Institutional Repository - Research Portal
Dépôt Institutionnel - Portail de la Recherche

researchportal.unamur.be

University of Namur

THESIS / THÈSE

DOCTOR OF SCIENCES

Genome Landscapes
A Window into the Evolution of Human Viruses

Poulain, Florian

Award date:
2023

Awarding institution:
University of Namur

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 16. May. 2024

UNIVERSITÉ
DE NAMUR

Genome Landscapes: A Window into the Evolution of Human Viruses.

Florian Poulain

Juillet 2023

Thesis Presented for the Degree of Doctor of Biological Science

Abstract

The genomic sequences of human viruses are the product of long-term host-virus coevolution. Exploring the nucleotide composition of these genomes offers the opportunity to unravel the dynamics involved in their evolution, which can be influenced by the cellular environment, like for instance by the APOBEC3 innate effectors.

The APOBEC3 cytidine deaminase family plays a crucial role in the human innate immune system by restricting the life cycle of viruses through viral genome editing at 5'-TC-3' dinucleotides. The resulting selective pressure can be observed in the genomes of human viruses. A constant but incomplete restriction by APOBEC3 leads to an underrepresentation of the APOBEC3 5’-TC-3’ target site. To identify the viral species targeted by APOBEC3, we explored the presence of the APOBEC3 footprint (i.e. TC depletion) among 33,500 human virus sequences. This extensive investigation revealed that at least 22% of the tested human virus species are targeted by APOBEC3. Importantly, we observed strong APOBEC3 footprint on a wide range of virus species, including ssDNA, dsDNA, ssRNA+, and retro-transcribed viruses. Additionally, by exploring the footprint at the genic level, we made a novel discovery of APOBEC3-mediated editing in the EBV herpesvirus and mastadenovirus. This investigation highlights the significant evolutionary constraints imposed by innate immune factors on the genomes of numerous human viruses. It provides a comprehensive view of the broad range of APOBEC3's action and highlights its importance in shaping viral evolution.

APOBEC3 editing is one of many mutational processes shaping virus evolution. We next intended to identify which are the other mutational processes using an approach without a priori. Albeit this task has not been completed yet, we laid the foundations for such analysis. Practically, we reconstructed 487 phylogenetic trees from 55 viral species spanning 23 families and the 7 Baltimore groups. Ancestral sequences were predicted for each node of the phylogenetic trees. By systematically comparing sequences to their ancestor, we generated a collection of over 2.4 million substitutions. For each of the 12 substitution types, the immediate 5’ and 3’ bases were taken into account, dividing the substitution types into 192 subclasses, the so-called the substitution landscape. For most of the viruses, we observed a high degree of symmetry within the substitution landscapes, where each substitution class appears to be canceled out by its opposite (e.g. the C>T substitutions are as frequent as the T>C). Recent zoonotic viruses, like the MERS-CoV, display an asymmetric landscape suggesting that their sequence did not reach equilibrium yet. We also observed that a significant proportion of the substitutions are back-and-forth, i.e. a succession of a first mutation followed by its reversion at a later time-point along the same branch. Surprisingly, the sole feature that distinguishes back-and-forth substitutions from non back-and-forth substitutions (called uncompensated substitutions) is their substitution rate (higher for the back-and-forth). We propose that reversion is a frequent phenomenon in viral history and may contribute to long-term viral sequence stability.

Taken together, these investigations contribute to a better understanding of the forces driving virus evolution and pave the way for the identification of yet unknown mutational processes.

President of the thesis jury :

Professor Benoît Muylkens,
Université de Namur (UNAMUR)

Members of the thesis jury :

Professor Simon Dellicour,
Université Libre de Bruxelles (ULB)

Professor Jean-François Flot,
Université Libre de Bruxelles (ULB)

Professor Michaël Herfs,
Université de Liège (ULG)

Professor Philippe Lemey,
Katholieke Universiteit Leuven (KUL)

Thesis promoter :

Professor Nicolas Gillet,
Université de Namur (UNAMUR)

This thesis was supported by PhD fellowship from F.R.S-FNRS Televie and Unamur NARILIS.

Remerciements

Débuter une thèse, c'est toujours un départ pour une longue et incertaine aventure. Il y a au tout début une direction, quelques contours qui se dessinent, mais comme pour tout périple, on ne peut contempler le chemin parcouru qu'à la fin du voyage. Toute cette histoire aura avant tout été possible grâce à ma compagne **Aurélie Van Tongelen** qui n'a cessé de m'accompagner et de me soutenir. Un périple c'est d'autant plus rigolo qu'on le partage à deux et puis à quatres.

Amateur de jeux de rôle , j'aime à m'imaginer cette aventure comme une quête dans un donjon. Dans cet univers fantastique, les personnages sont caractérisés par leur fonction. On y trouve des magiciens, des voleurs, des paladins .Ils ont également un niveau de personnage qui témoigne de leur expérience.

Il y a ici comme dans toute bonne aventure un commanditaire, un maître voleur déjà expérimenté, de niveau très supérieur, qui m'a recruté pour mener à bien cette quête. Ce maître voleur nommé **Nicolas Gillet**, possède une longue expérience des donjons, et comme tout bon voleur est très rusé et agile pour faire de la science.

Lorsqu'on se promène dans le donjon, il y a souvent plusieurs chemins possibles et il est tentant de vouloir en explorer le plus grand nombre pour y trouver des trésors ou des objets intéressants. Peut-être un peu trop attiré par la possibilité d'explorer de nouvelles questions de biologie, j'ai eu souvent tendance à m'y égarer. Heureusement le maître voleur n'est jamais bien loin, et tempère mes égarements, parfois non sans mal. Certains chemins s'avèrent tout de même parfois être des impasses et il est préférable de les abandonner.

On croise beaucoup de personnages dans un donjon. Durant ma quête j'ai notamment croisé un vieux guerrier paladin, très expérimenté et toujours prêt à faire le bien autour de lui, j'ai nommé **Benoît Muylkens**. Il m'a beaucoup aidé dans des moments d'incertitudes et représente pour moi un réel modèle.

Il y a aussi d'autres aventuriers encore novices, avec qui partager les galères mais également les réjouissances. Je pense ici notamment à **Laura Rubiano, Sarah Mathieu** et bien d'autres **Alexandra Decloux, Maxence Collard**, …, la liste est longue et non exhaustive.

N'oublions pas quelques incontournables du donjon tel que le tavernier de l'auberge, **Kévin Willemart,** toujours présent pour aider un aventurier en détresse, aussi bien dans sa quête que pour se désaltérer.

Laetitia Wiggers, la gardienne des clefs et tenancière du magasin. Elle à toujours tout ce dont vous avez besoin et même ce dont vous n'avez pas besoin.

La guérisseuse **Hélène Dumont**, toujours prête tout comme le paladin à faire le bien autour d'elle.

La gardienne de la chambre forte, **Maite Raffaele.**

Le maître d'arme, un indispensable à rencontrer pour améliorer son équipement, **Pascal Vanbel.**

Il y a également l'herboriste, **Damien Coupeau**. Constamment occupé à réaliser d'étranges et insolites expériences dans sa tour reculée. Il propose aux aventuriers des coups de main et des objets magiques capables de faire la différence au moment propice. Tout comme le tavernier, il est toujours prompt à partager un petit verre avec un aventurier, mais l'issue est parfois chaotique.

En tant qu'apprenti magicien, j'ai eu la chance dans mon épopée de croiser la route d'un certain **Simon Dellicour**, mage niveau 34 qui bien que pratiquant une magie de nature légèrement différente de la mienne m'a énormément appris.

Je tiens à remercier les membres de mon jury, **Michaël Herfs** avec qui il fut passionnant de collaborer, **Philippe Lemey** et **Jean-François Flot.**

Difficile de tenir le compte de tous les personnages qui m'ont tant apporté durant ces 5 années d'aventure. Je remercie **l'ensemble des membres de l'URVI**. Je livre ici une heureuse pensée pour **Chantal Ippersiel.**

Du soutien il en aura fallu. Je désire terminer ces remerciements comme je les ai commencés, en évoquant les personnes qui me sont les plus chères. Il y a bien sûr toute **ma famille de Fayence**, mes parents, mes sœurs et mon frère ainsi que tous mes neveux. Bien que peu familiers pour la plupart, avec mes travaux, ils ont toujours eu beaucoup de patience pour m'écouter, me comprendre et m'encourager.

Je remercie également **ma famille de Belgique**, toujours là pour partager de beaux moments. Merci notamment à **Eliane et André Van Tongelen,** qui ont beaucoup compté pour moi au cours de cette aventure.

Table of contents

1. Introduction

1.1. What is a virus?

1.1.1. From the discovery to the contemporary definition

Like viruses themselves, their definition has evolved over time and continues to evolve in the present day. In Latin, the term '*virus*' referred to poison or disease in a generic sense. The term was first used by Martinus W. Beijerinck in 1898 to name a '*contagium vivum fluidum*' - a contagious living fluid [1]. At that time, the dogma that diseases are caused by infectious agents such as bacteria had already been well demonstrated. But, observations made on mosaic disease of tobacco and tulip highlighted the presence of disease without observable bacteria. Dmitry I. Ivanovsky's work on tobacco mosaic disease also showed that an extract of infected plants remained contagious even after being filtered through minute filters, which did not allow the passage of bacteria. Ivanovsky initially proposed that this substance was a toxin produced by bacteria, but Martinus W. Beijerinck observed that it was able to be amplified during successive steps of inoculation and filtration. This observation proved that the infectious agent was a microorganism that he named a virus [2].

This definition was quickly refined by Friedrich Loeffler and Paul Frosch, who concluded that a virus is not a fluid but a particle. They studied foot-and-mouth disease in cattle and observed that inoculums lost infectivity when passed through filters of a certain size [3].

The crystal structure of the tobacco mosaic virus was observed by electron microscopy in 1939, revealing a complex virion structure. At that time, the question of whether proteins or nucleic acids served as the genetic material was central. In 1953, research on the T2 bacteriophage established that nucleic acids, but not proteins, served as the genetic material. This led David Baltimore to classify viruses into six (now seven) groups based on the support of their genetic information, including double-stranded DNA viruses (dsDNA), single-stranded DNA viruses (ssDNA), single-stranded positive RNA viruses (ssRNA+), single-stranded negative RNA viruses (ssRNA-), double-stranded RNA viruses (dsRNA), and retro-transcribed RNA viruses (rt-RNA), as well as retro-transcribed DNA viruses (rt-DNA) [4]. Baltimore's classification, proposed in 1971, is still used in combination with the International Committee on Taxonomy of Viruses classification [5], which organizes viruses based on their virion structure, phylogenetic distances, and biological characteristics at the family, genus, and species taxonomic levels [6].

All of these historical explorations allow us to build our contemporary definition of a virus :

A virus is an infectious, obligate intracellular parasite with DNA or RNA genetic information support. This genetic information hightjack the host cell's cellular systems to produce viral components. Virions, which are infectious viral particles, are formed through the assembly of newly created components. These progeny virions, created during the infection process, serve as carriers for transmitting the viral genetic material to the subsequent host cell or organism. Upon disassembly, they initiate the next infectious cycle [7]. This current definition does not for instance englobe the viroids, which are plant infectious nucleic acid molecules without protein coat [8]. A virus can be also defined by its life cycle phase. The viral life cycle can be split into two phases: an intracellular phase during which the viral genome is replicated and virion particles are produced and assembled, and a virion phase during which the genomic material is transported to a new host cell.

1.1.2. Evolving the definition of viruses ?

Despite this large consensual definition, its interpretation is still a matter of debate. The question of what a virus is refers to two other important questions: what is the origin of viruses, and are viruses alive? Historically, the observation and characterization of viruses have been focused essentially on a virion-centric view. Virions can be isolated, allowing virus identification and easy access to virus genomes. This has led some authors to define a virus as a coated genetic element (a virion) [9]. According to this definition, viruses are considered as non-living structures because they lack self-metabolic activity and do not produce the adenosine triphosphate, an essential energy source inside cells, in their virions. This virion-centric view is still widely accepted and defended [10,11].

In opposition to the virion-centric view, an intracellular-centric view has emerged. This point of view compares the virion stage to a seedling stage and the virion itself to a gamete or a tree seed where the virus is inanimate. In contrast, during the intracellular stage, viruses exhibit significant activity by producing proteins, hijacking cellular metabolism and machinery, and ultimately ensuring their replication. Consequently, some authors consider viruses to be alive under this definition. The intracellular-centric virus definition varies slightly among authors, but can be defined as:

- "A Molecular Organism": Claudiu I. Bandea proposed that viruses derived from ancient intracellular parasites like endobacteria that progressively lost their membrane and metabolic functions, keeping only their genetic material and propagating it by using host machinery [12,13].
- "A Replication Factory": Jean-Michel Claverie proposed a virus definition based on the detection of replication factories inside cells, which defines a virus as a replication factory. Besides this, a virus is an absolute parasite that replicates using the macromolecular machinery of other biological entities and can encapsulate and disseminate genomes in metabolically inert structures [14,15]. Based on these characteristics, Claverie suggested that the virus definition should also encompass viroids and plasmids.
- "A Virion-encoding Organism": Patrick Forterre proposed a unique definition of a virus as "A Virion-encoding Organism," which comprises both the virion phase and a

phase of virocells. A virocell refers to a host cell that has been infected and reprogrammed to produce virions. Therefore, according to Forterre's definition, a virus is considered a living organism that inhabits both the cellular and non-cellular world [16,17].

The definition of viruses has progressively evolved throughout history and is still a subject of ongoing discussion and refinement. As with many scientific concepts, the aim of these debates is to deepen our understanding and stimulate the development of new hypotheses and ideas. For instance, in the field of oncogenic virus research, there is a growing interest in viewing viruses as reprogrammed cells, with an analogy drawn between virus-infected cells and cancerous cells. This approach allows researchers to explore the transition from one cellular state to another, and to generate new hypotheses that may lead to new insights and treatments.

1.2. Virus origins

1.2.1. The possible origins of life

The origins of viruses are inextricably linked with the origins of life itself. Therefore, in order to fully comprehend the history and evolution of viruses, it is important to have a good understanding of the origins of life and how the first living cells emerged. By understanding the origin of life, we can gain insights into how viruses co-evolved with living cells. The currently accepted theory explaining the origin of life is the Oparin-Haldane theory, which postulates that life originated from a prebiotic soup of basic building blocks that progressively increased in complexity. In the primordial environment of early Earth, high concentrations of inorganic molecules such as CO_2 and ammonium, combined with high temperatures and energy from the sun, led to the production of the first organic molecules, such as amino acids and nucleotides, in this soup. These basic organic molecules were then further complexified to produce polymer macromolecules like proteins and nucleic acids. The polymers may have assembled into units or structures capable of sustaining and replicating themselves. Oparin suggested that these might have been "colonies" of proteins clustered together to carry out metabolism, while Haldane suggested that macromolecules became enclosed in membranes to form cell-like structures [18].

There are various hypotheses regarding the formation of the first soup of amino acids, nucleic acids, and lipid molecules that led to the origin of life on Earth. In 1953, Stanley Miller conducted a famous experiment that demonstrated the possibility of producing amino acids in a laboratory setting by subjecting a medium mimicking the primordial Earth's atmosphere to an electrical energy shock [19]. This experiment provided evidence that amino acids could have been formed in the primordial ocean. Later on, nucleic acids and lipids were also produced in the laboratory from inorganic molecules. Another hypothesis, not necessarily exclusive to the first, is the panspermia origin of amino acids and nucleic acids. It suggests that these building blocks of life could have been brought to Earth by comets, asteroids, and meteorites during large earth bombardments in primordial ages. The

Earth's water could have also been a part of this bombardment. There are several observations supporting this theory, such as the detection of around 80 amino acids inside an Australian meteorite in 1962 [20], which was later confirmed by successive space missions. Glycine was also isolated from samples returned to Earth in 2006 from Comet Wild-2 by NASA's Stardust mission [21] and were identified on comet 67P/Churyumov-Gerasimenko by ESA Rosetta mission in 2016 [22]. The most remarkable results were obtained in 2022 by the analysis of the samples of asteroid 162173 RYUGU by JAXA's Hayabusa 2 mission, where amino acids and other soluble organic compounds were extracted.

From these molecular soups containing amino acids, nucleic acids, and lipid molecules, more complex macromolecules, such as RNA, DNA, and proteins, were gradually assembled. There is a vast debate about the nature of the first macromolecule, but RNA is believed to be the most likely candidate. Indeed, the first macromolecule should be able to be self-amplifying and therefore have enzymatic activity and support genetic information. Only RNA molecules could perform such functions as far as we know. RNA can self-replicate thanks to its base complementarity, and mRNA is well known to support information. For example, ribozymes, like ribosomes [23] or RNase P [24], have enzymatic activity. In 1998, other *in vitro* experiments confirmed the large enzymatic potential of RNA. By observing a large number of randomly generated RNA molecules, some of them have been reported to harbor enzymatic activity [25]. These observations support the theory of an RNA world preceding the appearance of DNA.

The transition between the RNA world to cells could gradually occur by the introduction of self-replicating RNAs inside a lipid membrane. This fusion could create an intramembrane medium that helps metabolic reactions and promotes complexification. Before the emergence of DNA-based cells, an intermediate step could have been RNA-based cells. Then, DNA, which is more stable in terms of molecules and for genetic information preservation, was probably selected to build DNA-based cells. These cells, which later gave rise to the different life kingdoms of prokaryotes, eukaryotes, and archaea, are named the Last Universal Cellular Ancestor (LUCA) [26,27].

1.2.2. Possible origins of viruses

Since the origins of life remain a mystery with many possibilities, and given that the origin of viruses is closely linked to the origin of life, it is not surprising that various hypotheses have been proposed in the literature to explain the origins of viruses. There are three main hypotheses regarding virus origins[28].

“The virus-first hypothesis”: The first hypothesis is the "virus-first" hypothesis, which suggests that viruses descended directly from self-replicating RNA molecules from the RNA world. They could have used the primordial soup as a host. These molecules progressively developed the capability to create a capsid and then became able to infect primordial cells. This hypothesis is supported by the fact that viruses encode proteins like capsid proteins which have no equivalent in other living kingdoms [30]. Another argument is the apparent simplicity of plant viroids that are infectious non-enveloped RNA molecules. These molecules do not encode protein, self-replicate by ribozyme activity, and use the host RNA polymerase and RNase H. Some authors propose that this apparent simplicity is a relic of the RNA world [31], but there is a matter of debate about whether simplicity is necessarily linked to evolutionary inferiority.

“The escape hypothesis” : The second hypothesis is the "escaping" hypothesis, which suggests that viruses could have resulted from escaped nucleic acid from cells which also acquired the ability to self-replicate. This theory reposed first on the observation of host bacterial DNA in the sequence of phage. Potential primordial RNA cells or RNA fragments could have acquired the replication mechanism from RNA cell chromosomes, and the acquisition of viral genes could have resulted from insertion of genetic material. However, there is no homology between viral and cellular genes that could indicate such an escaped event, except in rare cases such as the human Delta virus (HDV) that has a ribozyme RNA structure on its genome that is closely related to the CPEB3 ribozyme present in human introns [32]. HDV could represent a potential example of viruses produced by a cellular RNA escape [33].

“The reduction hypothesis” : The third hypothesis is the "reduction" hypothesis, which proposes that viruses could result from symbiotic or parasitic primordial cells that progressively lost their own components and began to use the components of their host cell [12]. This hypothesis is supported by the fact that positive single-strand viruses all share a common viral hallmark, and giant mimivirus viruses with a 1.2 mbp genome could represent an intermediate state in the virus simplification process, although there is no homology between mimivirus genes and eukaryotic, prokaryotic, or archaeal genes [34,35].

These three main hypotheses for explaining the origin of viruses have been combined by different authors to produce hybrid models. One such model is the chimeric hypothesis, which combines aspects of the virus-first and escape hypotheses [36]. This model suggests that nucleic acids may have originated from the RNA world, while envelope proteins were acquired from primordial cells. Another hypothesis is the symbiogenic hypothesis, which is an extension of the reduction models [37]. This hypothesis suggests that virocells and primordial cells may have coexisted until the virocells became parasites of the ancestral host cells.

In summary, there is ongoing debate and research on the origins of viruses, and while these hypotheses provide potential explanations, the true origin of viruses remains elusive. Currently, modern viruses could be polyphyletic, meaning that they have different origins [38]. Therefore, it is possible that there is not just one single scenario that explains the origins of viruses, but rather multiple events that have contributed to their emergence.

1.3. Virus evolution

As with other organisms, the principles governing viral genome evolution have been increasingly elucidated over the past century. Since the research presented in this manuscript aim to enhance the understanding of virus evolution, it is essential to initially introduce various principles and concepts related to virus evolution, as well as the broader evolution of organisms.

1.3.1. Principles of molecular evolution

The principles of molecular evolution suggest that protein and nucleic acid chains undergo constant sequence modifications, a phenomenon referred to as molecular evolution. These principles are based on the theory of neo-Darwinism, which emerged in the 1940s and unifies the work of Charles Darwin and Gregor Mendel. There are three main principles in molecular evolution according to this theory [39] [40]: (1) the mutation flux is the driving force of evolution, (2) natural selection favors mutations that promote genome transmission, and (3) isolation can modulate the evolutionary process and lead to speciation. In simpler terms, these principles state that mutations occur constantly and drive evolution, some mutations are advantageous and will be kept while others will be lost, and changes in the environment can affect this evolutionary process.

The evolution rate, also called the substitution rate, is used to measure the mutation flux. The terms mutation and substitution are often confused or considered synonyms. However, to understand the research presented in this manuscript, it is important that the subtlety of their difference is clear. A substitution occurs when a mutation is fixed in a population's genome. Typically, the substitution rate is expressed as the number of substitutions per nucleotide and per year. This rate cannot be zero because genes have a natural optimum level that has been optimized by selection [41]. The rate must be high enough to allow genome plasticity and enable organisms to adapt to environmental changes, but it cannot be too high, as it could destabilize the genome [42].

Alfred Sturtevant observed in 1937 that the substitution rate can vary among drosophila species and strains [41]. A few years later, in the early 1960s, Emile Zuckerkandl and Linus Pauling proposed the concept of the molecular clock which suggests that the rate of evolutionary change in DNA sequences over time is relatively constant. They conducted studies on the hemoglobin proteins of human, gorilla, and horse sequences and found that the rate of substitution is relevant to the estimated species divergences based on fossil dating [43–45]. The molecular clock concept proposed two main ideas that are fundamental to molecular phylogeny:

Firstly, the genetic distance between two sequences from different individuals is directly correlated with real time. Secondly, the evolution rate being approximately constant over time and among evolutionary lineages.

This second idea has been found to be inaccurate for some genes and species sequences where the molecular clock is violated due to environmental factors. To address this problem, phylogenetic reconstructions nowadays use a relaxed clock model, which allows in a phylogenetic tree reconstruction to have different rates of substitutions in different branches of the tree [45].

To explain the constancy of substitution accumulation on protein sequences, Zuckerkandl and Pauling proposed that most of them do not affect protein function. Their work led to another theory called the neutral theory which was proposed in 1968 by Motoo Kimura [46]. Based on the calculation of the nucleotide substitution rate, Kimura observed that the accumulation of DNA substitution is too high by comparison to the estimation of the apparition in human population of new variant forms of a gene that determines a particular trait or characteristic (alleles). He concluded that this difference can only be explained by the fact that most of the mutations are neutral for natural selection [46]. Based on the development of this theory, Kimura and his colleagues proposed five predictions [47] :

- *For each protein, the rate of evolution in terms of amino acid substitutions is approximately constant/site per year for various lines, as long as the function and tertiary structure of the molecule remain essentially unaltered.*
- *Functionally less important molecules or parts of a molecule evolve (in terms of mutant substitutions) faster than more important ones.*
- *Those mutant substitutions that disrupt less the existing structure and function of a molecule (conservative substitutions) occur more frequently in evolution than more disruptive ones.*
- *Gene duplication must always precede the emergence of a gene having a new function.*
- *Selective elimination of definitely deleterious mutants and random fixation of selectively neutral or very slightly deleterious mutants occur far more frequently in evolution than positive Darwinian selection of definitely advantageous mutants.*“ [47].

Those theories agree on the stochasticity of evolution meaning that evolution implies a large degree of randomness, but which still follows certain rules or patterns over time. Then, the stochasticity of evolution can be modeled or analyzed statistically. Stochasticity can affect evolution on three different scales: the stochasticity of mutations at the sequence level, the stochasticity of life history that impacts the individual, and the stochasticity of the environment that impacts the population [48].

The description of these principles of molecular evolution has paved the way for deciphering the evolution of genome sequences and understanding the complex evolutionary history of organisms.

1.3.2. Measuring the virus evolution

Quantifying intangible phenomena such as evolution remains a key question to understanding the importance and role of the involved mechanisms. Part of evolution is driven by single nucleotide variation (SNV) also called substitution. Other types of mutation, the insertion and deletion can also alter the genome sequence. The rate of genome variation at a single position can be observed using two different measures: the substitution rate and the mutation rate. These measures refer to the sum of captured SNVs *in vivo* over a defined period of time or *in vitro* after several rounds of replication or infection cycles, respectively.

1.3.3. The mutation rate

Because the measure of the mutation rate corresponds to the number of nucleotide sequence mutations per nucleotide and per genome replication (m/n/r), it requires a complete description of the virus replication cycle, which is not the case for all viruses. Alternatively, the mutation rate can be measured by reporting the number of mutations per nucleotide and per cycle of infection (or cell burst for lytic viruses) (m/n/c). Some authors also report the mutation rate by unit of time, but this sometimes corresponds to a non-differentiation between the mutation rate and the substitution rate, which is also reported as the number of substitutions per nucleotide and per year (s/n/y). In most studies, the mutation rate is measured in laboratory assays on individual clones or on an entire population after a defined number of passages.

Historically the virus mutation rate has been first assessed by site specific tests like the Luria–Delbrück fluctuation tests. This test follows the mutation at specific sites that give phenotypic modification and then allow the report of the mutation per genome replication or per infection cycle. This type of test has been used for the first time in viruses in 1976 to determine the bacteriophage Qβ spontaneous mutation rate by the fluctuation follow-up of a unique site [49,50]. This value of $1x10^{-3}$ m/n/c was later corrected to $1.4×10^{-4}$ m/n/r by Luria–Delbrück fluctuation test based on three genomic positions, and to $1.3×10^{-4}$ m/n/c by full genome sequencing [51]. This second measure illustrates another approach, which is based on mutation rate calculation through the measurement of base exchange frequency among a sequence of virus genomes. A third category of test corresponds to cell-free *in vitro* replication error measurement by replicase polymerization assay.

All three methods have advantages and limitations. In the case of the first method, the fluctuation test, the use of a limited number of sites can bias the measurement and limit the interpretation for the type and context of substitution. It is interesting to note that recently, *Pauly et al.* developed an extended Luria-Delbrück fluctuation test approach by using twelve constructs of a green fluorescent protein reporter gene [52]. In these different gene constructs, the mutation of the reporter site leads to the restoration of GFP fluorescence. Each of the twelve GFP constructs allows for the tracking of a specific type of substitution among the twelve possible types (A>C, A>T, A>G, C>A, C>T, C>G, T>A, T>C, T>G, G>A, G>C, G>T). Since the reporting sites are neutral, this test does not have a bias towards underestimating lethal or disadvantageous mutations, resulting in a higher number of mutations being captured. For example, *Pauly et al.* measured a mutation rate of $1.8×10^{-4}$ s/n/r for H1N1 and $2.5×10^{-4}$ s/n/r for Hong Kong 2014 H3N2 influenza using this test, which is higher than the previously estimated rate of $2.7×10^{-6}$-$3.0×10^{-5}$ substitutions per nucleotide per strand copied (s/n/r) obtained through the mutation frequency approach[52].

For the second method, the full genome mutation frequency approach is calculated on a population of virus sequences. It is therefore subject to sequencing error bias. It is also limited to the capture of non lethal and low deleterious mutations. Even if the randomization of the culture plate passage can reduce this selection bias, an underestimation of the mutations will remain. Finally, the *in vitro* polymerase error essay is unbiased but limited to be only able to report the error due to the polymerase.

Thanks to the progressive accumulation of mutation rates measurement, Drak's studies in the nineties started to make comparisons between organisms and viruses [53]. These observations revealed the incredible speed of the virus mutation rates by comparison to other entities and especially by comparison to their host. For instance, there are 9 orders of magnitude between mutation rates of humans and HIV1 [54]. Drake also observed that RNA viruses have higher mutation rates than DNA viruses. These observations were next confirmed and completed thanks to the development of deep sequencing approaches, the constant accumulation of virus sequences and the development of statistical inference approaches to reconstruct virus history [55–57].

1.3.4. The substitution rate

The substitution rate captures the overall pattern of modifications occurring naturally within a population or organism. It is expressed as the number of substitutions by nucleotide and by year (s/n/y). Substitution rate variation is essentially modulated by four parameters: (1) The amount of sequence punctual modification, (2) The fitness effect of a mutation determines whether it is neutral or advantageous, which increases the probability of it being retained, or deleterious or lethal, which leads to the purging of the mutation.(3) The generation time and (4) the effective population size of the virus [55].

1.3.5. Influence of the generation time and the effective population size on the substitution rate variation.

The generation time of a virus is the time from virion production to the production of new virions. It encompasses the host-to-host transmission time, cellular adsorption time, and replication time, representing one complete cycle of the viral life cycle [55]. In epidemiology, the measure of the generation time is based on the observation of transmission pairs like it has been largely applied during the SARS2 coronavirus pandemic for example [58]. Thanks to the observation of virus and bacteria dynamics, the generation time has been correlated with the substitution rate, indicating that the two are linked.

The link between the effective population size (Ne) to the substitution rate is more versatile. Ne is the minimum representation of the real population size that evolved with the same dynamics [55]. Changes in Ne reflect the environmental history of the virus. A decrease in Ne increases the probability of a disadvantageous substitution being fixed in the population, while an advantageous mutation is less likely to be fixed due to increased genetic drift [59]. The impact of this effect is minimal on a large population but is amplified during a bottleneck period, such as during between-host transmission as in the case of HIV1 viruses [60].

1.3.6. The phylogenetic tree inference

The Bayesian approach is one of the most commonly used statistical methods for inferring virus phylogenetic trees due to its ability to handle uncertainty and missing data, as well as its capacity to address the complex and multifactorial processes involved in virus evolution. This approach notably allows the reconstruction of dated phylogenetic trees by the use of sample collection dates as prior.

Briefly, a bayesian phylogenetic tree reconstruction can be summarized in four steps.

1. Model Specification: The first step is to specify a model of nucleotide or amino acid substitution that best describes the evolutionary processes that produced the observed sequences. This model includes assumptions about mutation rates, base composition, and other factors that influence the evolution of the sequences.
2. Prior Specification: Next, prior probability distributions are specified for the tree topology and branch lengths. These priors represent the researcher's beliefs about the likely values of the parameters before taking into account the observed data.
3. Markov Chain Monte Carlo (MCMC) Sampling: A Bayesian MCMC algorithm is used to generate a large number of samples from the posterior probability distribution of the tree topology and branch lengths given the observed data and the prior distributions. The MCMC algorithm starts at an initial tree topology and branch length and proposes changes to the tree topology and/or branch lengths, accepting or rejecting each change based on the posterior probability of the new state compared to the current state.
4. Posterior Probability Calculation: The samples generated by the MCMC algorithm are used to estimate the posterior probability distribution of the tree topology and branch lengths. This distribution provides information about the uncertainty in the estimates of the tree topology and branch lengths.

The strength of the Bayesian approach is its capacity to work on a large distribution of possible phylogenetic trees.

1.3.7. The influence of the measurement timescale on the substitution rate : The TDRP

Thanks to the constant accumulation of virus sequences and notably to the collection of ancient sequences, a Time Dependent Rate Phenomenon (TDRP) has been reported. This phenomenon corresponds to an invert correlation between the time scales of the measurements and the substitution rate (**Figure 1**). In other words, the longer is the time between the different virus collection times used for an inference, the lower is the substitution rate estimated from this phylogenetic tree. The observation of this phenomenon is independent of the method used for the tree inference and of the group or family of the considered viruses [61,62]. Interestingly, the log-transformed rates linearly decrease with the log-transformed measurement timescales with a slope around -0.65 for all virus groups. Such a phenomenon is not specific for viruses. It has also been observed for the bacteria and other eukaryotes [63].

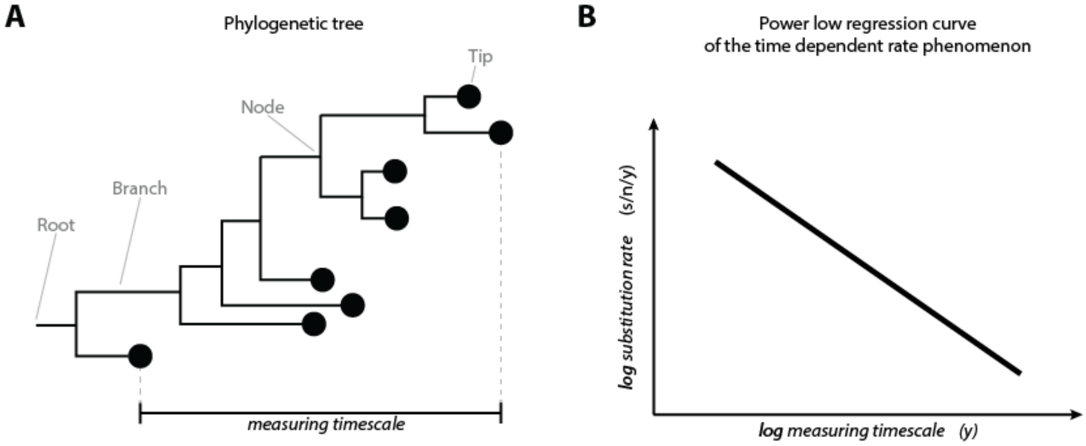

A
Phylogenetic tree
Tip
Node
Branch
Root
measuring timescale
B
Power low regression curve
of the time dependent rate phenomenon
log substitution rate (s/n/y)
log measuring timescale (y)

Figure1: Schematic representation of the time-dependent rate phenomenon. A. On a phylogenetic tree, the measuring time scale corresponds to the period between sample date of collection. **B.** The time-dependent rate phenomenon is observed by a negative correlation between substitution rate and the measuring time scale.

Different mechanisms have been proposed to explain this phenomenon such as sequence purification. It corresponds to the progressive removing of deleterious mutations along successive viral generations. Because deleterious mutations are essentially non synonymous, sequence purification can be evidenced by comparing the relative proportions of non-synonymous substitutions and synonymous substitutions along time. The TDRP effect has been reported to shape the first month substitution rate evolution in the case of H1N1 2009 and the SARS2 2020 pandemic. In these two datasets, sequence purification has been observed by the quantification of the proportion of non-neutral sites (i.e. non synonymous) [64]. It is a direct observation of the sequence purification taking place on short evolutionary time frame.

Site saturation is another factor that can contribute to generating the TDRP, which refers to the accumulation of successive substitutions at a single site. Due to this phenomenon, only a few of the changes will be captured by sequence comparison, leading to an underestimation of the actual number of substitutions. This underestimation will become more pronounced with longer measurement timescales, resulting in the underestimation of genetic distance and substitution rates. In their report of the TDRP, Duchêne *et al* observed a correlation between HIV1 POL and ENV genes sequence saturation and the measurement timescale [61]. Saturation was estimated by the ratio between the observed and the expected entropy of the sequence. Another method to quantify sequence saturation is based on the comparison between transition and transversion mutations. A transition mutation is a nucleotide substitution between two nucleotides of the same class (i.e., purine to purine or pyrimidine to pyrimidine), while a transversion is a substitution between two nucleotides of different classes (i.e., purine to pyrimidine or vice versa). Empirical observation showed that, in most data sets, transitions happen more frequently than transversions. The proportion of transitions in a dataset can vary when it is affected by sequence saturation, as the under-detection of substitutions causes the proportion of transitions to progressively decrease. Eventually, transversions will outnumber transitions indicating saturation occurred [65,66].

The virus substitution rate is determined by the sum of many processes that each represent a fragment of the virus's history. One source of variation in the substitution rate observed for RNA viruses is the cellular tropism of the viral species [68]. Differences in cellular environments suggest that the substitution rate and mutation rate could be dependent on various mutational processes that vary across different types of cells.

1.4. The Sources of viral mutations

Genome alterations are the motor of virus evolution. They encompass several processes that can be resumed by four categories: replication error, genome editing, t environment alterations, and large-scale genome remodeling. The following sections focus on the description of single-position mutations.

1.4.1. Polymerase error

The efficiency of the replication is directed by the nucleotide selectivity, the proofreading activity, and the mismatch repair [67]. The nucleotide selectivity is the ability of a replicase to discriminate between correct and incorrect nucleotides. Tested by *in vitro* measurement, this selectivity appears to be similar between DNA virus polymerase, Reverse transcriptase and RNA virus polymerase, with a range of one mutation each 10^4 to 10^5 replicated nucleotides [68,69].

Despite this similar selectivity, RNA virus polymerases have a much lower fidelity due to the lack of a 3' exonuclease proofreading activity. One reported exception is the coronavirus replication complex that possesses a proofreading activity via the NSP14 protein [70].Thanks to this component, the mutation rate of coronavirus is around 10^{-7}s/n/c whereas the other RNA viruses mutate at a rate between 10^{-3} and 10^{-5} s/n/c [71].

The rate of genome mutation induced by replication can be modulated by the presence and access to the DNA damage repair machinery of the host cell. This effect is specific to DNA viruses that are able to interact with damage sensing and repair machinery and modulate its activity [74]. For instance, papillomaviruses are able to upregulate DNA damage response factors and activate ATM- and ATR-dependent signaling pathways, two distinct pathways involved in the sensing of DNA damage, during their vegetative viral replication [72]. On the contrary, adenoviruses are able to inhibit ATR activation during infection by the action of their E4orf6 viral gene [73].

