Identification of Novel Viral Interacting Proteins through the Detection of Genetic INNovations (DGINN) combined with functional assays

Léa Picard

▶ **To cite this version:**

Léa Picard. Identification of Novel Viral Interacting Proteins through the Detection of Genetic INNovations (DGINN) combined with functional assays. Molecular biology. Université de Lyon, 2020. English. NNT : 2020LYSEN031 . tel-03510332

HAL Id: tel-03510332
https://theses.hal.science/tel-03510332

Submitted on 4 Jan 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Numéro National de Thèse : 2020LYSEN031

THESE de DOCTORAT DE L'UNIVERSITE DE LYON
opérée par
l'Ecole Normale Supérieure de Lyon

Ecole Doctorale N°340
Biologie Moléculaire, Intégrative et Cellulaire

Discipline : Sciences de la vie et de la santé

Soutenue publiquement le 04/09/2020, par:
Léa PICARD

Identification of novel Viral Interacting Proteins through the Detection of Genetic INNovations (DGINN) combined with functional assays

Identification de nouvelles protéines interagissant avec les virus par la Détection d'INNovations Génétiques (DGINN) combinée à des essais fonctionnels

---

Devant le jury composé de :

Dr Bravo, Ignacio	Directeur de recherche MIVEGEC	Rapporteur
Dre Sironi, Manuela	Professeure IRCCS Eugenio Medea	Rapporteure
Dr Daubin, Vincent	Directeur de recherche LBBE	Examinateur
Dre Margottin-Goguet, Florence	Directrice de recherche Institut Cochin	Examinatrice
Pr Volff, Jean-Nicolas	Professeur des universités IGFL	Examinateur
Dre Etienne, Lucie	Chargée de recherche (HDR) CIRI	Directrice de thèse
Dr Guéguen, Laurent	Maître de conférences (HDR) LBBE	Co-encadrant de thèse

Abstract

The identification of cellular proteins that interfere with virus replication is a key challenge in virology. Amongst them, finding those engaged in long-term virus-host interaction and co-evolution is of particular interest. In the host, such selective pressures induce diverse genetic innovations, such as site-specific positive selection, gene copy number variation, recombination, etc. Under the hypothesis that genetic innovations in innate immunity may particularly occur in viral interacting proteins, we developed a pipeline for retrieving orthologous sequences, aligning them and reconstructing their phylogeny, followed by the detection of genetic innovations. This streamlined procedure uniquely allows for the detection of paralogous genes, recombination breakpoints, and signatures of positive selection with several widely-used methods.

We validated this evolutionary and predictive pipeline on genes with known selection profiles. Furthermore, we screened two datasets of candidate genes. The first one was composed of 56 genes which knock-downs impact the interferon response to viral infection. The second one was composed of 60 genes upregulated in macrophages resistant to HIV infection. We found numerous genes presenting important marks of genetic conflict, thus potentially encoding for novel Viral Interacting Proteins. Two of these candidates are undergoing functional characterization for their role in the HIV replicative cycle, and others are pending further investigation.

Overall, we designed a complete and highly-flexible pipeline, available to the public, that can screen large datasets and allow researchers to rank candidate genes in order to prioritize their wet-lab experiments.

Acknowledgements

I would like to thank the jury members for accepting to evaluate my work: Dr Manuela Sironi and Dr Ignacio Bravo for the time they took to review my manuscript in a period that was made all the more intense by the COVID crisis, Dr Florence Margottin-Goguet, Dr Vincent Daubin and Pr Jean-Nicolas Volff for taking part as examinators.

Obviously I thank my PhD supervisors, Dr Lucie Etienne and Dr Laurent Guéguen, for giving me the opportunity to undertake this PhD journey, back when I was applying for an engineer position. Their scientific guidance and the knowledge they have imparted me will no doubt prove invaluable in my future endeavours. I would also like to thank them for their support and their patience during the writing of this manuscript, which has proven to be a difficult exercise.

I thank Dr Andrea Cimarelli for his support, both in terms of lab space and scientific comments on my work, but also for providing the various datasets on which I worked. I would also like to acknowledge the members of my thesis committee, Dr Bastien Boussau, Dr Hélène Dutartre and Dr Xavier Morelli, for their input accross the three meetings spanning those four years. This helped me to prioritize projects more efficiently.

Of course I would like to thank all the members of the team at the CIRI, both past and present, for being great support whether it was for bitching, whining, discussing science or drinking beers (lots of drinking beers): Xuan-Nhi, Stéphanie, Anuj, Li, Yuxin, Federico, Clara, Mathilde, Mégane, Aftab, Giulia, Thomas, Julie, Margaux, Clément, Alexandre, Véro, Fabien, Stefania, Romain, Mathilde. Spending those four years with you has been a great experience.

On a personal level, I would like to thank my family: my mom who I know thought that at some point I would be done with university (no such luck Mam!), my sisters for being in turn extremely annoying and extremely great, as sisters should, but always supportive during hard times, and my dad for always being enthusiastic, whether for the

tattoos or the science he didn't understand. I would also like to thank my best friends, Max and Clem, for dealing with me those past four years and a half, from Montpellier to Lyon, with a rather dubiously healthy dose of beers and discussions carrying on to the sunrise (and sometimes even a bit later than that). I thank the Bermuda Triangle for being one of our greatest, albeit rather dangerous for overall sobriety, ideas. I thank Cam for being Cam, who might not follow me on all my harebrained ideas nowadays, but is still ready to laugh at me and with me whenever I share my misadventures with her. Finally, I'd like to thank Davide for being a great internship supervisor all those years ago, and then taking such an active interest in my subsequent career: you have been, without a doubt, an important part of what motivated me to stick to science.

I would not forget all my others friends as well, met during high school or university, through a velvet beret, the Foyer, representing the doctoral school, militantism, or karate. All you guys have been great, for various reasons which I have not time to detail here, but for which I feel no less.

Table of Contents

List of Figures

List of Tables

Introduction

Chapter 1

Genetic conflict in the context of viral-host interactions

1.1 Natural selection and adaptation to the environment

1.1.1 The theory of natural selection

In 1859, Charles Darwin formulated his theory of natural selection, based on his observations that animals living in different geographical regions shared characteristics (Darwin 1859). He also noted that they differ on specific traits, which seems to have participated in their adaptation to their environment. From there he formulated his theory that species originated from a single population which then spontaneously acquired different characteristics in response to said environment.

Elaborating on the works of Thomas Malthus, Darwin formulated the theory of natural selection. The constant reproduction of individuals will cause a strain on the resources of the environment they share with other individuals of their own species and with other species. This leads to a competition for access to those resources amongst the different individuals. By natural selection, only the individuals best adapted to their environment will survive, reproduce, and transmit their traits to their descendants. Thus, natural selection acts on traits which vary within the population, are transmitted to the next generation, and are linked to the reproductive success of the individuals bearing them, what we now call fitness.

This theory was unified in the 20th century with the works of other scientists such as Gregor Mendel to give birth to the synthetized theory of evolution. It posits genes as the units supporting genetic information and proposes an action of natural selection on the random mutations appearing in the population.

1.1.2 Types of natural selection

There are three types of natural selection: purifying or negative selection, positive or diversifying selection, and neutral or balancing selection (Figure 1) (Quintana-Murci and Clark 2013; Nielsen 2005).

Purifying selection acts on alleles that have deleterious effects on the individuals in the environment they live in. It is based on the assumption that any bearer of mutations inducing unfavorable genetic variants for life expectancy or reproductivity will produce much less descendants, if any, than non-bearers (Figure 1). The frequency of the mutation is thus rapidly reduced in the population. If the mutation is deleterious, the variant will be completely eliminated, otherwise it might persist at a low frequency in the population. This type of selection is expected for genes with essential functions for the organism, such as the so-called housekeeping genes.

