2010. Multiclass relevance vector machines: sparsity and accuracy. IEEE Trans Neural Netw 21:1588-1598.
Querido, E, MR Morrison, H Chu-Pham-Dang, SWL Thirlwell, D Boivin, PE Branton. 2001. Identification of three functions of the adenovirus E4orf6 protein that mediate p53 degradation by the E4orf6-E1B55K complex. Journal of Virology 75:2508-2508.
Raki, M, M Sarkioja, S Escutenaire, et al. 2011. Switching the fiber knob of oncolytic adenoviruses to avoid neutralizing antibodies in human cancer patients. J Gene Med 13:253-261.
Robinson, CM, J Rajaiya, MP Walsh, D Seto, DW Dyer, MS Jones, J Chodosh. 2009. Computational analysis of human adenovirus type 22 provides evidence for recombination among species D human adenoviruses in the penton base gene. Journal of Virology 83:8980-8985.
Robinson, CM, D Seto, MS Jones, DW Dyer, J Chodosh. 2011a. Molecular evolution of human species D adenoviruses. Infection Genetics and Evolution 11:1208-1217.
Robinson, CM, G Singh, C Henquell, MP Walsh, H Peigue-Lafeuille, D Seto, MS Jones, DW Dyer, J Chodosh. 2011b. Computational analysis and identification of an emergent human adenovirus pathogen implicated in a respiratory fatality. Virology 409:141-147.
Robinson, CM, G Singh, JY Lee, et al. 2013. Molecular evolution of human adenoviruses. Sci Rep 3:1812.
Robinson, DM, DT Jones, H Kishino, N Goldman, JL Thorne. 2003. Protein evolution with dependence among codons due to tertiary structure. Molecular Biology and Evolution 20:1692-1704.
Rubenwolf, S, H Schutt, M Nevels, H Wolf, T Dobner. 1997. Structural analysis of the adenovirus type 5 E1B 55-kilodalton-E4orf6 protein complex. Journal of Virology 71:1115-1123.
Russell, WC. 2009. Adenoviruses: update on structure and function. J Gen Virol 90:1-20.
Saitou, N, M Nei. 1987. The Neighbor-Joining method - a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4:406-425.
Salminen, MO, JK Carr, DS Burke, FE Mccutchan. 1995. Identification of Breakpoints in Intergenotypic Recombinants of Hiv Type-1 by Bootscanning. AIDS Research and Human Retroviruses 11:1423-1425.
Schneider-Brachert, W, V Tchikov, O Merkel, et al. 2006. Inhibition of TNF receptor 1 internalization by adenovirus 14.7K as a novel immune escape mechanism. Journal of Clinical Investigation 116:2901-2913.
Selva, EM, L New, GF Crouse, RS Lahue. 1995. Mismatch correction acts as a barrier to homeologous recombination in Saccharomyces cerevisiae. Genetics 139:1175-1188.
Seto, D, J Chodosh, JR Brister, MS Jones, C Members of the Adenovirus Research. 2011. Using the whole-genome sequence to characterize and name human adenoviruses. Journal of Virology 85:5701-5702.
Sharp, PA. 1994. Split genes and RNA splicing. Cell 77:805-815.
Shull, VL, AP Vogler, MD Baker, DR Maddison, PM Hammond. 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. Systematic Biology 50:945-969.
Singh, G, CM Robinson, S Dehghan, T Schmidt, D Seto, MS Jones, DW Dyer, J Chodosh. 2012. Overreliance on the hexon gene, leading to misclassification of human adenoviruses. Journal of Virology 86:4693-4695.
Smith, JM. 1992. Analyzing the mosaic structure of genes. Journal of Molecular Evolution 34:126-129.

Sonnhammer, EL, G von Heijne, A Krogh. 1998. A hidden Markov model for predicting transmembrane helices in protein sequences. Proc Int Conf Intell Syst Mol Biol 6:175-182.