1.4.2. Mutational processes triggered by the host

Historically mutational processes have been discovered through *in vivo* or *in vitro* exposition to mutagenic factors. Thanks to the development of deep sequencing and to the large amount of available sequences, a new non a priori approach from the field of cancerology has been developed to decipher mutational spectra [74]. By comparing the mutational landscapes of thousands of tumor genomes with their matched normal genomes, researchers have identified tens of different mutational processes acting on the tumor genome. Importantly, each mutational process can be characterized by a specific mutational signature, which is a distinctive combination of mutation types within specific contexts. To date, a catalog of 67 mutational signatures for single-base substitutions has been extracted, of which 49 were considered likely to be of biological origin [75].The mutational process associated with the 18 remaining signatures has not yet been identified.

Since, this approach has also been applied for deciphering the mutational signatures present in virus genomes [79–84]. For instance, three mutational signatures related to virus editing or virus genome oxidation have been identified from the SARS-CoV-2 substitutional landscape. **Figure 2** presents these three different mutational signatures. There are 192 substitution subclasses corresponding to the twelve types of substitutions (A>T, A>C, T>A, T>G, C>A, C>G, G>T, G>C) with four possible nucleotides in 5' and with four possible nucleotides in 3'.

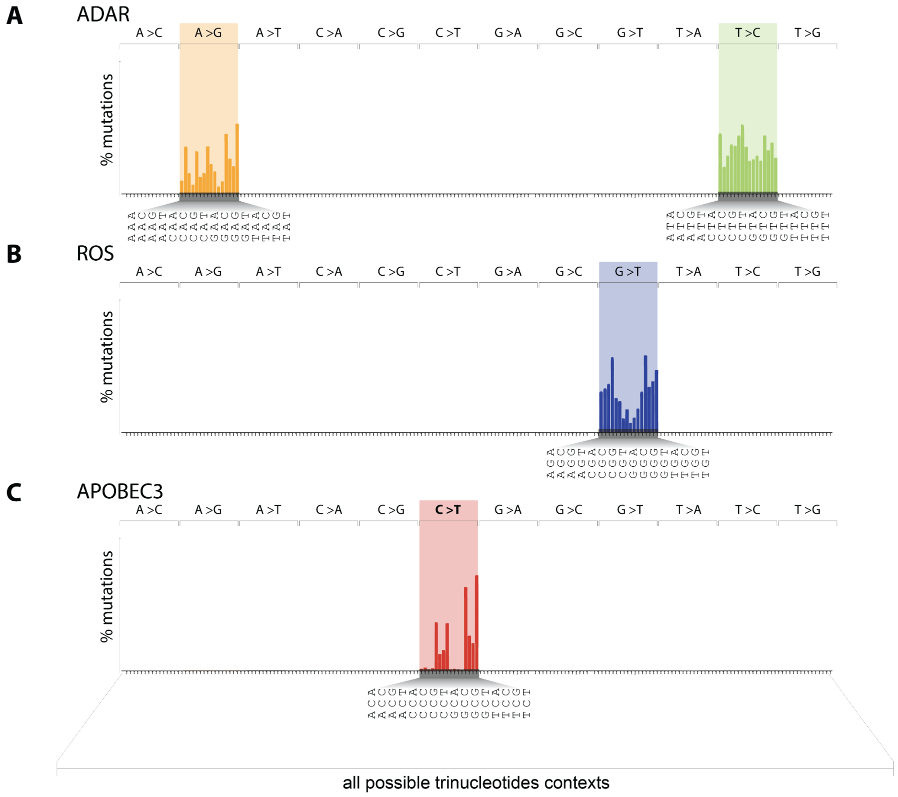


Figure 2: Schematic representation of three virus mutational signatures. These representations show the proportion of the 192 substitution subclasses for the ADAR (**A**), ROS (**B**) and APOBEC3 (**C**) mutational processes. The colors correspond to the type of substitution: yellow for A>G, green for T>C, blue for G>T and red for C>T.

1.4.2.1. ADAR signature

ADARs, or double-stranded RNA-dependent adenosine deaminases, are another component of the innate immune system that can edit and cause hyper-mutation in RNA virus genomes[85,86]. These enzymes work by deaminating adenosines in double-stranded RNA template, which transforms them into inosines. After viral replication, inosines will be replaced by guanosine. This process leads to A > G base substitutions. Due to ADAR editing on dsRNA template, the resulting ADAR mutational signature, as illustrate **figure 2A**, is a combination of A>G and complementary T>C substitutions.

1.4.2.2. ROS-associated mutational signature

Alternatively, the virus genomes can be altered by chemical metabolites from the cellular environment. Among these mechanisms, the reactive oxygen species (ROS) has been reported to cause mutations in viruses and have been associated with a mutational signature for example in the genome of SARS2 viruses [79,87,88]. One of the most common forms of oxidative damage is the oxidation of guanine to 8-oxoguanine (8-oxoG). This damaged base can pair with adenine instead of cytosine during viral replication, leading to a G > T transversion mutation [88,89]. According to the literature, the ROS associated mutational signature is composed of G>T substitutions in A[G>T]X and T[G>T]X substitutions subclasses (**Figure 2B**).

1.4.2.3. APOBEC3 signature and APOBEC3-editing mechanism

The APOBEC3 mutational signature has been observed in tumor genomes and viral genomes [76–78]. The **figure 2C** gives a schematic illustration of this signature. The APOBEC3 mutational signature is characterized by of C>T substitutions with a T or a C upstream the mutated C (T[C>T]X and C[C>T]X subclasses).

The Apolipoprotein B mRNA editing catalytic polypeptide-like enzyme (APOBEC) genes are a subgroup of the human cytidine deaminase family, which also includes APOBEC1, APOBEC2, APOBEC3, APOBEC4, and activation-induced deaminase (AID)[90–92]. These different proteins have diverse functions, with APOBEC1 and APOBEC2 having implications on lipid metabolism [79] and AID being involved in antibody diversification. APOBEC3 proteins are effectors of cellular innate immunity through the restriction of retrotransposons and virus genomes by deaminase and deaminase-independent activity [80,81]. The function of APOBEC4 remains elusive, despite its possible implication in spermatogenesis [82]. All these members of the APOBEC family have one or two zinc-coordinating (Z) catalytic domains [83].

Among this family, the APOBEC3 proteins are a subfamily of seven cytidine deaminases (APOBEC3 A, B, C, DE, F, G, and H) [80,83,84]. These enzymes are able to induce the deamination of cytidine to uracil in ssDNA or ssRNA sequences. All APOBEC3 members, except APOBEC3G, have a favored target site composed of the dinucleotide 5’-T<u>C</u>-3’ (where the underlined letter is the edited cytosine), while APOBEC3G favors the 5’-C<u>C</u>-3’ dinucleotide. The editing induced C > U mutations, which in the case of a DNA virus will result in C > T mutations after genome replication [85]. In case of the editing of the minus strand of the viral genome, like it is the case for HIV1, the mutations detected on the positive strand genome will be G > A.

The APOBEC3s can interrupt virus life cycle by genome hyper-edition [86]. This restriction mechanism has been widely reported during HIV1 infection, where approximately 25% of the detected HIV1 virions in the plasma of infected patients are hyper-edited [87]. To illustrate the magnitude of this effect, it has been reported that APOBECs are responsible for 98% of new HIV1 mutations, versus only 2% by reverse transcriptase, which does not have proofreading capabilities [88].

Alternatively, APOBEC3 editing can remain moderate. The low level of mutations allows the viral replication and then the fixation of the substitutions which are not too deleterious. The explanation for this sub-lethal activity of APOBEC3 is due to the virus's ability to counteract the editing through APOBEC3 degradation or sequestration.

The capacity of APOBEC3 to restrict viral infection have been described in many virus species such as hepatitis B virus (HBV)[89–92], polyomaviruses (JC PyV and BK PyV) [93,94], human T-cell leukemia virus-1 (HTLV-1) [95,96], human papillomavirus (HPV) [97–99] and herpesviruses [100,101] including the Epstein–Barr virus (EBV) [102,103], herpes simplex virus-1 (HSV-1)[104], and Kaposi's sarcoma-associated herpesvirus (KSHV) [105].

1.5. The APOBEC3, a source of virus evolution

As obligate parasites, viruses' evolution is constrained by the host. Host cells have innate immunity factors and notably APOBEC3, to restrain virus replication. Meanwhile, viruses acquired countermeasures to evade host immunity. This constant adaptation between viruses and host cells has been described as an ongoing arms race.

1.5.1. Evolutionary history of the APOBEC gene family

The ongoing arms race between viruses and cells is exemplified by the long evolutionary history of the APOBEC family. The origins of this family of genes have been linked to bacterial, yeast, or plant deaminases, thanks to the presence of similar amino acid motifs found in the catalytic site of APOBEC cytidine deaminase.[106,107]. Based on sequence analysis, it has been suggested that the APOBEC4 gene is the founder of the APOBEC gene family [108]. AID gene was formed through duplication of APOBEC4 and was then duplicated to form APOBEC1 genes. Another subsequent duplication, from AID to APOBEC3, appears to have occurred in the placental mammal infraclass. [106]. The ancestors of these placental mammals likely possessed three types of cytidine deaminase domains named Z1, Z2, and Z3 (for zinc-coordinating catalytic domains 1, 2, and 3), which are also present in the contemporary human APOBEC3 proteins [109]. The presence of homologs of the Z1, Z2, and Z3 domains in vertebrate genomes enables tracing the complex evolution of APOBEC3 throughout mammalian diversification.

The history of the APOBEC3 is paved by larged steps of gene duplication or locus contraction which seems to have been driven by conflicts with ancient viruses. This led to the representation of 18 copies of the Z halodomain in *Chiroptera* genomes. In contrast, mice have only one reported APOBEC3 gene and marsupials have no APOBEC3 orthologs [110]. In primates, the APOBEC3 family counts 7 members. By comparing APOBEC genes in primates, it was found that APOBEC2 was under purifying selection and APOBEC3C was under positive selection. When comparing humans and chimpanzees, AID and APOBEC3A were found to be under purifying selection, while APOBEC3B, APOBEC3D, and APOBEC3G and some portions of APOBEC3F were under positive selection [111]. The presence of positive selection on the sequence of host genes, such as APOBEC3, indicates an ongoing coevolution between viruses and their hosts. Notably, the emergence of APOBEC3G, specific to primates, may be a result of the selective pressure exerted by lentiviruses [112].

The complex history of APOBEC genes duplication illustrates the constant adaptation of the host to its pathogens.

1.5.2. Viral countermeasures

In response to the acquisition of a repertoire of APOBEC innate factors, viruses have also acquired the ability to counteract the restrictions imposed by APOBEC3s.

For example, HIV developed a specific countermeasure to APOBEC3s with its protein Viral Infectivity Factor (VIF). The viral restriction activity of APOBEC3s has historically been described for HIV1 viruses depleted for the VIF protein [80]. Indeed, when the HIV-1 virus is depleted for VIF (HIV-1 ΔVIF), the APOBEC3s are able to drastically restrict the virus life cycle compared to the wild-type virus. Many reports on APOBEC3 restriction of HIV1 allow to depict the following model [113]:

- HIV1 is a reverse transcribed virus. Inside the virion, the viral genome is in the form of two RNA+ molecules. During cell infection and release of the HIV1 genome into the cytoplasm, the RNA+ genome is reverse transcribed into a double-stranded DNA genome molecule, with an intermediate phase as a single-stranded negative DNA. The double-stranded DNA is then integrated into the cellular genome and will produce genomic RNA+ copy for the next generation of virions.
- In absence of functional VIF protein, APOBEC3s members (A3G, A3H, A3D and A3F) are able to edit the ssDNA- intermediate that produces C > U mutations **(figure 3)**. These mutations correspond to G>A mutations in the positive DNA strand.
- In the wild-type (WT) virus, VIF binds to APOBEC3 enzymes and mediates their ubiquitination by the cellular Cullin5 E3-ubiquitin ligase, followed by their proteasomal degradation. VIF efficiently degrades APOBEC3G and H haplotype II, and with less efficiency, APOBEC3D and F, via the CUL2 proteasome pathway.

In addition to the examples of HIV with the VIF protein, many other viruses have acquired mechanisms to counteract APOBEC3 activity. For example, the EBV virus encodes inhibitors for APOBEC3B [103]. Similarly, HPV16 and BK PyV have fewer occurrences of the

5’TC-3’ APOBEC3 target site in their genomes, which could allow them to resist hyper-editing by APOBEC3 enzymes [94,99].

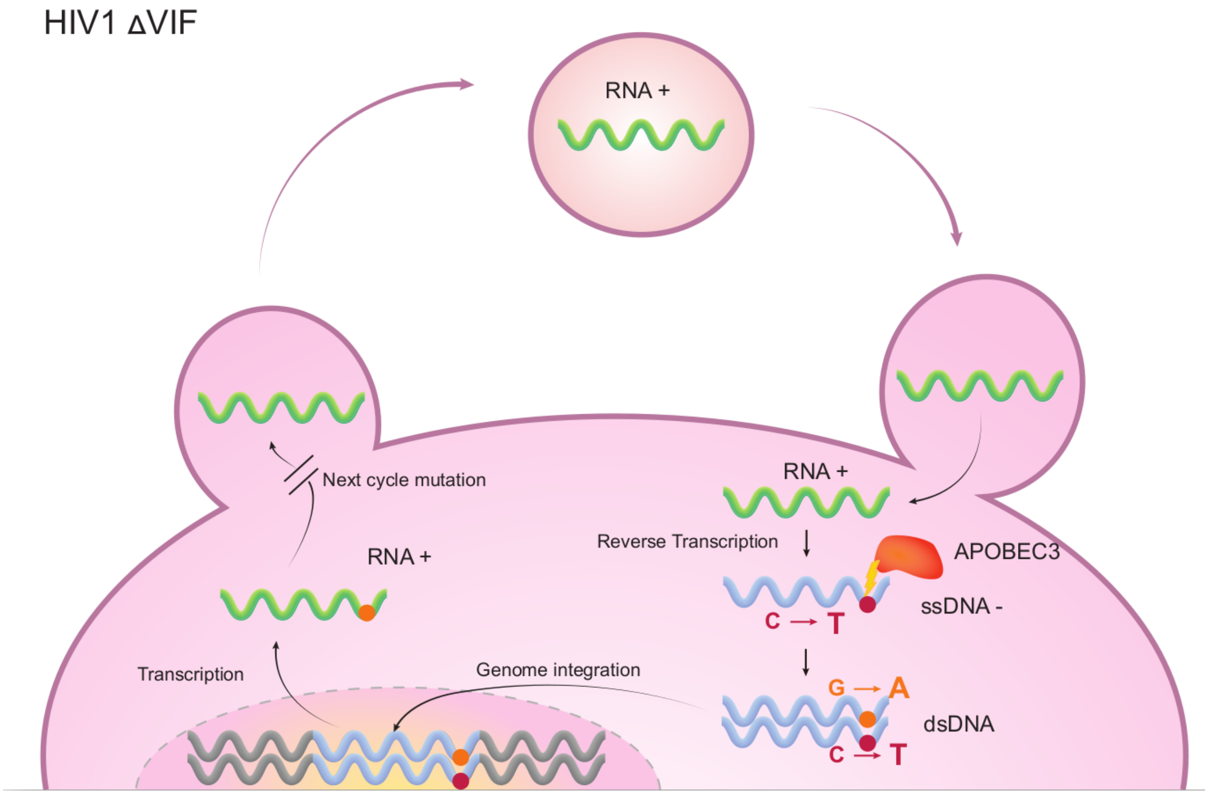


Figure 3: The APOBEC3 ssDNA- genome edition during HIV1 *ΔVIF* replication cycle. HIV-1 virions, containing RNA+ genomes, initiate a new cell infection through attachment, fusion and uncoating processes. During these steps, the RNA+ genome is reverse transcribed into ssDNA-. In *ΔVIF* HIV-1 strain, the produced ssDNA- genome undergoes editing by APOBEC3, resulting in the accumulation of C to T substitutions. Subsequent polymerization leads to the formation of a dsDNA genome. Due to base-pair complementation, the APOBEC3-induced substitutions results in G to A mutations on the positive strand of the genome. If the level of deamination is not lethal, the virus proceeds with the viral cycle by integrating the dsDNA genome, transcribing a new RNA+ genome, and packaging it into a new virion, which initiates the next cycle of infection.

2. Objectives

In addition to their role in restricting human viruses, APOBEC3 enzymes are a significant source of genome mutations in cancerous cells. This deregulation of the proteins can occur during HPV and HPyV infections. Understanding the relationship between viruses and host factors that shape their genomes may provide evidence of potential viral origins of cancer.

In this context, we initially investigated the influence of APOBEC3 on the genomes of human viruses to identify viruses targeted by these enzymes, which could be potential oncogenic viruses. As a result, we identified parvovirus B19 and adenoviruses as viruses that most probably trigger APOBEC3-mediated innate response. Going beyond this initial goal, we gained a comprehensive understanding of the broad spectrum of APOBEC3's viral targets and also elucidated editing dynamics.

The development of novel bioinformatic approaches in the field of cancer research, enabling the deciphering of different mutational processes on large datasets, led me to adapt such a pipeline for human viruses. These approaches provide a unique opportunity to explore the diverse mutational mechanisms involved in virus evolution. Furthermore, they also enable the observation of mutational processes, analogous to the APOBEC3 example that may be dysregulated by the virus and shared with cancer cells. Unlike mutation calling in human cancers, extracting virus substitutions required reconstructing ancestor sequences to serve as reference sequences. Additionally, the search for mutational signatures necessitated a large quantity of sequences from genomes exposed to diverse sources of mutations. These constraints motivated the development of a systematic automated pipeline capable of collecting virus substitutions across a wide range of viral species. This extensive collection of substitutions allows us to explore the dynamics of virus substitution landscapes. The discovery of a mirror effect within the substitution landscapes further led to the exploration of the back-and-forth substitutions.

The substantial substitution collection generated during this investigation will serve as a valuable resource for future explorations of the mutational processes involved in virus evolution, potentially connecting them to the mutational processes active in cancer cells.

3. Results

Footprint of the host restriction factors APOBEC3 on the genome of human viruses

The APOBEC3 proteins play a crucial role in the human innate immune system by specifically targeting viral genomes at 5’-TC-3’ sites to restrict the virus life cycle. One notable indication of recent or persistent exposure to APOBEC3 activity is the depletion of the 5’-TC-3’ dinucleotide target in neutral positions within the viral genome. This APOBEC3 footprint has been observed in certain examples such as HPV and PyV viruses, human retroelements, and somewhat ambiguously in the case of HIV-1. However, while a few examples of viruses targeted by APOBEC3 have already been reported, the status of the majority of human virus species remains unexplored. Until now, a comprehensive overview of the APOBEC3 footprint across all virus species has been lacking whereas such a global overview could provide a deeper and interesting understanding of this mechanism when observed on a broader scale. Therefore, our study aimed to investigate the presence of the APOBEC3 footprint in a large dataset of human virus sequences.

Through our investigation, we aimed to encompass the range of APOBEC3's activity across the majority of currently annotated viral species. This approach, based directly on the information present in virus genome sequences, has the potential to identify new targets for APOBEC3. In addition to uncovering new APOBEC3 targets, our investigation also delved into viruses where the impact of APOBEC3 editing remained uncertain. This was for example particularly the case for RNA viruses. In addition, we also conducted an in-depth analysis of the APOBEC3 footprint presence in the genome of HIV-1, which has been a topic of debate in the scientific literature. Indeed, various authors reported contradictory observations. Our analysis was able to provide new insights into these different points and helped to clarify the situation. A second interesting analysis we conducted was a comparison between different animal viruses. This analysis offered the possibility to better understand the dynamic of the APOBEC3 footprint, with notably the comparison between endemic and emergent coronaviruses. Thanks to the report of APOBEC3 footprint at the scale of virus genes, we were also interested in the exploration of local editing of virus genomes.

Considering these different aspects, we were able to give a comprehensive overview of the APOBEC3 footprint across a wide range of virus species and enhance our understanding of the cross-interactions between viruses and host cells.

PLOS PATHOGENS

RESEARCH ARTICLE

Footprint of the host restriction factors APOBEC3 on the genome of human viruses

Florian Poulain*, Noémie Lejeune, Kévin Willemart, Nicolas A. Gillet*

Namur Research Institute for Life Sciences (NARILIS), Integrated Veterinary Research Unit (URVI), University of Namur, Namur, Belgium

* poulainflorian@gmail.com (FP); nicolas.gillet@unamur.be (NAG)

Abstract

APOBEC3 enzymes are innate immune effectors that introduce mutations into viral genomes. These enzymes are cytidine deaminases which transform cytosine into uracil. They preferentially mutate cytidine preceded by thymidine making the 5'TC motif their favored target. Viruses have evolved different strategies to evade APOBEC3 restriction. Certain viruses actively encode viral proteins antagonizing the APOBEC3s, others passively face the APOBEC3 selection pressure thanks to a depleted genome for APOBEC3-targeted motifs. Hence, the APOBEC3s left on the genome of certain viruses an evolutionary footprint.

The aim of our study is the identification of these viruses having a genome shaped by the APOBEC3s. We analyzed the genome of 33,400 human viruses for the depletion of APOBEC3-favored motifs. We demonstrate that the APOBEC3 selection pressure impacts at least 22% of all currently annotated human viral species. The *papillomaviridae* and *polyomaviridae* are the most intensively footprinted families; evidencing a selection pressure acting genome-wide and on both strands. Members of the *parvoviridae* family are differentially targeted in term of both magnitude and localization of the footprint. Interestingly, a massive APOBEC3 footprint is present on both strands of the B19 erythroparvovirus; making this viral genome one of the most cleaned sequences for APOBEC3-favored motifs. We also identified the endemic *coronaviridae* as significantly footprinted. Interestingly, no such footprint has been detected on the zoonotic MERS-CoV, SARS-CoV-1 and SARS-CoV-2 coronaviruses. In addition to viruses that are footprinted genome-wide, certain viruses are footprinted only on very short sections of their genome. That is the case for the *gamma-herpesviridae* and *adenoviridae* where the footprint is localized on the lytic origins of replication. A mild footprint can also be detected on the negative strand of the reverse transcribing HIV-1, HIV-2, HTLV-1 and HBV viruses.

Together, our data illustrate the extent of the APOBEC3 selection pressure on the human viruses and identify new putatively APOBEC3-targeted viruses.

OPEN ACCESS

Citation: Poulain F, Lejeune N, Willemart K, Gillet NA (2020) Footprint of the host restriction factors APOBEC3 on the genome of human viruses. PLoS Pathog 16(8): e1008718. https://doi.org/10.1371/journal.ppat.1008718

Editor: Paul Francis Lambert, University of Wisconsin Madison School of Medicine and Public Health, UNITED STATES

Received: March 2, 2020

Accepted: June 19, 2020

Published: August 14, 2020

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.ppat.1008718

Copyright: © 2020 Poulain et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: S1 Table, S2 Table and S3 Table files are available from the Dryad database via the link: https://datadryad.org/stash/

Author summary

APOBEC3 cytidine deaminases are enzymes that restrict many viruses by mutating their genomes. In doing so, they exert a selection pressure and leave onto these viruses an

share/VikUp3LVVUInFa88SbRwP4JGQuPpehlvcqdZSBevHwY.

Funding: This study was supported by F.R.S.-FNRS (Fonds de la Recherche Scientifique de Belgique, https://www.frs-fnrs.be) grant CDR No. 31270116 and by the University of Namur. F.P. is a PhD fellow supported by a NARILIS (https://www.narilis.be) research grant and by F.R.S.-FNRS Télévie (https://www.frs-fnrs.be) grant PDR-TLV No. 34972507. N.L. is a PhD fellow supported by FRIA (Fonds pour la Formation à la Recherche dans l'Industrie et dans l'Agriculture, https://www.frs-fnrs.be) grant No. 31454280. Computational resources have been provided by the Consortium des Équipements de Calcul Intensif (CÉCI), funded by the F.R.S.-FNRS under grant No. 2.5020.11 and by the Walloon Region (https://recherche-technologie.wallonie.be). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

evolutionary footprint. In addition to their antiviral role, APOBEC3s have also been identified as a major source of mutations in cancer, wrongly targeting the cell genome. For example, high-risk papillomaviruses, whose viral genomes carry an APOBEC3 footprint, indirectly promote cell transformation due to the sustained APOBEC3 mutagenic activity. In this study, we perform for the first time a general screening for the APOBEC3 footprint in all currently annotated human viruses. We show that approximately 22% of human viral species are shaped by the APOBEC3 selection pressure and extend the list of APOBEC3-footprinted viruses with adenoviruses and autonomous parvoviruses. Knowing which virus is restricted by the APOBEC3 mutagenic activity could lead to the identification of new viruses associated with cancer.

Introduction

The APOBEC3s (apolipoprotein B mRNA-editing enzyme, catalytic subunit 3 or A3s) are innate immune effectors restricting many exogenous viruses and endogenous retroelements [1–3]. The human genome encodes for seven A3 genes (namely A3A, B, C, D, F, G and H), with several spliced transcripts and allelic variants for each. These seven genes originate from gene duplications and rearrangements that have occurred during mammalian evolution and represent a classic example of the virus-host arms race [4]. The A3s are cytidine deaminases that convert cytosine to uracil present in single stranded DNA or RNA. Such editing on viral genomes generally results in C to T (or U) transition after replication of the genome. The A3s preferably mutate cytosine in a 5'TC dinucleotide context with the notable exception of A3G that favors a C before the mutated C [5].

The antiviral activity of the A3s has been first reported for the reverse transcribing viruses HIV-1 (human immunodeficiency virus-1), HTLV-1 (human T-lymphotropic virus-1) and HBV (Hepatitis B virus) [6–8]. Editing occurs during reverse transcription on the negative strand leading to G to A mutations on the positive strand [9–14]. Subsequently, A3-introduced mutations have been reported on various double-stranded DNA (EBV for Epstein-Barr virus, HSV-1 for herpes simplex virus-1, α-HPVs for alpha human papillomaviruses, BK PyV for BK polyomavirus), single-stranded DNA (TT virus) and single-stranded RNA (HCoV-NL63 for human coronavirus NL63) viruses [15–19]. It is important to note that the antiviral action of the A3 proteins is not based solely on their deaminase activity. Deaminase-independent restriction has been demonstrated against endogenous retroelements, reverse transcribing viruses, adeno-associated viruses and many RNA viruses (HCV for hepatitis C virus, RSV for respiratory syncytial virus, HCoV-NL63, mumps virus and measles virus) [20–25].

The co-evolution between virus and host leads to the selection of viral proteins capable of countering the restriction effect of A3s. HIV-1 encodes for the Vif protein which promotes A3G degradation [26]. HTLV-1 evades A3G restriction by excluding A3G from virions [27]. BORF2 protein from EBV inhibits A3B deaminase activity and re-locates it far from viral replication centers [28]. Besides these active viral mechanisms targeting A3 proteins, some viruses have evolved passive strategies to limit A3 restriction. One such strategy is the depletion of A3-favored motifs from the viral genome. By repetitive exposure to A3 activity, non-lethal mutations can accumulate in the genomic sequence leading to the under-representation of the motifs favored by A3s. This under-representation of A3-favored motifs is called A3 evolutionary footprint. Thus, the 5'TC dinucleotide motif is under-represented in the genome of α-HPVs and BK polyomavirus [17,29]. Similarly, 5'TC and 5'CC (favored by A3G) motifs are under-represented in the negative strand of LTR (long terminal repeat) and non-LTR endogenous retroelements [30]. Conflicting data have been reported regarding evidence of an A3 footprint on the HIV-1 genome [31,32]. Recently, an under-representation of certain A3 motifs has been shown in the genome of the γ-herpesviruses EBV and KSHV (Kaposi sarcoma herpes virus) [33]. Finally, codon usage in coronaviruses suggests that cytosine deamination is an important biochemical force which shapes the evolution of these viruses [34].

Different bioinformatics approaches have been used to search for evidence of an A3 evolutionary footprint in viral genomes [17,29–33,35,36]. In this study, we adapted and extended the Warren *et al*. approach [29] to carry out a general screening for the A3 footprint of the genomes of all currently annotated human viruses. We first demonstrate the sensitivity and specificity of our approach: i. an A3 footprint is detected in viruses which have already been shown to be depleted for A3-targeted motifs; ii. no A3 footprint has been reported in viruses from animals lacking A3 genes. We showed that as much as 22% of currently annotated human viral species are shaped by the A3 selection pressure. We confirmed previous reports showing that papillomaviruses and polyomaviruses are generally strongly footprinted by the A3s. Among the most A3-footprinted viruses, we identified autonomous parvoviruses and in particular the B19 erythroparvovirus as deeply cleansed for A3-favored motifs. Importantly, we showed that the A3 footprint observed in coronaviruses is limited to the endemic viruses and absent from the zoonotic MERS-CoV (middle east respiratory syndrome coronavirus), SARS-CoV-1 (severe acute respiratory syndrome coronavirus) and SARS-CoV-2 viruses. We also carried out a gene-specific A3 footprint search and identified local footprint on EBV and adenoviruses consistent with genome targeting during the initiation of replication.

Results

Definition of the A3 footprint

The A3 footprint is defined as the under-representation of A3-targeted motifs. Different approaches have been devised to estimate the over/under-representation of a given motif [17,29–33,35,36].

Firstly, and because most of the A3 proteins (A3A, A3B, A3C, A3F and A3H) favors deamination of cytosine to uracil in a 5'TC dinucleotide context, we have chosen to look for the under-representation of the 5'TC motif. Moreover, as originally developed by Warren *et al*., we refined our analysis by distinguishing the position of the 5'TC motif relative to the coding sequence. Namely, we differentiate three K-mers containing the TC motif; one K-mer having the C in the first position of the codon (NNTCNN), one K-mer having the C in the second position of the codon (TCN) and one K-mer having the C in the third position of the codon (NTC). A3-introduced deamination of cytosine in viral genome produces an uracil that can be fixed in the form of thymidine after genome replication. This transition will have different impacts depending on the position of the mutated C. The C to T mutation will be non-synonymous if the C is at the first or second position of the codon (Fig 1A, NNTCNN and TCN K-mers). However, if the mutated C occupies the third position of the codon, the C to T mutation will always be synonymous (Fig 1A, NTC K-mer). Therefore, A3-driven natural selection should deplete more intensively NTC codons than TCN or NNTCNN motifs (as in those cases the C to U mutation will impact the encoded amino acid). Obviously, A3 editing can also target the template strand where a C to T mutation will translate into G to A transition in the coding strand. Again, this transition will have different impacts depending on the position of the mutated G. The G to A mutation will be non-synonymous if the G is at the first or second position of the codon (Fig 1A, GAN and NGA K-mers). However, if the mutated G occupies the third position of the codon the mutation will be most of the time synonymous (Fig 1A, NNGANN K-mer). Because synonymous mutations are presumably more likely to be retained

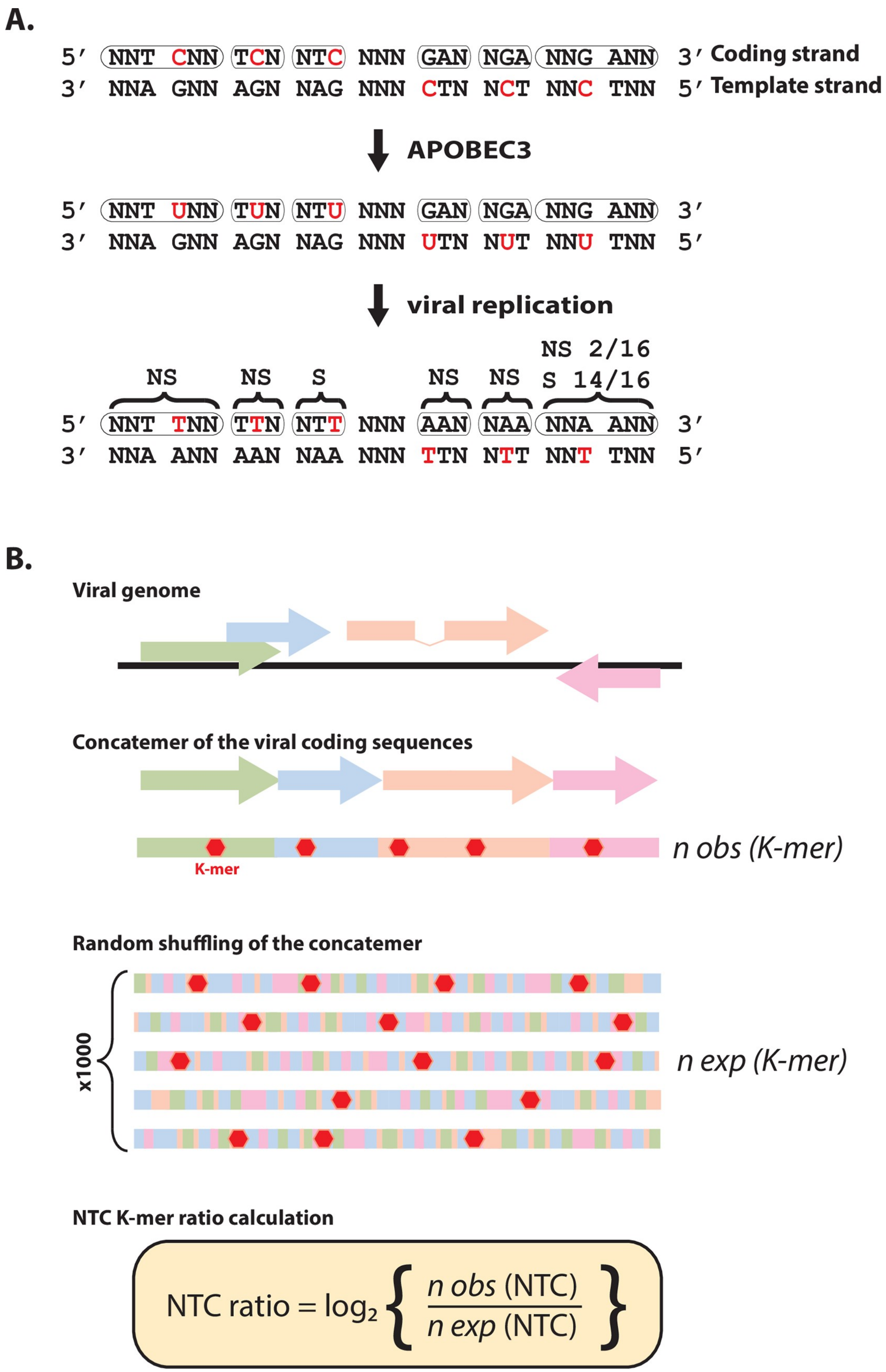

A.
5′ NNT CNN TCN NTC NNN GAN NGA NNG ANN 3′ Coding strand
3′ NNA GNN AGN NAG NNN CTN NCT NNC TNN 5′ Template strand
APOBEC3
5′ NNT UNN TUN NTU NNN GAN NGA NNG ANN 3′
3′ NNA GNN AGN NAG NNN UTN NUT NNU TNN 5′
viral replication
NS 2/16
S 14/16
NS NS S NS NS
5′ NNT TNN TTN NTT NNN AAN NAA NNA ANN 3′
3′ NNA ANN AAN NAA NNN TTN NTT NNT TNN 5′
B.
Viral genome
Concatemer of the viral coding sequences
K-mer
n obs (K-mer)
Random shuffling of the concatemer
x1000
n exp (K-mer)
NTC K-mer ratio calculation
$\text{NTC ratio} = \log_2 \left\{ \frac{n\ obs\ (\text{NTC})}{n\ exp\ (\text{NTC})} \right\}$

Fig 1. Definition and estimation method of the A3 footprint. A. A3-induced cytidine deamination followed by viral replication leads to C to T mutations (in red). Most of the A3 enzymes favor deamination in a 5'TC context. The TC dinucleotide motif is depicted in three possible codon contexts on both coding and template strand. Depending on the position of the mutated C, the C to T transition can be synonymous (S) or non-synonymous (NS). Proportion of S and NS mutations is reported when the two types of mutation can be produced. Because synonymous mutations are more likely to be retained, the A3 footprint can be defined as the depletion of the NTC and/or NNGANN codons. B. Depletion or enrichment of a given K-mer (e.g. NTC) is calculated as the log2 ratio of the observed occurrence of that K-mer (n obs) divided by its expected occurrence (n exp). For each human virus, its coding sequences (colored arrows) are concatenated to generate a synthetic coding genome from which we obtain the n obs of a given K-mer. The synthetic coding genome is then shuffled a thousand times and the n exp is calculated as the average count for that K-mer.

https://doi.org/10.1371/journal.ppat.1008718.g001

than non-synonymous, we define the A3 footprint (with the exception of A3G-induced footprint) as the depletion of NTC or NNGANN K-mers. Calculation of observed vs expected K-mer ratio has been adapted from Warren *et al*. and detailed in the material and methods section. Briefly, a synthetic coding genome was generated by concatenating the different coding sequences allowing the counting of the occurrence of a given K-mer (Fig 1B, n obs (K-mer)). Each synthetic coding genome has been randomly shuffled a thousand times. The expected count is calculated as the average of the occurrences of this K-mer over the thousand iterations (Fig 1B, n exp (K-mer)). A negative K-mer ratio indicates depletion of that K-mer. The observed vs expected ratio of the NTC K-mer will be compared to those of the NNTCNN and TCN K-mers. Similarly, the observed vs expected ratio of the NNGANN K-mer will be compared to those of the GAN and NGA K-mers. Of note, for the sake of clarity and simplicity, we have chosen to stick with a DNA genetic code throughout the manuscript. The reader will read a T as a U in the context of RNA viruses.