Positive selection acts on alleles giving a selective advantage to its bearers (Figure 1). It induces a rapid rise in frequency for such alleles in the population, which can lead to fixation, and is a sign of individuals adapting to their environment.

Lastly, balancing selection acts on multiallelic sites to maintain multiple alleles within a population. This would be the case if heterozygosity is more advantageous than homozygosity. Otherwise, it may result from a phenotype which advantage depends on its relative frequency to the other phenotypes. The selection then depends on frequency, and can be either positive or negative according to the selective advantage of the variant varying with its frequency in the population. The last type of balancing selection is based on oscillations of the most advantageous genotype in time (during the life of the bearer) or space (depending on the environment). This selection decreases differences between populations/species, but will increase the genetic diversity within a population.

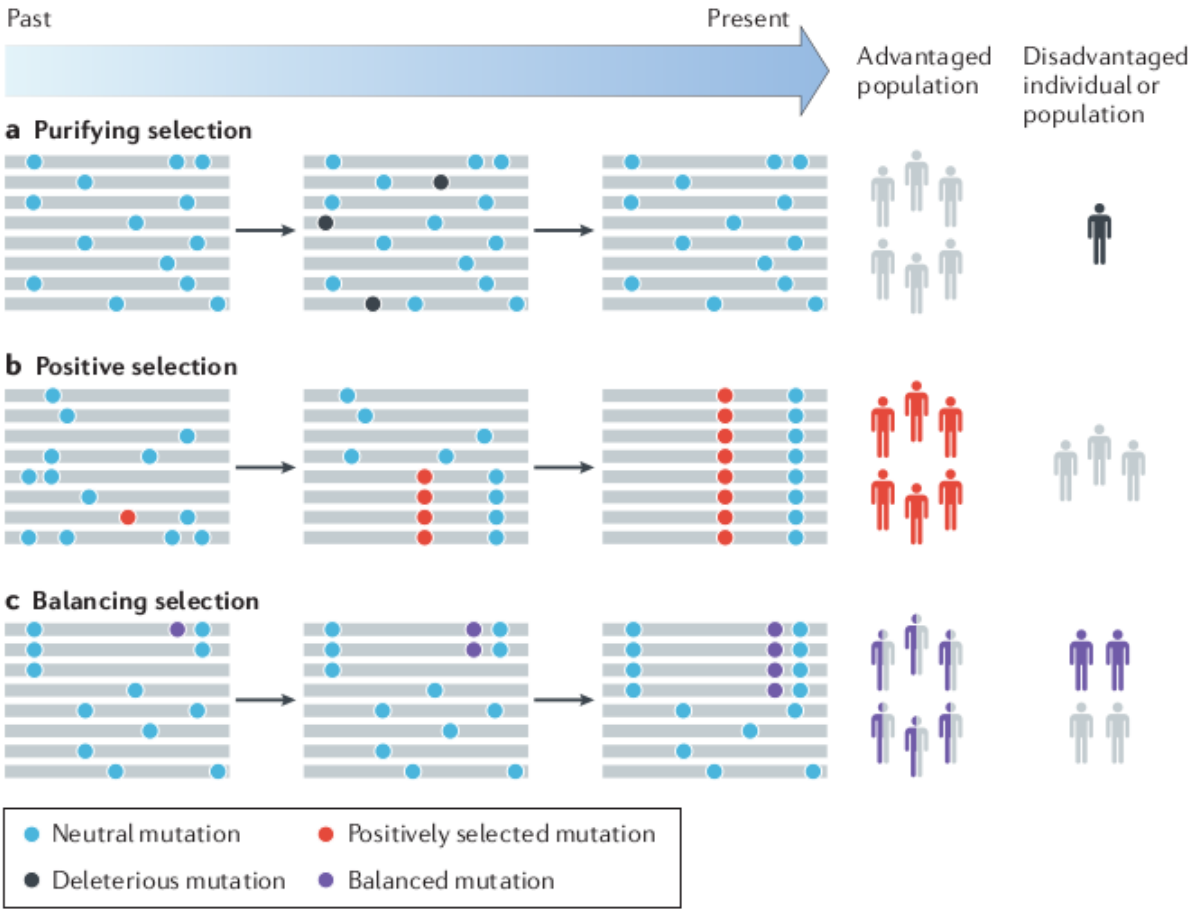


Figure 1: Types of natural selection and their legacy on the genome.

The evolutionary fate of different types of mutations is represented in a sample of eight chromosomes. Blue circles, neutral polymorphisms. **a** — Purifying selection removes deleterious alleles (indicated by black circles) from the population. Lethal mutations are immediately removed from the population while mildly deleterious are tolerated but kept at low population frequencies. These mutations tend to be associated with rare, severe disorders. **b** — Positive selection increases the frequency of an advantageous mutation (indicated by a red circle) in the population. Advantageous mutations can be fixed (completed selective sweep) or polymorphic (ongoing selective sweep; not shown) in the population. Positively selected mutations are often associated with common traits which present complex modes of inheritance. **c** — Balancing selection maintains polymorphism in the population as a result of heterozygote advantage and frequency-dependent advantage (not shown). In the illustrated example, a mutation (indicated by a purple circle) confers a selective advantage at the heterozygote state, so individuals who are heterozygous at this particular position have a greater fitness than homozygous individuals.
From Quintana-Murci and Clark 2013.

1.2 Genetic conflicts and the Red Queen hypothesis

1.2.1 Antagonistic relationships between species drive genetic conflicts

While organisms constantly adapt to abiotic variations, such as temperature or rainfall, in their environment, those adaptations are not countered: the environment does not adapt itself to the new phenotype in the population, and this adaptation remains unidirectional.

However, other factors can provide the impetus for adaptation; for example, interactions with other groups of organisms. Species sharing the same environment in a manner in which a change in one species' fitness impacts that of the other are co-evolving. When this relationship is antagonistic (i.e., increase of fitness on one side causes a decrease on the other side), those two species are said to be in conflict.

Such conflicts lead to rounds of adaptation and counter-adaptation between the two species, as they cannot both adapt to an optimal fitness at the same time. This evolutionary arms-race is expected to continue unless one of the species becomes extinct. This concept has been formalized by Van Valen 1973 in the Red Queen's hypothesis. It is based on Lewis Caroll's character, the Red Queen, running with Alice across a chessboard to only stay in the same place (Carroll and Tenniel 1871). Van Valen considers it a metaphor for the constant adaptation of species.

Genetic conflicts are intrinsically linked to natural selection. Let us consider two species involved in a genetic conflict as "prey" and "predator". An allele enhancing the prey's fitness will be favored within the prey's population: its frequency will augment as it is positively selected through the selection pressure applied by the predator (Figure 1). Meanwhile, this rise in fitness in the prey will be accompanied by a corresponding drop in fitness in the predator. Any allele allowing the predator's fitness to rise again will be favored in the predator's population, following the same schematic as previously described.

These observations can be translated to the molecular scale, as genetic conflicts can be viewed as series of molecular interactions between two partners: one for which function the interaction is deleterious, the prey, and another which benefits from the interaction, the predator.The predator tries to recognize the prey and maintain the interaction, while the preyaims to evade the predator. This leads to rounds of adaptation to evade the interaction and counter-adaptations at the interface between the two proteins.