Stewart, PL, CY Chiu, S Huang, T Muir, YM Zhao, B Chait, P Mathias, GR Nemerow. 1997. Cryo-EM visualization of an exposed RGD epitope on adenovirus that escapes antibody neutralization. EMBO Journal 16:1189-1198.
Takeuchi, S, N Itoh, E Uchio, K Aoki, S Ohno. 1999. Serotyping of adenoviruses on conjunctival scrapings by PCR and sequence analysis. Journal of Clinical Microbiology 37:1839-1845.
Tamura, K, M Nei. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial-DNA in humans and chimpanzees. Molecular Biology and Evolution 10:512-526.
Tamura, K, G Stecher, D Peterson, A Filipski, S Kumar. 2013. MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. Molecular Biology and Evolution 30:2725-2729.
Tollefson, AE, K Toth, K Doronin, M Kuppuswamy, OA Doronina, DL Lichtenstein, TW Hermiston, CA Smith, WSM Wold. 2001. Inhibition of TRAIL-induced apoptosis and forced internalization of TRAIL receptor 1 by adenovirus proteins. Journal of Virology 75:8875-8887.
Toogood, CI, J Crompton, RT Hay. 1992. Antipeptide antisera define neutralizing epitopes on the adenovirus hexon. Journal of General Virology 73 ( Pt 6):1429-1435.
Vulic, M, RE Lenski, M Radman. 1999. Mutation, recombination, and incipient speciation of bacteria in the laboratory. Proc Natl Acad Sci U S A 96:7348-7351.
Waddell, PJ, H Kishino, R Ota. 2007. Phylogenetic methodology for detecting protein interactions. Molecular Biology and Evolution 24:650-659.
Walsh, MP, A Chintakuntlawar, CM Robinson, et al. 2009. Evidence of molecular evolution driven by recombination events influencing tropism in a novel human adenovirus that causes epidemic keratoconjunctivitis. PLoS One 4:e5635.
Walsh, MP, J Seto, EB Liu, et al. 2011. Computational analysis of two species C human adenoviruses provides evidence of a novel virus. Journal of Clinical Microbiology 49:3482-3490.
Webster, A, IR Leith, J Nicholson, J Hounsell, RT Hay. 1997. Role of preterminal protein processing in adenovirus replication. Journal of Virology 71:6381-6389.
Windheim, M, JH Southcombe, E Kremmer, et al. 2013. A unique secreted adenovirus E3 protein binds to the leukocyte common antigen CD45 and modulates leukocyte functions. Proc Natl Acad Sci U S A 110:E4884-E4893.
Wohl, BP, P Hearing. 2008. Role for the L1-52/55K protein in the serotype specificity of adenovirus DNA packaging. Journal of Virology 82:5089-5092.
Wold, WSM, AE Tollefson. 1998. Adenovirus E3 Proteins: 14.7K, RID, and gp19K Inhibit Immune-Induced Cell Death; Adenovirus Death Protein Promotes Cell Death. Seminars in Virology 8:515-523.
Yang, Z. 2006. Computational Molecular Evolution: OUP Oxford.
Zhang, Y, JM Bergelson. 2005. Adenovirus receptors. Journal of Virology 79:12125-12131.
Zhang, Z, S Schwartz, L Wagner, W Miller. 2000. A greedy algorithm for aligning DNA sequences. Journal of Computational Biology 7:203-214.
Zheng, ZM. 2010. Viral oncogenes, noncoding RNAs, and RNA splicing in human tumor viruses. Int J Biol Sci 6:730-755.
Zhou, X, CM Robinson, J Rajaiya, S Dehghan, D Seto, MS Jones, DW Dyer, J Chodosh. 2012. Analysis of human adenovirus type 19 associated with epidemic keratoconjunctivitis and its reclassification as adenovirus type 64. Invest Ophthalmol Vis Sci 53:2804-2811.
Zhou, X, DP Tuck. 2007. MSVM-RFE: extensions of SVM-RFE for multiclass gene selection on DNA microarray data. Bioinformatics 23:1106-1114.