Secondly, and because A3G favors deamination of cytidine when preceded by another cytidine, we have chosen to look for the under-representation of the 5'CC motif. Following the same rationale, A3G-footprinted viruses should display to a stronger depletion of NCC codons compared to CCN or NNCCNN motifs (or a depletion of NNGGNN motifs versus the GNN and NGG motifs) (S1A Fig). The NCC ratio will be compared to those of the NNCCNN and CCN K-mers. Similarly, the NNGGNN ratio will be compared to those of the GGN and NGG K-mers.

An A3 footprint is detected in viruses known to induce A3 expression

Given that the BK polyomavirus has recently been demonstrated to induce A3B expression and that it was depleted in 5'TC motifs [17], we considered this virus as a positive control to validate our approach. Fig 2A shows a strong depletion of the NTC motif. On the contrary, the dinucleotide 5'TC in the context of the NNTCNN and TCN K-mers are neither over- nor under-represented. The significant differences between the NTC ratio and the TCN (or NNTCNN) ratios reveal that the frequency of the TC motif is dependent on its position within a codon. It suggests that NTC depletion can be tolerated because of the degeneration of the genetic code. Importantly, the absence of NNTCNN and TCN depletion infers that this virus is still vulnerable to A3 restriction because deamination of those cytidines would lead to changes in the amino acid sequence. Fig 2B highlights the fact that NTC depletion is genome-wide and can be observed in each gene. Similarly, the 5'GA motif is significantly less abundant in the NNGANN context than in the GAN or NGA codons. The 5'GA depletion is also genome-wide (Fig 2B). We read these observations as the consequence of an A3 activity acting on both coding and template strands.

Extending our analysis to other polyomaviruses shows that both JC polyomavirus and Merkel cell polyomavirus bear an A3 footprint; footprint that is also genome-wide and present on both strands (Fig 2C–2F). The magnitude of the A3 footprint appears lighter on the Merkel

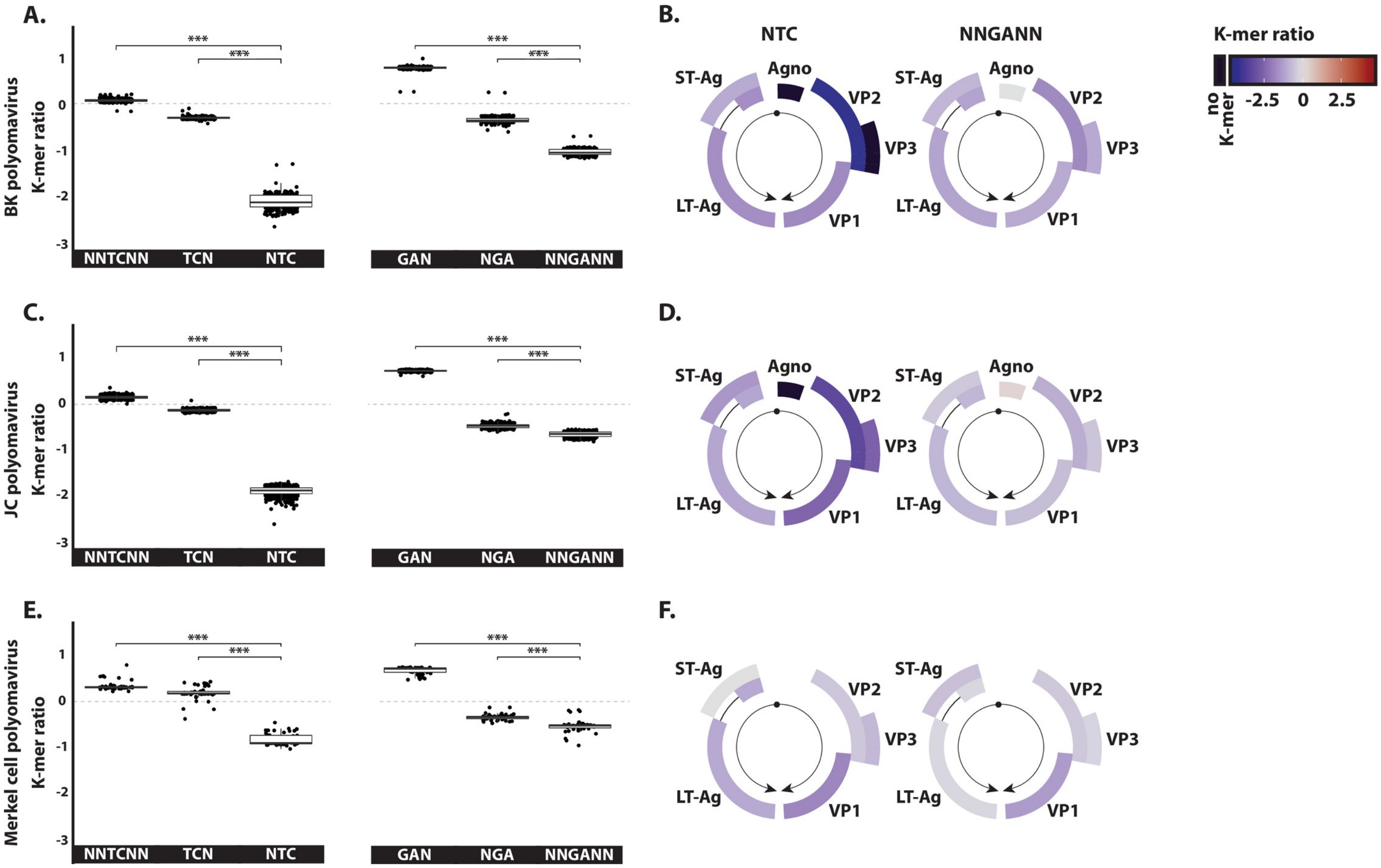


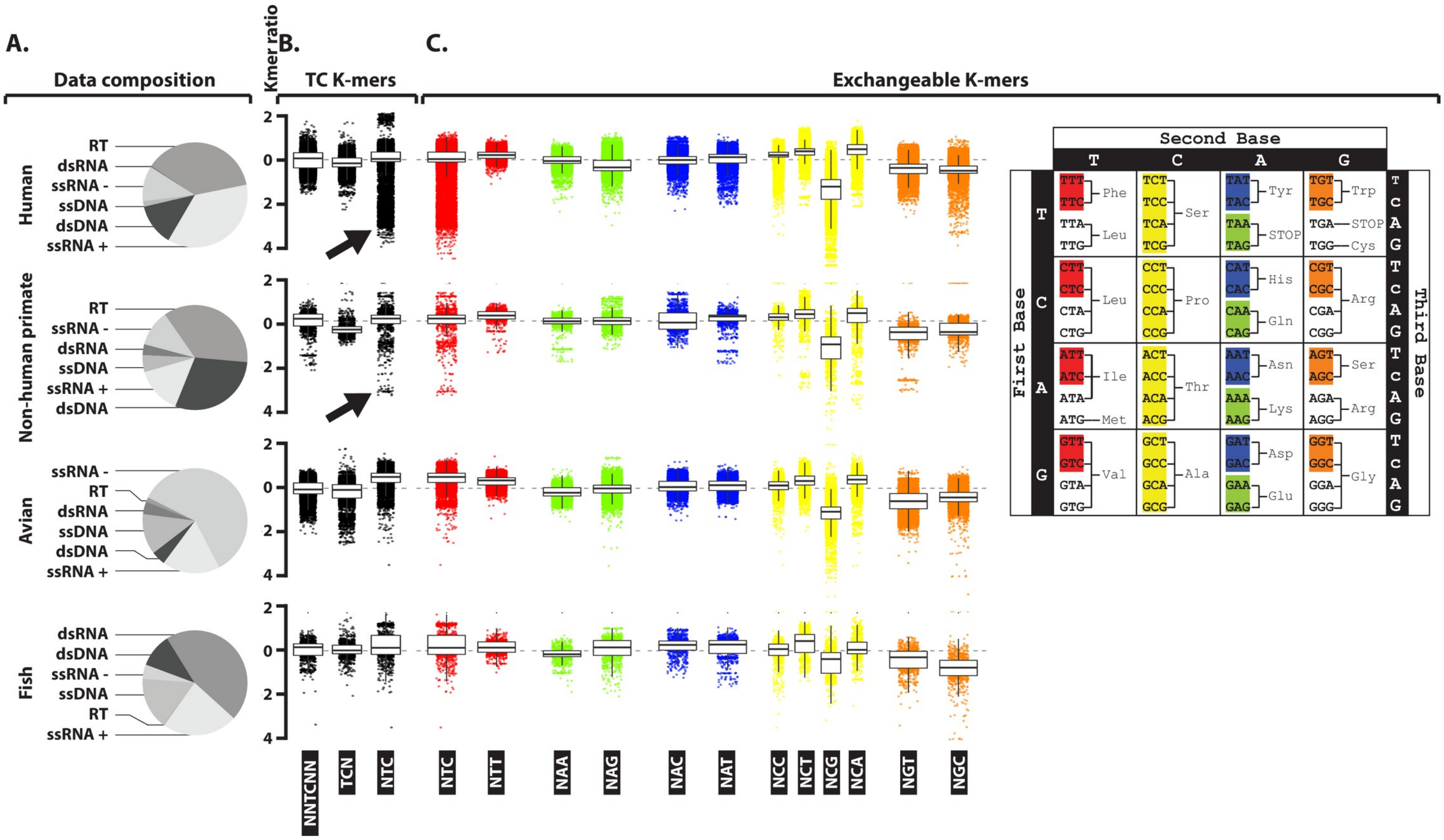
Fig 2. Evidence of an A3 footprint in human polyomaviruses. The observed/expected ratios of TC dinucleotide at various codon positions and on both strand (i.e. NNTCNN, TCN, NTC, GAN, NGA and NNGANN) were calculated for BK polyomavirus (panels A-B), JC polyomavirus (panels C-D) and Merkel cell polyomavirus (panel E-F). For the dot plots, each point stands for a unique full-length viral genome. Median and quartile are depicted by a boxplot. P-values were calculated by Student's unpaired, two-tailed t-test (NS for not significant, * p< 0.05, ** p< 0.01, *** p< 0.001). Panels B, D and F illustrate NTC and NNGANN ratios for the different viral coding sequences. A colored scale with increasing shades of blue indicating depletion and increasing shades of red indicating enrichment. Replication origin is illustrated by a black dot and gene transcriptional orientation is symbolized by black arrows.

https://doi.org/10.1371/journal.ppat.1008718.g002

cell polyomavirus. Analysis of a larger number of polyomavirus species shows a stronger NTC depletion in *beta-polyomaviridae* by comparison to *alpha-polyomaviridae* (S2 Fig). The *delta-polyomaviridae* are affected by an A3 footprint of variable intensity depending on the species considered (S2 Fig).

The A3 footprint is limited to viruses infecting hosts endowed with A3 genes

We showed that our approach sensitively detects an NTC depletion in viruses known to promote A3 expression. We then wondered to what extent this depletion is widespread among viruses. We therefore downloaded genomic sequences of 33,400 human, 1,397 non-human primate, 9,160 avian, and 570 fish viruses and calculated NNTCNN, TCN and NTC ratios (Fig 3B). With NNTCNN, TCN and NTC median ratios close to zero, most of the sequences are not A3-footprinted (Fig 3B, box plots). However, the distribution of the NTC ratios is bimodal in human and non-human primate viruses with a subpopulation of sequences strongly depleted for the NTC codon (Fig 3B, arrows and 4A red part of the distribution plot).



First Base	T	C	A	G	Third Base
T	TTT Phe TTC Phe TTA Leu TTG Leu	TCT Ser TCC Ser TCA Ser TCG Ser	TAT Tyr TAC Tyr TAA STOP TAG STOP	TGT Trp TGC Trp TGA—STOP TGG—Cys	T C A G
C	CTT Leu CTC Leu CTA Leu CTG Leu	CCT Pro CCC Pro CCA Pro CCG Pro	CAT His CAC His CAA Gln CAG Gln	CGT Arg CGC Arg CGA Arg CGG Arg	T C A G
A	ATT Ile ATC Ile ATA Ile ATG—Met	ACT Thr ACC Thr ACA Thr ACG Thr	AAT Asn AAC Asn AAA Lys AAG Lys	AGT Ser AGC Ser AGA Arg AGG Arg	T C A G
G	GTT Val GTC Val GTA Val GTG Val	GCT Ala GCC Ala GCA Ala GCG Ala	GAT Asp GAC Asp GAA Glu GAG Glu	GGT Gly GGC Gly GGA Gly GGG Gly	T C A G

Fig 3. A sub-population of Human and non-human primate viruses is depleted in NTC codon. Four datasets including Human viruses (n = 33,400), non-human primate viruses (n = 1,397), avian viruses (n = 9,160) and fish viruses (n = 570) have been analyzed for their observed/expected K-mer ratios. A. The composition of each data set regarding the breakdown into viral groups is illustrated by pie charts. B. The observed/expected ratios of TC dinucleotide at various codon positions for Human, non-human primate, bird and fish viruses are illustrated by dot plots (one point represents one unique viral sequence). C. K-mers are grouped and colored according to their capacity to encode a common amino-acid (in red for NTT/C, in yellow for NCC/G/T/A, in orange for NGT/C, in blue for NAC/T and in green for NAA/G).

https://doi.org/10.1371/journal.ppat.1008718.g003

Crucially, such footprinted subpopulation is not detected in avian or fish viruses (Fig 3B). Hence, the absence of a detectable A3 footprint in avian and fish viral sequences is consistent with the restriction of the A3 genes family to the mammalian class [37]. It is worth mentioning that each Baltimore's group is represented in the human, non-human primate, avian and fish viral sequence data sets, albeit in different proportions (Fig 3A).

Due to the redundancy of the genetic code, different codons can encode for the same amino acid. The third position of the codons is highly reiterated (redundant) and allows synonymous substitutions. This is notably the case for the NTC and NTT codons where the C or T at the third position are perfectly interchangeable (Fig 3C, colored in red). While we observe a subpopulation of sequences depleted for NTC, such depletion is not observed for NTT. Hence, the NTC depletion cannot simply be explained by the under-representation of an amino acid. The pair NTC/NTT is not the only interchangeable pair. The NAA/NAG, NAC/NAT, NGT/NGC duos and the NCC/NCT/NCG/NCA quartet are also interchangeable. The distribution of NTC ratios remains the sole being bimodal with a subpopulation of strongly depleted sequences. The general NCG depletion (monomodal distribution with a median significantly less than zero) is the result of the well characterized CG dinucleotide under-representation in viral genomes [38]. This depletion is shared in all viral datasets while the NTC depletion is specific to a subpopulation of human and non-human primate viral sequences.

By breaking down the human viruses into their respective Baltimore's group (S3 Fig), we observed that NTC depletion is not present in reverse transcribing nor in negative sense single strand RNA viruses. A mild general depletion is present in double strand RNA viruses. Importantly, a strong general depletion can be observed in double strand DNA viruses. Finally, in single strand DNA and positive sense single strand RNA viruses, certain specific viral sequences appear also significantly depleted. We also observed a mild general NCC depletion in single strand DNA and double strand RNA viruses, justifying further investigation for a possible A3G-induced footprint (S1 Fig). No NCC depletion is observed in double strand DNA, single strand RNA nor in reverse transcribing viruses.

Screening for human viruses' genomes marked by an A3 footprint

In order to identify A3-footprinted viruses, we detailed the NTC and NNGANN ratios for 870 human viral species (Fig 4A). We observed that the NTC and NNGANN distributions are bimodal with a subpopulation of depleted sequences in each case. We consider a viral species as footprinted by A3s when its NTC or NNGANN median ratio is inferior to the population median by at least two times the standard deviation. Hence, about 17% of the viral species are depleted for NTC (143 species over 870) and about 16% are depleted for NNGANN motifs (136 species over 870). In total, 175 species (22%) present an A3 footprint on either one or both strands. This subgroup is essentially composed by double-stranded DNA viruses with numerous *alpha-*, *beta-* and *gamma- papillomaviridae* (αPV, βPV and γPV) but also *beta-polyomaviridae* (BKPyV, JCPyV, KIPyV, WUPyV and HPyV9) and the delta-polyomavirus MWPyV (Fig 4B). These viruses show a strong depletion for both the NTC and NNGANN motifs by comparison to NNTCNN/TCN and GAN/NGA (Fig 4B). Of note, NTC depletion generally goes with a mild to a significative NTT enrichment (S4 Fig). In the strongly NTC-depleted viruses HPV16, HPV18 and HPV31, a TC depletion is also observed in the non-coding region of the genome regardless of the analyzed motif (S5 Fig). To recapitulate, the A3 footprint on the *papillomaviridae* and *polyomaviridae* is genome-wide and on both strands (Fig 2, Fig 4B and S6 Fig). Importantly, we also identified the single-stranded DNA virus erythroparvovirus B19 and the single-stranded RNA virus HKU1 beta-coronavirus as strongly footprinted by A3s (Fig 4B, highlighted). We will further detail the A3 footprint of these viruses in the following sections.

In order to specifically look for A3G-footprinted viruses, we calculated the NCC and NNGGNN ratios for 870 human viral species (S1B Fig). We observed that the NCC and NNGGNN ratios are mostly narrowly distributed around the zero value. Viruses depleted for NCC are generally single stranded DNA viruses (S1B Fig). However, in many of them we observed a concomitant depletion of the NNCCNN motif casting doubts on the causal link between the observed NCC depletion and A3G editing (S1C Fig).

B19 erythroparvovirus genome bears a strong A3 footprint

One of the most A3-footprinted virus is the B19 erythroparvovirus. Among the *parvoviridae* family, the TC ratio analysis showed a strong NTC depletion for erythroparvovirus B19 and to a lesser extent for parvovirus 4 and bocavirus 4 (Fig 5A). We also observed a significant depletion of the NNGANN K-mer for each autonomous parvovirus (erythroparvovirus, parvovirus 4 and bocaparvovirus 1, 2, 3 and 4) (Fig 5A). Thus, autonomous parvoviruses appear to be footprinted either on both strands as for the erythroparvovirus, the parvovirus 4 and bocaparvovirus 4 or only on the template strand for the bocaparvovirus 1, 2 and 3. It is interesting to note that the AAV-1 (adeno-associated dependoparvovirus 1) shows a totally different pattern with even a slight enrichment in NTC codons. This dependoparvovirus is not footprinted by the A3s.

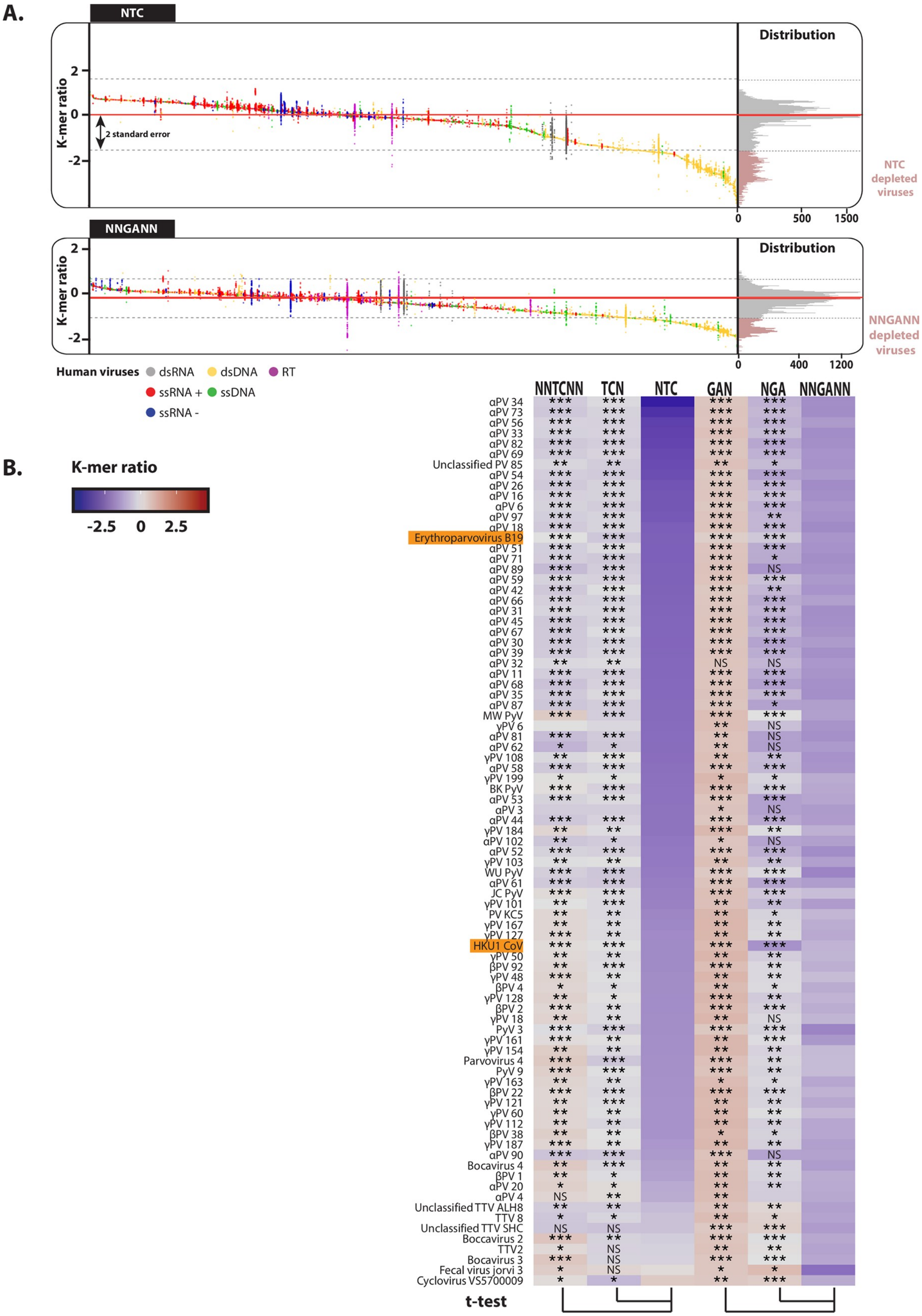

A.
NTC
K-mer ratio
2
0
-2
2 standard error
Distribution
0
500
1500
NTC depleted viruses
NNGANN
K-mer ratio
2
0
-2
Distribution
0
400
1200
NNGANN depleted viruses
Human viruses ● dsRNA ● dsDNA ● RT
● ssRNA + ● ssDNA
● ssRNA -
B.
K-mer ratio
-2.5 0 2.5
NNTCNN TCN NTC GAN NGA NNGANN
αPV 34
αPV 73
αPV 56
αPV 33
αPV 82
αPV 69
Unclassified PV 85
αPV 54
αPV 26
αPV 16
αPV 6
αPV 97
αPV 18
Erythroparvovirus B19
αPV 51
αPV 71
αPV 89
αPV 59
αPV 42
αPV 66
αPV 31
αPV 45
αPV 67
αPV 30
αPV 39
αPV 32
αPV 11
αPV 68
αPV 35
αPV 87
MW PyV
γPV 6
αPV 81
αPV 62
γPV 108
αPV 58
γPV 199
BK PyV
αPV 53
αPV 3
αPV 44
γPV 184
αPV 102
αPV 52
γPV 103
WU PyV
αPV 61
JC PyV
γPV 101
PV KC5
γPV 167
γPV 127
HKU1 CoV
γPV 50
βPV 92
γPV 48
βPV 4
γPV 128
βPV 2
γPV 18
PyV 3
γPV 161
γPV 154
Parvovirus 4
PyV 9
γPV 163
βPV 22
γPV 121
γPV 60
γPV 112
βPV 38
γPV 187
αPV 90
Bocavirus 4
βPV 1
αPV 20
αPV 4
Unclassified TTV ALH8
TTV 8
Unclassified TTV SHC
Boccavirus 2
TTV2
Bocavirus 3
Fecal virus jorvi 3
Cyclovirus VS5700009
t-test

Fig 4. Search for the A3-footprinted human viruses. A. The NTC and NNGANN observed/expected ratios for 33,400 human viruses' genomes (from 870 unique species) were calculated, grouped by species and colored according to the Baltimore classification. Each point represents a unique viral genome. Abundance distribution is depicted by a histogram on the right-hand side of the panel. Viral species with an NTC or NNGANN ratio below two times the standard deviation (dotted grey line) from the population median (red line) are the putative A3-footprinted viruses. B. The observed/expected ratios of TC dinucleotide at various codon positions and on both strands (i.e. NNTCNN, TCN, NTC, GAN, NGA and NNGANN) were calculated for the putative A3-footprinted viral species and depicted by a heatmap. A colored scale with increasing shades of blue indicating depletion and increasing shades of red indicating enrichment. P-values were calculated by Student's unpaired, two-tailed t-test (NS for not significant, * p< 0.05, ** p< 0.01, *** p< 0.001). (PV stands for papillomavirus, PyV for polyomavirus).

https://doi.org/10.1371/journal.ppat.1008718.g004

Detailed analysis of the erythroparvovirus B19 sequences shows a nearly complete NTC cleansing along the whole genome (Fig 5B, red marks). On the contrary, NTT codons are distributed all along (Fig 5B, green marks). Some NTC codons remain present in a short, discrete section of the NS1 gene. This region also encodes for the 7.5k protein in another coding frame. Hence, the remaining TC motifs in the NS1 gene are TCN codon context in the 7.5k protein

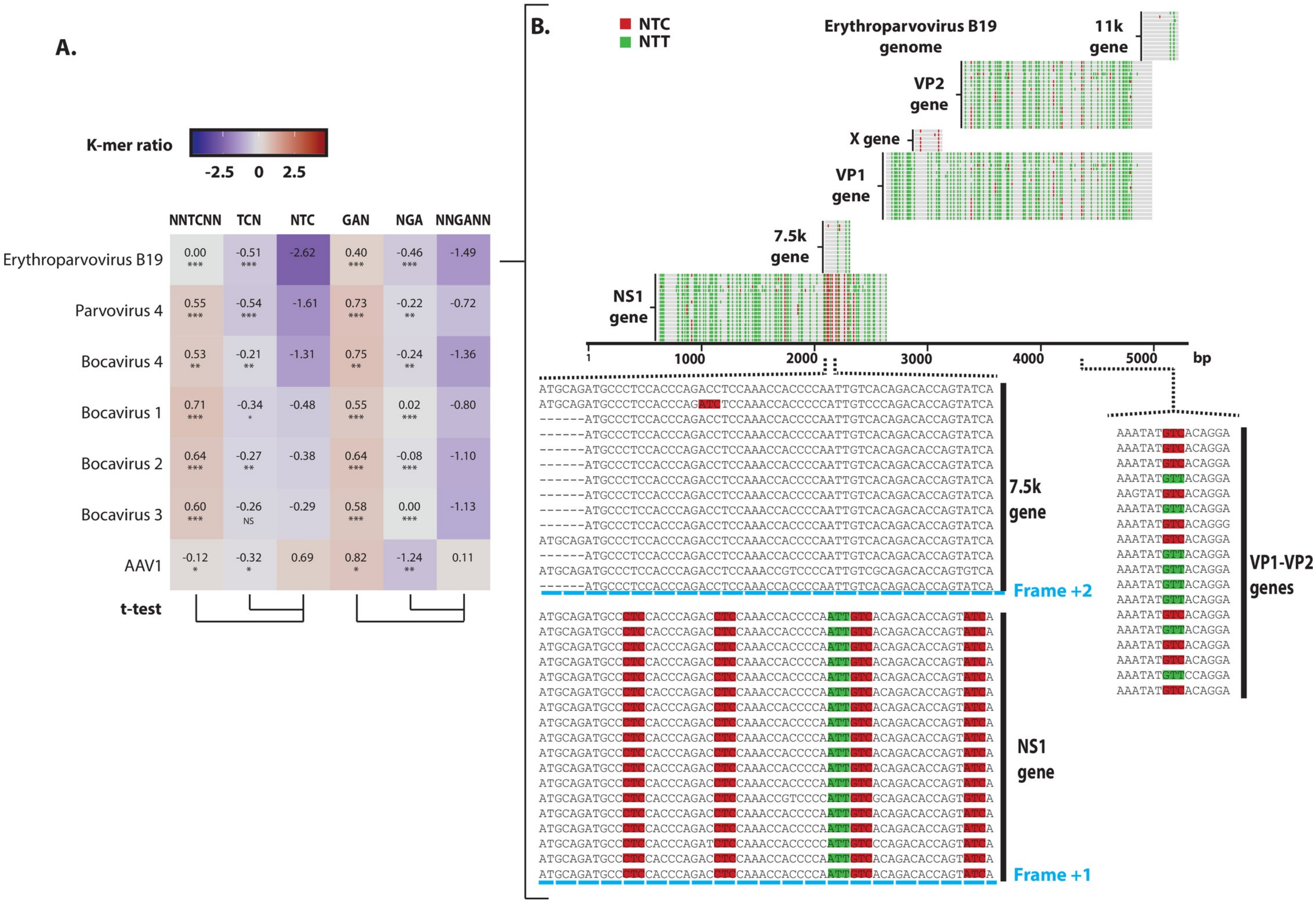


Fig 5. Intensive A3 footprint on both strands of the B19 Erythroparvovirus genome. A. The observed/expected ratios of TC dinucleotide at various codon positions for the B19 Erythroparvovirus were compared to those of the other human members of the *parvoviridae* family and depicted by a heatmap. A colored scale with increasing shades of blue indicating depletion and increasing shades of red indicating enrichment. P-values were calculated by Student's unpaired, two-tailed t-test (NS for not significant, * p< 0.05, ** p< 0.01, *** p< 0.001). B. Coding sequences (NS1, 7.5k, VP1, X, VP2 and 11k) from 18 full-length B19 erythroparvirus were depicted by grey lines overlaid by red marks to symbolize NTC and green marks to position NTT codons. Zoom-in detailed a 60 bp-long sequence from the NS1 and 7.5k genes (from nucleotide 1723 to 1783). A second zoom-in detailed a 15 bp-long sequence from the VP1-VP2 genes (from nucleotide 3973 to 3987).

https://doi.org/10.1371/journal.ppat.1008718.g005

gene. The mutation of those TCs would introduce non-synonymous mutation in the 7.5k protein. This probably explains the conservation of those TC motifs.

Moreover, among the 18 sequences illustrated in Fig 5B, some locations harbor a mix of NTC and NTT codons, suggesting that C to T transition is still an active process (zoom for VP1-VP2 sequence). An A3 footprint can also be observed in the template strand as shown by the NNGANN depletion (Fig 5A and S7 Fig). These observations show that the B19 erythroparvovirus is submitted to an ongoing and strong A3 selection pressure acting on both strands of the virus.

Human endemic coronaviruses but not zoonotic coronaviruses carry an A3 footprint

By comparing the NTC ratio to the NNTCNN and TCN ratios, we observed a common NTC depletion in the NL63, 229E, HKU1 and OC43 coronaviruses; the HKU1-CoV being the most strongly deleted in NTC codons (Fig 6). In coronaviruses, all viral genes are encoded by the positive strand. In other words, the coding strand of each gene is on the positive strand. Therefore, the depletion of NTC codons is indicative of an A3 activity on the positive strand. These observations corroborate the *in vitro* detection of a soft rate of A3C, A3F and A3H editing on the NL63 genome and the NNU/NNC codon bias previously reported for the HKU1 coronavirus [19,34].

Next, we investigated the presence of an A3 footprint on the template strand (corresponding to the negative strand in coronaviruses) by comparing the 5'GA ratios in different codon contexts. However, we did not observe a progressive depletion of the GA motif (i.e. NNGANN ratio $<$ NGA ratio $\leq$ GAN ratio) which would be expected in the presence of a GA to AA mutational pressure. For that reason, we cannot rule on the presence of an A3 footprint on the negative strand (Fig 6).

Finally, unlike endemic viruses, the zoonotic viruses MERS-CoV, SARS-CoV-1 and SARS-CoV-2 and their animal ancestors camel-MERS, bat-MERS and bat-SARS are not depleted for NTC codons (Fig 6).

Looking for an A3 footprint at the gene level

Since a non-random distribution for A3 mutations has been reported for some viruses, we then looked for spatially circumscribed A3 footprint; i.e. A3 footprint limited to certain viral genes To limit our screening on genes which are depleted compared to the whole genome, we subtracted to the genic K-mer ratio, the corresponding genomic K-mer ratio to define the differential ratio for NTC and NNGANN K-mers (Fig 7A). In other words, we looked for viruses harboring local A3 footprint amongst an otherwise non-footprinted genome. Thus, differences between genic and genomic K-mer ratios were calculated for 252,766 viral genes. Fig 7B shows the viral genes having an NTC (or NNGANN) differential ratio inferior to the median by at least two times the standard deviation. Thus, we identified many genes being footprinted by A3 among otherwise unaffected viral genomes. Most of these genes belongs to two families of double-stranded DNA viruses; i.e. *herpesviridae* (HHV-1, 2, 3, 4, 5 and 8) and *adenoviridae* (AdV A, B, C, D, E and F). We also observed A3-footprinted genes in the reverse transcribing HBV, HIV-1 and HTLV-1. In order to better shed light on the possible mechanisms responsible for such editing, we position the identified genes along the corresponding viral genomes and detailed these analyses in the following sections.

An equivalent analysis was performed to report a local A3G footprint amongst an otherwise non-footprinted genome. NCC and NNGGNN-depleted genes are listed in S1D Fig. Similarly to what has been observed at the genome level, many of the NCC-depleted genes are also depleted for the NNCCNN motif making difficult to ascribe the observed NCC depletion to the sole A3G editing activity.

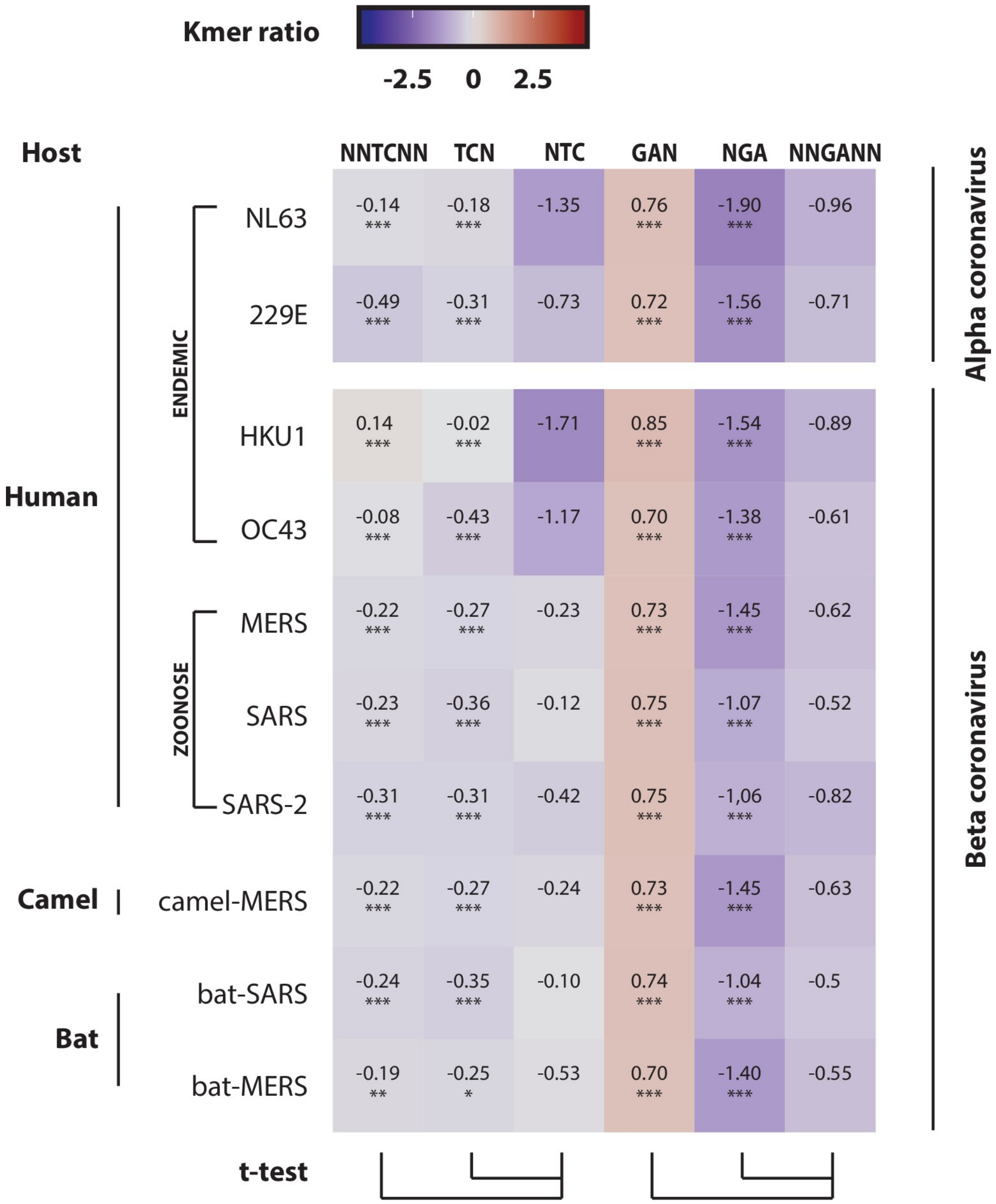
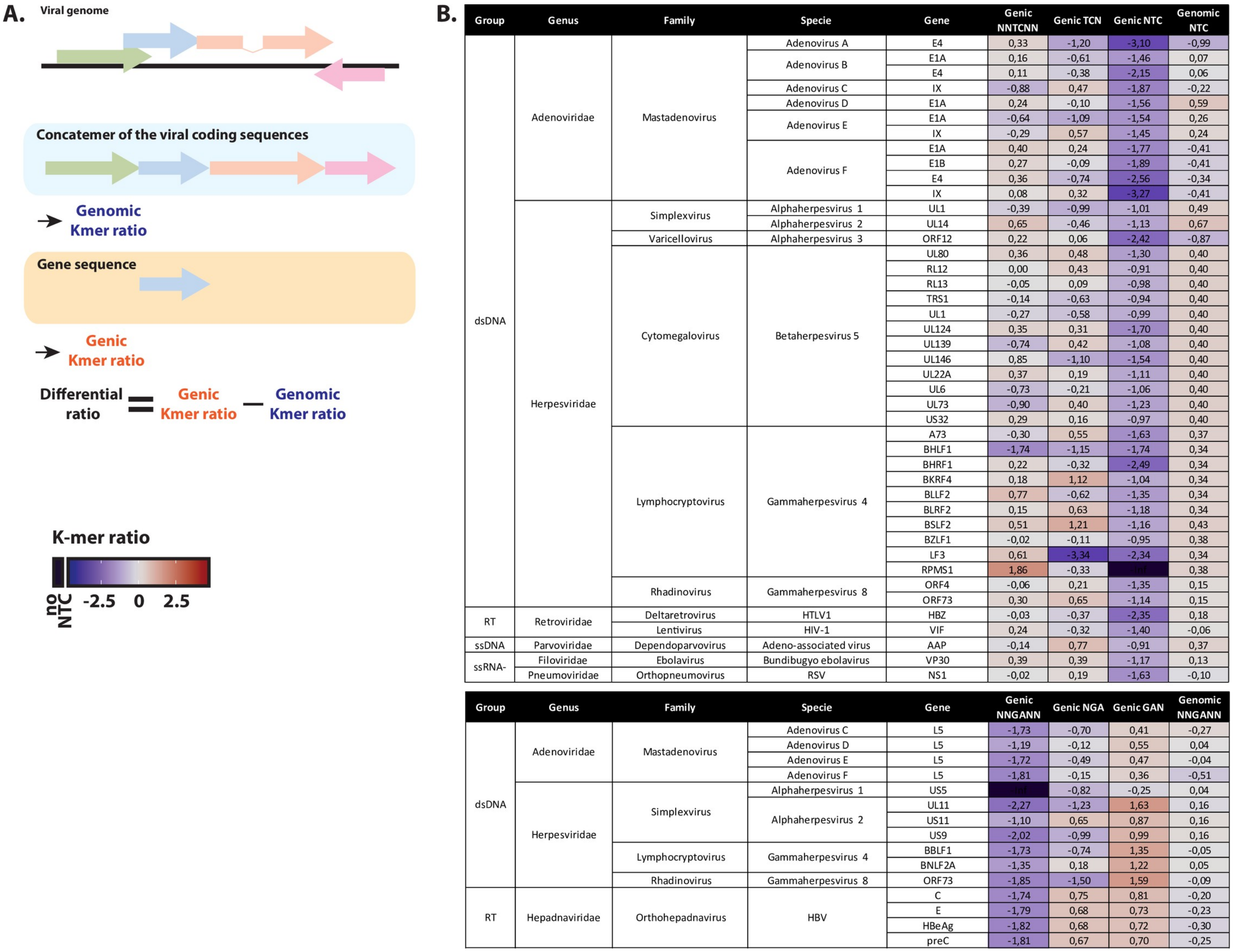


Fig 6. A3 footprint on endemic but not on zoonotic coronaviruses. The observed/expected ratios of TC dinucleotide at various codon positions were calculated for endemic human coronaviruses (229E, NL63, OC43 and HKU1) and compared to those of zoonotic coronaviruses (MERS-CoV, SARS-CoV-1 and SARS-CoV-2) and their ancestors (camel-MERS and bat-SARS). A colored scale with increasing shades of blue indicating depletion and increasing shades of red indicating enrichment. P-values were calculated by Student's unpaired, two-tailed t-test (NS for not significant, * p< 0.05, ** p< 0.01, *** p< 0.001).

https://doi.org/10.1371/journal.ppat.1008718.g006

Identification of an A3 footprint at the replication origins of adenoviruses and EBV

Adenoviruses A and B present a strong NTC depletion for the E1A and E4 genes (Fig 8A); genes localized at both ends of the linear genome (S8 Fig). The same trend can be observed in Adenovirus C, D and E, although to a lesser extent (S9 Fig). Importantly, these E1A and E4 genes are being strongly depleted for NTC but not for the NNGANN motif (Fig 8B). In other


A.
Viral genome
Concatemer of the viral coding sequences
Genomic Kmer ratio
Gene sequence
Genic Kmer ratio
Differential ratio = Genic Kmer ratio − Genomic Kmer ratio
K-mer ratio
no NTC
-2.5
0
2.5
B.