1.2.2 Genetic conflicts also occur at the intra genomic level

Intragenomic conflict is observable with transposable elements. These elements contain the information necessary to replicate within their host genome, but are susceptible to elimination through mutation if they do not replicate enough. However, both their transposition and their over-representation in the genome can negatively affect the host's fit-

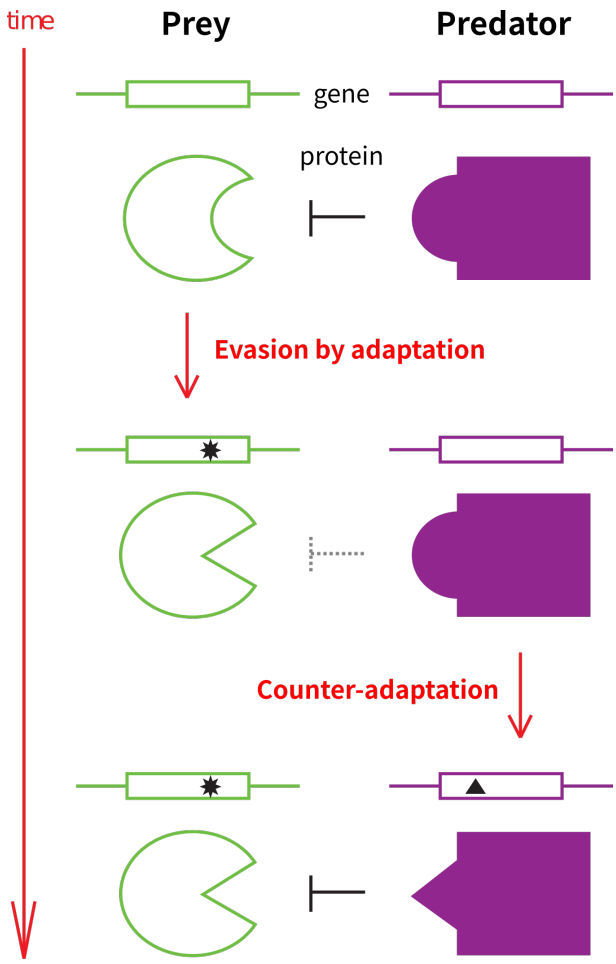


Figure 2: Long term genetic conflict.

Two proteins are engaged into a "Red Queen" conflict inducing rapid adaptive evolution of the gene encoding them over evolutionary time. This leaves signatures of positive selection in the genomes of both prey and predator, in particular at the interface between the two proteins. The ability of the predator's protein to interact with the prey's protein puts a selective pressure on the prey that will evolve and evade the interaction. Such selective pressure will, in turn, induce a rapid evolution in the predator based on their ability to regain their interaction with the prey protein. These interactions therefore set up unceasing evolutionary arms races. Black stars and triangles within genes represent amino acid changes in the genes. Adapted from Etienne 2015.

ness, by affecting genome stability and fertility, for example. In the germline especially, hosts use a variety of strategies to silence transposable elements (such as DNA and histone modifications). The KRAB-Zinc Finger (KZNF) genes present marks of genetic adaptation directly related to the regulation of transposable elements activity: they are part of an incredibly extent repertoire of 400 duplicated genes in humans, many of which are targeted to individual transposon families (Bruno et al. 2019).

Another example of intragenomic conflict is meiotic drive genes. The parasitic nature of these genes lies in their ability to skew the transmission rate to the next generation in their favor, from 50% normally to up to 100%. This effect is mediated through a toxic

effect on gametes that do not bear them, and has been described in both fungi spores and animal sperm (reviewed in McLaughlin and Malik 2017). This negatively impacts the fitness of the host, who loses numerous gametes from the action of such genes, and hosts have evolved drive-suppressor systems to prevent their killing effect (Zanders and Unckless 2019). For example, the large wtf multigenic family in *Schizosaccharomyces pombe* comprises more than 20 genes, some of which are known to be killer meiotic drivers. Various members of the family show various marks of rapid evolution, consistent with the adaptation pattern induced by genetic conflict (Eickbush et al. 2019).

Other intragenomic conflicts, as well as conflicts that have occured between two groups within a species (such as between two sexes), have been reviewed in McLaughlin and Malik 2017.

1.3 Viruses as evolutionary shapers of host genomes

Host-pathogen interactions are antagonistic in nature. Therefore, it comes as no surprise that virus-host interactions set up major genetic conflicts that have affected the host genomes over evolutionary times.

1.3.1 The prey and the predator

All Viral Interacting Proteins (VIPs) come in two flavors, provirals and antivirals. Provirals enhance viral replication, such as entry receptors and co-receptors. Antivirals, also called restriction factors, are specialized genes that have evolved to target target various steps of the viral cycle and block viral replication. Viruses, for their part, have evolved means to either evade or antagonize the host's defenses. Depending on the context, the identity of the prey and predator can thus be exchanged, depending on which organism is on offense.

When the virus evades the host restriction factor that negatively targets one of its protein, the viral protein is the prey. It adapts by avoiding recognition through rapid evolution of the targeted interface. This is the "strategy" used in the case of the capsid protein of lentiviruses to escape recognition by the TRIM5 protein (reviewed in Daugherty and Malik 2012; Etienne 2015). If, however, the virus uses an antagonist to counteract the action of the host restriction factor, then the viral protein acts as the predator. This latter

"strategy" can be particularly useful against sensors/PRRs (see Introduction - Chapter III) and restriction factors with broad antiviral actions, which target a replication step that is common across several virus families. For example, the Protein Kinase R (PKR) recognizes double-strand DNA and mediates an inhibition of translation initiation through phosphorylation of eIF2α, inhibiting various viruses (Elde et al. 2009; Rothenburg et al. 2009). As such, several of those viruses have evolved antagonists or evasion mechanisms to avoid PKR inhibition, such as poxviruses, influenza A virus, herpesviruses, etc. (Gal-Ben-Ari et al. 2018).

This complex relationship, in which either of the players can alternatively be the offender or the defender (Daugherty and Malik 2012), proves the inextricable influence of both host and virus on each other's evolution.

1.3.2 Widespread influence of viruses on host genomes

Viruses are strong candidates to explain the pervasive presence of adaptive mutations in their host genomes. In mammals, a number of proteins involved in antiviral defense display high rates of adaptation, but overall VIPs appear to evolve unusually slowly rather than unusually fast in both plants (Mukhtar et al. 2011; Weßling et al. 2014) and animals (Jäger et al. 2012; Davis et al. 2015). However, a study of 1,300 proteins reported as VIPs in the literature found out that they were under much stronger evolutionary constraint than other proteins, but also displayed much higher rates of adaptation (Enard et al. 2016). The authors estimate that viruses are responsible for at least 30% of all adaptive mutations in the human genome, though they point out this number is likely underestimated. This is due to the following reasons: the probable existence of many yet-undiscovered VIPs, the possibility of viruses driving adaptation on proteins they do not directly interact with, and the bias towards human VIPs in their dataset (Enard et al. 2016).

In primates, the potential role of viruses, especially lentiviruses, as strong drivers of adaptation had already been theorized (Cagan et al. 2016). It was further confirmed in vervet monkeys (Svardal et al. 2017). Their analysis showed that the co-evolution of vervet monkeys with their Simian Immunodeficiency Virus (SIV) was responsible for part of this adaptation, though other viruses may have also driven some of it. Similarly, viruses, likely SIVs, were found to be partly responsible for strong signatures of adaptation in chimpanzees (Schmidt et al. 2019). And these large scale genomic studies were

confirmed by the analyses of polymorphism in specific lentiviral VIP-encoding genes; e.g. APOBEC3G (Compton et al. 2012; Krupp et al. 2013), TRIM5 (Cagliani et al. 2010; McCarthy et al. 2015), CD4 (Meyerson et al. 2014), amongst others.

Recent evidence also points out that viruses may have influenced the adaptation of birds (Shultz and Sackton 2019), as well as invertebrates (Palmer et al. 2018). This highlights the widespread influence of viruses on the evolutionary history of their hosts.

1.4 Mechanisms of genetic adaptation in host-virus arms-races

Adaptation can be mediated through diverse mechanisms. Here, we highlight some of them.

1.4.1 Single point mutations and indels

Single-point mutations of proteins can have diverse effects: acquisition of new functions, loss of function, conformational changes, etc. These mutations lead to the appearance of multiple alleles within the population. A major mechanism of adaptation is by the fixation of mutations inducing favorable variants in the population by positive selection. However, balancing selection can itself be a feature of adaptation, by maintaining different haplotypes in order to deal with multiple selection pressures from the environment. This is the case of the Major Histocompatibility Complex (MHC) genes (Quintana-Murci 2019; Radwan et al. 2020) which are able to recognize a multitude of pathogens. This complex is a major actor of adaptive immunity, illustrating the impact that natural selection can have at different timescales on either innate or adaptive immunity.

Short insertions and deletions within protein-coding genes can also play a role in adaptation. However, the changes they induce on protein structure and function are much more dramatic than with single-point mutations, as they can easily change the open reading frame of the gene, and even abolish its expression entirely (Rokas and Holland 2000). As such, they are rarer and more easily eliminated in the population by purifying selection, though insertions are slightly better tolerated than deletions (Ajawatanawong and Baldauf 2013).

1.4.2 Recombination

Recombination is an important role-player in adaptation, and can act at various levels. Meiotic recombination plays a major role in sexual species by generating new genetic combinations every generation, upon which natural selection can operate. However, though sex demonstrably speeds up adaptive changes (McDonald et al. 2016), the overall impact of meiotic recombination on adaptation is debated (Ritz et al. 2017; Ortiz-Barrientos et al. 2016). Non-meiotic recombination, on the other hand, has been shown to have a strong impact on adaptation in various organisms, such as the evolution of the V(D)J system in jawed vertebrates (Carmona and Schatz 2017) and bacterias (Didelot and Maiden 2010). It has also been shown to be a major driver of viral evolution (Pérez-Losada et al. 2015).

Most notably, recombination is the driver of gene conversion. Gene conversion occurs when a donor sequence at one locus of the genome is copied, completely or partially, to another allele's locus, as a consequence of recombination to repair DNA double strand breaks. It can have different outcomes: generation of duplicate genes through segmental duplication, homogeneization of duplicates causing reduction of genetic diversity, and variant shuffling through recombination to generate novel alleles from the pool of existing ones (Daugherty and Zanders 2019).

1.4.3 Duplication and gene family expansion

Another major source of adaptation is gene duplication and expansion. While this can happen through gene conversion, it is not the only mechanism that can lead to gene duplication. These events can also result from unequal crossing overs during meiotic recombination, through retrotransposition or through whole genome duplications, sometimes followed by massive gene loss and specialization (reviewed in Crow 2006; Magadum et al. 2013). Whether they originate from gene conversion events or from the three other possibilities listed here, the fate of duplicated genes can be diverse, and either be functionally maintained or diverge (Figure 3).

The functions of the gene can be conserved and favor the organism's fitness through a simple dose effect, with more copies of the same gene increasing its level of expression, as is observed with the high number of genes encoding histone proteins or ribosomal RNAs (Hurst and Smith 1998). Functional can be insured either by homogeneization through

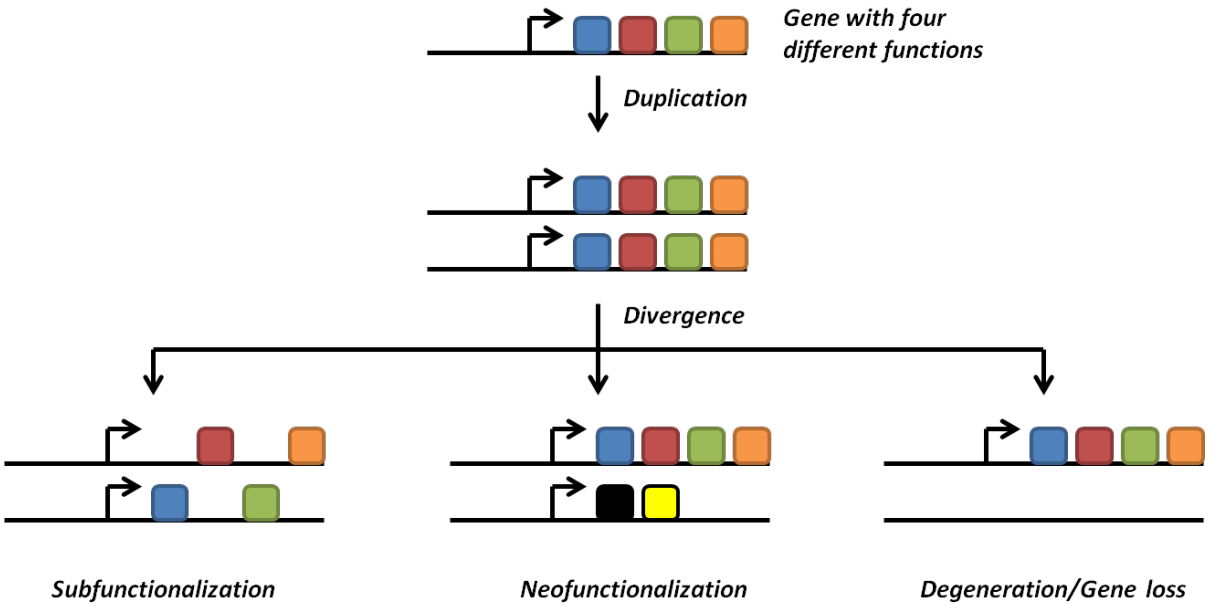


Figure 3: The possible fates of duplicated genes.

All colored squares represent different domains/functions of a gene. Following duplication, a gene can evolve to different fates: either both copies can be maintained as they are, each conserving all four original functions, or they can diverge over evolutionary times. This divergence can lead to subfunctionalization (where the original functions are shared between the two different copies), neofunctionalization (with one copy maintaining the original functions while the other one acquires new ones) and gene loss or degeneration of one of the copies.

gene conversion or by strong purifying selection. In the absence of such mechanisms, the duplicated copies will undergo divergence.

Divergence can lead to subfunctionalization, in which the duplicated genes share the various functions of the original gene amongst themselves. An alternative outcome is neofunctionalization, in which the duplicated copy acquires novel functions, expanding the organism's repertoire. These processes have been shown to mediate adaptation in humans through recent duplications of genes involved in development of the brain (Dennis and Eichler 2016).

Finally, harboring two copies of the same gene is generally not advantageous to the organism, as it produces functional redundancy. In that case, the surplus copies of the gene will slowly undergo pseudogeneization through the accumulation of random mutations in their sequences under the influence of neutral selection (degeneration), or be lost entirely. Pseudogenes are unexpressed or functionless, and can diverge so much from their parental gene as to be unidentifiable, and up to 60% of a gene family can be constituted of pseudogenes in humans (Magadum et al. 2013).

Chapter 2

Tools and pipelines to study genetic adaptation on protein-coding genes

2.1 Independent tools

2.1.1 Retrieving homologs

Identifying events of genetic adaptation on protein-coding genes implies to first retrace their evolutionary history. The very first step of that process lies in the retrieval of all the evolutionary related sequences, i.e. homologs, of the gene in the species of interest.

Retrieving homologs is often performed through searching for similar sequences in databanks (Hu and Kurgan 2019; Pearson 2013). It can be performed with tools such as BLAST (Camacho et al. 2009), BLAT (Bina 2008), FASTA (Pearson and Lipman 1988), HMMER (Johnson et al. 2010) or MMseqs2 (Steinegger and Söding 2017). Of those, BLAST is arguably the most widely used, either through its web implementation or its command-line interface. Large and diverse sequence databanks can be searched through, such as NCBI's GenBank (Benson et al. 2018), EMBL's Nucleotide Sequence Database for nucleotide sequences (Kulikova et al. 2007), or UniProt for protein sequences (Boutet et al. 2016), amongst others. Those databanks contain massive datasets, allowing for the retrieval of homologous sequences from many diverse species, and are continuously growing with the introduction of new sequences from high-throughput sequencing experiments.