Group	Genus	Family	Specie	Gene	Genic NNTCNN	Genic TCN	Genic NTC	Genomic NTC
dsDNA	Adenoviridae	Mastadenovirus	Adenovirus A	E4	0,33	-1,20	-3,10	-0,99
			Adenovirus B	E1A	0,16	-0,61	-1,46	0,07
				E4	0,11	-0,38	-2,15	0,06
			Adenovirus C	IX	-0,88	0,47	-1,87	-0,22
			Adenovirus D	E1A	0,24	-0,10	-1,56	0,59
			Adenovirus E	E1A	-0,64	-1,09	-1,54	0,26
				IX	-0,29	0,57	-1,45	0,24
			Adenovirus F	E1A	0,40	0,24	-1,77	-0,41
				E1B	0,27	-0,09	-1,89	-0,41
				E4	0,36	-0,74	-2,56	-0,34
				IX	0,08	0,32	-3,27	-0,41
	Herpesviridae	Simplexvirus	Alphaherpesvirus 1	UL1	-0,39	-0,99	-1,01	0,49
			Alphaherpesvirus 2	UL14	0,65	-0,46	-1,13	0,67
		Varicellovirus	Alphaherpesvirus 3	ORF12	0,22	0,06	-2,42	-0,87
		Cytomegalovirus	Betaherpesvirus 5	UL80	0,36	0,48	-1,30	0,40
				RL12	0,00	0,43	-0,91	0,40
				RL13	-0,05	0,09	-0,98	0,40
				TRS1	-0,14	-0,63	-0,94	0,40
				UL1	-0,27	-0,58	-0,99	0,40
				UL124	0,35	0,31	-1,70	0,40
				UL139	-0,74	0,42	-1,08	0,40
				UL146	0,85	-1,10	-1,54	0,40
				UL22A	0,37	0,19	-1,11	0,40
				UL6	-0,73	-0,21	-1,06	0,40
				UL73	-0,90	0,40	-1,23	0,40
				US32	0,29	0,16	-0,97	0,40
		Lymphocryptovirus	Gammaherpesvirus 4	A73	-0,30	0,55	-1,63	0,37
				BHLF1	-1,74	-1,15	-1,74	0,34
				BHRF1	0,22	-0,32	-2,49	0,34
				BKRF4	0,18	1,12	-1,04	0,34
				BLLF2	0,77	-0,62	-1,35	0,34
				BLRF2	0,15	0,63	-1,18	0,34
				BSLF2	0,51	1,21	-1,16	0,43
				BZLF1	-0,02	-0,11	-0,95	0,38
				LF3	0,61	-3,34	-2,34	0,34
				RPMS1	1,86	-0,33	-Inf	0,38
		Rhadinovirus	Gammaherpesvirus 8	ORF4	-0,06	0,21	-1,35	0,15
				ORF73	0,30	0,65	-1,14	0,15
RT	Retroviridae	Deltaretrovirus	HTLV1	HBZ	-0,03	-0,37	-2,35	0,18
		Lentivirus	HIV-1	VIF	0,24	-0,32	-1,40	-0,06
ssDNA	Parvoviridae	Dependoparvovirus	Adeno-associated virus	AAP	-0,14	0,77	-0,91	0,37
ssRNA-	Filoviridae	Ebolavirus	Bundibugyo ebolavirus	VP30	0,39	0,39	-1,17	0,13
	Pneumoviridae	Orthopneumovirus	RSV	NS1	-0,02	0,19	-1,63	-0,10

Group	Genus	Family	Specie	Gene	Genic NNGANN	Genic NGA	Genic GAN	Genomic NNGANN
dsDNA	Adenoviridae	Mastadenovirus	Adenovirus C	L5	-1,73	-0,70	0,41	-0,27
			Adenovirus D	L5	-1,19	-0,12	0,55	0,04
			Adenovirus E	L5	-1,72	-0,49	0,47	-0,04
			Adenovirus F	L5	-1,81	-0,15	0,36	-0,51
	Herpesviridae	Simplexvirus	Alphaherpesvirus 1	US5	-Inf	-0,82	-0,25	0,04
			Alphaherpesvirus 2	UL11	-2,27	-1,23	1,63	0,16
				US11	-1,10	0,65	0,87	0,16
				US9	-2,02	-0,99	0,99	0,16
		Lymphocryptovirus	Gammaherpesvirus 4	BBLF1	-1,73	-0,74	1,35	-0,05
				BNLF2A	-1,35	0,18	1,22	0,05
		Rhadinovirus	Gammaherpesvirus 8	ORF73	-1,85	-1,50	1,59	-0,09
RT	Hepadnaviridae	Orthohepadnavirus	HBV	C	-1,74	0,75	0,81	-0,20
				E	-1,79	0,68	0,73	-0,23
				HBeAg	-1,82	0,68	0,72	-0,30
				preC	-1,81	0,67	0,70	-0,25

Fig 7. Search for an A3 footprint at the gene level. A. Alongside the observed/expected K-mer ratios calculated from the synthetic coding genomes (named genomic K-mer ratios), K-mer ratios were also computed for each viral coding sequence individually (named genic K-mer ratios). Differential ratio is defined as the subtraction of genic K-mer ratio to the corresponding genomic K-mer ratio. B. List of the putative A3-footprinted viral genes and belonging to an otherwise non-depleted viral genome (having at least five reported sequences).

https://doi.org/10.1371/journal.ppat.1008718.g007

words, these two genes are A3-footprinted on their coding strand only. Due to the relative position of those genes and the strand-displacement strategy used for genome replication, we propose a model where A3 editing would occur specifically during the initiation of genome replication on the displaced strands (Fig 8C). Indeed, at the beginning of DNA replication, the displaced strand corresponds to the coding strand of E1A at one end of the linear genome and to the coding strand of E4 at the other extremity. One might also notice an NNGANN depletion for most of the L5 genes (Fig 8B and S9 Fig). Considering the position and orientation of that gene, such footprint might also reflect an A3 activity on the displaced strand (Fig 8C).

Among the 172 EBV genes, only 12 are significantly depleted for NTC (Fig 7B). Interestingly, the five most depleted are localized around the lytic origins of replication. BHLF1 and BHRF1 are localized on both sides of the first lytic origin of replication and LF3, RPMS1 and

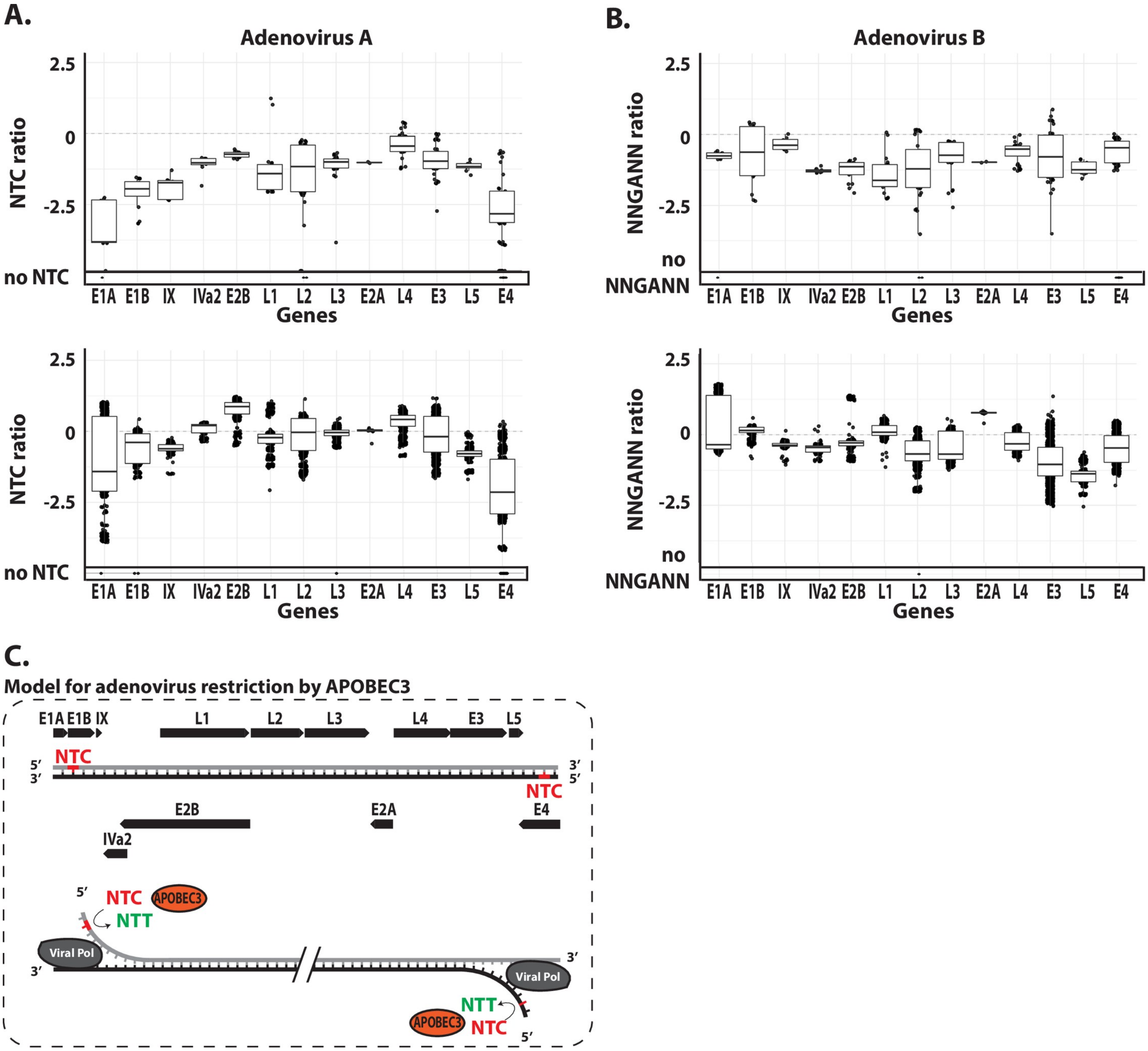


Fig 8. A3 footprint at the genomic ends of adenoviruses. NTC observed/expected ratios (panel A) and NNGANN observed/expected ratios (panel B) were calculated for the different genes of the Adenovirus A and B (each point represents a unique coding sequence). C. Proposed model for A3-editing activity on the adenovirus genome. Genes are represented by black arrows. A3-favored NTC sequence is represented in red and the NTT edited product in green.

https://doi.org/10.1371/journal.ppat.1008718.g008

A73 are on both sides of the second lytic origin of replication (Fig 9B). Similar to the adenoviruses, this local A3 footprint is very much strand-specific and present on the lagging strands of the replication forks surrounding the lytic origins (Fig 9B). Thus, the EBV specific footprint is pointing toward A3 editing during the beginning of replication at the lytic origins. We summarize these observations by a scheme in Fig 9C.

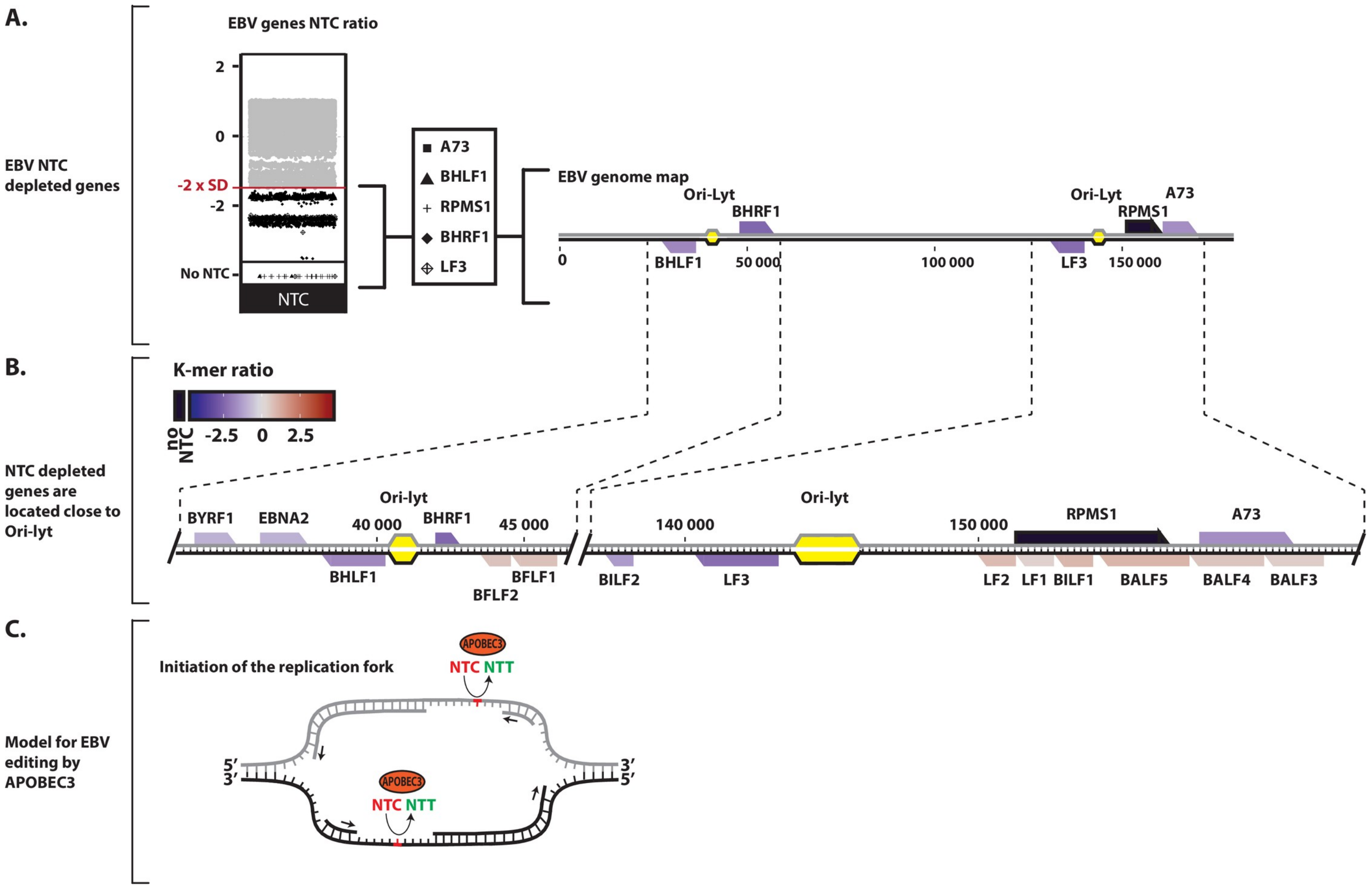


Fig 9. A3 footprint at the lytic replication origins of EBV. A. NTC observed/expected ratios were calculated for the different genes of EBV (each point represents a unique coding sequence) and the five most A3-footprinted genes were highlighted and positioned on the EBV genome map. B. Zoom-in detailing the NTC ratios of the genes surrounding the Ori-Lyt (lytic origin of replication) of EBV. A colored scale with increasing shades of blue indicating NTC depletion and increasing shades of red indicating NTC enrichment. C. Proposed model for A3-editing activity favoring the lagging strand at the EBV lytic origin of replication. A3-favored NTC sequence is represented in red and the NTT edited product in green.

https://doi.org/10.1371/journal.ppat.1008718.g009

The footprint on the HTLV-1, HBV, HIV-1 and HIV-2 genomes fits with editing during reverse transcription

We observed that the HBZ gene of the HTLV-1 virus is depleted for NTC codons (Fig 7B). Because HBZ is an antisense transcript, its coding strand corresponds to the genomic negative strand. Hence, the NTC depletion of the HBZ gene is indicative of an A3 editing activity on the negative strand (Fig 10A and 10B). We then wondered whether such A3 footprint was restricted to the HBZ coding region or rather extend further. We observed that the coding sequences of the sense transcripts Gag, Pro, Pol and Tax are depleted for the NNGANN motif (Fig 10A and 10B). These observations suggest that A3s left an evolutionary footprint on the HTLV-1 virus through editing during reverse transcription.

Depletion for the NNGANN motif has been observed for the C, preC/HBeAg coding sequences of HBV (Fig 7B). These observations support the involvement of an A3 editing activity on the DNA negative strand during the reverse transcription process. However, it is interesting to report that nor the Pol neither the S and X coding sequences are being footprinted (Fig 10C and 10D)

Conflicting data were reported concerning the presence of an A3 evolutionary footprint on the HIV-1 genome [31,32]. Fig 11 reports ratios of the TC and GA-containing K-mers for the HIV-1, HIV-2 and SIV genomes. The HIV-1 genomes were spread out into their respecting groups (M, N, O) and subtypes (group M subtypes A, B, C, D and E). No NTC depletion was observed on the HIV-1, HIV-2 and SIV genomes (Fig 11). We concluded that A3s did not leave a footprint on the plus strand. Importantly, a mild but consistent NNGANN depletion was observed in HIV-1, HIV-2 and SIV genomes, depletion compatible with A3-editing during reverse transcription.

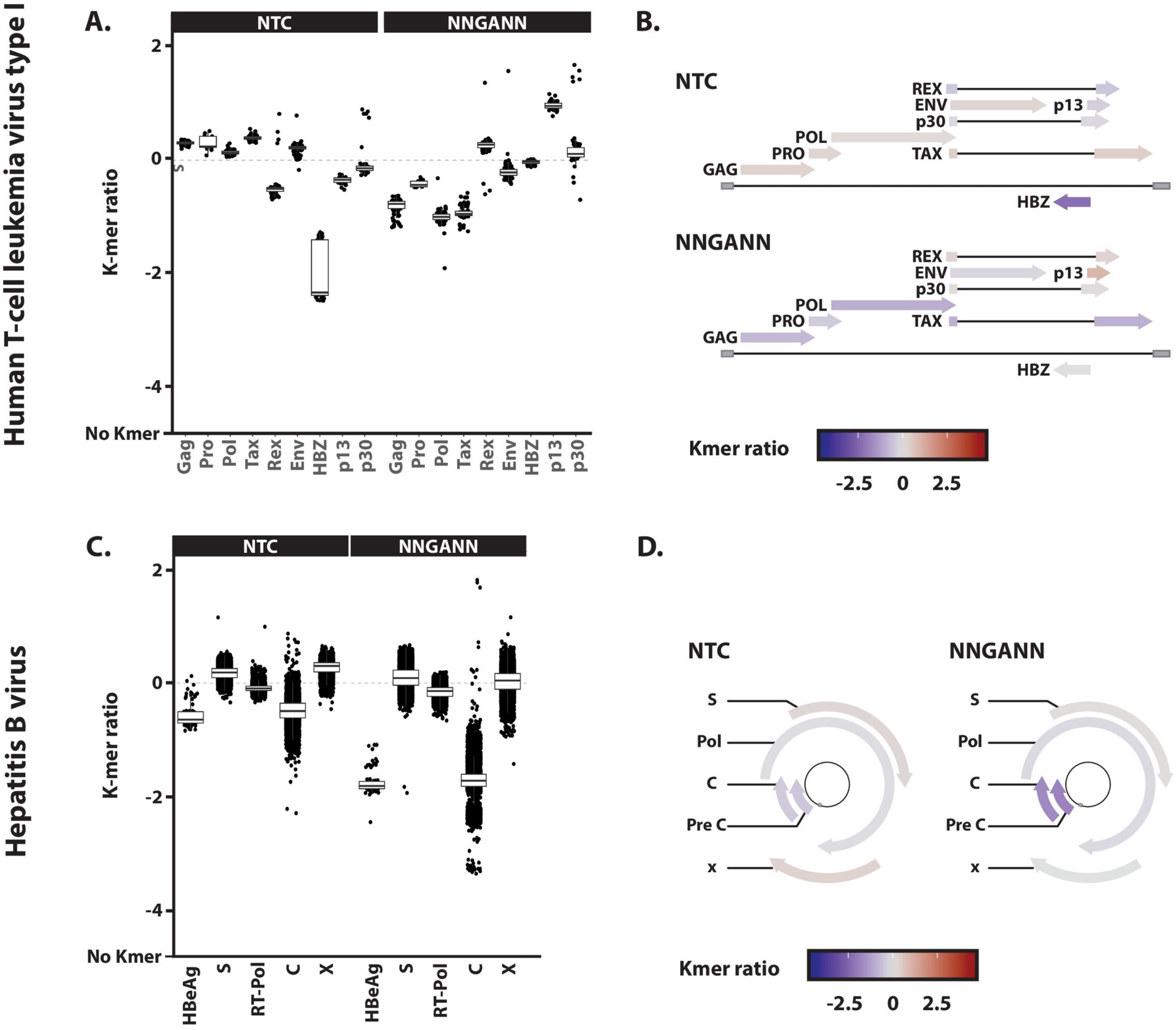


Fig 10. A3 footprint on the negative strand of HTLV-1 and HBV. A. NTC and NNGANN observed/expected ratios were calculated for the different genes of HTLV-1. B. Each gene specific NTC and NNGANN ratio median values were reported on HTLV-1 genome map by a colored scale. C. NTC and NNGANN observed/expected ratios were calculated for the different genes of HBV. D. Each gene specific NTC and NNGANN ratio median values were reported on HBV genome map by a colored scale.

https://doi.org/10.1371/journal.ppat.1008718.g010

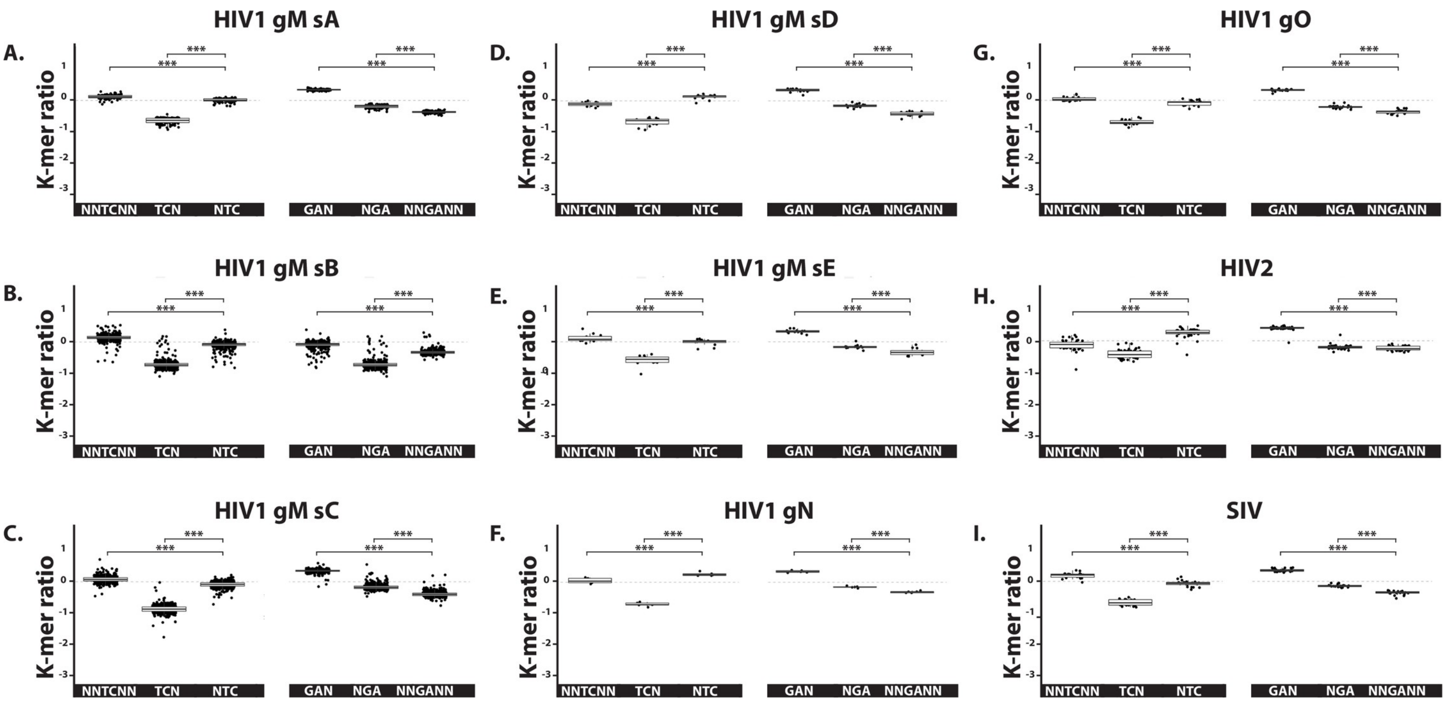


Fig 11. A3 footprint on the negative strand of HIV-1, HIV-2 and SIV. The observed/expected ratios of TC dinucleotide at various codon positions and on both strand (i.e. NNTCNN, TCN, NTC, GAN, NGA and NNGANN) were calculated for the genomes of HIV-1 (distributed into their respective groups and subtypes, panels A to G), HIV-2 (panel H) and SIV (panel I). Each point stands for a unique full-length viral genome. Median and quartile are depicted by a boxplot. P-values were calculated by Student's unpaired, two-tailed t-test (NS for not significant, * $p< 0.05$, ** $p< 0.01$, *** $p< 0.001$).

https://doi.org/10.1371/journal.ppat.1008718.g011

Search for AID or APOBEC1-footprinted viruses

The APOBEC family of genes also counts the AID (or AICDA), APOBEC1, APOBEC2 and APOBEC4 genes. AID is critical for somatic hypermutation and class switch recombination by editing the immunoglobulin loci in B cells [39]. APOBEC1 plays an important role in lipid metabolism by editing the apolipoprotein B pre-mRNA [40,41]. Importantly, APOBEC1 and AID appear also to participate to the restriction of viruses and retroelements [1,42]. Evidence for AID and APOBEC1 evolutionary footprints were investigated by looking for the depletion of their favored motifs, respectively WRC for AID [43] and WCW for APOBEC1 [44].

The distributions of the WRC and NNGYWN ratios do not point towards viruses significantly footprinted by AID at the whole genome level (S10B Fig). Nevertheless, putatively AID-footprinted genes were identified in several double strand DNA viruses, notably the B-cell-infecting virus EBV (S10C Fig).

The distributions of the NWCWNN and NWGWNN ratios show evidence of genome-wide footprinted viruses (S11B Fig). However, it is not possible to disentangle the APOBEC1 footprint from the APOBEC3 footprint as the 6-mers NWCWNN contains the 3-mers NTC. The putatively APOBEC1-footprinted viruses are those that also bear the putative APOBEC3 footprint (Fig 4B and S11C Fig).

Of note, APOBEC2 and APOBEC4-favored motifs have not been described so far.

Discussion

In this study, we investigated the distribution of the A3 footprint along a large set of 33,400 human virus complete genomes. We first observed that no less than 22% of all referenced human viral species have a genome-wide A3 footprint. Among these, we mainly identified viruses from the *papillomaviridae*, *polyomaviridae*, *coronaviridae* and autonomous *parvoviridae* families. In addition to this category of viruses targeted over their entire sequence, we have identified viruses which have an A3 footprint spatially limited to a short section of their genome. This is notably the case for certain *herpesviridae* and *adenoviridae* where the A3 footprint is localized on genomic sequences used to initiate replication of viral DNA.

Our study is in line with previous publications reporting the presence of an A3 footprint on *papillomaviridae* [29] and on the BK polyomavirus [17]. Above all and because we analyzed all currently annotated human viruses with the same approach, we can compare the magnitude of the A3 selection pressure between different viral families. Thus, we show that the *papillomaviridae* and the *polyomaviridae* families are those whose footprint is most intensive (Fig 4A). Those viruses have evolved to thrive under an ongoing and strong A3 selection pressure. The strong NTC depletion reduces exposure of the viral genome to the introduction of uracil and consequently to the base excision repair-mediated DNA degradation. Importantly, not only do these viruses tolerate such pressure, but they even actively promote the expression of certain A3 proteins. Indeed, high risk α-HPVs have been shown to trigger and stabilize A3A and A3B via their oncoproteins E6 and E7 [16,45–47]. Likewise, it has recently been shown that BK and JC β-PyV upregulate A3B through their large T antigen [17,48]. In both the α-HPVs and β-PyVs, the induced A3 proteins are enzymatically active and therefore capable of deamination [17,49]. The selective advantage which would provide a sustained expression of A3A and/or A3B proteins is still debated. On the one hand, A3A has been shown to restrict HPV *in vitro* [45]. Those viruses are still susceptible to A3 restriction. Indeed, deamination of the remaining TC motifs are most of the time non-synonymous. On the other hand, the deaminase activity could positively impact viral fitness by participating to the genetic diversification of the virus or even by protecting the host cell against the reactivation of retroelements [50,51]. We speculate that the error rate of the host DNA polymerase could be too low for viruses with such small DNA genome, hence requiring the A3 editing activity to drive their evolution. Within the *polyomaviridae* family, the magnitude of the A3 footprint differs significantly between species; species of the *betapolyomavirus* genus appearing to be the most strongly footprinted (S2 Fig). Such differences could find their origin in the capacity of the large T Ag at inducing the A3 proteins. To draw a parallel with the *alpha-papillomaviridae*, E6 from high-risk α-HPVs were found to be more potent at inducing A3B than those of low-risk strains [49]. Besides, the cell type hosting the virus can also influence the level of A3 expression. The difference in tissue tropism between the *alpha-* and *beta-papillomaviridae* has been proposed to explain the stronger footprint on the former [29]. The full spectrum of tissue and cell tropism has not been clearly established for the *polyomaviridae*, making this type of correlative analysis tricky. Our analysis also shows that the A3 footprint is present on both strands of the *papillomaviridae* and *polyomavirirидae* genomes. This is compatible with an editing activity during viral DNA replication.

Among the *parvoviridae* family, the erythroparvovirus B19 exhibits an intensive footprint on both strands of its genome, the bocaparvoviruses being mainly footprinted on the negative strand and the dependoparvovirus adeno-associated virus-1 showing no evidence of A3 selection pressure. These dissimilarities might be explained by differences at the replicative and packaging levels. Thus, the *parvoviridae* family consists of viruses that package a single copy of their short linear single-stranded DNA genome into preformed capsids. The packaging takes place in the nucleus of the infected cell. While most can encapsidate DNA strands of either polarity with equal efficiency, some family members, predominantly package negative strand genome. In the case of the erythroparvovirus, there is an equivalent amount of positive and negative genome that is produced during replication and subsequently encapsidated (reviewed in [52]). For *bocaparvoviridae*, the replication produces 90% of negative ssDNA [53]. Such difference could explain the location of the A3 footprint in the negative strand of the *bocaparvoviridae*. Thus, we propose that A3 editing activity takes place inside the nascent virions of the autonomous *parvoviridae*.

Our screening also reported the *coronaviridae* as A3-footprinted. The canonical substrate for the A3 proteins is single stranded DNA and until recently, the viruses identified as being restricted by the A3 deaminase activity were either DNA viruses or viruses having a DNA intermediate (i.e. reverse transcribing viruses). However, recent reports demonstrated that A3A and A3G can deaminate ribocytidine within a single stranded RNA molecule [54–56]. *In vivo*, A3 mutational signature has been reported in the positive single strand RNA Rubella virus [57]. Importantly, Milewska and colleagues demonstrated that cytoplasmic A3s can restrict the NL63 coronavirus *in vitro* [19]. The A3-mediated restriction of the HCoV-NL63 appears to be both deaminase-dependent and independent. A3 restriction did not cause hypermutation on the viral genome, but C to T and G to A point mutations were observed in HCoV-NL63 viruses passaged in A3-expressing cells but not in wild-type cells. It is a matter of debate whether the hypermutated genomes could not be retrieved because of the high fitness cost of such mutations or because the A3 are less processive on coronaviral RNA. Additionally, A3 proteins have been shown to interact with the nucleoproteins of the HCoV-229E, HCoV-NL63 and SARS-CoV-1 viruses [19,58]. Finally, a recent report demonstrates the presence of APOBEC and ADAR editing on the SARS-CoV-2 transcriptome [59]. Thus, knowing that A3s can bind the nucleoprotein and that A3 footprint is present on the positive strand of the viral genome, we suggest that A3 editing occurs on the packaged genome. Two *beta-coronaviridae* are endemic to humans (HCoV-OC43 and HCoV-HKU1), they are widespread, have been circulating in human for at least several decades and may cause 10 to 15% of common colds (review in [60]). Both have an A3 footprint on the positive strand. In comparison, no evidence of footprint was observed on the zoonotic *beta-coronaviridae* SARS-CoV-1, MERS-CoV or SARS-CoV-2. The absence of an evolutionary footprint on SARS-CoV-1 and MERS-CoV could find its explanation in the relative low number of infected individuals and the short duration of viral circulation. According to the World Health Organization, SARS-CoV-1 infected about 8.000 people over a period of few months and have been declared eradicated in May 2004. The MERS-CoV infected so far less than 3.000 people by causing episodic outbreaks in the Middle East. The figures for the SARS-CoV-2 are radically different with more than 5 million confirmed cases as of May 2020. In that respect, it will be interesting to track the evolution of the pandemic SARS-CoV-2 regarding a possible introduction of an A3 footprint through its interhuman transmissions. It is worth reiterating that no footprint was detected on the SARS bat isolates, although the bat A3 locus is the largest and most diverse known repertoire of A3 genes in mammals [61]. Perhaps SARS-like viruses possess a yet unknown and unique A3 inhibiting mechanism. Interestingly, the SARS-CoV-2 genome contains a novel ORF, called ORF10. ORF10 encodes a protein that has been demonstrated to interact with the CUL2 complex [62]. This interaction is reminiscent of the interaction between the Vif of BIV (bovine immunodeficiency virus) and CUL2 [63]. One could speculate that ORF10 could play a Vif-like role in bat and also in Human. Also, a shorter version of SARS-CoV-2 ORF10 is present in all SARS-like viruses [64]. This could explain why the SARS-like viruses are not A3-footprinted.