Homologs are genes which share a common evolutionary ancestry. The divergence from this common ancestry can arise from speciation events, in which two paralogs will follow

different evolutionary histories as part of the genomes of different species. The alternative option is for two homologs to diverge through the consequences of a duplication events, in which case they are called paralogs, and are found in the genome of the same species. A highly efficient way to retrieve homologs is thus to compare their similarity, i.e. the degree to which they ressemble each other in terms of their sequence and composition in either nucleotide or amino-acids. However, sequence similarity is not the sole marker of homology: sometimes orthologs share more resemblance through structural similarity, for example (Pearson 2013).

Moreover, the distinction between orthologs and paralogs cannot be solely based on similarity: orthologs will sometimes share more similarity than paralogs, depending on the various selective pressures to which they have been confronted. For various methods of detecting marks of genetic adaptation based on phylogenetics, the most pertinent genes to study are orthologs, and not paralogs, in order to properly retrace the evolutionary history of a gene between different species.

Multiple approaches have been developed to both retrieve and robustly infer orthologous relationships across species, which results are available in different databases, either generalist (COG/KOG, Tatusov et al. 2003; HOGENOM, Dufayard et al. 2005; InParanoid, Ostlund et al. 2010) or taxonomically specialized (OPTIC for vertebrates, Heger and Ponting 2008; INVHOGEN for nonvertebrates, Paulsen and Haeseler 2006; GreenPhylDB for plants, Conte et al. 2008; OrthoMAM for mammals, Scornavacca et al. 2019; etc...). However, one of the main drawbacks of those databases is their ability to keep up to date with the ever increasing sequences and genomes contained in the previously described databanks from the NCBI and EMBL, which might impact the number of retrieved homologs at a given time.

2.1.2 Aligning coding sequences

One of the major endeavors of exploring genetic innovation at the level of protein-coding genes rests in the proper production of a Multiple Sequence Alignment (MSA). All analyses for molecular evolution rely on this MSA, such as phylogenetic inference or detection of positive selection. Those analyses can be severely impacted by the quality of the MSA on which they are performed (Loytynoja and Goldman 2008; Wong et al. 2008). Specifically, positive selection is detected by comparing the rates of non-synonymous mutations

(dN) over synonymous mutations (dS). This implies the necessity to consider the coding sequence at the nucleotide (NT) level, while retaining the ability to infer the corresponding amino-acid (aa) sequence. Thus, detecting genetic innovations on protein-coding sequences relies on the ability to produce correct codon alignments.

Historically, widely used MSA softwares such as CLUSTAL (Thompson et al. 1994), MUSCLE (Edgar 2004) or MAFFT (Katoh et al. 2002) proposed to either align nucleotide sequences or protein sequences, but did not take into account the codon unit. This led to the improper introduction of gaps in the middle of codons rather than between them, due to the inability to detect the translation frameshift those gaps caused. However, some of those softwares remain excellent solutions for nucleotide MSA and are widely used, as is the case for MAFFT. It presents a wide array of options to provide good quality nucleotide or protein alignments, and can automatically select the best ones for the sequences provided by the user.

A common strategy to address the problem of aligning coding sequences follows a three-step approach: translating the sequence from nt to aa, aligning the aa sequence and deriving the codon alignment from the protein one (translation-alignment-"back-translation"). Numerous tools automate this approach, such as revTrans (Wernersson and Pedersen 2003), transAlign (Bininda-Emonds 2005), PAL2NAL (Suyama et al. 2006) and TranslatorX (Abascal et al. 2010). However, this strategy fails to properly handle frameshift substitutions, which can lead to erroneous translated AA sequences and, in turn, codon alignments (Thompson et al. 2011).

To date, only two MSA softwares offer true codon alignment modes, PRANK (Loytynoja and Goldman 2008) and MACSE (Ranwez et al. 2011). PRANK, especially, has been shown to produce better alignments for positive selection analyses compared to aligners which did not offer a codon mode (Fletcher and Yang 2010; Jordan and Goldman 2012; Markova-Raina and Petrov 2011; Privman et al. 2012; Schneider et al. 2009).

2.1.3 Substitution models and reconstructing phylogenies

Inferring the phylogeny, i.e. the reconstruction of evolutionary relationships between individuals or groups of organisms (species, populations), is an essential step to identify underlying genetic traits that shape evolution (reviewed in Smith et al. 2020). This holds true for the detection of genetic adaptation, which necessarily requires knowledge of the

evolutionary relationships between homologous sequences to highlight the innovations that may have occurred.

The main statistical methods of reconstructing phylogenies rely on either Maximum-Likelihood (ML) or Bayesian inference (see Whelan and Morrison, 2017 for an extensive review). These methods account for uncertainty in the evolutionary history by assigning probabilities for sequence changes. Widely-used softwares for inference of phylogenetic trees include PhyML (Guindon et al. 2010) and RaxML (Stamatakis 2014) for maximum-likelihood, and BEAST (Bouckaert et al. 2019) and MrBayes (Ronquist et al. 2012) for Bayesian inference.

These methods rely on the modeling of substitutions of one nucleotide (or amino-acid) to another. This modeling is made through Markov models, matrices describing the conditional probabilities of changes from one nucleotide (or amino-acid) to another. Models differ through the assumptions on substitution rates: for example, the simplest DNA substitution model assumes that the rate is identical between all substitutions and equal nucleotide frequencies (JC, Jukes and Cantor 1969). However, models rapidly evolved to account a better description of changes, such as different rates of changes between transversions and transversions (K80, Kimura, 1980) or different nucleotide equilibrium frequencies (F81, Felsenstein 1981). Extensions to those originals models were also incorporated, for example in the HKY model (Hasegawa et al. 1985). One of the most complex DNA substitution model is the general time-reversible model (GTR, Tavaré 1986), which uses different rates for every change and different nucleotide frequencies. In addition, all models can account for a proportion of invariable sites (+I) (Shoemaker and Fitch 1989), as well as rates of variation across sites with a gamma distribution (+G) (Yang 1994).

The complete modeling of data includes other parameters, such as stationarity (nucleotide frequencies are constant), reversibility (probabilities are the same for a substitution in one direction and its reverse direction at the equilibrium of the model) and homogeneity (the rates of change are the same across the tree branches) of the evolutionary process at each site. More complex modelings have been developed to add even more parameters, as well as to model codon evolution for more realistic evolutionary inferences of protein-coding sequences (reviewed in Arenas 2015). The complexity of these models is raised by their accounting of the genetic code in their rates of substitution, including the position of the nucleotide within the codon. Importantly, they also differentiate

between substitutions that cause a change in the amino-acid encoded (non-synonymous mutations) and those that do not (synonymous mutations). This is of major relevance to the detection of positive selection at the inter-species level, as will be discussed later.

2.2 Identifying genetic innovations

2.2.1 Duplication

The identification of duplication events is of great importance in molecular evolution. The first issue stems from the distinction between orthologous and paralogous sequences. As only orthologs reflect the evolutionary history of their species, phylogeny inferences must be made after strictly separating orthologs from their paralogs when aiming to retrace the history of one gene. The difficulty of automating this separation is further compounded by gene deletions, variations in evolutionary rates and lateral gene transfer (LGT).

One of the main methods to resolve those relationships, when genomic sequences and reconstruction of synteny are not available solutions, can be done through phylogenetic analysis of the relevant gene families, in particular tree reconciliation following the Duplication-Loss(-Transfer) (DL(T)) model. This approach reconciles the phylogenetic tree of the sequences of interest (the so-called "gene" tree) with respect to a trusted species tree (Szöllősi et al. 2015). Several softwares are available to perform tree reconciliation, such as Notung (Stolzer et al. 2012), ecceTERA (Jacox et al. 2016), Ranger-dtl (Bansal et al. 2018), and the recently developed TreeRecs (Comte et al. 2019).

While those programs are useful to provide reconciled trees annotated with DL(T) events, their interpretation can sometimes be difficult and require in-depth analysis to properly discriminate orthologs from paralogs and reconstruct the evolutionary history of a gene family.

2.2.2 Recombination

Given the evolutionary impact and pervasiveness of recombination, the importance of detecting such events cannot be understated. Evidently, the first interest lies in the detection of genetic innovations arising from recombination and gene conversion events, which are prevalent in duplicated genes families, as outlined previously.

However, many methods pertaining to molecular evolution do not account for recombination, and assume nucleotide sequences evolve without recombination. Intuitively, not properly accounting for recombination can undermine the validity of such analyses. For example, phylogenetic inference based on nucleotide sequences, as described previously, assumes that all sites in the alignment share the same evolutionary history. This might not be the case when recombination occurs, and different parts of the alignment may have different histories. Indeed, it has been pointed out before that not accounting for recombination can lead to bias, or even errors, in the reconstruction of phylogenies (Posada and Crandall 2002). Methods subsequently based on such phylogenies, such as positive selection analyses, can in turn be affected, for example by inflating the number of sites detected as positively selected (Anisimova et al. 2003; Kosakovsky Pond et al. 2008).

Many softwares are available for the detection and characterization of recombination on nucleotide sequences, spanning a great variety of approaches, from testing for overall evidence of recombination to identifying breakpoints and recombinant sequences. The full landscape of such programs can be seen at a glance in Figure 4 (see Martin et al. 2011 for an extensive overview of those programs).

2.2.3 Positive selection

One of the major mechanisms driving evolutionary innovation is the accumulation of beneficial amino-acid changes. Over evolutionary time, this may lead to a signature of positive selection, i.e. an excess of non-synonymous substitution rate over synonymous rate.

To detect these signatures, the ratios dN and dS are calculated over a set of sequences. If dN/dS (sometimes called ω) is much smaller than 1, this denotes that non-synonymous changes are outweighted by synonymous changes, due to their elimination from sequences as they probably cause a fitness disadvantage: this means purifying selection. A dN/dS approximating 1 denotes that neither non-synonymous or synonymous changes are being favored and are unlikely to have much impact on the organism's fitness: such sites would be considered under neutral selection. However, when dN/dS is superior to 1, amino-acid changes have probably enhanced fitness, leading to their fixation in the population through positive selection.

As described previously, inference of ω is rendered possible by probabilistic codon

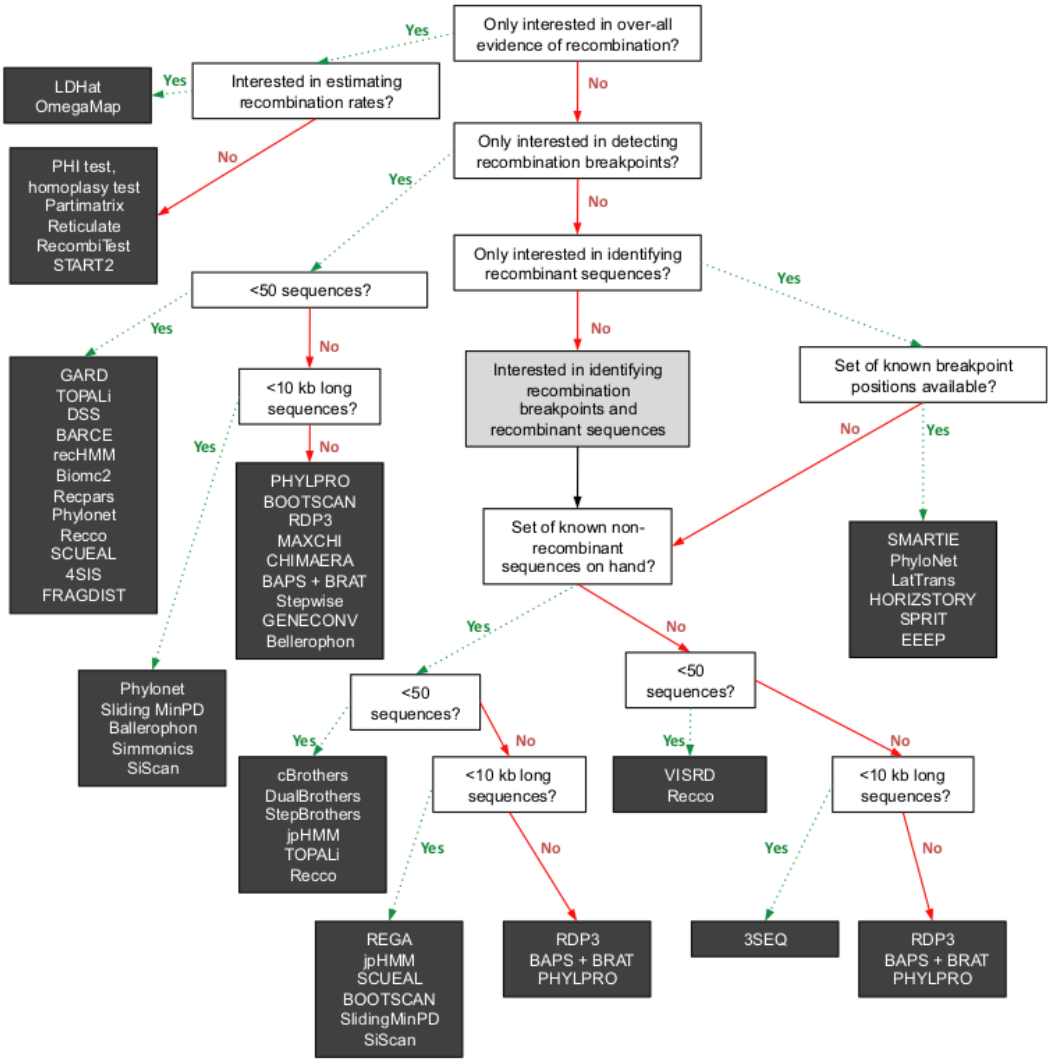


Figure 4: A decision tree outlining the landscape of available programs for recombination analyses.

This decision tree presents an overview of available programs for the detection of recombination, depending on the specific needs of the user. From Martin et al. 2011.

substitution models where it is a parameter (Nielsen and Yang 1998; Guindon et al. 2004; Kosakovsky Pond and Frost 2005; Yang 2007). Based on a codon alignment and its phylogenetic tree, the maximum likelihood estimate of such models is computed and allows to infer dN/dS (ie, ω) over the data. Indeed, genes that have undergone positive selection are unlikely, as a whole, to have an overall $\omega > 1$, as that would mean that most of their sites have undergone positive selection. On the contrary, ω is more likely to vary along the gene, with sites of functional importance in the interface involved in the genetic conflict being positively selected, while others remain under neutral or negative selection depending on their involvement in other functions of the protein. This heterogeneity can

be modelized by mixed models including several categories of ω, which calculate for each site the mean likelihood to belong to either of those categories.