In addition to this category of viruses that are footprinted on their entire sequence, we identified viruses that show an A3 footprint only on a very limited section of their genome. This is notably the case for the *gamma-herpesviridae* EBV and *adenoviridae*. The A3 footprint on EBV is spatially limited to the lagging strands around the lytic replication origins. Interestingly, both EBV and KSHV were recently demonstrated to encode viral proteins capable of inhibiting A3B activity [28,65]. Those viral proteins (EBV BORF2 and KSHV ORF61) are both the large subunit of the ribonucleotide reductase, expressed during lytic replication and providing the precursors necessary for viral DNA synthesis. We speculate that their expression could avoid the extension of the A3-editing further along the viral genome. Of course, these viral proteins inhibiting A3B may not be the sole actors protecting the viral genome. The coating of the viral DNA by the major DNA binding protein (DBP), the compartmentalization of the viral DNA replication [66] and the switch from theta to the rolling circle replication might also limit access to the viral single-stranded DNA. The fact that no footprint was detected around the latency origin of replication, ori-P, points toward an A3-editing acting during lytic replication. The strategy deployed by EBV and KSHV to cope with A3 restriction is somehow opposite to the one used by the papilloma and polyomaviruses. Indeed, EBV and KSHV actively protect their genome from the A3s as opposed to the papilloma and polyomaviruses which simply cope with a high mutational rate. We speculate here that the large size of the herpesvirus genome would not tolerate an unrestrained A3 activity. Finally, we identified several AID-footprinted genes in the EBV genome which supports recent observations made by Martinez et al. [33]. Those genes are not spatially clustered and further investigation would be necessary to link the detected footprint to AID activity.

Similarly, we identified an A3 footprint on *adenoviridae* localized at the ends of their linear genome where the origins of replication are located. The presence of an A3 footprint on the lagging strand of the origins of replication and its absence on the leading strand is striking. It parallels the footprint on the lytic origins of replication in EBV and it is also reminiscent of the A3 activity in cancer genomes. Thus, A3-related mutations in cancer genomes are strongly enriched on the lagging strand and early-replicating euchromatic regions [67–70]. The leading strand, being protected by the nascent complementary DNA, is less or even not accessible for deamination. Another circumstance where the viral DNA is transiently single-stranded is during transcription. Indeed, the coding strand within the transcriptional bubble is temporally single stranded. In cancer genomes, the observed distribution of APOBEC3-signature mutations is transcription independent [70]. Again, it is very similar to our observations on viral genomes where no evidence points toward A3-editing during viral transcription. Overall, the similarities between *adenoviridae* and *herpesviridae* regarding their relationship to A3s appear substantial. Both are large double-stranded DNA viruses replicating their genome in the nucleus and capable of lytic and latent/persistent phases. It would not be surprising to find that *adenoviridae* are also able to inhibit A3 activity. Along with these speculations, we found important to underline that the dependoparvovirus AAV-1 is the only member of the *parvoviridae* family which does not have an A3 footprint. It would be interesting to test whether it is an intrinsic characteristic of the satellite virus or whether it is mediated by the helper virus (usually an adenovirus or a herpes virus).

The detection of an A3 footprint on the negative strand of the C (Core) and preC/HBeAg coding sequences of HBV is compatible with A3-editing on the DNA negative strand during reverse transcription. A3-related mutations have been detected on the negative strand of the C and preC region and it has been proposed that these mutations could be beneficial for the virus [71]. Indeed, during the natural course of HBV infection, the HBeAg expression is being lost after production of antibodies against it. HBeAg is an accessory non-particulate protein encoded by the preC mRNA and displaying immunomodulatory properties. HBeAg is described as a tolerogen that allows the virus to establish infection. Seroconversion against the HBeAg leads to the selection of HBeAg-negative mutants. The ability to develop mutations, altering HBeAg expression, can influence the length of the HBeAg-positive phase, which is important for determining the clinical course (reviewed in [72]). Our observation of an A3 footprint in the preC and C region further supports the idea that A3 can positively participate to the immune escape. That would not be the first example of the hijacking an antiviral weapon for the benefit to the pathogen. Hepatitis D virus is a circular complementary single-stranded RNA virus that requires editing of its genome by a cellular adenosine deaminase (ADAR-1) to complete its life cycle (reviewed in [73]). Likewise, it is striking to note that only the C and preC/HBeAg coding regions are being footprinted and not the others coding sequences downstream. In fact, the mutational load introduced by the A3s is much stronger on the 5' end of the negative strand (corresponding to the Pol, S and X regions) because the newly synthesized double strand DNA eventually displaces the single strand DNA from the capsid walls, making it accessible to deaminase activity [74]. Thus, the 5' end of the negative strand is found to be frequently hypermutated; the term hypermutation referring to as mutations clustered on a short sequence. It is essential to underline that our observations reflect the A3-induced mutations which were conserved and not those which put an end to the viral cycle. Consequently, the phenomenon of hypermutation will not leave an evolutionary footprint. In this respect, the HBV mutation spectrum from *in vivo* cirrhotic samples shows that even though the majority of HBV genomes are strongly mutated by A3s and these virions are probably defective, a small fraction is slightly modified and may therefore still be infectious [71]. Finally, the Pol, S and X genes overlap on different reading frames, which implies that these coding regions are less permissive to mutations (a silent mutation in one frame may not be in the other frame).

We paid particular attention to HIV in our analysis as the APOBEC3 antiviral activity has been historically discovered in that field of research [6]. We observed a weak A3 evolutionary footprint on the minus strand of the HIV genomes in support of the observations made by Jern et al. [31]. The weakness of the footprint can be explained at least in part by the efficiency of the A3-inhibiting protein Vif. Also, the error-prone RT could be responsible for the reversion of some A3-induced mutations providing that the virus can still complete its life cycle. Finally, A3s are well known to also restrict HIV through a deaminase independent activity, therefore without leaving any footprint.

The intensity of the A3 footprint is very strong in many *papillomaviridae* and *polyomaviridae*. Several viruses of these families are well-known tumor viruses (HPV-16, HPV-18, Merkel cell polyomavirus, etc.) and a mechanistic link between A3 expression and the development of cancer has been established in HPV positive cervical and oesopharyngeal cancers [75,76]. S12 Fig shows oncogenic viruses (confirmed or suspected) and their respective A3 footprint. It illustrates that an A3 footprint is not present in every tumor virus. HCV shows no A3-footprint, is still definitely a tumor virus. Nevertheless, we wonder whether the presence of a strong footprint may suggest involvement in cancer. The BK and JC polyomaviruses have a footprint as intensive as HR-HPVs (S12 Fig). BK PyV infects the kidneys and the urinary tract and is suspected of playing a role in certain bladder cancer where viral expression and integration have been reported [77,78]. BK PyV triggers A3 expression *in vitro* and the A3-related mutations found in bladder tumors account for two-thirds of the total mutational load [17,79]. Similarly, along with the *alpha*, the *beta-papillomaviridae* show a similar A3 footprint (Fig 4B). Some members of the *beta-papillomavirus* genus are suspected to play a role in non-melanoma skin cancer (cutaneous squamous-cell carcinoma) [80,81]. Even so they do not seem to insert into the cell genome, they might promote carcinogenesis initiation. Some β-HPVs were demonstrated to potentiate the deleterious effect of UV radiations and to drive skin carcinogenesis in mice with a hit-and-run mechanism [82]. Finally, the erythroparvovirus B19 is one of the most footprinted virus. While its replication occurs primarily in erythroid tissues, the erythroparvovirus B19 commonly persists in a wide range of tissues [83]. A link between the erythroparvovirus B19 and thyroid cancer has been proposed but evidence is scarce to date (72–74). Of note, the A3-related mutations in thyroid tumors make about 40% of the total mutational load [79]. We think that further research should be carried out to rule in or out the involvement of these later viruses in cancer.

In conclusion, the present study represents the first global screening for the A3 selection pressure on all currently annotated human viruses. We demonstrate that many *papillomaviridae*, *polyomaviridae*, autonomous *parvoviridae* and *coronaviridae* can thrive despite being under the selective pressure of the A3 proteins. Those viruses cope with A3 editing activity thanks to a deep cleansing of A3-favored motifs in their genome. *Herpesviridae* and *adenoviridae* display a subtler A3 footprint limited to the lytic origins of replication, probably thanks to active mechanisms of A3 inhibition. A3 deamination appears to occur during replication of viral DNA (sometimes limited to the lagging strand) for the double-stranded DNA viruses and/or inside the capsid for the single-stranded DNA and RNA viruses. The causal link established between HPV infection and the A3 mutational signature in human cancer also lead us to propose to consider the *beta-papillomaviridae* and the erythroparvovirus B19 as potentially promoting A3 expression and therefore exposing the cell genome to a mutagenic activity.

Material and methods

Fasta sequences

We downloaded complete viral genomes from the "NCBI Virus" database (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/) as released in April 2020. We retrieved only full-length genomes by selecting "Complete" for the criterion "Nucleotide Completeness". We retrieved Human viruses by selecting "Humans" for the criterion "Host". We retrieved non-human primate viruses by selecting "Primate" for the criterion "Host" and by deducting the human viruses from this data set. We retrieved avian viruses by selecting "Aves (birds)" for the criterion "Host". We retrieved fish viruses by selecting "Actinopterygii (ray-finned fish)" for the criterion "Host". We also retrieved Camel MERS viruses by selecting "Camelus dromedaries (Arabian Camel)" for the criterion "Host" and "MERS-CoV" for criterion "Virus". We retrieved Bat-MERS viruses by selecting "Chiroptera (bats)" for the criterion "Host" and "MERS-CoV" for the criterion "Virus". We retrieved Bat-SARS viruses by selecting "Chiroptera (bats)" for the criterion "Host" and "Severe acute respiratory syndrome-related coronavirus" for criterion "Virus". The dataset of Human viruses was supplemented by manually curated human virus complete genome sequences from the "NCBI nucleotide" database. Using these criteria, 33,400 Human, 1,397 non-human primate, 9,160 avian, 570 fish, 259 Camel MERS, 5 Bat MERS and 33 Bat SARS full-length viral genomes were collected. GenBank accession ID's are treated as unique and listed in the S1, S2 and S3 Tables.

Calculation of the K-mer representation ratio

A K-mer encompasses a collection of sequences with a common motif. For instance, the NTC K-mer includes the ATC, CTC, GTC and TTC sequences. In addition, as we limit our analysis to coding sequences, we force our K-mers to be in the reading frame and therefore to correspond to codons. For example, the NTC K-mer actually includes the ATC, CTC, GTC and TTC codons. Following the same logic, the NNTCNN K-mer comprises the 256 pair of codons having a T at the end of the first codon and a C to start the second codon. We calculated the observed vs. expected K-mer representation ratio as described by Warren *et al.* [29]. Briefly, each coding sequence has been randomly shuffled a thousand times, retaining only the nucleotide composition. The expected count of a given K-mer is calculated as the average of the occurrences of this K-mer over the thousand iterations. The K-mer ratio is given as the log2 ratio of the observed occurrence of this K-mer to the expected occurrence. To calculate the ratio of a given K-mer for an entire viral genome, a "synthetic coding genome" was generated by concatenating the different coding sequences (Fig 1B). The synthetic coding sequence is then randomly shuffled a thousand times and K-mer ratio calculated as above. A K-mer ratio $<< 0$ indicates K-mer under representation and a K-mer ratio equal to zero means that no representation bias is observed.

Statistical analysis

Unpaired Student's t test has been used where appropriate. The results were considered statistically significant at a P-value of <0.05. All boxplot, heatmap and map representations have been generated using ggplot R package.

Supporting information

S1 Fig. Search for A3G-footprinted human viruses. A. A3G favors deamination of cytidine when preceded by another cytidine. The 5'CC dinucleotide motif is depicted in three possible codon contexts on both coding and template strand. Depending on the position of the mutated C, the C to T transition can be synonymous (S) or non-synonymous (NS). Proportion of S and NS mutations is reported when the two types of mutation can be produced. Because synonymous mutations are more likely to be retained, A3G-footprinted viruses should display to a stronger depletion of NCC codons compared to CCN or NNCCNN motifs (and/or a depletion of NNGGNN motifs versus the GNN and NGG motifs). B. The NCC and NNGGNN observed/expected ratios for 33,400 human viruses' genomes (from 870 unique species) were calculated, grouped by species and colored according to the Baltimore classification. Each point represents a unique viral genome. Viral species with an NCC or NNGGNN ratio below two times the standard deviation (dotted grey line) from the population median (red line) are retained for further analysis in panel C. C. The observed/expected ratios of 5'CC dinucleotide at various codon positions and on both strands (i.e. NNCCNN, CCN, NCC, GGN, NGG and NNGGNN) were calculated for the NCC and/or NNGGNN depleted viral species and depicted by a heatmap. A colored scale with increasing shades of blue indicating depletion and increasing shades of red indicating enrichment. P-values were calculated by Student's unpaired, two-tailed t-test (NS for not significant, * $p< 0.05$, ** $p< 0.01$, *** $p< 0.001$). D. List of the viral genes displaying NCC or NNGGNN depletion and belonging to an otherwise non-depleted viral genome.
(PDF)

S2 Fig. NTC depletion among several *polyomaviridae* family members. The observed/expected ratios of TC dinucleotide at various codon positions (i.e. NNTCNN, TCN, NTC) were calculated for several polyomaviruses and the corresponding genus (alpha, beta and delta) is reported for each virus.
(PDF)

S3 Fig. K-mer ratios of the human viruses split according to Baltimore's groups. Human viruses were broken down into their respective Baltimore's group and analyzed for their observed/expected K-mer ratios.
(PDF)

S4 Fig. NTT, NTA and NTG K-mers ratios of the A3-footprinted viruses. The observed/expected ratios of NTC, NTT, NTA and NTG K-mers were calculated for the putative A3-footprinted viral species and depicted by a heatmap. A colored scale with increasing shades of blue indicating depletion and increasing shades of red indicating enrichment. P-values were calculated by Student's unpaired, two-tailed t-test (NS for not significant, * $p< 0.05$, ** $p< 0.01$, ***

$p < 0.001$).
(PDF)

S5 Fig. TC depletion in HPV non-coding sequences. The observed/expected ratios of TC dinucleotide at various "codon" positions (i.e. NNTCNN, TCN, and NTC) were calculated for the non-coding sequences of human papillomavirus 16, 18 and 31.
(PDF)

S6 Fig. A3 footprint on HPV16, HPV18 and HPV31. NTC and NNGANN observed/expected ratios were calculated for the different genes of the HPV16, HPV18 and HPV31 and were reported on their genomic maps using a colored scale with increasing shades of blue indicating NTC depletion and increasing shades of red indicating NTC enrichment. Replication origin is illustrated by a black dot and gene transcriptional orientation is symbolized by black arrow.
(PDF)

S7 Fig. B19 erythroparvovirus genome is depleted for NNGANN K-mer. Coding sequences (NS1, 7.5k, VP1, X, VP2 and 11k genes) from 18 full-length B19 erythroparvoviruses were depicted by grey lines overlaid by red marks to symbolize NNGANN and green marks to position NNAANN codons.
(PDF)

S8 Fig. NTC depletion of the E1A and E4 genes of the Adenovirus A and B. NTC observed/expected ratios were calculated for the different genes of the Adenovirus A and B and were reported on their genomic maps using a colored scale with increasing shades of blue indicating NTC depletion and increasing shades of red indicating NTC enrichment.
(PDF)

S9 Fig. A3 footprint on Adenovirus C, D, E, F and G. NTC and NNGANN observed/expected ratios were calculated for the different genes of the Adenoviruses C, D, F and G (each point represents a unique coding sequence).
(PDF)

S10 Fig. Search for AID-footprinted viruses. A. AID favors cytidine deamination in a 5' WRC context. The WRC trinucleotide motif is depicted in three possible codon contexts on both coding and template strand. Depending on the position of the mutated C, the C to T transition can be synonymous (S) or non-synonymous (NS). Proportion of S and NS mutations is reported when the two types of mutation can be produced. B. The WRC and NNGYWN observed/expected ratios for 33,400 human viruses' genomes (from 870 unique species) were calculated, grouped by species and colored according to the Baltimore classification. Each point represents a unique viral genome. C. List of the putative AID-footprinted viral genes (displaying WRC or NNGYWN depletion) and belonging to an otherwise non-depleted viral genome.
(PDF)

S11 Fig. Search for APOBEC1-footprinted viruses. A. APOBEC1 favors cytidine deamination in a 5' WCW context. The WCW trinucleotide motif is depicted in three possible codon contexts on both coding and template strand. Depending on the position of the mutated C, the C to T transition can be synonymous (S) or non-synonymous (NS). Proportion of S and NS mutations is reported when the two types of mutation can be produced. B. The NWCWNN and NWGWNN observed/expected ratios for 33,400 human viruses' genomes (from 870 unique species) were calculated, grouped by species and colored according to the Baltimore

classification. Each point represents a unique viral genome. C. The observed/expected ratios of WCW trinucleotide at various codon positions and on both strands (i.e. NWCWNN, WCW, NNWCWN, NWGWNN, WGW and NNWGWN) were calculated for the NWCWNN and/or NWGWNN depleted viral species and depicted by a heatmap. A colored scale with increasing shades of blue indicating depletion and increasing shades of red indicating enrichment. P-values were calculated by Student's unpaired, two-tailed t-test (NS for not significant, * $p< 0.05$, ** $p< 0.01$, *** $p< 0.001$). D. List of the putative APOBEC1-footprinted viral genes (displaying NWCWNN or NWGWNN depletion) and belonging to an otherwise non-depleted viral genome.
(PDF)

S12 Fig. A3 footprint in human oncogenic viruses. NTC and NNGANN observed/expected ratios were calculated for each available coding sequence of eleven well-known cancer-related viruses. Each point represents a unique viral coding sequence. The coding sequences are grouped and colored according to gene name.
(PDF)

S1 Table. Genomic K-mer ratios for human viruses. Observed/expected K-mer ratios for each genomic human viral sequence (available for download at https://doi.org/10.5061/dryad.n8pk0p2sd).
(TXT)

S2 Table. Genic K-mer ratios for human viruses. Observed/expected K-mer ratios for each genic human viral sequence (available for download at https://doi.org/10.5061/dryad.n8pk0p2sd).
(ZIP)

S3 Table. Genomic K-mer ratios for non-human viruses. Observed/expected K-mer ratios for each genomic and genic non-human viral sequence (available for download at https://doi.org/10.5061/dryad.n8pk0p2sd).
(TXT)

Acknowledgments

We thank members of the URVI lab for valuable discussions.

Author Contributions

Conceptualization: Florian Poulain, Noémie Lejeune, Kévin Willemart, Nicolas A. Gillet.

Data curation: Florian Poulain.

Formal analysis: Florian Poulain.

Funding acquisition: Nicolas A. Gillet.

Investigation: Florian Poulain.

Methodology: Florian Poulain, Nicolas A. Gillet.

Project administration: Nicolas A. Gillet.

Resources: Nicolas A. Gillet.

Software: Florian Poulain.

Supervision: Nicolas A. Gillet.

Validation: Florian Poulain.

Visualization: Florian Poulain.

Writing – original draft: Florian Poulain.

Writing – review & editing: Nicolas A. Gillet.

References

1. Harris RS, Dudley JP. APOBECs and virus restriction. Virology. 2015;479–480: 131–145. https://doi.org/10.1016/j.virol.2015.03.012 PMID: 25818029
2. Willems L, Gillet NA. APOBEC3 Interference during Replication of Viral Genomes. Viruses. 2015; 7: 2999–3018. https://doi.org/10.3390/v7062757 PMID: 26110583
3. Salter JD, Bennett RP, Smith HC. The APOBEC Protein Family: United by Structure, Divergent in Function. Trends Biochem Sci. 2016; 41: 578–594. https://doi.org/10.1016/j.tibs.2016.05.001 PMID: 27283515
4. Münk C, Willemsen A, Bravo IG. An ancient history of gene duplications, fusions and losses in the evolution of APOBEC3 mutators in mammals. BMC Evol Biol. 2012; 12: 71. https://doi.org/10.1186/1471-2148-12-71 PMID: 22640020
5. Taylor BJ, Nik-Zainal S, Wu YL, Stebbings LA, Raine K, Campbell PJ, et al. DNA deaminases induce break-associated mutation showers with implication of APOBEC3B and 3A in breast cancer kataegis. Stamatoyannopoulos J, editor. eLife. 2013; 2: e00534. https://doi.org/10.7554/eLife.00534 PMID: 23599896
6. Sheehy AM, Gaddis NC, Choi JD, Malim MH. Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. Nature. 2002; 418: 646–650. https://doi.org/10.1038/nature00939 PMID: 12167863
7. Sasada A, Takaori-Kondo A, Shirakawa K, Kobayashi M, Abudu A, Hishizawa M, et al. APOBEC3G targets human T-cell leukemia virus type 1. Retrovirology. 2005; 2: 32. https://doi.org/10.1186/1742-4690-2-32 PMID: 15943885
8. Turelli P, Mangeat B, Jost S, Vianin S, Trono D. Inhibition of Hepatitis B Virus Replication by APOBEC3G. Science. 2004; 303: 1829–1829. https://doi.org/10.1126/science.1092066 PMID: 15031497
9. Lecossier D, Bouchonnet F, Clavel F, Hance AJ. Hypermutation of HIV-1 DNA in the Absence of the Vif Protein. Science. 2003; 300: 1112–1112. https://doi.org/10.1126/science.1083338 PMID: 12750511
10. Mangeat B, Turelli P, Caron G, Friedli M, Perrin L, Trono D. Broad antiretroviral defence by human APOBEC3G through lethal editing of nascent reverse transcripts. Nature. 2003; 424: 99–103. https://doi.org/10.1038/nature01709 PMID: 12808466
11. Mahieux R, Suspène R, Delebecque F, Henry M, Schwartz O, Wain-Hobson S, et al. Extensive editing of a small fraction of human T-cell leukemia virus type 1 genomes by four APOBEC3 cytidine deaminases. J Gen Virol. 2005; 86: 2489–2494. https://doi.org/10.1099/vir.0.80973-0 PMID: 16099907
12. Fan J, Ma G, Nosaka K, Tanabe J, Satou Y, Koito A, et al. APOBEC3G Generates Nonsense Mutations in Human T-Cell Leukemia Virus Type 1 Proviral Genomes In Vivo. J Virol. 2010; 84: 7278–7287. https://doi.org/10.1128/JVI.02239-09 PMID: 20463074
13. Suspène R, Guétard D, Henry M, Sommer P, Wain-Hobson S, Vartanian J-P. Extensive editing of both hepatitis B virus DNA strands by APOBEC3 cytidine deaminases in vitro and in vivo. Proc Natl Acad Sci U S A. 2005; 102: 8321–8326. https://doi.org/10.1073/pnas.0408223102 PMID: 15919829
14. Henry M, Guétard D, Suspène R, Rusniok C, Wain-Hobson S, Vartanian J-P. Genetic Editing of HBV DNA by Monodomain Human APOBEC3 Cytidine Deaminases and the Recombinant Nature of APOBEC3G. PLoS ONE. 2009; 4. https://doi.org/10.1371/journal.pone.0004277 PMID: 19169351
15. Suspène R, Aynaud M-M, Koch S, Pasdeloup D, Labetoulle M, Gaertner B, et al. Genetic Editing of Herpes Simplex Virus 1 and Epstein-Barr Herpesvirus Genomes by Human APOBEC3 Cytidine Deaminases in Culture and In Vivo. J Virol. 2011; 85: 7594–7602. https://doi.org/10.1128/JVI.00290-11 PMID: 21632763
16. Vartanian J-P, Guétard D, Henry M, Wain-Hobson S. Evidence for Editing of Human Papillomavirus DNA by APOBEC3 in Benign and Precancerous Lesions. Science. 2008; 320: 230–233. https://doi.org/10.1126/science.1153201 PMID: 18403710
17. Verhalen B, Starrett GJ, Harris RS, Jiang M. Functional Upregulation of the DNA Cytosine Deaminase APOBEC3B by Polyomaviruses. J Virol. 2016; 90: 6379–6386. https://doi.org/10.1128/JVI.00771-16 PMID: 27147740

18. Peretti A, Geoghegan EM, Pastrana DV, Smola S, Feld P, Sauter M, et al. Characterization of BK polyomaviruses from kidney transplant recipients suggests a role for APOBEC3 in driving in-host virus evolution. Cell Host Microbe. 2018; 23: 628–635.e7. https://doi.org/10.1016/j.chom.2018.04.005 PMID: 29746834

19. Milewska A, Kindler E, Vkovski P, Zeglen S, Ochman M, Thiel V, et al. APOBEC3-mediated restriction of RNA virus replication. Sci Rep. 2018; 8: 5960. https://doi.org/10.1038/s41598-018-24448-2 PMID: 29654310

20. Fehrholz M, Kendl S, Prifert C, Weissbrich B, Lemon K, Rennick L, et al. The innate antiviral factor APOBEC3G targets replication of measles, mumps and respiratory syncytial viruses. J Gen Virol. 2012; 93: 565–576. https://doi.org/10.1099/vir.0.038919-0 PMID: 22170635

21. Peng Z-G, Zhao Z-Y, Li Y-P, Wang Y-P, Hao L-H, Fan B, et al. Host apolipoprotein B messenger RNA-editing enzyme catalytic polypeptide-like 3G is an innate defensive factor and drug target against hepatitis C virus. Hepatol Baltim Md. 2011; 53: 1080–1089. https://doi.org/10.1002/hep.24160 PMID: 21480314

22. Stenglein MD, Harris RS. APOBEC3B and APOBEC3F inhibit L1 retrotransposition by a DNA deamination-independent mechanism. J Biol Chem. 2006; 281: 16837–16841. https://doi.org/10.1074/jbc.M602367200 PMID: 16648136

23. Tsuge M, Noguchi C, Akiyama R, Matsushita M, Kunihiro K, Tanaka S, et al. G to A hypermutation of TT virus. Virus Res. 2010; 149: 211–216. https://doi.org/10.1016/j.virusres.2010.01.019 PMID: 20138932

24. Chen H, Lilley CE, Yu Q, Lee DV, Chou J, Narvaiza I, et al. APOBEC3A Is a Potent Inhibitor of Adeno-Associated Virus and Retrotransposons. Curr Biol. 2006; 16: 480–485. https://doi.org/10.1016/j.cub.2006.01.031 PMID: 16527742

25. Narvaiza I, Linfesty DC, Greener BN, Hakata Y, Pintel DJ, Logue E, et al. Deaminase-independent inhibition of parvoviruses by the APOBEC3A cytidine deaminase. PLoS Pathog. 2009; 5: e1000439. https://doi.org/10.1371/journal.ppat.1000439 PMID: 19461882

26. Yu X, Yu Y, Liu B, Luo K, Kong W, Mao P, et al. Induction of APOBEC3G ubiquitination and degradation by an HIV-1 Vif-Cul5-SCF complex. Science. 2003; 302: 1056–1060. https://doi.org/10.1126/science.1089591 PMID: 14564014

27. Derse D, Hill SA, Princler G, Lloyd P, Heidecker G. Resistance of human T cell leukemia virus type 1 to APOBEC3G restriction is mediated by elements in nucleocapsid. Proc Natl Acad Sci. 2007; 104: 2915–2920. https://doi.org/10.1073/pnas.0609444104 PMID: 17299050

28. Cheng AZ, Yockteng-Melgar J, Jarvis MC, Malik-Soni N, Borozan I, Carpenter MA, et al. Epstein–Barr virus BORF2 inhibits cellular APOBEC3B to preserve viral genome integrity. Nat Microbiol. 2019; 4: 78–88. https://doi.org/10.1038/s41564-018-0284-6 PMID: 30420783

29. Warren CJ, Van Doorslaer K, Pandey A, Espinosa JM, Pyeon D. Role of the host restriction factor APOBEC3 on papillomavirus evolution. Virus Evol. 2015; 1. https://doi.org/10.1093/ve/vev015 PMID: 27570633

30. Anwar F, Davenport MP, Ebrahimi D. Footprint of APOBEC3 on the Genome of Human Retroelements. J Virol. 2013; 87: 8195–8204. https://doi.org/10.1128/JVI.00298-13 PMID: 23698293

31. Jern P, Russell RA, Pathak VK, Coffin JM. Likely Role of APOBEC3G-Mediated G-to-A Mutations in HIV-1 Evolution and Drug Resistance. PLoS Pathog. 2009; 5. https://doi.org/10.1371/journal.ppat.1000367 PMID: 19343218

32. Ebrahimi D, Anwar F, Davenport MP. APOBEC3 Has Not Left an Evolutionary Footprint on the HIV-1 Genome▽. J Virol. 2011; 85: 9139–9146. https://doi.org/10.1128/JVI.00658-11 PMID: 21697498

33. Martinez T, Shapiro M, Bhaduri-McIntosh S, MacCarthy T. Evolutionary effects of the AID/APOBEC family of mutagenic enzymes on human gamma-herpesviruses. Virus Evol. 2019; 5: vey040. https://doi.org/10.1093/ve/vey040 PMID: 30792902

34. Woo PCY, Wong BHL, Huang Y, Lau SKP, Yuen K-Y. Cytosine deamination and selection of CpG suppressed clones are the two major independent biological forces that shape codon usage bias in coronaviruses. Virology. 2007; 369: 431–442. https://doi.org/10.1016/j.virol.2007.08.010 PMID: 17881030

35. Chen J, MacCarthy T. The preferred nucleotide contexts of the AID/APOBEC cytidine deaminases have differential effects when mutating retrotransposon and virus sequences compared to host genes. PLoS Comput Biol. 2017; 13. https://doi.org/10.1371/journal.pcbi.1005471 PMID: 28362825

36. Shapiro M, Meier S, MacCarthy T. The cytidine deaminase under-representation reporter (CDUR) as a tool to study evolution of sequences under deaminase mutational pressure. BMC Bioinformatics. 2018; 19: 163. https://doi.org/10.1186/s12859-018-2161-y PMID: 29716522

37. LaRue RS, Jónsson SR, Silverstein KA, Lajoie M, Bertrand D, El-Mabrouk N, et al. The artiodactyl APOBEC3 innate immune repertoire shows evidence for a multi-functional domain organization that existed

in the ancestor of placental mammals. BMC Mol Biol. 2008; 9: 104. https://doi.org/10.1186/1471-2199-9-104 PMID: 19017397

38. Karlin S, Doerfler W, Cardon LR. Why is CpG suppressed in the genomes of virtually all small eukaryotic viruses but not in those of large eukaryotic viruses? J Virol. 1994; 68: 2889–2897. https://doi.org/10.1128/JVI.68.5.2889-2897.1994 PMID: 8151759

39. Muramatsu M, Kinoshita K, Fagarasan S, Yamada S, Shinkai Y, Honjo T. Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. Cell. 2000; 102: 553–563. https://doi.org/10.1016/s0092-8674(00)00078-7 PMID: 11007474

40. Powell LM, Wallis SC, Pease RJ, Edwards YH, Knott TJ, Scott J. A novel form of tissue-specific RNA processing produces apolipoprotein-B48 in intestine. Cell. 1987; 50: 831–840. https://doi.org/10.1016/0092-8674(87)90510-1 PMID: 3621347

41. Davidson NO, Innerarity TL, Scott J, Smith H, Driscoll DM, Teng B, et al. Proposed nomenclature for the catalytic subunit of the mammalian apolipoprotein B mRNA editing enzyme: APOBEC-1. RNA N Y N. 1995; 1: 3.

42. Moris A, Murray S, Cardinaud S. AID and APOBECs span the gap between innate and adaptive immunity. Front Microbiol. 2014; 5. https://doi.org/10.3389/fmicb.2014.00534 PMID: 25352838

43. Larijani M, Frieder D, Basit W, Martin A. The mutation spectrum of purified AID is similar to the mutability index in Ramos cells and in ung(-/-)msh2(-/-) mice. Immunogenetics. 2005; 56: 840–845. https://doi.org/10.1007/s00251-004-0748-0 PMID: 15650878

44. Rosenberg BR, Hamilton CE, Mwangi MM, Dewell S, Papavasiliou FN. Transcriptome-wide sequencing reveals numerous APOBEC1 mRNA-editing targets in transcript 3' UTRs. Nat Struct Mol Biol. 2011; 18: 230–236. https://doi.org/10.1038/nsmb.1975 PMID: 21258325

45. Warren CJ, Xu T, Guo K, Griffin LM, Westrich JA, Lee D, et al. APOBEC3A functions as a restriction factor of human papillomavirus. J Virol. 2015; 89: 688–702. https://doi.org/10.1128/JVI.02383-14 PMID: 25355878

46. Mori S, Takeuchi T, Ishii Y, Yugawa T, Kiyono T, Nishina H, et al. Human Papillomavirus 16 E6 Upregulates APOBEC3B via the TEAD Transcription Factor. J Virol. 2017; 91. https://doi.org/10.1128/JVI.02413-16 PMID: 28077648

47. Westrich JA, Warren CJ, Klausner MJ, Guo K, Liu C-W, Santiago ML, et al. Human Papillomavirus 16 E7 Stabilizes APOBEC3A Protein by Inhibiting Cullin 2-Dependent Protein Degradation. J Virol. 2018; 92. https://doi.org/10.1128/JVI.01318-17 PMID: 29367246

48. Starrett GJ, Serebrenik AA, Roelofs PA, McCann JL, Verhalen B, Jarvis MC, et al. Polyomavirus T Antigen Induces APOBEC3B Expression Using an LXCXE-Dependent and TP53-Independent Mechanism. mBio. 2019; 10. https://doi.org/10.1128/mBio.02690-18 PMID: 30723127

49. Vieira VC, Leonard B, White EA, Starrett GJ, Temiz NA, Lorenz LD, et al. Human Papillomavirus E6 Triggers Upregulation of the Antiviral and Cancer Genomic DNA Deaminase APOBEC3B. mBio. 2014; 5. https://doi.org/10.1128/mBio.02234-14 PMID: 25538195

50. Warren CJ, Westrich JA, Van Doorslaer K, Pyeon D. Roles of APOBEC3A and APOBEC3B in Human Papillomavirus Infection and Disease Progression. Viruses. 2017; 9. https://doi.org/10.3390/v9080233 PMID: 28825669

51. Wallace NA, Münger K. The curious case of APOBEC3 activation by cancer-associated human papillomaviruses. PLOS Pathog. 2018; 14: e1006717. https://doi.org/10.1371/journal.ppat.1006717 PMID: 29324878

52. Qiu J, Söderlund-Venermo M, Young NS. Human Parvoviruses. Clin Microbiol Rev. 2017; 30: 43–113. https://doi.org/10.1128/CMR.00040-16 PMID: 27806994

53. Chen KC, Shull BC, Lederman M, Stout ER, Bates RC. Analysis of the termini of the DNA of bovine parvovirus: demonstration of sequence inversion at the left terminus and its implication for the replication model. J Virol. 1988; 62: 3807–3813. https://doi.org/10.1128/JVI.62.10.3807-3813.1988 PMID: 2843676

54. Sharma S, Patnaik SK, Taggart RT, Kannisto ED, Enriquez SM, Gollnick P, et al. APOBEC3A cytidine deaminase induces RNA editing in monocytes and macrophages. Nat Commun. 2015; 6: 1–15. https://doi.org/10.1038/ncomms7881 PMID: 25898173

55. Sharma S, Patnaik SK, Kemer Z, Baysal BE. Transient overexpression of exogenous APOBEC3A causes C-to-U RNA editing of thousands of genes. RNA Biol. 2017; 14: 603–610. https://doi.org/10.1080/15476286.2016.1184387 PMID: 27149507

56. Sharma S, Wang J, Alqassim E, Portwood S, Cortes Gomez E, Maguire O, et al. Mitochondrial hypoxic stress induces widespread RNA editing by APOBEC3G in natural killer cells. Genome Biol. 2019; 20: 37. https://doi.org/10.1186/s13059-019-1651-1 PMID: 30791937

57. Perelygina L, Chen M, Suppiah S, Adebayo A, Abernathy E, Dorsey M, et al. Infectious vaccine-derived rubella viruses emerge, persist, and evolve in cutaneous granulomas of children with primary immunodeficiencies. PLOS Pathog. 2019; 15: e1008080. https://doi.org/10.1371/journal.ppat.1008080 PMID: 31658304

58. Wang S-M, Wang C-T. APOBEC3G cytidine deaminase association with coronavirus nucleocapsid protein. Virology. 2009; 388: 112–120. https://doi.org/10.1016/j.virol.2009.03.010 PMID: 19345973

59. Giorgio SD, Martignano F, Torcia MG, Mattiuz G, Conticello SG. Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2. Sci Adv. 2020; eabb5813. https://doi.org/10.1126/sciadv.abb5813 PMID: 32596474

60. van der Hoek L. Human coronaviruses: what do they cause? Antivir Ther. 2007; 12: 651–658. PMID: 17944272

61. Hayward JA, Tachedjian M, Cui J, Cheng AZ, Johnson A, Baker ML, et al. Differential Evolution of Antiretroviral Restriction Factors in Pteropid Bats as Revealed by APOBEC3 Gene Complexity. Mol Biol Evol. 2018; 35: 1626–1637. https://doi.org/10.1093/molbev/msy048 PMID: 29617834

62. Gordon DE, Jang GM, Bouhaddou M, Xu J, Obernier K, White KM, et al. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. Nature. 2020; 1–13. https://doi.org/10.1038/s41586-020-2286-9 PMID: 32353859

63. Zhang W, Wang H, Li Z, Liu X, Liu G, Harris RS, et al. Cellular Requirements for BIV Vif-Mediated Inactivation of Bovine APOBEC3 Proteins. J Virol. 2014. https://doi.org/10.1128/JVI.02072-14 PMID: 25142583

64. Cagliani R, Forni D, Clerici M, Sironi M. Coding potential and sequence conservation of SARS-CoV-2 and related animal viruses. Infect Genet Evol. 2020; 83: 104353. https://doi.org/10.1016/j.meegid.2020.104353 PMID: 32387562

65. Cheng AZ, Moraes SN de, Attarian C, Yockteng-Melgar J, Jarvis MC, Biolatti M, et al. A Conserved Mechanism of APOBEC3 Relocalization by Herpesviral Ribonucleotide Reductase Large Subunits. bioRxiv. 2019; 765735. https://doi.org/10.1101/765735

66. Nagaraju T, Sugden AU, Sugden B. Four-dimensional analyses show that replication compartments are clonal factories in which Epstein–Barr viral DNA amplification is coordinated. Proc Natl Acad Sci. 2019; 116: 24630–24638. https://doi.org/10.1073/pnas.1913992116 PMID: 31744871

67. Hoopes JI, Cortez LM, Mertz TM, Malc EP, Mieczkowski PA, Roberts SA. APOBEC3A and APOBEC3B Preferentially Deaminate the Lagging Strand Template during DNA Replication. Cell Rep. 2016; 14: 1273–1282. https://doi.org/10.1016/j.celrep.2016.01.021 PMID: 26832400

68. Seplyarskiy VB, Soldatov RA, Popadin KY, Antonarakis SE, Bazykin GA, Nikolaev SI. APOBEC-induced mutations in human cancers are strongly enriched on the lagging DNA strand during replication. Genome Res. 2016; 26: 174–182. https://doi.org/10.1101/gr.197046.115 PMID: 26755635

69. Haradhvala NJ, Polak P, Stojanov P, Covington KR, Shinbrot E, Hess JM, et al. Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair. Cell. 2016; 164: 538–549. https://doi.org/10.1016/j.cell.2015.12.050 PMID: 26806129

70. Kazanov MD, Roberts SA, Polak P, Stamatoyannopoulos J, Klimczak LJ, Gordenin DA, et al. APOBEC-Induced Cancer Mutations Are Uniquely Enriched in Early-Replicating, Gene-Dense, and Active Chromatin Regions. Cell Rep. 2015; 13: 1103–1109. https://doi.org/10.1016/j.celrep.2015.09.077 PMID: 26527001

71. Vartanian J-P, Henry M, Marchio A, Suspène R, Aynaud M-M, Guétard D, et al. Massive APOBEC3 editing of hepatitis B viral DNA in cirrhosis. PLoS Pathog. 2010; 6: e1000928. https://doi.org/10.1371/journal.ppat.1000928 PMID: 20523896

72. Kramvis A, Kostaki E-G, Hatzakis A, Paraskevis D. Immunomodulatory Function of HBeAg Related to Short-Sighted Evolution, Transmissibility, and Clinical Manifestation of Hepatitis B Virus. Front Microbiol. 2018; 9. https://doi.org/10.3389/fmicb.2018.02521 PMID: 30405578

73. Casey JL. Control of ADAR1 Editing of Hepatitis Delta Virus RNAs. In: Samuel CE, editor. Adenosine Deaminases Acting on RNA (ADARs) and A-to-I Editing. Berlin, Heidelberg: Springer; 2012. pp. 123–143. https://doi.org/10.1007/82_2011_146 PMID: 21732238

74. Nair S, Zlotnick A. Asymmetric Modification of Hepatitis B Virus (HBV) Genomes by an Endogenous Cytidine Deaminase inside HBV Cores Informs a Model of Reverse Transcription. J Virol. 2018; 92. https://doi.org/10.1128/JVI.02190-17 PMID: 29491156

75. Henderson S, Chakravarthy A, Su X, Boshoff C, Fenton TR. APOBEC-Mediated Cytosine Deamination Links PIK3CA Helical Domain Mutations to Human Papillomavirus-Driven Tumor Development. Cell Rep. 2014; 7: 1833–1841. https://doi.org/10.1016/j.celrep.2014.05.012 PMID: 24910434

76. Smith NJ, Fenton TR. The APOBEC3 genes and their role in cancer: insights from human papillomavirus. J Mol Endocrinol. 2019; 62: R269–R287. https://doi.org/10.1530/JME-19-0011 PMID: 30870810

77. Tang K-W, Alaei-Mahabadi B, Samuelsson T, Lindh M, Larsson E. The landscape of viral expression and host gene fusion and adaptation in human cancer. Nat Commun. 2013; 4: 1–9. https://doi.org/10.1038/ncomms3513 PMID: 24085110

78. Starrett GJ, Buck CB. The case for BK polyomavirus as a cause of bladder cancer. Curr Opin Virol. 2019; 39: 8–15. https://doi.org/10.1016/j.coviro.2019.06.009 PMID: 31336246

79. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature. 2013; 500: 415–421. https://doi.org/10.1038/nature12477 PMID: 23945592

80. Howley PM, Pfister HJ. Beta genus papillomaviruses and skin cancer. Virology. 2015;479–480: 290–296. https://doi.org/10.1016/j.virol.2015.02.004 PMID: 25724416

81. Tommasino M. The biology of beta human papillomaviruses. Virus Res. 2017; 231: 128–138. https://doi.org/10.1016/j.virusres.2016.11.013 PMID: 27856220

82. Viarisio D, Müller-Decker K, Accardi R, Robitaille A, Dürst M, Beer K, et al. Beta HPV38 oncoproteins act with a hit-and-run mechanism in ultraviolet radiation-induced skin carcinogenesis in mice. PLOS Pathog. 2018; 14: e1006783. https://doi.org/10.1371/journal.ppat.1006783 PMID: 29324843

83. Adamson-Small LA, Ignatovich IV, Laemmerhirt MG, Hobbs JA. Persistent parvovirus B19 infection in non-erythroid tissues: Possible role in the inflammatory and disease process. Virus Res. 2014; 190: 8–16. https://doi.org/10.1016/j.virusres.2014.06.017 PMID: 24998884

The prevalence of back-and-forth substitutions in the evolutionary landscape of human viruses

The APOBEC3 proteins play a crucial role in the human innate immune system by targeting viral genomes at 5’-TC-3’ sites to restrict the virus life cycle. One notable indication of recent or persistent exposure to APOBEC3 activity is the depletion of the 5’-TC-3’ dinucleotides within the viral genome. This APOBEC3 footprint has been observed in certain viruses such as HPV and PyV viruses, human retroelements, and somewhat ambiguously in the case of HIV-1. However, while a few examples of viruses targeted by APOBEC3 have already been reported, the status of the majority of human virus species remains unexplored. Until now, a comprehensive overview of the APOBEC3 footprint across all virus species has been lacking. Therefore, our study aimed to investigate the presence of the APOBEC3 footprint in the largest possible set of human viruses.

We analyzed the genome of 33,400 human viruses for the depletion of APOBEC3-favored motifs. We demonstrate that the APOBEC3 selection pressure impacts at least 22% of all currently annotated human viral species. The *papillomaviridae* and *polyomaviridae* are the most intensively footprinted families, evidencing a selection pressure acting genome-wide and on both strands. Members of the *parvoviridae* family are differentially targeted in terms of both magnitude and localization of the footprint. Interestingly, a massive APOBEC3 footprint is present on both strands of the B19 erythroparvovirus; making this viral genome one of the most cleaned sequences for APOBEC3-favored motifs. We also identified the endemic *coronaviridae* as significantly footprinted. Interestingly, no such footprint has been detected on the zoonotic MERS-CoV, SARS-CoV-1 and SARS-CoV-2 coronaviruses. In addition to viruses that are footprinted genome-wide, certain viruses are footprinted only on very short sections of their genome. That is the case for the *gamma-herpesviridae* and *adenoviridae* where the footprint is localized on the lytic origins of replication. A mild footprint can also be detected on the negative strand of the reverse transcribing HIV-1, HIV-2, HTLV-1 and HBV viruses.

Together, our data illustrate the extent of the APOBEC3 selection pressure on the human viruses and identify new putatively APOBEC3-targeted viruses.

The prevalence of back-and-forth substitutions in the evolutionary landscape of human viruses

Florian Poulain [1], Simon Dellicour [2,3] and Nicolas A. Gillet [1*]

[1] Namur Research Institute for Life Sciences (NARILIS), Integrated Veterinary Research Unit (URVI), University of Namur, Namur, Belgium

[2] Department of Microbiology, Laboratory of Clinical and Epidemiological Virology, Immunology and Transplantation, Rega Institute, KU Leuven, Leuven, Belgium

[3] Spatial Epidemiology Lab (SpELL), Université Libre de Bruxelles, Brussels, Belgium

* Corresponding authors: florian.poulain@unamur.be; nicolas.gillet@unamur.be

Abstract

Due to their large population size and high mutation rate, viruses have a strong capacity to evolve. They are constantly adapting to their environment, and in some cases, this adaptation leads to the reversion of substitutions to restore a higher fitness level. To capture the dynamics of virus substitution, a large quantity of observations was required. To construct a comprehensive collection of substitutions, we developed a bioinformatic pipeline that enables systematic clustering of virus sequences into operational taxonomic units, reconstruction of their evolutionary history through phylogenetic tree inference, and extraction of substitution context and dynamics. Through the clustering of 487 OTUs, we collected approximately 2.5 million substitutions from 55 viral species belonging to different groups of the Baltimore classification. The exploration of different classes of substitutions through substitution landscapes highlights the presence of a mirror effect, where compensatory substitutions exist for each other. This mirror effect appears to be a result of frequent and persistent events of substitution reversion. The investigation of nucleotide reversion events revealed that approximately 20% of the detected substitutions undergo back-and-forth exchanges. The presence of such reversion events at specific sites appears to be primarily associated with a high position-specific substitution rate. As no differences were found between back-and-forth and non-compensated reversions, we propose that reversion affects a wide range of substitution rates. We also report the presence of fast-evolving positions on the genomes of all groups of human viruses, including the dsDNA viruses. This observation suggests a potentially higher capacity for evolution than previously reported. Together, these findings shed new light on the evolutionary dynamics of human viruses.

Introduction

Viruses are the fastest evolving entities thanks to a high mutation rate combined with a short life cycle, with each cycle generating a large progeny. The rate of virus evolution is typically expressed as the number of substitutions per nucleotide per year. This rate has been reported to be significantly different between viral families with values greater than 10^{-3} substitution per nucleotide per year (s/n/y) for the most rapidly evolving viruses and less than 10^{-8} s/n/y for the most stable. The substitution rate is notably influenced by the type of nucleic acids composing the viral genome (DNA or RNA, single-stranded or double-stranded), by the length of the viral genome, and by the polymerase(s) responsible for viral replication (with or without proofreading capabilities)[1].

The rate of substitution of a given virus has also been shown to depend on the time scale between sample collections. The longer the time period over which different sequences are collected, the lower the measured substitution rate [2,3]. The time-dependence of the rates (or TDRP for time-dependence rate phenomenon) has been explained by processes such as sequence site saturation, purifying selection, and methodological artifacts causing overestimation of short-term rates and underestimation of long-term rates [2,4].

Among the processes that affect the dynamics of virus evolution, reversion remains poorly characterized. A reversion describes a population that returns to an ancestral state [5]. At the level of a single substitution, it refers to a return to the initial nucleotide or amino acid. Escape mutations at antigenic epitopes followed by their reversions have been reported for several viruses; for example, during HIV1 [6,7] or HCV evolution [8]. Reversions to the consensus sequence appear to be positively selected in HIV1 and may explain part of the time-dependence of the substitution rate [9].

In this study, we reconstructed 487 phylogenetic trees from 55 viral species spanning 23 families and the 7 Baltimore groups. Ancestral sequences were predicted for each node of the phylogenetic trees. Thus, by systematically comparing sequences to their ancestor, we generated a collection of over 2.4 million substitutions. For each substitution type, the immediate 5’ and 3’ bases were taken into account, dividing the substitution types into 192 subclasses, the so-called the substitution landscape. We observed a high degree of symmetry within the substitution landscapes, where each substitution class appears to be canceled out by its opposite. Along those lines, we observed that a significant proportion of the substitutions are back-and-forth, i.e. a succession of a first mutation followed by its reversion at a later time-point along the same branch.

Materials and methods

Sequence collection

Viral sequences were downloaded from NCBI or GISAID databases. Only full-length genomes from human-hosted viruses with a collection date were retained (Fig 1A). In the case of segmented viruses, each segment was treated as an independent viral species. All sequence IDs are reported in Table S1.

Sequence clustering into operational taxonomic units (OTU)

Viral genome sequences have been assigned into different pre-OTUs based on their taxonomic information. We used the taxonomic levels of species and subspecies (subspecies, group, type, subtype, serotype, genotype, or taxon) to group the sequences into preOTUs (Fig 1A). For some viruses, different taxonomic levels have been used to generate the preOTUs. For example, in the case of HIV1, some preOTUs gather sequences grouped according to the group (M, N, O or P) while other preOTUs gather sequences grouped according to the subtype (n.b. group M has 12 different subtypes).

PreOTUs counting fifty sequences or less, were directly tested for temporal signal. Detection of temporal signal qualifies the preOTU as an OTU. Failure to detect a temporal signal led to the discard of that preOTU. For the preOTUs containing more than fifty sequences, fifty sequences were randomly selected and tested for temporal signal. Detection of temporal signal qualifies the preOTU as an OTU. Failure to detect a temporal signal led to the random redraw of fifty sequences. A maximum of one hundred and twenty iterations were performed to obtain a maximum of 3 OTUs with a valid temporal signal. Temporal signal was tested by a root-to-tip regression approach [10]. Briefly, sequences of a given preOTU were aligned by Kalign multiple sequence aligner, a phylogenetic tree was then inferred by FastTree and rooted by the Reroot package. The resulting tree was finally tested for temporal signal by the Temporal_signal_functions R package and retained when the test p-value was below 0.01 (Fig 1A).

Bayesian phylogenetic inference of the sequences of the OTUs

Time-scale phylogenetic trees were generated for each OTU by Bayesian phylogenetic inference using the BEAST1.10 software package (Fig 1A). We used a general time-reversible substitution model permitting variation in substitution rate among sites and the presence of invariable sites (GTR+Γ+I model) [11–13]. We assumed constant population size. We chose an uncorrelated relaxed molecular clock model to allow the substitution rate to vary from branch to branch [14].

Substitution extraction from OTU phylogenetic trees

For each OTU, the Bayesian phylogenetic inference generated a posterior distribution of trees. To remain in a Bayesian framework and to take into account the uncertainty associated with the phylogenetic inference, subsequent analyses were systematically performed on 100 trees sampled from each posterior distribution. To extract the substitutions, each sequence has been compared to its direct ancestor. Counting of substitutions was done on 100 trees per OTU and median values were reported (Fig 1C).

Substitution landscapes similarity

The percentage of similarity (*Psim)* between two substitution landscapes X and Y was calculated as follow:

$$Psim = 100 - \sum_{i=1}^{i=j} \left| x_i - y_i \right|$$

Where x_i is the percentage of substitution of the i^{th} subclass of landscape X

Where y_i is the percentage of substitution of the i^{th} subclass of landscape Y

Where *j* is the total number of subclasses; *j*=96 when comparing landscape symmetry and *j*=192 when comparing back-and-forth versus uncompensated substitution landscapes.

Figure formatting and statistical analysis:

All figure formatting and statistics have been performed by R and ggplot collection packages.

Results

Comprehensive catalog of substitutions occurring during virus evolution

To build a catalog of viral genomes substitutions, we developed a bioinformatics pipeline that systematically retrieves sequences from human viruses, classifies them, reconstructs phylogenetic trees and calls for the substitutions occurring along those trees. The pipeline consists of four distinct steps: sequence download, preOTU (pre-operational taxonomic units) assembly, OTU curation with temporal signal and Baysian phylogenetic tree reconstruction for each OTU (Fig 1A).

Viral sequences were downloaded from the NCBI database [15]. In addition, influenza sequences were downloaded from the GISAID database [16]. Only the full-length genome sequences of human-hosted viruses with a collection date were retained. Sequences were then clustered in preOTU using the lowest taxonomic level available (e.g. class, group, serotype, subtype, type, or subspecies). For instance, HIV1 sequences were clustered by subtype, whereas papillomavirus sequences were clustered by type. The taxonomic data for each sequence was obtained through automated reading of the GenBank files. At this point, 13,000 preOTUs were aggregated.

The second step consisted of testing the preOTUs for the presence of a temporal signal among the sequences by a root-to-tip regression approach [10]. At this stage, 487 OTUs were retained and processed further for phylogenetic tree reconstruction (Table S1).

Time-scaled phylogenetic trees for the retained OTUs were generated by Bayesian phylogenetic inference. We used a relaxed molecular clock model to allow substitution rate to vary among branches [14], and a general time-reversible substitution model permitting variation in substitution rate among sites and the presence of invariable sites (GTR+Γ+I model) [11–13].

Substitutions were extracted from OTU phylogenetic trees. For each OTU, the Bayesian phylogenetic inference generated a posterior distribution of trees. To remain in a Bayesian framework and to take into account the uncertainty associated with the phylogenetic inference, substitutions calling were performed on 100 trees sampled from each posterior distribution. To extract the substitutions, each sequence was compared to its direct ancestor (Fig 1A).

We finally generated phylogenetic trees for 487 OTUs covering 55 viral species and spanning 23 families and 7 groups of the Baltimore classification (Fig 1B). Substitution calling identified around 2.4 million substitutions with a bias for transitions over transversions (Fig 1C). The 5’ and 3’ immediate base contexts were taken into account to draw the substitution landscapes counting 192 subclasses made by the 12 substitution classes × 4 types of the 5′ immediate upstream base × 4 types of the 3′ immediate downstream base (Fig 1C). Figure 1D displays the substitution rates computed for each of the 487 OTUs and allocated them into 55 different viral species. The substitution rates are normalized relative to the genome size; expressed as number of substitutions per nucleotide and per year (s/n/y). The substitution rates estimated from our dataset corroborate the rates reported in the literature (Fig S1).

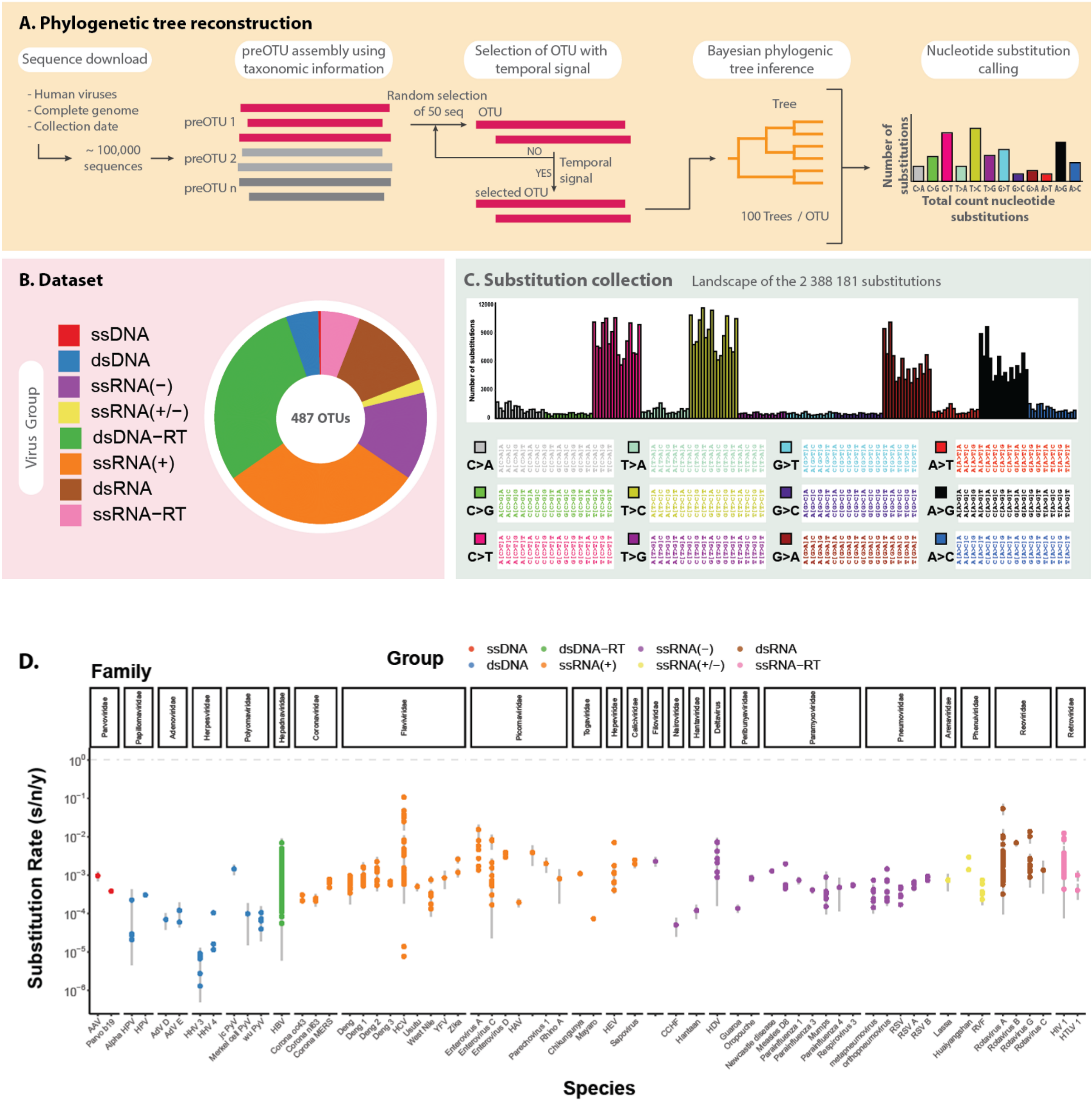


Figure 1: Implementation of a systematic pipeline of viral substitutions calling to construct a large substitution collection. A. The developed pipeline includes five steps. Genomic sequences of human viruses were first downloaded from the NCBI or GISAID databases. The sequences were then grouped according to their taxonomic information, forming preOTUs (pre-operational taxonomic units). Fifty sequences were chosen at random in each preOTU and retained to constitute a definitive OTU on the condition of detecting a temporal signal. Bayesian phylogenetic tree reconstructions were done on 487 different OTUs and nucleotide substitutions were called on a selection of 100 trees per OTU. **B.** This approach produced a dataset of 487 OTUs belonging to 55 viral species of the four viral groups. **C.** The full substitution collection can be represented by a bar plot also called substitution landscape. The substitutions were divided into 12 classes (colored), further subdivided according to the type of the upstream and downstream base. **D.** The substitution collection can also be reported by their substitutions rates. The 487 values for OTUs are grouped in 55 virus species and colored according to the viral group. The rates were calculated based on a selection of 100 trees per OTU. Each dot represents the median substitution rate of a given OTU with an error bar indicating standard deviation.

Viral substitution landscapes display a high degree of symmetry

As shown in Figure 1, the pipeline allows the splitting of the 12 types of substitutions into a total of 192 subclasses by considering the bases at the 5’ and 3’ positions of the mutation. Thus, for each OTU, we generated a substitution landscape by reporting the proportion of each of the 192 substitution subclasses. In Figure 2A, each type of substitution is placed tail-to-tail with its opposite substitution. For example, the subclass C[C>T]G is placed tail-to-tail to C[T>C]G. Using this representation, we observed symmetry within the substitution landscapes. Such a mirror effect is obvious for numerous viruses, like HIV1 or HTLV1. The percentage of “mirror effect” was calculated for each OTU by comparing the upper and lower parts of the substitution landscape as described in the material and methods (percentage of similarity). High percentages of mirror effect were detected in every viral group. The mirror effect is positively correlated with the number of substitutions (Fig. 2B). The higher the number of substitutions per tree, the higher the percentage of symmetry (Fig. 2B).

The GTR substitution model used for the phylogenetic tree inference used equal proportions for opposite types of substitution (C>T = T>C). Because the mirror effect could be generated by the substitution model, the presence of substitution landscape symmetry was assessed before and after inference on artificially generated OTU (Fig S2). These OTUs sequenced were initially produced by the introduction of a control substitution landscape according to a real phylogenetic tree topology. The analysis of fourteen test OTUs from diverse viral groups and species revealed that the phylogenetic tree inference did not introduce any mirror effect. Furthermore, an OTU with a high proportion of mirror effect on the input landscape maintained the same mirror effect after pipeline inference and substitution calling (Fig S2).

Considering that the observed symmetry between opposite subclasses of substitution could potentially be explained by reversion events, we further investigated the presence of such reversions in the substitution collection.

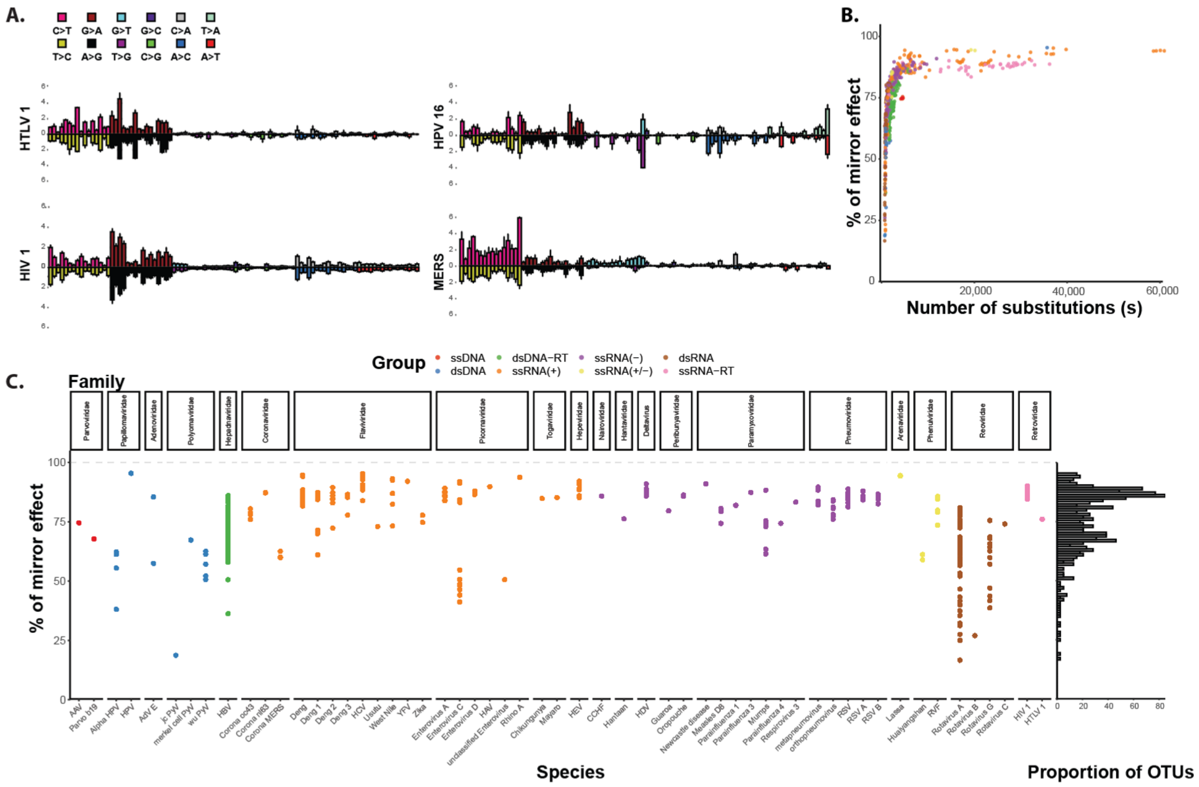


Figure 2: Viral substitution landscapes display a high degree of symmetry also called mirror effect. A. The reorganized substitution landscapes, with opposite subclasses of substitution place tail-to-tail revealed symmetry between the upper part and the lower part of the landscape. The five different viruses (HPV16, HIV1, HTLV-1 and MERS) are depicted by bar plots reporting the substitution frequencies for 192 subclasses (12 substitution classes × 4 types of the 5′ immediate upstream base × 4 types of the 3′ immediate downstream base). Error bar indicates standard error of the mean (HPV16 substitution landscape was averaged from the substitution landscapes of 4 different OTUs, HIV1 landscape was averaged from 35 OTUs, HTLV-1 from 2 OTUs and MERS from 3 OTUs). **B.** Percentage of symmetry is calculated by the comparison of the upper and lower part of each OTU substitution landscapes. The percentage of symmetry asymptotically approaches 100% with the increase of the total number of substitutions detected along the different phylogenetic trees. **C.** The detail of this percentage of symmetry for each given OTU is reported by a dot plot. One dot corresponds to one OTU, grouped by species and colored by viral group. Histogram on the right-hand side reports the distribution of the percentage of symmetry.

Back-and-forth substitutions are frequent during virus evolution

We define as back-and-forth substitution, the succession of a first mutation followed by its reversion at a later time-point along the same branch (Fig 3A). The second substitution will be reversion to the initial state. On average, 20% of the observed substitutions are back-and-forth mutations (Fig 3B). Up to 50% of the substitutions are back-and-forth in HCV phylogenetic trees (Fig 3B). Back-and-forth substitutions were detected in 487 among the 473 OTUs.

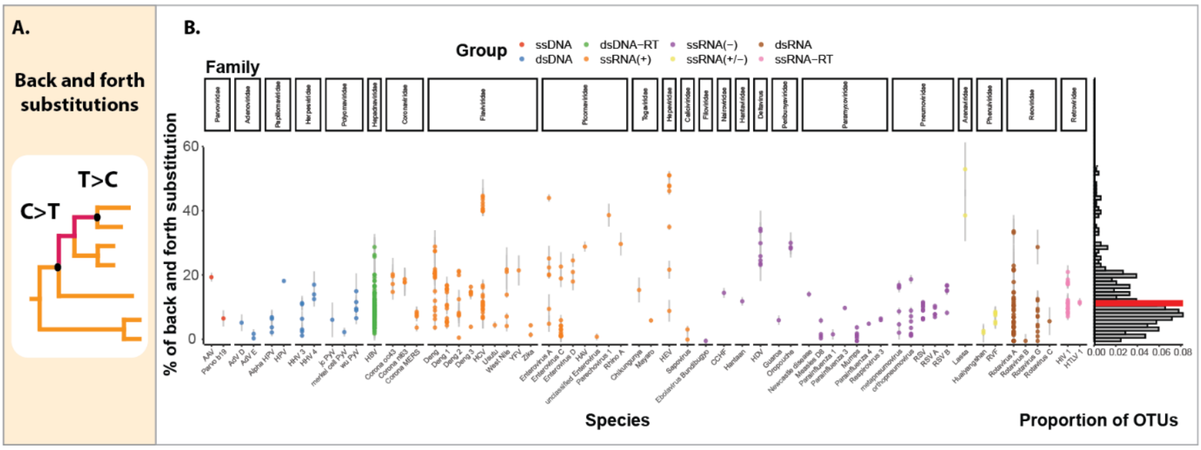


Figure 3: Back-and-forth substitutions are frequent during virus evolution. A. We define as back-and-forth substitution, a substitution that will undergo reversion in either descendant sequence. **B.** Proportion of back-and-forth substitutions for the 487 OTUs. Each dot represents the median proportion of back-and-forth substitutions of a given OTU with an error bar indicating standard deviation. Histogram on the right-hand side reports the aggregated distribution of back-and-forth substitutions. The red mine reports the median value of 20.6 % of back and forth substitution on the full dataset substitutions.

As opposed to the back-and-forth substitutions (BF), the uncompensated substitutions (UC) include substitutions that did not revert later in descending sequences (Fig 4A). For each OTU, we divide the substitution landscape into a landscape for back-and-forth substitutions and a landscape for uncompensated substitutions. One example is illustrated in Figure 4B for HCV. To test landscape similarity for the entire dataset, we calculated, for each OTU, a percentage of similarity between the landscapes of back-and-forth and uncompensated substitutions. Figure 4C shows that the percentage of similarity increases with the total number of substitutions and approaches perfect similarity when several thousand of substitutions can be detected along the tree.

We then tested whether BF and UC substitutions differ regarding their synonymous over non-synonymous ratios. Figures 4D and Table S2 display the proportion of synonymous and non-synonymous mutations for the BF and UC categories for 166 OTUs. Among the 166 tested OTUs, only 15 showed a statistical enrichment for synonymous in the back-and-forth subgroup (Table S2).

While the BF and UC substitutions did not appear to differ in their landscape nor their synonymous-to-non-synonymous ratios, they seem to differ in their rate of evolution. Thus, the sites (genome coordinate) with back-and-forth events (black dots) have a systematically higher substitution rate than the sites where no back-and-forth have been detected (gray dots) (Fig 4E).

Fast-evolving positions are present in all virus groups

Thanks to the reconstitution of ancestral sequences, we were able to calculate the substitution rates for each position (i.e. each nucleotide) of the viral genome. Figure 5A displays what we called the single position substitution rates; expressed as number of substitutions per year (s/y).

By utilizing this metric, dsDNA viruses no longer exhibit distinct characteristics compared to viruses of other groups. To compare the distribution of single position substitution rates without bias arising from the larger number of substitution rates in bigger sequences, we calculated the median value of non-zero single position substitution rates for each OTU. These median values were found to be negatively correlated with the measurement timescale (Fig 5B). Importantly, there was no apparent correlation between single position substitution rates and genome size (Fig 5C). Both small and large viruses display positions that change at seemingly elevated rates.

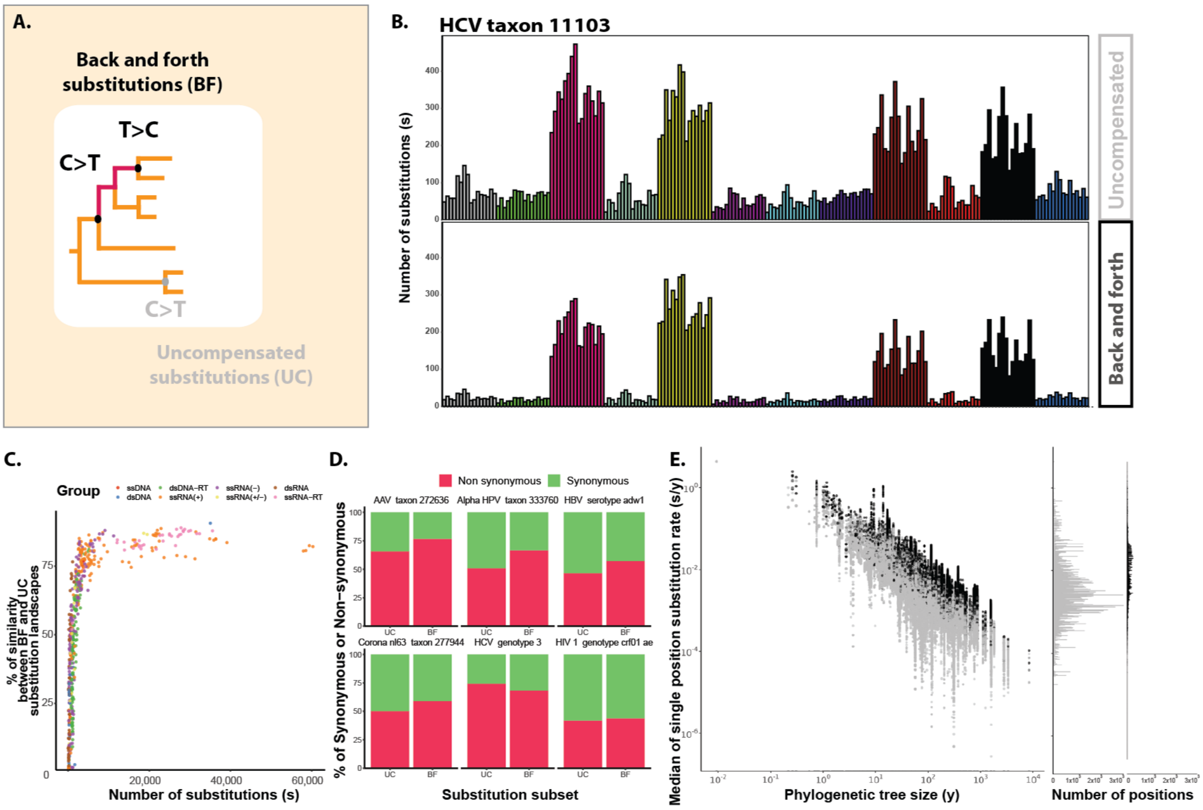


Figure 4: Back-and-forth substitutions and uncompensated substitutions display similar landscapes and similar synonymous to non-synonymous ratios. A. In contrast to back-and-forth substitutions, we define an uncompensated substitution as a substitution that has not reverted in subsequent descendants. **B.** The comparison between the uncompensated (upper plots) and the back-and-forth (lower plots) substitution landscapes for the OTU HCV NCBI taxon id 11103 do not reveal differences. Bar plot is colored according to the substitution type. **C.** The calculation for proportion of similarity between the back-and-forth and uncompensated substitution landscapes allows analysis of all the OTU dataset. These values asymptotically approach 100% with the increase of the total number of substitutions detected along the different phylogenetic trees. **D.** Differences for the proportion of non-synonymous (red) and synonymous (green) substitutions within the back-and-forth and uncompensated substitutions were also tested. Bar plots report the proportion observed for six OTUs from different virus species, **E.** The relationship between the substitution rate at individual positions and the occurrence of back-and-forth events was also examined. Positions undergoing back-and-forth substitutions (represented by black dots) exhibited higher substitution rates compared to positions experiencing uncompensated substitutions (represented by gray dots).