To ascertain whether positive selection has occurred on a given gene, the codon alignment of this gene is fitted to a model allowing for positive selection (with a category of $\omega > 1$) and its neutral alternative (without this $\omega > 1$ category). For example, in widely-used site model M1a (Figure 5), the likelihood is computed on the hypothesis that ω is allowed to vary along sites (between two categories: $\omega < 1$, in this example 0.06, and $\omega = 1$) but no positive selection is allowed, while the M2a model has one more category allowing for positive selection (in the example $\omega = 3$). Likewise, the M7 (resp. M8) model disallows (resp. allows) for positive selection, but differs from the previous models by modeling the $\omega < 1$ through a beta distribution (discretized in a given number of categories for computation purposes). The latter leads to a more realistic modeling of the selective pressure on the different sites, but is more computationally intense. The likelihoods of nested models are then compared through a likelihood-ratio test to ascertain whether allowing positive selection (models M2 or M8) provides a better fit to the data over models not allowing positive selection (models M1 or M7, respectively).

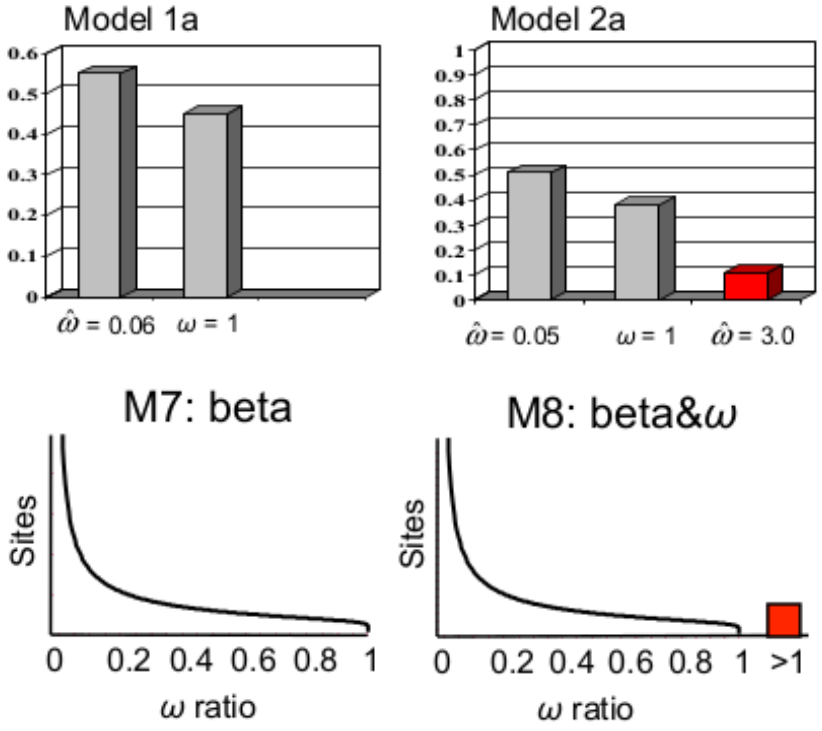


Figure 5: An overview of codeml site models to explore positive selection.

Graphical representation of ω categories in site models. Adapted from Yang, 2007, figures by Joseph Bielawski.

This can be done at different levels by allowing the ω to vary depending on the question explored. The site models we used as example will allow ω to vary along the sites and can therefore be used to identify which specific sites are under positive selection. In branch models, ω is assumed to vary amongst lineages, i.e. across branches of the phylogenetic tree. The branch models can therefore be used to identify which branches have been subjected to positive selection. Finally, branch-site models combine both, and ω will be allowed to vary across both sites and branches. The branch-site models can therefore be used to determine which sites in specific lineages have been subjected to positive selection. However, the latter is highly parameter rich and infers positive selection from very little information (one site and one branch), which make them more liable to bias. The most widely-used software to compute such analyses is PAML codeml (Yang 2007). Other software suites for the such analyses include Bio++ (Guéguen et al. 2013) and HyPhy (Pond et al. 2005). The codon substitution models included in Bio++ are the same as those used in PAML. However, Bio++ allows for far more parametrization, for example making it possible to assume non-stationarity and non-homogeneity of the evolutionary process. HyPhy for its part, uses completely different models, though the spirit of the approach remains the same.

In this part we focused strictly on positive selection during inter-species evolution, which is the main topic of this thesis. Other statistic tests are available for the study of intra-species evolution and analysis of polymorphisms in populations, reviewed in Quintana-Murci and Clark 2013.

2.3 Pipelines

Due to the widespread interest in identifying the evolutionary mechanisms leading to adaptation on protein-coding genes, a number of pipelines for the identification of such mechanisms have been made available over the years. These tools aimed at automating the analysis of large datasets in order to provide reliable and replicable genome-scale analyses, and mostly focused on the detection of positive selection.

As outlined previously (Lee et al. 2017), a successful pipeline devoted to this goal requires the automation of essential steps. Similarly, the developers of PosiGene (Sahm et al. 2017) highlighted some of the tasks and challenges for such pipelines, reviewing

Table 1: Features of existing pipelines for the detection of positive selection on large datasets.

	Datamonkey	Selecton	JcoDA	IDEA	PSP	POTION	PoSeiDon	PosiGene
automatic homolog retrieval	-	-	-	-	±	±	-	±
detection of ORFs	-	-	-	-	+	-	-	-
high quality codon alignment	-	+	-	-	-	+	±	+
phylogenetic tree reconstruction	+	+	+	+	+	+	+	+
detection of duplication events	-	-	-	-	±	±	-	±
detection of recombination events	+	-	-	-	+	+	+	-
multiple methods to detect positive selection	±	±	-	-	-	-	-	-
flexible	+	-	-	-	-	±	-	-
user-friendly	+	+	±	+	±	-	+	±

existing tools in 2017 for their ability to answer such challenges. Based on their observations and our own, we established a list of features we needed for a complete automation of our workflow in the lab, and observed that none of the existing tools perfectly fulfilled our expectations (Table 1).

Primarily, searching for sequences and identifying orthologous relationships represent a time-consuming and difficult process. Few tools include this step: softwares such as the Hyphy suite (Pond et al. 2005), Selecton (Stern et al. 2007), IDEA (Egan et al. 2008), JcoDa (Steinway et al. 2010) or even the more recent PoSeiDon (Fuchs et al. 2017), require the user to provide his own set of hand-curated sequences or alignment. Amongst those that do automate the retrieval of homologous sequences, PhyleasProg (Busset et al. 2011) and PSP (Su et al. 2013) are limited to vertebrate species and prokaryotic species respectively. PosiGene (Sahm et al. 2017) does assign ortholog relationships but relies on annotations, mostly based on HomoloGene, which can make it challenging to use with non-model species. Overall, no existing tool include the automatic search of homologous sequences and their subsequent assignment to orthologous groups on a wide array of species, while including the possibility to work on non-model species.

The second step of importance requires an accurate multiple sequence alignment and corresponding phylogenetic tree. The necessity of high-quality codon alignments for the accurate detection of residues under positive selection has been highlighted previously (Chapter 2, section 1.2). Overall, previous pipelines tend to rely on aligners without a codon alignment option and with high penalties for gaps, such as ClustalW (Thomp-

son et al. 1994) or MUSCLE (Edgar 2004), and often used the three-steps approach of translation-alignment-"back-translation" to provide codon alignments which was described previously, rather than aligners which perform true codon alignments.

Finally, van der Lee and colleagues identify the ability to access parameterization of the Maximum Likelihood models used for the detection of positive selection. For this, existing tools rely extensively, and almost exclusively, on PAML codeml (Yang 2007), which has allowed the identification of numerous genes under positive selection. However, the extent to which codeml is parameterizable is limited, and the addition of different models could help confirm and expand on the results obtained through it. Very few tools include different models than the ones present in PAML: Selecton (Stern et al. 2007) includes a Mechanistic and Empirical Codon model and PSP (Su et al. 2013) fitmodel (Guindon et al. 2004). Bio++ (Guéguen et al. 2013) libraries, which are very flexible as to parameterization, or the different models from HyPhy (Pond et al. 2005), have not been used in those pipelines, though they can provide other perspectives on the detection of positive selection, and be used in conjunction with other models to confirm or infirm results. This appeared to be major points for an innovative tool.