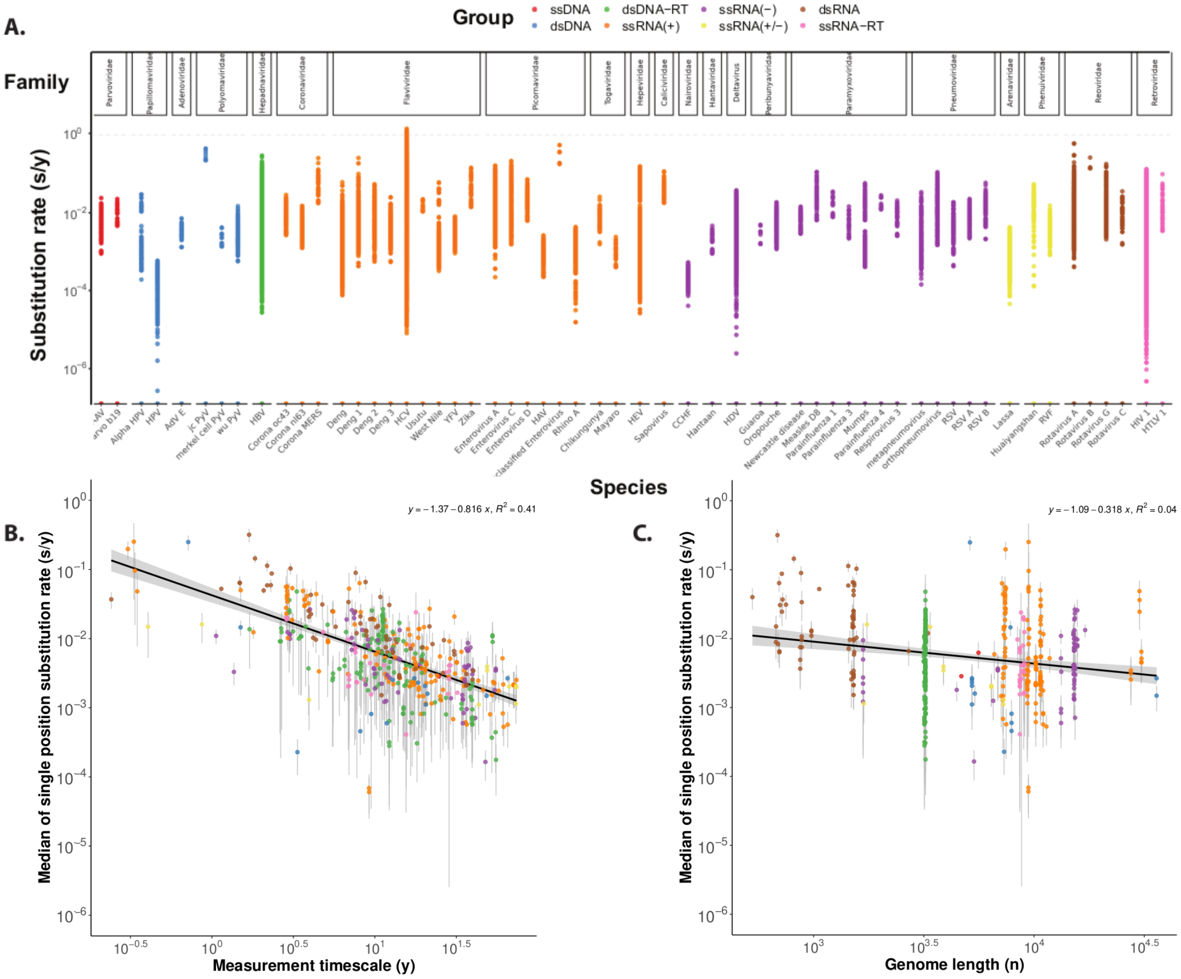


Figure 5: Fast-evolving positions are present in all virus groups. A. A detailed representation of the single position substitution rates for each genome of the 487 OTUs is presented in a dot plot. The OTUs are categorized into 55 virus species and color-coded based on their viral groups. For each position, the rates were calculated using a selection of 100 trees per OTU. Each dot on the plot represents the median substitution rate at a specific genome position for a given OTU. **B.** The median values of the single position substitution rates across the genome, calculated for each OTU, exhibit a negative correlation with the OTU measurement timescales. The relationship is illustrated by a black line indicating the best-fit linear regression. **C.** In contrast, the median values of the single position substitution rates across the genome for the OTUs show no correlation with the genome size. This is demonstrated by a black line representing the best-fit linear regression.

Discussion

Could the mirror effect and the back-and-forth of substitutions explain the apparent evolutionary stasis observed over long periods of time?

The description of the substitution landscapes has revealed an interesting mirror effect between pairs of substitution subclasses that can potentially compensate for each other. The mirror effect seems to shape all families of viruses. Importantly, the mirror effect is positively correlated with the number of substitutions identified along the inferred tree. Thus, providing that we analyze a tree large enough to yield several thousand of substitutions, we might observe a perfect mirror effect for each viral species (with, perhaps, the exception of recent zoonotic viruses as discussed later).

Back-and-forth substitutions would be the driving force responsible for the mirror effect. In our dataset, we observed that, on average, 20% of the substitutions are back-and-forth. But, maybe most importantly, the substitutions that are not back-and-forth (the uncompensated substitutions) appear barely different from the back-and-forth. While we expected the back-and-forth substitutions to show lower non-synonymous/synonymous ratios than the uncompensated substitutions, we found no significant difference between them. Nor did we find significant differences between their landscapes. This led us to propose that we only captured a fraction of the back-and-forth substitutions and that most of the uncompensated substitutions could be mislabeled.

Our observations could support the idea that genomic sequences of all viruses (not only the dsDNA but also the RNA viruses) are more stable over time than is generally thought. Indeed, the observation of the TDRP and the estimation of evolution rates as low as 10^{-9} s/n/y when measured over long period of time (millions of years) lead to the proposal that, over long timescales, the evolution rates of viruses would approach those of their host [17]. Thus, indicating the rate of evolution of a given virus only by a number of substitutions per nucleotide per year only makes sense if we associate this value with the period from which the rate is estimated. Like the consensus sequence used to describe the average sequence of a quasi-species cloud, it might be of interest to define the “time-traveling consensus sequence” as an average of the viral sequences over a given timeframe.

Are the back-and-forth substitutions “loose screws” of the viral genomes or key sites for immune escape?

We observed that the back-and-forth substitutions are not enriched for synonymous changes compared to the uncompensated mutations. In fact, for some viruses, the back-and -forth substitutions are more frequently non-synonymous than synonymous (Fig 4D and Table S2). It will be necessary to study further the impact of the BF substitutions on the viral proteins, but they might very well contribute to the immune escape. Escape mutations at antigenic epitopes followed by their reversions (i.e. back-and-forth substitutions) have been reported for several viruses [6–8]. Thus, we could describe the virus as a dice that will expose, at a given time point, one of its faces to the immune response. Once the host's immune response has adapted to a given face (herd immunity achieved), the die will be re-rolled to escape the response and a new face will be exposed (selection of a new variant).

The shape of the virus (i.e. the conformation of its structural proteins and its repertoire of antigenic peptides) will oscillate around a long-term consensus sequence. The short-term substitution rates will constitute the metric of these oscillations (frequency of dice rolls).

Asymmetry within the substitution landscapes of recent zoonotic viruses

We observed a remarkable asymmetry of the substitution landscape for the MERS coronavirus (Fig 2A and Fig S3) despite the collection of a high number of substitutions. This substitution landscape is very similar to the landscape reported for the SARS-CoV-2 coronavirus with the C to T and G to T substitutions that are not totally compensated by the T to C and T to G changes. On the contrary, the substitution landscapes of the endemic coronavirus OC43 and NL63 show a high level of symmetry (Fig. 2C) with a percentage of mirror effect above 75%. Although based on circumstantial evidence, it is tempting to propose that the lack of mirror effect observed for MERS and for SARS-2 is the result of still incomplete adaptation to its host (Fig S3). The shape of the virus continues to change to adapt to its new ecological niche.

Short term evolution of dsDNA viruses may not be so slow after all

The evolution rates of dsDNA viruses were consistently reported as lower compared to the other viral groups. This has been explained by the fact that those viruses rely on host or viral DNA polymerases which are highly faithful and capable of proofreading activity. The first group of Baltimore, those of the dsDNA viruses, comprises viruses with very different genome lengths. The papilloma- and-polyomaviruses have short genomes of between 5 to 8kb, adenoviruses are more medium-length with genomes about 35kb-long and finally the herpes- and poxviruses have large genomes of 120 to 240kb. Small genome viruses tend to display higher rates than large genome viruses [19]. To exclude the parameter of genome size, we computed single position substitution rates (those rates can only be computed for the sites where we detected one or several substitutions along the inferred trees). We observed that fast evolving sites (rates of 10^{-2} substitution per year for a time frame of 1 year) were detected in every viral species, and also in dsDNA viruses. With this observation we would like to reconsider the common assumption that dsDNA viruses adapt less rapidly than viruses from the other groups. Indeed, modification of a few sites are sufficient to allow immune evasion.

In this study, we reconstructed 487 phylogenetic trees from 55 viral species spanning 23 families and the 7 Baltimore groups. The estimated substitution rates were in accordance with the numerous studies conducted previously. Importantly, by looking at the substitution rate at the level of a unique site, we observed that fast-evolving positions are present in every viral group, even in dsDNA viruses. We also observed that, for most viruses, the substitution landscape displays a high degree of symmetry where substitutions appear to cancel each other. Along those lines, we observed that back-and-forth substitutions are very common during virus evolution. This leads us to propose that viral genome sequences are actually rather conserved over long periods of time despite fluctuating rapidly over short time frames.

References

1. Duffy S, Shackelton LA, Holmes EC. Rates of evolutionary change in viruses: patterns and determinants. Nat Rev Genet. 2008;9: 267–276. doi:10.1038/nrg2323

2. Duchêne S, Holmes EC, Ho SYW. Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates. Proc R Soc B Biol Sci. 2014;281: 20140732. doi:10.1098/rspb.2014.0732

3. Aiewsakun P, Katzourakis A. Time-Dependent Rate Phenomenon in Viruses. J Virol. 2016;90: 7184–7195. doi:10.1128/JVI.00593-16

4. Ghafari M, du Plessis L, Raghwani J, Bhatt S, Xu B, Pybus OG, et al. Purifying Selection Determines the Short-Term Time Dependency of Evolutionary Rates in SARS-CoV-2 and pH1N1 Influenza. Mol Biol Evol. 2022;39: msac009. doi:10.1093/molbev/msac009

5. Porter ML, Crandall KA. Lost along the way: the significance of evolution in reverse. Trends Ecol Evol. 2003;18: 541–547. doi:10.1016/S0169-5347(03)00244-1

6. Leslie AJ, Pfafferott KJ, Chetty P, Draenert R, Addo MM, Feeney M, et al. HIV evolution: CTL escape mutation and reversion after transmission. Nat Med. 2004;10: 282–289. doi:10.1038/nm992

7. Allen TM, Altfeld M, Yu XG, O'Sullivan KM, Lichterfeld M, Le Gall S, et al. Selection, Transmission, and Reversion of an Antigen-Processing Cytotoxic T-Lymphocyte Escape Mutation in Human Immunodeficiency Virus Type 1 Infection. J Virol. 2004;78: 7069–7078. doi:10.1128/JVI.78.13.7069-7078.2004

8. Timm J, Lauer GM, Kavanagh DG, Sheridan I, Kim AY, Lucas M, et al. CD8 Epitope Escape and Reversion in Acute HCV Infection. J Exp Med. 2004;200: 1593–1604. doi:10.1084/jem.20041006

9. Druelle V, Neher RA. Reversions to consensus are positively selected in HIV1 and bias substitution rate estimates. Virus Evol. 2022;9: veac118. doi:10.1093/ve/veac118

10. Rambaut A, Lam TT, Max Carvalho L, Pybus OG. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol. 2016;2: vew007. doi:10.1093/ve/vew007

11. Miura RM. Some Mathematical Questions in Biology: DNA Sequence Analysis. American Mathematical Soc.; 1986.

12. Shoemaker JS, Fitch WM. Evidence from nuclear sequences that invariable sites should be considered when sequence divergence is calculated. Mol Biol Evol. 1989;6: 270–289. doi:10.1093/oxfordjournals.molbev.a040550

13. Yang Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. J Mol Evol. 1994;39: 306–314. doi:10.1007/BF00160154

14. Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. Relaxed Phylogenetics and Dating with Confidence. PLOS Biol. 2006;4: e88. doi:10.1371/journal.pbio.0040088

15. NCBI Virus. [cited 24 Jun 2022]. Available: https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/

16. GISAID - Initiative. [cited 24 Jun 2022]. Available: https://www.gisaid.org/

17. Simmonds P, Aiewsakun P, Katzourakis A. Prisoners of war — host adaptation and its constraints on virus evolution. Nat Rev Microbiol. 2019;17: 321. doi:10.1038/s41579-018-0120-2

18. Yi K, Kim SY, Bleazard T, Kim T, Youk J, Ju YS. Mutational spectrum of SARS-CoV-2 during the global pandemic. Exp Mol Med. 2021;53: 1229–1237. doi:10.1038/s12276-021-00658-z

19. Sanjuán R. From Molecular Genetics to Phylodynamics: Evolutionary Relevance of Mutation Rates Across Viruses. PLOS Pathog. 2012;8: e1002685. doi:10.1371/journal.ppat.1002685

Supplementary figure

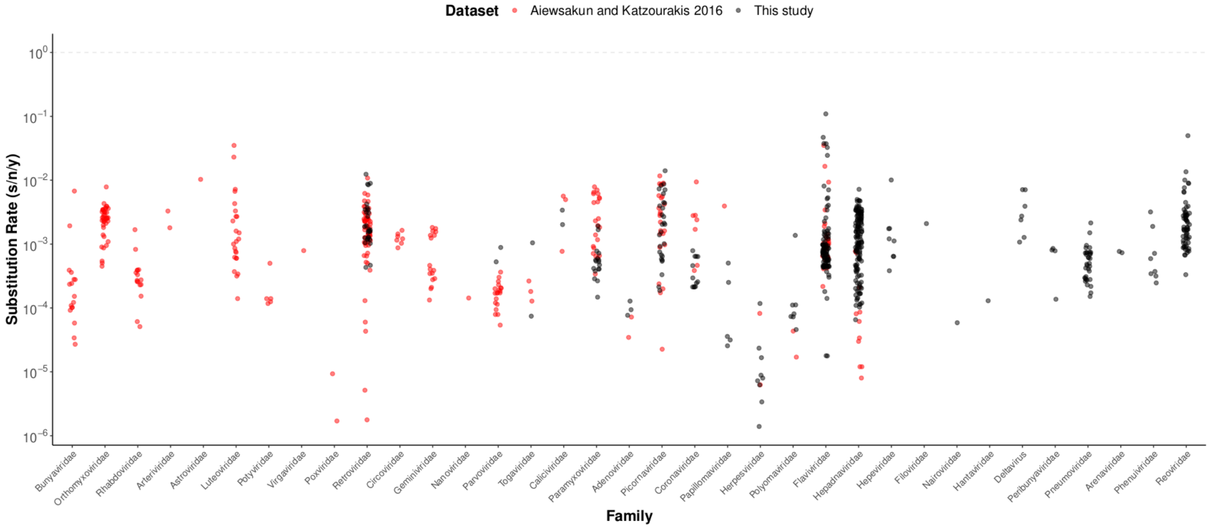
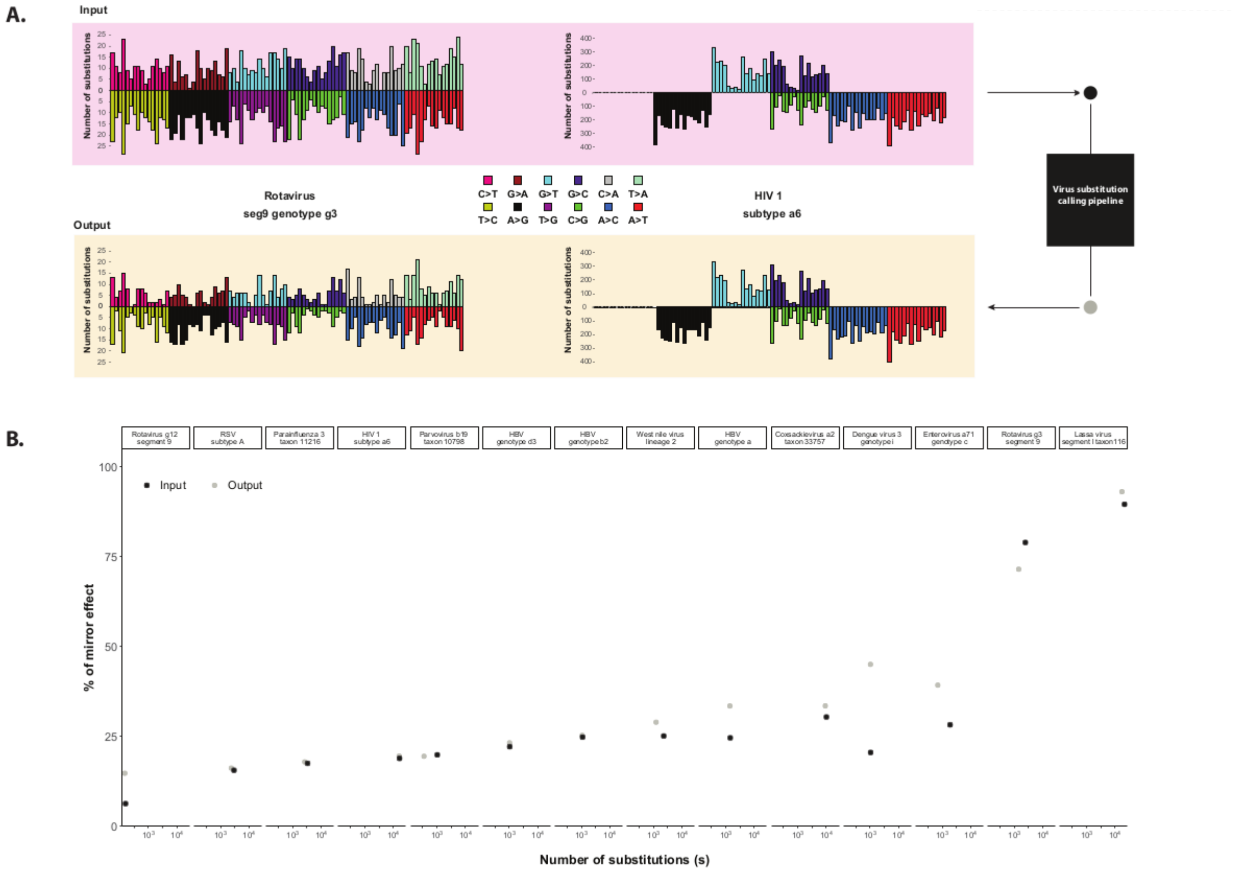


Figure S1 : The substitution rates of the dataset corroborate literature values . The substitution rates reported in the literature [3] are represented by red dots and those of the present work by black dots.



;

Figure S2: Control for potential artifactual mirror effect induced by phylogenetic tree inference. A. Test OTUs were generated by inserting a controlled substitution landscape (pink area) into virus genome sequences, based on the topology of real phylogenetic trees. The resulting OTUs fasta file was then processed using the virus substitution calling pipeline, which includes the inference of new phylogenetic trees. The output substitution landscape (yellow area) was extracted from these trees. This representation showcases the input and output substitution landscapes of the Test OTU Rotavirus segment 9 genotype g3 and the test OTU HIV1 subtype A6 for qualitative comparison. **B.** Comparison between the input (black dots) and output (gray dots) landscapes assess the proportion of the mirror effect for each of the fourteen test OTUs. This comparison confirms the absence of artifactual mirror effect resulting from OTU phylogenetic tree inference.

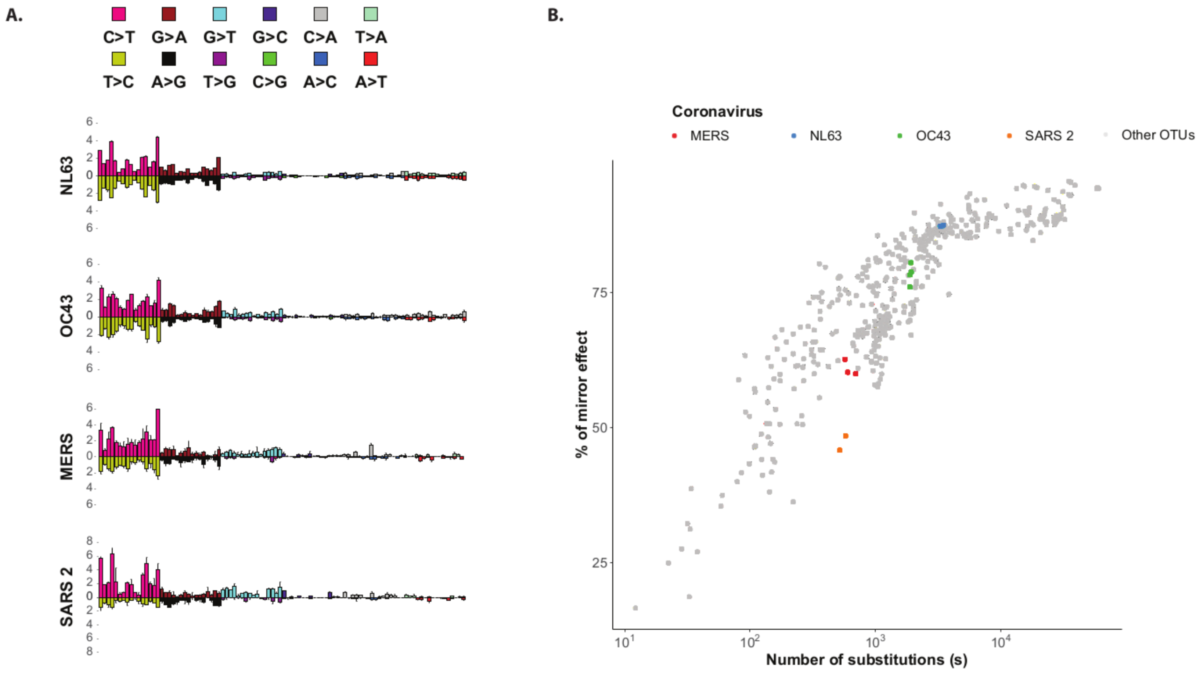


Figure S3 : Comparison of the mirror effect of endemic and emergent coronavirus substitution landscapes. A. Endemic coronavirus NL63 and OC43 substitution landscapes present a symmetry between opposite subclasses of substitutions which is not observed for the emergent coronavirus MERS and SARS-2. **B.** The percentage of mirror effect for the SARS-2 OTUs (orange dots) are in the lower range by comparison with the other OTUs (gray dots).

4. Discussion and Perspectives

The original objective of my thesis was the identification of the human viruses that evolve under the selection pressure of the APOBEC3 innate effectors. As it was well known that the APOBEC3 enzymes preferentially target the TC motifs and turn them into TT, the search for the APOBEC3 footprint was reduced to the search for TC depletion in viral genomes (or more precisely NTC codon depletion as explained previously).

The first part of the discussion will elaborate on the interesting case of the coronaviruses and on the difference of APOBEC3 footprint between endemic and zoonotic species. We will also propose several ideas to improve the identification of the APOBEC3 footprint (or indeed other mutational footprint) on viral genomes.

The search for the APOBEC3 fingerprint was possible because we had an a priori on the effect of these enzymes on the viral genomes (i.e. a depletion of TC motifs). The second part of this thesis was to identify mutational processes acting on viruses, shaping viral evolution, without any a priori. The idea was to take advantage of the tools developed by researchers in the field of cancer research to analyze tumor genomes. The first step of this project was to gather substitution landscapes for tens or hundreds of human viruses. This led to the second paper drafted in this manuscript. This task is still ongoing, and we will discuss the different steps that will be performed in the next coming months.

4.1. Footprint of the APOBEC3 enzymes on viruses

4.1.1. APOBEC3 footprint on endemic and zoonotic viruses: the interesting case of coronaviruses.

During our exploration of the APOBEC3 footprint on viruses, we observed a clear footprint on the four endemic coronaviruses NL63, HKU1, OC43 and 229E whereas no footprint on the recent zoonotic SARS-1, SARS-2 and MERS-CoV. We also failed to detect an APOBEC3 footprint (i.e. depletion of NTC codons) on the SARS and MERS viruses isolated from bats (the bats being the natural host) nor on the MERS viruses isolated from camels.

The lack of a footprint in bat coronaviruses is particularly surprising, considering the presence of 18 homologous APOBEC3 subdomains Z, with at least four of them possessing cytidine deaminase activity favoring the 5'-TC-3' dinucleotide, the same favored site as the human APOBEC3s [114]. Moreover, some of these domains can restrict HIV1-ΔVIF [114]. It is possible that, like the Vif protein for HIV, the bat coronaviruses possess antagonization mechanisms against the bat APOBEC3 proteins.

Although we did not observe an APOBEC3 footprint on SARS-2, several papers reported APOBEC3-related mutations on the SARS-2 genome based on inter and intra host sequencing data [76–78,115]. Although these two observations may seem contradictory, they remain compatible. The APOBEC3 editing may be too recent to significantly shape the viral genome with an APOBEC3 footprint (depletion of NTC codons). After many years of residing in the human host and becoming endemic, the virus will experience constant APOBEC3 restriction that may lead to a depletion of NTC k-mer and a detectable APOBEC3 footprint. The detection of mutations attributed to an APOBEC3 activity on the SARS-2 genome may be explained by an imperfect antagonization of the human APOBEC3 proteins.

Since the dates of origin of the zoonotic transmissions of the NL63, HKU1 and 229E coronaviruses are unknown, it is difficult to assess the dynamics of footprint appearance. However, insights can be gained by observing the OC43 coronavirus. This virus is believed to have originated from a rodent coronavirus ancestor that was transmitted to cattle, pigs, or other animals before infecting humans [116]. This human transmission appears to have occurred during the 1889-1890 pandemic of respiratory disease [117]] This suggests that an APOBEC3 footprint can shape the genome of an endemic virus with high circulation within a period of one hundred years. However, since we lack information about the presence or absence of APOBEC3 footprints in rodent and bovine coronaviruses, we cannot definitively interpret the presence of this footprint as resulting from adaptation to the human species.

We also observed a remarkable difference between the substitution landscapes of the endemic and zoonotic coronaviruses. The four endemic coronaviruses display a highly symmetric substitution landscape suggesting that those viruses reached an equilibrium with the host. On the contrary, the MERS and SARS-2 substitution landscapes are asymmetric with notably more C>T than T>C. This asymmetry may reflect the ongoing process of adaptation to its new ecological niche, i.e. the human host.

A recent model proposed by Simmonds et al., named the 'Prisoner of War' model, suggests that the mutational space available to the virus is strongly constrained by the host [118]. The virus is shaped by constraints imposed by the intra-host environment, particularly the need for the virus to utilize and manipulate the cellular machinery for its replication, as well as the requirement to evade or counteract the host immune response. Thus, over the long term (i.e. thousands to millions of years), the rate of evolution of the virus approximates the rate of evolution of its host.

When a virus jumps into a new species, it will face new constraints and will evolve by exploring its new mutational space. After a prolonged period of residence in its new host, the virus becomes adapted and its sequence becomes more stable with time. This model explains the lower substitution rates observed over long measurement timescales.

The observation of an asymmetric substitution landscape for the recent zoonotic viruses might suggest that they did not reach equilibrium with the host. They are still exploring their new ecological niche. If confirmed our results could be an illustration of the Prisoner Of War model.

4.1.2. Improved approaches to evidence of mutational footprints.

During virus evolution, a mutational process can induce selective pressure that shapes the genome with fixation of neutral substitutions, and loss of lethal or deleterious substitutions. If the exposure is constant, this process can progressively change the codon composition at synonymous sites, leaving a footprint on the genome. Our study focused on the detection of the APOBEC3 footprint among human virus species highlighted the possibility of using the codon composition from a large number of sequences to search for specific mutational pressure.

To investigate codon bias, we grouped codons into k-mers, such as the NTC k-mer, which contains the ATC, GTC, CTC, and TTC codons. This k-mer contains the specific APOBEC3 target site 5’-TC-3’, with the C at the third codon position. In future investigations, in contrast to the specific investigation we are currently conducting, we could develop an optimized, non-a priori approach for k-mer depletion. Some observations lead us to think that other mutational footprint can be detectable with such a modified approach. For instance, based on another model for randomization, Martinez *et al.* detected an AID footprint on the genome of EBV virus [105], although we were not able to report the same observation. Moreover, it is interesting to note that a bias in codon usage has been reported in many viral genomes. In some cases, viruses preferentially use codons with underrepresented tRNAs, which seems to be a suboptimal usage. However, this observation could possibly be explained by evolutionary constraints induced by mutational processes [119].

The optimization of a new k-mer depletion research approach should include the following implementations.

1. Firstly, capturing a potential k-mer depletion bias with a lower magnitude than the one observed for the APOBEC3 footprint may require a more sensitive approach with an improved randomization method. In our study, the calculation of the k-mer ratio corresponds to the ratio between the observed proportion of a specific k-mer (for example, NTC) and the expected proportion of that k-mer. The expected number of k-mers was determined using a simple nucleotide sequence randomization approach, which neutralized possible bias in nucleotide proportion but did not consider bias arising from codon usage, GC content, or dinucleotide composition. In the literature, three different elegant sequence randomization approaches have been proposed and tested by *Shapiro et al.* [120]. These approaches strictly conserve the sequence's amino acid composition and are based on the replacement of the third position nucleotide, according to the proportion of the purine or pyrimidine nucleotide in the codon's third position (gc3 model), the nucleotide type in the codon's third position (n3 model), or the dinucleotide content in positions 2 and 3 of the codons (dn23 model). *Shapiro et al* used these models to estimate k-mer bias and described an APOBEC3 target site under-representation in AAV2 and HPV16 genes [120] as well as in EBV genomes [105]. In our investigation, we also detected APOBEC3 footprints in HPV16, AAV1 (we did not test sequences for AAV2 in our analysis), and EBV. This observation suggests that our results are consistent with theirs.
2. Secondly, although the full genome approach was useful for investigating the APOBEC3 footprint, we found that our investigations at the gene scale revealed an even higher number of sequences shaped by mutational effects, as seen in the Adenovirus species and the Epstein-barr gamma-herpesvirus. To perform our research with greater precision at a more “sub-gene level”, we can use sliding windows to progressively capture the footprint on different fragments of the genome. However, we must pay particular attention to the size of the windows and perform multiple tests to determine the optimal size. To reduce the uncertainty of our measurements, we should use small sequence fragments coupled with a high number of iterations to determine the expected k-mer number. With this approach, we expect to observe local APOBEC3 footprints, similar to what we found on the EBV and Adenovirus genomes close to the origin of replication.
3. Lastly, to conduct a non-a priori study, all possible combinations of codon k-mers should be considered for analysis. This could include not only the 16

possible NXX k-mers (where X can be A, T, C, or G) that we have already explored, but also more extended positions such as the, 4 NNX, the 16 NNX-XNN and the 16 NNX-NNX k-mers.

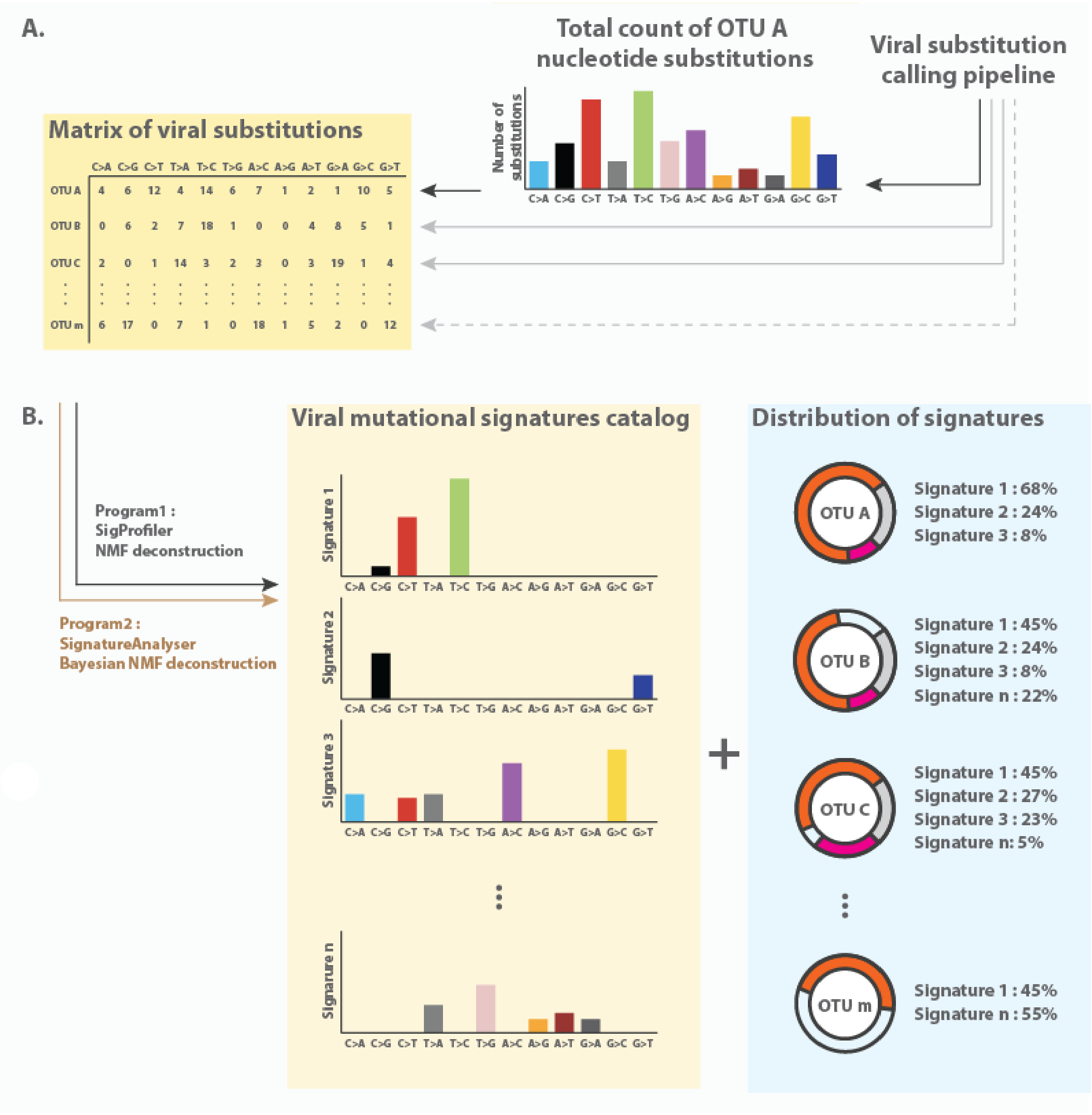
By applying those three proposed approaches and utilizing a larger dataset available now compared to the one used in our research (2020 dataset), we could perform a new investigation to identify new mutational processes. This kind of analysis could help us to discover new footprints, in addition to the APOBEC3 footprint, that may have left their evolutionary mark in viral genomes and establish new links between specific footprints and mutational processes. In a broader perspective, the aim of obtaining such results is to improve our overall understanding of the host-directed human viruses evolution.

4.2. Deciphering the mutational signatures shaping viruses' evolution

4.2.1. Catalog of mutational signatures acting on viruses' genomes

As already described in the introduction, computational packages originally developed for cancer mutation analysis allow the deconstruction of the substitution landscape into different mutational signatures [74]. Such analysis led to the discovery of a catalog of 68 cancer mutational signatures [75].

It is possible to adapt and apply the packages developed for cancer genome analysis to decipher the mutational signatures involved in virus evolution. Thus, deconstruction of the substitution landscapes of the HIV1, MERS and SARS-2 into different mutational signatures has been done recently [76,77,115,121,122].


A.
Total count of OTU A nucleotide substitutions
Viral substitution calling pipeline
Number of substitutions
C>A C>G C>T T>A T>C T>G A>C A>G A>T G>A G>C G>T
Matrix of viral substitutions
B.
Program1 : SigProfiler NMF deconstruction
Program2 : SignatureAnalyser Bayesian NMF deconstruction
Viral mutational signatures catalog
Signature 1
Signature 2
Signature 3
Signature n
C>A C>G C>T T>A T>C T>G A>C A>G A>T G>A G>C G>T
+
Distribution of signatures
OTU A
Signature 1 : 68%
Signature 2 : 24%
Signature 3 : 8%
OTU B
Signature 1 : 45%
Signature 2 : 24%
Signature 3 : 8%
Signature n : 22%
OTU C
Signature 1 : 45%
Signature 2 : 27%
Signature 3 : 23%
Signature n: 5%
OTU m
Signature 1 : 45%
Signature n : 55%

Matrix of viral substitutions

	C>A	C>G	C>T	T>A	T>C	T>G	A>C	A>G	A>T	G>A	G>C	G>T
OTU A	4	6	12	4	14	6	7	1	2	1	10	5
OTU B	0	6	2	7	18	1	0	0	4	8	5	1
OTU C	2	0	1	14	3	2	3	0	3	19	1	4
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
OTU m	6	17	0	7	1	0	18	1	5	2	0	12

Figure 4: Deciphering the mutational signatures of human virus processes from the matrix of viral substitutions. A. The results of the substitution counts will be aggregated into a unique matrix. A simulated example of the total number of nucleotide substitutions for the Operative Taxonomic Unit (OTU) A is represented by a colored barplot. The viral substitution matrix includes all counts for the context of the twelve substitutions for m OTUs. **B.** A Non-negative Matrix Factorization (NMF) will be applied to the viral substitution matrix by the SigProfiler software and in parallel by the SignatureAnalyser software. A viral mutational catalog of n signatures will be extracted. Simulated mutational signatures are represented by a colored bar plot. The distribution of signatures for each OTU will be also extracted by the programs. Simulated signature distribution is represented by colored pie charts.

Thanks to the large collection of more than 2.5 million substitutions that we gathered in our second study, a comprehensive search for viral mutational signatures can be conducted. In practice, the different substitution landscapes generated for each viral species (each OTU) will be combined into a single matrix (**Figure 4A**). This matrix will be decomposed using a Non-negative Matrix Factorization (NMF) method. To optimize the search for mutational signatures, similar to the approach used in cancer signature investigations, two algorithms will be utilized: SigProfiler [123] and SignatureAnalyzer [75], which use NMF with a likelihood or a Bayesian approach, respectively. The results from these two approaches will be compared and will lead to the production of a catalog of viral mutational signatures (**Figure 4B**).

By deconstructing the virus genomes' substitution landscape, several mutational signatures will be extracted. Some signatures might be attributable to already known and defined mutational processes, while others may not. This approach should identify the implication of the already described APOBEC3, ADAR and ROS mutational signatures in the evolution of human viruses and allow the discovery of new mutational signatures. Above all, it will allow us to test whether certain mutational signatures are similar to those identified in human tumors.

4.2.2. Mutational signatures common between viruses and human tumors

Viral infections can trigger mutational processes, which can induce changes in the host cell genome. One example is the HPV16 virus, which interferes with the host's APOBEC3 cytidine deaminases, which can unfortunately induce mutations in the host genome. This effect can be detected through the presence of APOBEC3-related mutations in both the host cell and the associated viral genome. Besides APOBEC3 editing, other mechanisms may also have the ability to target both the virus and its host. Further investigation could uncover shared mutational processes that impact both viruses and cellular genomes. Identifying such shared mutational processes may help reveal the unknown roles of certain viruses in the development of tumors.

To test this hypothesis, substitution landscapes extracted from human tumors could be deconstructed alongside the substitution landscapes of viruses. A library of cancer mutations is available in the Pan-Cancer Analysis of Whole Genome (PCAWG) consortium database[124]. Therefore, it is possible to explore this question using different strategies:

1. The matrix of viral substitutions landscapes can be first deconstructed using the catalog of human cancer mutational signatures. We expect to retrieve certain cancer mutational signatures in some specific viruses. For instance, the APOBEC3 mutational signature (consisting of C to T mutations in a 5‘TC context) should significantly contribute to the substitution landscapes of the HPV16 or HBV viruses as these viruses are well known to be targeted by the APOBEC3 enzymes. Importantly, in the catalog of cancer mutational signatures, 18 of them are still

unexplained, meaning that they are still not associated with a known mutational process. Identifying such mutational signatures in virus genomes might help their understanding.

2. The matrix of substitutions in human cancer and viruses can be combined, and de novo signature extraction can be performed. It would be possible to test whether some de novo mutational signatures are common between tumor and virus genomes. For example, it would be worth paying particular attention to human tumors positive for certain viruses, such as liver tumors positive for HBV [125] and to test if mutational signatures in those HBV-positive tumor genomes would also be identified in the HBV genome.

An interesting theory named “the hit and run” theory proposes that viruses could facilitate accumulation of mutations and therefore tumorigenesis but be dispensable for cancer maintenance [126]. The detection of viral mutational signature(s) in virus-negative tumors could suggest the involvement of a virus by a “hit and run” mechanism. Further studies would then be needed to identify the causative virus. As with the HPV16/18 vaccine, the discovery of virus involvement in cancer could allow the development of prophylactic treatments or specific therapies developed for virally-induced cancer [127].

5. Conclusion

The genome sequences of human viruses serve as witnesses to the restriction induced by the host factors, and their dynamic substitutions indicate a constant adaptation to the environment. In two separate investigations, I explored the landscape of genomes to describe this complex history.

In a first study, we investigated the evolutionary pressure induced by a family of cellular host effectors of the innate immune system known as APOBEC3. Due to this constant strong evolutive pressure, there is a bias in the codon composition of the genome that shows an under-representation of APOBEC3 target sites, which is indicative of a long history of deamination of the genome sequence. Thanks to research on the APOBEC3 footprint in the genomes of a large number of human viruses, we were able to describe a portion of the virus restriction spectrum of APOBEC3. We observed genome shaping in 22% of the tested species. We also reported the details of the APOBEC3 footprint in the different virus species, which allowed us to describe a part of the history of APOBEC3's genome targeting. We notably detected a strong under-representation of the APOBEC3 editing site in Parvovirus B19, endemic Coronaviruses, and Polyomaviruses species and we also confirmed this one in the genome HIV1 and Papillomaviruses. By examining gene-specific footprints, we were able to observe a potential link between genome editing of Adenovirus and gammaherpesvirus during the initial stages of their replication. We also report local footprint for HTLV1 and HBV viruses. All of these observations provide a broad perspective on the influence of APOBEC3 in human viruses, further advancing the field of APOBEC3 research.

In a second study, we investigated the substitution landscape of human viruses. Thanks to the creation of a large collection of 2.5 million substitutions for 55 viral species from all groups of the Baltimore classification, we were able to reconstruct virus genome dynamics by inferring 587 phylogenetic trees. Such a large dataset allows us to capture back-and-forth substitution events. Surprisingly, upon comparison of the subclasses of back-and-forth and non-compensated substitutions, we did not observe any different patterns. We also observed the same proportion of synonymous and non-synonymous substitutions. The presence of reversion at genome-specific sites appears to be mainly dependent on the substitution rate of these sites. We also described a mirror effect on the virus substitution landscape, which could illustrate a significant phenomenon of substitution compensation in the history of the virus.Taken together, all of these observations lead us to propose that the back-and-forth captures that account for 20% of total substitutions are just a fraction of the true reverted events. This implies that genome virus reversion, which has thus far been reported in the literature for only a few site-specific events, may actually occur much more frequently during genome evolution.

Through these different investigations, we have collected a large amount of data that paves the way for future exploration of virus substitution dynamics. The collection of substitutions in human viruses could be notably used to explore the mutational signatures of the mutational processes involved in genome evolution.

6. References

1. Beijerinck MW, 1851-1931. Ueber ein Contagium vivum fluidum als Ursache der Fleckenkrankheit der Tabaksblätter. J. Müller; 1898. Available: https://scholar.google.com/scholar_lookup?title=Ueber+ein+Contagium+vivum+fluidum+als+Ursache+der+Fleckenkrankheit+der+Tabaksbla%CC%88tter&author=Beijerinck%2C+M.+W.+%28Martinus+Willem%29&publication_year=1898
2. Taylor MW. Introduction: A Short History of Virology. In: Taylor MW, editor. Viruses and Man: A History of Interactions. Cham: Springer International Publishing; 2014. pp. 1–22. doi:10.1007/978-3-319-07758-1_1
3. Burrell CJ, Howard CR, Murphy FA. History and Impact of Virology. Fenner Whites Med Virol. 2017; 3–14. doi:10.1016/B978-0-12-375156-0.00001-1
4. Baltimore D. Expression of animal virus genomes. Bacteriol Rev. 1971;35: 235–241.
5. Lefkowitz EJ, Dempsey DM, Hendrickson RC, Orton RJ, Siddell SG, Smith DB. Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV). Nucleic Acids Res. 2018;46: D708–D717. doi:10.1093/nar/gkx932
6. Fauquet CM. Taxonomy, Classification and Nomenclature of Viruses. Encycl Virol. 2008; 9–23. doi:10.1016/B978-012374410-4.00509-4
7. Flint SJ, Racaniello VR, Rall GF, Skalka AM, Enquist LW. Principles of virology. 2015.
8. Kovalskaya N, Hammond RW. Molecular biology of viroid–host interactions and disease control strategies. Plant Sci. 2014;228: 48–60. doi:10.1016/j.plantsci.2014.05.006
9. Jacob F, Wollman EL. Viruses and genes. Sci Am. 1961;204: 93–107.
10. Moreira D, López-García P. Ten reasons to exclude viruses from the tree of life. Nat Rev Microbiol. 2009;7: 306–311. doi:10.1038/nrmicro2108
11. van Regenmortel MHV. The metaphor that viruses are living is alive and well, but it is no more than a metaphor. Stud Hist Philos Sci Part C Stud Hist Philos Biol Biomed Sci. 2016;59: 117–124. doi:10.1016/j.shpsc.2016.02.017
12. Baˆndea CI. A new theory on the origin and the nature of viruses. J Theor Biol. 1983;105: 591–602. doi:10.1016/0022-5193(83)90221-7
13. Bandea C. The Origin and Evolution of Viruses as Molecular Organisms. Nat Preced. 2009; 1–1. doi:10.1038/npre.2009.3886.1
14. Claverie J-M. Viruses take center stage in cellular evolution. Genome Biol. 2006;7: 110. doi:10.1186/gb-2006-7-6-110
15. Nasir A, Romero-Severson E, Claverie J-M. Investigating the Concept and Origin of Viruses. Trends Microbiol. 2020;28: 959–967. doi:10.1016/j.tim.2020.08.003
16. Forterre P. Manipulation of cellular syntheses and the nature of viruses: The virocell concept. Comptes Rendus Chim. 2011;14: 392–399. doi:10.1016/j.crci.2010.06.007
17. Forterre P. To be or not to be alive: How recent discoveries challenge the traditional definitions of viruses and life. Stud Hist Philos Biol Biomed Sci. 2016;59: 100–108. doi:10.1016/j.shpsc.2016.02.013
18. Origin Of Life: Oparin-Haldane Hypothesis. [cited 8 Aug 2022]. Available: http://www.simsoup.info/Origin_Landmarks_Oparin_Haldane.html
19. Miller SL. A Production of Amino Acids Under Possible Primitive Earth Conditions. Science. 1953;117: 528–529. doi:10.1126/science.117.3046.528
20. Degens ET, Bajor M. Amino acids and sugars in the bruderheim and Murray meteorite. Naturwissenschaften. 1962;49: 605–606. doi:10.1007/BF01178050
21. Elsila JE, Glavin DP, Dworkin JP. Cometary glycine detected in samples returned by Stardust. Meteorit Planet Sci. 2009;44: 1323–1330. doi:10.1111/j.1945-5100.2009.tb01224.x
22. Altwegg K, Balsiger H, Bar-Nun A, Berthelier J-J, Bieler A, Bochsler P, et al. Prebiotic chemicals—amino acid and phosphorus—in the coma of comet 67P/Churyumov-Gerasimenko. Sci Adv. 2016;2: e1600285. doi:10.1126/sciadv.1600285
23. Schulze H, Nierhaus KH. Minimal set of ribosomal components for reconstitution of the

peptidyltransferase activity. EMBO J. 1982;1: 609–613. doi:10.1002/j.1460-2075.1982.tb01216.x

24. Guerrier-Takada C, Gardiner K, Marsh T, Pace N, Altman S. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. Cell. 1983;35: 849–857. doi:10.1016/0092-8674(83)90117-4
25. Lorsch JR, Szostak JW. In vitro evolution of new ribozymes with polynucleotide kinase activity. Nature. 1994;371: 31–36. doi:10.1038/371031a0
26. The universal ancestor | PNAS. [cited 3 May 2023]. Available: https://www.pnas.org/doi/10.1073/pnas.95.12.6854
27. Lombard J, López-García P, Moreira D. The early evolution of lipid membranes and the three domains of life. Nat Rev Microbiol. 2012;10: 507–515. doi:10.1038/nrmicro2815
28. Durzyńska J, Goździcka-Józefiak A. Viruses and cells intertwined since the dawn of evolution. Virol J. 2015;12: 169. doi:10.1186/s12985-015-0400-7
29. Witzany G. The Viral Origins of Telomeres and Telomerases and their Important Role in Eukaryogenesis and Genome Maintenance. Biosemiotics. 2008;1: 191–206. doi:10.1007/s12304-008-9018-0
30. Abrescia NGA, Bamford DH, Grimes JM, Stuart DI. Structure Unifies the Viral Universe. Annu Rev Biochem. 2012;81: 795–822. doi:10.1146/annurev-biochem-060910-095130
31. Flores R, Gago-Zachert S, Serra P, Sanjuán R, Elena SF. Viroids: Survivors from the RNA World? Annu Rev Microbiol. 2014;68: 395–414. doi:10.1146/annurev-micro-091313-103416
32. A Genomewide Search for Ribozymes Reveals an HDV-Like Sequence in the Human CPEB3 Gene | Science. [cited 7 May 2023]. Available: https://www.science.org/doi/10.1126/science.1129308
33. Holmes EC. What Does Virus Evolution Tell Us about Virus Origins? J Virol. 2011;85: 5247–5251. doi:10.1128/JVI.02203-10
34. Raoult D, Audic S, Robert C, Abergel C, Renesto P, Ogata H, et al. The 1.2-Megabase Genome Sequence of Mimivirus. Science. 2004;306: 1344–1350. doi:10.1126/science.1101485
35. Nasir A, Kim KM, Caetano-Anolles G. Giant viruses coexisted with the cellular ancestors and represent a distinct supergroup along with superkingdoms Archaea, Bacteria and Eukarya. BMC Evol Biol. 2012;12: 156. doi:10.1186/1471-2148-12-156
36. Krupovic M, Dolja VV, Koonin EV. Origin of viruses: primordial replicators recruiting capsids from hosts. Nat Rev Microbiol. 2019;17: 449–458. doi:10.1038/s41579-019-0205-6
37. Nasir A, Kim KM, Caetano-Anollés G. Viral evolution. Mob Genet Elem. 2012;2: 247–252. doi:10.4161/mge.22797
38. Forterre P, Krupovic M, Prangishvili D. Cellular domains and viral lineages. Trends Microbiol. 2014;22: 554–558. doi:10.1016/j.tim.2014.07.004
39. The Evolutionary Synthesis — Ernst Mayr, William B. Provine. [cited 1 Sep 2022]. Available: https://www.hup.harvard.edu/catalog.php?isbn=9780674272262
40. Arber W. Molecular mechanisms driving Darwinian evolution. Math Comput Model. 2008;47: 666–674. doi:10.1016/j.mcm.2007.06.003
41. Sturtevant AH. Essays on Evolution. I. On the Effects of Selection on Mutation Rate. Q Rev Biol. 1937 [cited 16 Sep 2022]. doi:10.1086/394543
42. Duffy S. Why are RNA virus mutation rates so damn high? PLOS Biol. 2018;16: e3000003. doi:10.1371/journal.pbio.3000003
43. Zuckerkandl E, Pauling L. Molecular Disease, Evolution, and Genic Heterogeneity. Academic Press; 1962.
44. Zuckerkandl: In Evolving Genes and Proteins, ed.... - Google Scholar. [cited 13 Sep 2022]. Available: https://scholar.google.com/scholar_lookup?title=Evolving%20Genes%20and%20Proteins&pages=97-166&publication_year=1965&author=Zuckerkandl%2CE&author=Pauling%2CL
45. dos Reis M, Donoghue PCJ, Yang Z. Bayesian molecular clock dating of species

divergences in the genomics era. Nat Rev Genet. 2016;17: 71–80. doi:10.1038/nrg.2015.8

46. Kimura M. Evolutionary Rate at the Molecular Level. Nature. 1968;217: 624–626. doi:10.1038/217624a0
47. On Some Principles Governing Molecular Evolution*. [cited 7 Sep 2022]. doi:10.1073/pnas.71.7.2848
48. Lenormand T, Roze D, Rousset F. Stochasticity in evolution. Trends Ecol Evol. 2009;24: 157–165. doi:10.1016/j.tree.2008.09.014
49. Domingo E, Flavell RA, Weissmann C. In vitro site-directed mutagenesis: generation and properties of an infectious extracistronic mutant of bacteriophage Qbeta. Gene. 1976;1: 3–25. doi:10.1016/0378-1119(76)90003-2
50. Batschelet E, Domingo E, Weissmann C. The proportion of revertant and mutant phage in a growing population, as a function of mutation and growth rate. Gene. 1976;1: 27–32. doi:10.1016/0378-1119(76)90004-4
51. Correlation Between Mutation Rate and Genome Size in Riboviruses: Mutation Rate of Bacteriophage Qβ | Genetics | Oxford Academic. [cited 28 Apr 2023]. Available: https://academic.oup.com/genetics/article/195/1/243/5935437
52. Pauly MD, Procario MC, Lauring AS. A novel twelve class fluctuation test reveals higher than expected mutation rates for influenza A viruses. Kirkegaard K, editor. eLife. 2017;6: e26437. doi:10.7554/eLife.26437
53. Drake JW, Charlesworth B, Charlesworth D, Crow JF. Rates of spontaneous mutation. Genetics. 1998;148: 1667–1686. doi:10.1093/genetics/148.4.1667
54. Drake JW. The distribution of rates of spontaneous mutation over viruses, prokaryotes, and eukaryotes. Ann N Y Acad Sci. 1999;870: 100–107. doi:10.1111/j.1749-6632.1999.tb08870.x
55. Duffy S, Shackelton LA, Holmes EC. Rates of evolutionary change in viruses: patterns and determinants. Nat Rev Genet. 2008;9: 267–276. doi:10.1038/nrg2323
56. Peck KM, Lauring AS. Complexities of Viral Mutation Rates. J Virol. 92: e01031-17. doi:10.1128/JVI.01031-17
57. Jenkins GM, Rambaut A, Pybus OG, Holmes EC. Rates of Molecular Evolution in RNA Viruses: A Quantitative Phylogenetic Analysis. J Mol Evol. 2002;54: 156–165. doi:10.1007/s00239-001-0064-3
58. Chen D, Lau Y-C, Xu X-K, Wang L, Du Z, Tsang TK, et al. Inferring time-varying generation time, serial interval, and incubation period distributions for COVID-19. Nat Commun. 2022;13: 7727. doi:10.1038/s41467-022-35496-8
59. Woolfit M. Effective population size and the rate and pattern of nucleotide substitutions. Biol Lett. 2009;5: 417–420. doi:10.1098/rsbl.2009.0155
60. Gijsbers EF, Schuitemaker H, Kootstra NA. HIV-1 transmission and viral adaptation to the host. Future Virol. 2012;7: 63–71. doi:10.2217/fvl.11.134
61. Duchêne S, Holmes EC, Ho SYW. Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates. Proc R Soc B Biol Sci. 2014;281: 20140732. doi:10.1098/rspb.2014.0732
62. Aiewsakun P, Katzourakis A. Time-Dependent Rate Phenomenon in Viruses. J Virol. 2016;90: 7184–7195. doi:10.1128/JVI.00593-16
63. Ho SYW, Lanfear R, Bromham L, Phillips MJ, Soubrier J, Rodrigo AG, et al. Time-dependent rates of molecular evolution. Mol Ecol. 2011;20: 3087–3101. doi:10.1111/j.1365-294X.2011.05178.x
64. Ghafari M, du Plessis L, Raghwani J, Bhatt S, Xu B, Pybus OG, et al. Purifying Selection Determines the Short-Term Time Dependency of Evolutionary Rates in SARS-CoV-2 and pH1N1 Influenza. Mol Biol Evol. 2022;39: msac009. doi:10.1093/molbev/msac009
65. Lemey P, Salemi M, Vandamme A-M. The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing. In: Cambridge Core [Internet]. Cambridge University Press; Mar 2009 [cited 7 May 2023]. doi:10.1017/CBO9780511819049
66. Duchêne S, Ho SY, Holmes EC. Declining transition/transversion ratios through time

reveal limitations to the accuracy of nucleotide substitution models. BMC Evol Biol. 2015;15: 36. doi:10.1186/s12862-015-0312-6

67. Washington MT. DNA Polymerase Fidelity: Beyond Right and Wrong. Struct Lond Engl 1993. 2016;24: 1855–1856. doi:10.1016/j.str.2016.10.003
68. Campbell PJ, Getz G, Korbel JO, Stuart JM, Jennings JL, Stein LD, et al. Pan-cancer analysis of whole genomes. Nature. 2020;578: 82–93. doi:10.1038/s41586-020-1969-6
69. Menéndez-Arias L. Mutation Rates and Intrinsic Fidelity of Retroviral Reverse Transcriptases. Viruses. 2009;1: 1137–1165. doi:10.3390/v1031137
70. Eckerle LD, Becker MM, Halpin RA, Li K, Venter E, Lu X, et al. Infidelity of SARS-CoV Nsp14-Exonuclease Mutant Virus Replication Is Revealed by Complete Genome Sequencing. PLOS Pathog. 2010;6: e1000896. doi:10.1371/journal.ppat.1000896
71. Smith EC, Sexton NR, Denison MR. Thinking Outside the Triangle: Replication Fidelity of the Largest RNA Viruses. Annu Rev Virol. 2014;1: 111–132. doi:10.1146/annurev-virology-031413-085507
72. Moody CA, Laimins LA. Human papillomavirus oncoproteins: pathways to transformation. Nat Rev Cancer. 2010;10: 550–560. doi:10.1038/nrc2886
73. Blackford AN, Patel RN, Forrester NA, Theil K, Groitl P, Stewart GS, et al. Adenovirus 12 E4orf6 inhibits ATR activation by promoting TOPBP1 degradation. Proc Natl Acad Sci. 2010;107: 12251–12256. doi:10.1073/pnas.0914605107
74. Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. Deciphering Signatures of Mutational Processes Operative in Human Cancer. Cell Rep. 2013;3: 246–259. doi:10.1016/j.celrep.2012.12.008
75. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The repertoire of mutational signatures in human cancer. Nature. 2020;578: 94–101. doi:10.1038/s41586-020-1943-3
76. Graudenzi A, Maspero D, Angaroni F, Piazza R, Ramazzotti D. Mutational signatures and heterogeneous host response revealed via large-scale characterization of SARS-CoV-2 genomic diversity. iScience. 2021;24: 102116. doi:10.1016/j.isci.2021.102116
77. Yi K, Kim SY, Bleazard T, Kim T, Youk J, Ju YS. Mutational spectrum of SARS-CoV-2 during the global pandemic. Exp Mol Med. 2021;53: 1229–1237. doi:10.1038/s12276-021-00658-z
78. Kieran D. Lamb, Martha M. Luka, Megan Saathoff, Richard Orton, View ORCID ProfileMy Phan, Matthew Cotten, View ORCID ProfileKe Yuan, View ORCID ProfileDavid L. Robertson. SARS-CoV-2's evolutionary capacity is mostly driven by host antiviral molecules.
79. Sato Y, Probst HC, Tatsumi R, Ikeuchi Y, Neuberger MS, Rada C. Deficiency in APOBEC2 Leads to a Shift in Muscle Fiber Type, Diminished Body Mass, and Myopathy. J Biol Chem. 2010;285: 7111–7118. doi:10.1074/jbc.M109.052977
80. Harris RS, Bishop KN, Sheehy AM, Craig HM, Petersen-Mahrt SK, Watt IN, et al. DNA Deamination Mediates Innate Immunity to Retroviral Infection. Cell. 2003;113: 803–809. doi:10.1016/S0092-8674(03)00423-9
81. Dutko JA, Schäfer A, Kenny AE, Cullen BR, Curcio MJ. Inhibition of a Yeast LTR Retrotransposon by Human APOBEC3 Cytidine Deaminases. Curr Biol. 2005;15: 661–666. doi:10.1016/j.cub.2005.02.051
82. Rogozin IB, Basu MK, Jordan IK, Pavlov YI, Koonin EV. APOBEC4, a New Member of the AID/APOBEC Family of Polynucleotide (Deoxy)Cytidine Deaminases Predicted by Computational Analysis. Cell Cycle. 2005;4: 1281–1285. doi:10.4161/cc.4.9.1994
83. Salter JD, Bennett RP, Smith HC. The APOBEC Protein Family: United by Structure, Divergent in Function. Trends Biochem Sci. 2016;41: 578–594. doi:10.1016/j.tibs.2016.05.001
84. Sheehy AM, Gaddis NC, Choi JD, Malim MH. Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. Nature. 2002;418: 646–650. doi:10.1038/nature00939
85. Pathak VK, Temin HM. Broad spectrum of in vivo forward mutations, hypermutations,

and mutational hotspots in a retroviral shuttle vector after a single replication cycle: substitutions, frameshifts, and hypermutations. Proc Natl Acad Sci. 1990;87: 6019–6023. doi:10.1073/pnas.87.16.6019
86. Lecossier D, Bouchonnet F, Clavel F, Hance AJ. Hypermutation of HIV-1 DNA in the Absence of the Vif Protein. Science. 2003;300: 1112–1112. doi:10.1126/science.1083338
87. Delviks-Frankenberry KA, Nikolaitchik OA, Burdick RC, Gorelick RJ, Keele BF, Hu W-S, et al. Minimal Contribution of APOBEC3-Induced G-to-A Hypermutation to HIV-1 Recombination and Genetic Variation. PLOS Pathog. 2016;12: e1005646. doi:10.1371/journal.ppat.1005646
88. Sanjuán R, Domingo-Calap P. Genetic Diversity and Evolution of Viral Populations. Encycl Virol. 2021; 53–61. doi:10.1016/B978-0-12-809633-8.20958-8
89. Turelli P, Liagre-Quazzola A, Mangeat B, Verp S, Jost S, Trono D. APOBEC3-Independent Interferon-Induced Viral Clearance in Hepatitis B Virus Transgenic Mice. J Virol. 2008;82: 6585–6590. doi:10.1128/JVI.00216-08
90. Turelli P, Mangeat B, Jost S, Vianin S, Trono D. Inhibition of Hepatitis B Virus Replication by APOBEC3G. Science. 2004;303: 1829–1829. doi:10.1126/science.1092066
91. Chen Y, Hu J, Cai X, Huang Y, Zhou X, Tu Z, et al. APOBEC3B edits HBV DNA and inhibits HBV replication during reverse transcription. Antiviral Res. 2018;149: 16–25. doi:10.1016/j.antiviral.2017.11.006
92. Kanagaraj A, Sakamoto N, Que L, Li Y, Mohiuddin M, Koura M, et al. Different antiviral activities of natural APOBEC3C, APOBEC3G, and APOBEC3H variants against hepatitis B virus. Biochem Biophys Res Commun. 2019;518: 26–31. doi:10.1016/j.bbrc.2019.08.003
93. Peretti A, Geoghegan EM, Pastrana DV, Smola S, Feld P, Sauter M, et al. Characterization of BK polyomaviruses from kidney transplant recipients suggests a role for APOBEC3 in driving in-host virus evolution. Cell Host Microbe. 2018;23: 628-635.e7. doi:10.1016/j.chom.2018.04.005
94. Verhalen B, Starrett GJ, Harris RS, Jiang M. Functional Upregulation of the DNA Cytosine Deaminase APOBEC3B by Polyomaviruses. J Virol. 2016;90: 6379–6386. doi:10.1128/JVI.00771-16
95. Sasada A, Takaori-Kondo A, Shirakawa K, Kobayashi M, Abudu A, Hishizawa M, et al. APOBEC3G targets human T-cell leukemia virus type 1. Retrovirology. 2005;2: 32. doi:10.1186/1742-4690-2-32
96. Fan J, Ma G, Nosaka K, Tanabe J, Satou Y, Koito A, et al. APOBEC3G Generates Nonsense Mutations in Human T-Cell Leukemia Virus Type 1 Proviral Genomes In Vivo. J Virol. 2010;84: 7278–7287. doi:10.1128/JVI.02239-09
97. Endogenous APOBEC3B overexpression characterizes HPV-positive and HPV-negative oral epithelial dysplasias and head and neck cancers - Modern Pathology. [cited 8 May 2023]. Available: https://mp.uscap.org/article/S0893-3952(22)00670-6/fulltext
98. Zhu B, Xiao Y, Yeager M, Clifford G, Wentzensen N, Cullen M, et al. Mutations in the HPV16 genome induced by APOBEC3 are associated with viral clearance. Nat Commun. 2020;11: 886. doi:10.1038/s41467-020-14730-1
99. Warren CJ, Xu T, Guo K, Griffin LM, Westrich JA, Lee D, et al. APOBEC3A Functions as a Restriction Factor of Human Papillomavirus. J Virol. 2014;89: 688–702. doi:10.1128/JVI.02383-14
100. Cheng AZ, Moraes SN, Shaban NM, Fanunza E, Bierle CJ, Southern PJ, et al. APOBECs and Herpesviruses. Viruses. 2021;13: 390. doi:10.3390/v13030390
101. Cheng AZ, Moraes SN de, Attarian C, Yockteng-Melgar J, Jarvis MC, Biolatti M, et al. A Conserved Mechanism of APOBEC3 Relocalization by Herpesviral Ribonucleotide Reductase Large Subunits. bioRxiv. 2019; 765735. doi:10.1101/765735
102. Suspène R, Aynaud M-M, Koch S, Pasdeloup D, Labetoulle M, Gaertner B, et al. Genetic Editing of Herpes Simplex Virus 1 and Epstein-Barr Herpesvirus Genomes by Human APOBEC3 Cytidine Deaminases in Culture and In Vivo. J Virol. 2011;85: 7594–7602. doi:10.1128/JVI.00290-11
103. Cheng AZ, Yockteng-Melgar J, Jarvis MC, Malik-Soni N, Borozan I, Carpenter MA, et

al. Epstein–Barr virus BORF2 inhibits cellular APOBEC3B to preserve viral genome integrity. Nat Microbiol. 2019;4: 78–88. doi:10.1038/s41564-018-0284-6

104. Suspène R, Guétard D, Henry M, Sommer P, Wain-Hobson S, Vartanian J-P. Extensive editing of both hepatitis B virus DNA strands by APOBEC3 cytidine deaminases in vitro and in vivo. Proc Natl Acad Sci U S A. 2005;102: 8321–8326. doi:10.1073/pnas.0408223102

105. Martinez T, Shapiro M, Bhaduri-McIntosh S, MacCarthy T. Evolutionary effects of the AID/APOBEC family of mutagenic enzymes on human gamma-herpesviruses. Virus Evol. 2019;5. doi:10.1093/ve/vey040

106. Conticello SG, Thomas CJF, Petersen-Mahrt SK, Neuberger MS. Evolution of the AID/APOBEC Family of Polynucleotide (Deoxy)cytidine Deaminases. Mol Biol Evol. 2005;22: 367–377. doi:10.1093/molbev/msi026

107. Evolutionary effects of the AID/APOBEC family of mutagenic enzymes on human gamma-herpesviruses | Virus Evolution | Oxford Academic. [cited 8 May 2023]. Available: https://academic.oup.com/ve/article/5/1/vey040/5316048

108. Krishnan A, Iyer LM, Holland SJ, Boehm T, Aravind L. Diversification of AID/APOBEC-like deaminases in metazoa: multiplicity of clades and widespread roles in immunity. Proc Natl Acad Sci. 2018;115: E3201–E3210. doi:10.1073/pnas.1720897115

109. LaRue RS, Andrésdóttir V, Blanchard Y, Conticello SG, Derse D, Emerman M, et al. Guidelines for Naming Nonprimate APOBEC3 Genes and Proteins. J Virol. 2009;83: 494–497. doi:10.1128/JVI.01976-08

110. Ito J, Gifford RJ, Sato K. Retroviruses drive the rapid evolution of mammalian APOBEC3 genes. Proc Natl Acad Sci. 2020;117: 610–618. doi:10.1073/pnas.1914183116

111. Sawyer SL, Emerman M, Malik HS. Ancient Adaptive Evolution of the Primate Antiviral DNA-Editing Enzyme APOBEC3G. PLOS Biol. 2004;2: e275. doi:10.1371/journal.pbio.0020275

112. Uriu K, Kosugi Y, Ito J, Sato K. The Battle between Retroviruses and APOBEC3 Genes: Its Past and Present. Viruses. 2021;13: 124. doi:10.3390/v13010124

113. Sadeghpour S, Khodaee S, Rahnama M, Rahimi H, Ebrahimi D. Human APOBEC3 Variations and Viral Infection. Viruses. 2021;13: 1366. doi:10.3390/v13071366

114. Hayward JA, Tachedjian M, Cui J, Cheng AZ, Johnson A, Baker ML, et al. Differential Evolution of Antiretroviral Restriction Factors in Pteropid Bats as Revealed by APOBEC3 Gene Complexity. Mol Biol Evol. 2018;35: 1626–1637. doi:10.1093/molbev/msy048

115. Giorgio SD, Martignano F, Torcia MG, Mattiuz G, Conticello SG. Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2. Sci Adv. 2020; eabb5813. doi:10.1126/sciadv.abb5813

116. Vijgen L, Keyaerts E, Lemey P, Maes P, Van Reeth K, Nauwynck H, et al. Evolutionary History of the Closely Related Group 2 Coronaviruses: Porcine Hemagglutinating Encephalomyelitis Virus, Bovine Coronavirus, and Human Coronavirus OC43. J Virol. 2006;80: 7270–7274. doi:10.1128/JVI.02675-05

117. Vijgen L, Keyaerts E, Moës E, Thoelen I, Wollants E, Lemey P, et al. Complete Genomic Sequence of Human Coronavirus OC43: Molecular Clock Analysis Suggests a Relatively Recent Zoonotic Coronavirus Transmission Event. J Virol. 2005;79: 1595–1604. doi:10.1128/JVI.79.3.1595-1604.2005

118. Simmonds P, Aiewsakun P, Katzourakis A. Prisoners of war — host adaptation and its constraints on virus evolution. Nat Rev Microbiol. 2019;17: 321. doi:10.1038/s41579-018-0120-2

119. Jitobaom K, Phakaratsakul S, Sirihongthong T, Chotewutmontri S, Suriyaphol P, Suptawiwat O, et al. Codon usage similarity between viral and some host genes suggests a codon-specific translational regulation. Heliyon. 2020;6. doi:10.1016/j.heliyon.2020.e03915

120. Shapiro M, Meier S, MacCarthy T. The cytidine deaminase under-representation reporter (CDUR) as a tool to study evolution of sequences under deaminase mutational pressure. BMC Bioinformatics. 2018;19: 163. doi:10.1186/s12859-018-2161-y

121. Azgari C, Kilinc Z, Turhan B, Circi D, Adebali O. The Mutation Profile of SARS-CoV-2 Is Primarily Shaped by the Host Antiviral Defense. Viruses. 2021;13: 394. doi:10.3390/v13030394

122. Kieran D. Lamb, Martha M. Luka, Megan Saathoff, Richard Orton, My Phan, Matthew Cotten, Ke Yuan,David, L. Robertson. SARS-CoV-2's evolutionary capacity is mostly driven by host antiviral molecules. bioRxiv. doi:https://doi.org/10.1101/2023.04.07.536037

123. Islam SMA, Díaz-Gay M, Wu Y, Barnes M, Vangara R, Bergstrom EN, et al. Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. Cell Genomics. 2022;2: 100179. doi:10.1016/j.xgen.2022.100179

124. Goldman MJ, Zhang J, Fonseca NA, Cortés-Ciriano I, Xiang Q, Craft B, et al. A user guide for the online exploration and visualization of PCAWG data. Nat Commun. 2020;11: 3400. doi:10.1038/s41467-020-16785-6

125. Álvarez EG, Demeulemeester J, Jolly C, García-Souto D, Otero P, Pequeño A, et al. Aberrant integration of Hepatitis B virus DNA promotes major restructuring of human hepatocellular carcinoma genome architecture. bioRxiv; 2021. p. 2021.04.19.440412. doi:10.1101/2021.04.19.440412

126. Ferreira DA, Tayyar Y, Idris A, McMillan NAJ. A “hit-and-run” affair – A possible link for cancer progression in virally driven cancers. Biochim Biophys Acta BBA - Rev Cancer. 2021;1875: 188476. doi:10.1016/j.bbcan.2020.188476

127. Tashiro H, Brenner MK. Immunotherapy against cancer-related viruses. Cell Res. 2017;27: 59–73. doi:10.1038/cr.2016.153

7. List of scientific productions

Footprint of the host restriction factors APOBEC3 on the genome of human viruses
Poulain, F., Lejeune, N., Willemart K., & Gillet, N. A
14 août 2020, Plos Pathogens, 16, 8.
DOI: 10.1371/journal.ppat.1008718

Human papillomavirus E6/E7 oncoproteins promote radiotherapy-mediated tumor suppression by globally hijacking host DNA damage repair
Bruyere, D., Roncarati, P., Lebeau, A., Lerho, T., Poulain, F., Hendrick, E., Pilard, C., Reynders, C., Ancion, M., Luyckx, M., Renard, M., Jacob, Y., Twizere, J-C., Peiffer, R., Peulen, O., Delvenne, P., Hubert, P., McBride, A., Gillet, N., Masson, M., & Michael Herfs ,
2023, Theranostics. 13, 3, p. 1130-1149
DOI: 10.7150/thno.78091

The APOBEC3B cytidine deaminase is an adenovirus restriction factor
Lejeune, N., Mathieu, S., Decloux, A., Poulain, F., Blockx, Z., Raymond, K. A., Willemart, K., Vartanian, J-P., Suspène, R. & Gillet, N. A
févr. 2023, Plos Pathogens. 19, 2, p. e1011156 e1011156.
DOI: 10.1371/journal.ppat.1011156

Infection of bronchial epithelial cells by the human adenoviruses A12, B3 and C2 differently regulates the innate antiviral effector APOBEC3B
Lejeune, N., Poulain, F., Willemart, K., Blockx, Z., Mathieu, S. & Gillet, N. A
juin 2021, Journal of Virology. 95, 13, e02413-20.
DOI: 10.1128/jvi.02413-20

Susceptibility of neuroblastoma and glioblastoma cell lines to SARS-CoV-2 infection
Bielarz, V., WILLEMART, KEVIN., Avalosse, N., De Swert, K., Lotfi, R., Lejeune, N., Poulain, F., Ninanne, N., GILLOTEAUX, J., GILLET, NICOLAS. & Nicaise, C
1 mai 2021, Brain research. 1758, 147344.
DOI: 10.1016/j.brainres.2021.147344

SARS-CoV-2 Detection for Diagnosis Purposes in the Setting of a Molecular Biology Research Lab
Damien Coupeau, Nicolas Burton, Noémie Lejeune, Suzanne Loret, Astrid Petit, Srđan Pejaković, Florian Poulain, Laura Bonil, Gabrielle Trozzi, Laetitia Wiggers, Kévin Willemart, Emmanuel André, Lies Laenen, Lize Cuypers ,Marc Van Ranst ,Pierre Bogaerts, Benoît Muylkens, Nicolas Albert Gillet
Methods and Protocols. 3, 3
DOI: 10.3390/mps3030059