While those three points are indeed of prime importance, we also identified thatadaptation to environmental constraints is not solely solved through site-specific positive selection. Other potential sources of genetic innovations have been ignored in previously available tools. For example, recombination events are widely ignored in existing tools. Only PSP (Su et al. 2013) and PoSeiDon (Fuchs et al. 2017) accounted for them in their workflow.

Pipelines also tend not to retain indels and information about splice variants as they often rely on the use of filtering softwares such as Gblocks (Talavera and Castresana 2007) to remove regions with low alignment coverage. However, this frequently negatively impact the reconstruction of phylogenies (Tan et al. 2015) and removes information from the alignment which can be pertinent to the evolutionary history of the genes. Importantly, studies focusing on identifying genes under positive selection generally ignore duplication events (such as Kosiol et al. 2008; Hawkins et al. 2019; Cooper et al. 2019). Tools that automate homolog retrival and perform orthologous assignments rely on annotations, which means they cannot be used to detect either recent duplications, or even ancient ones on non-model species (Su et al. 2013; Sahm et al. 2017). This could lead to an

underestimation of the number of genes engaged in genetic conflicts.

Chapter 3

Innate immunity in the context of viral infection: pathogen sensing, the interferon response and interferon stimulated genes

Cellular organisms have evolved different strategies to protect themselves against infections from pathogens, be it viruses, bacteria or parasites. These strategies can be broadly categorized in two types of immunity. First, the innate immunity recognizes common patterns in pathogens and provokes a broad immune response designed to eliminate a wide array of pathogens. Second, the adaptive immunity will elicit a memory response tailored for the specific agent responsible for the infection when it occurs. Both are linked to each other, as innate immunity activates the adaptive one.

3.1 Pathogen recognition upon infection

Upon infection, common pathogen patterns, designated as PAMPs (Pathogen Associated Molecular Patterns), are recognized by PRRs (Pattern Recognition Receptors) (Seong and Matzinger 2004). They consist of small molecular motifs that have been evolutionarily conserved within a group of pathogens as they are essential to their life cycle. This has led the host to evolve a common immune response for that group.

PAMPs are a wide array of molecule types, such as flagellin, lyposaccharides and pep-

tidoglycan for bacteria. These molecules are not produced by the host and are thus easier to identify as exogenous. In the case of viral infection, however, PAMPs are often their RNA, DNA or DNA:RNA hybrid forms, molecules which are also found in the organism. As such, it is of paramount importance for the organism to be able to discriminate between its own nucleic acids and the ones brought by the virus through their PRRs. These PRRs belong to three main categories: TLRs (Toll-Like Receptors), RLRs (RIG-I-Like Receptors) and DNA sensors such as cGAS (Figure 6).

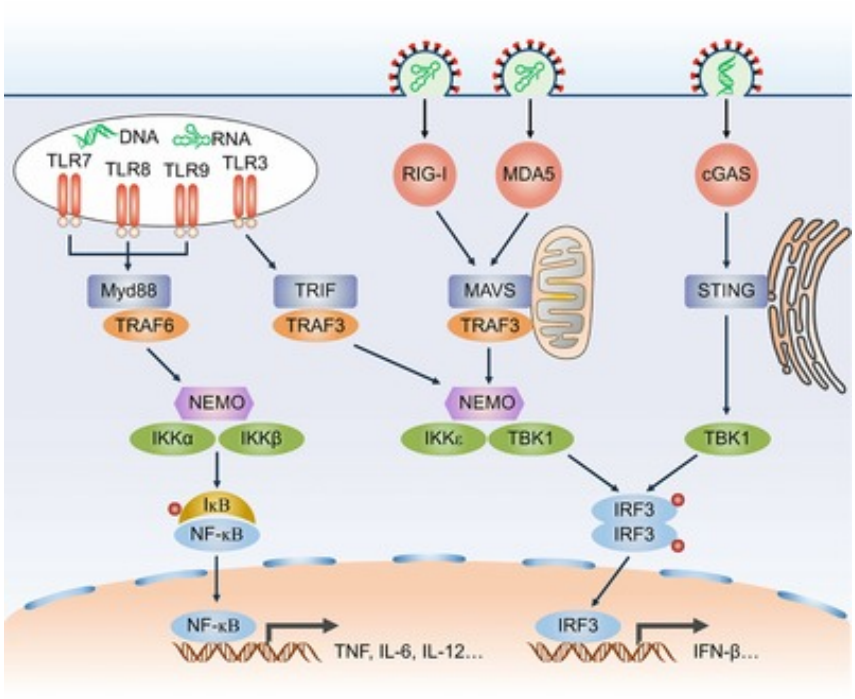


Figure 6: Antiviral signaling pathways.

Once activated by their respective ligands (viral RNA or DNA, endosomal or cytosolic), PRRs (TLRs, RLRs, DNA sensors, in red) recruit signal transduction adaptators (such as Myd88 or STING, in blue). For TLRs and RLRs, this then activates either TRAF6 or TRAF3 (in orange) which will induce the formation of complexes involving NEMO (in purple) in conjunction with kinases (in green), which phosphorylate transcription factors (in light blue) or their inhibitor (in yellow). This allows their translocatation into the nucleus where they will drive proinflammatory cytokines and type I IFNs expression. DNA sensors such as cGAS induce a similar signaling cascade, though without the involvment of TRAF proteins and NEMO. From Zhou et al. 2017

3.1.1 Toll-Like Receptors

Toll-Like Receptors are transmembrane proteins found at the surface of immune cells or the membrane of their intracellular vesicles. They constitute a multigenic family of 10 members in humans (named TLR1-10) and present the same domains: an ectodomain

recognizing a specific PAMP, a transmembrane domain, and a cytosolic TIR (Toll-IL-1 Receptor) domain for adaptators' binding.

TLR1 to 9 recognize specific PAMPs which allow for the recognition of a large variety of pathogens. TLR10, on the other hand, has an opposite effect to the rest of its paralogs and promotes anti-inflammatory processes (Hess et al. 2017). Once that PAMP has been recognized, the TLR dimerizes and recruits the specific adaptator molecules recognized by its TIR domain. Such adaptators can be: myeloid differentiation primary-response protein 88 (MYD88), TIR domain-containing adaptor protein inducing IFNβ (TRIF), MYD88-adaptor-like protein (MAL) and TRIF-related adaptor molecule (TRAM).

TLRs found at the surface of cells target bacterial PAMPs, while virus sensing happens mostly in intracellular vesicles and endosomes through detection of their nucleic acids. Out of the 10 human TLRs, four focuses on viral nucleic acid sensing: - TLR3 recognizes viral dsRNA (Alexopoulou et al. 2001) - TLR7 and TLR8 both recognize ssRNA from RNA viruses (Heil et al. 2004) - TLR9 recognizes the CpG (Cytosine – phosphate – Guanine) motif found in bacterial and viral DNA but quite rare in mammalian genomes (Hemmi et al. 2000; Ohto et al. 2015).

Once engaged, TLR3 will activate a TRIF-dependent pathway, while the other three will activate their signalisation through MyD88. However, their responses all converge in the activation of diverse transcription factors in order to induce production of type I interferon and inflammatory cytokines (O'Neill et al. 2013).

Interestingly, TLRs have been shown to be under pervasive purifying selection in primates, especially great apes, suggesting less redundant functions than in humans. In the same vein, endosomal TLRs, mostly devoted to viral sensing, appear to be under greater purifying selection than cell-surface TLRs. Yet, some sites were detected under positive selection, notably one mediating recognition of the Herpes Simplex Virus 1 (HSV-1), to which human are naturally resistant but non human primates are fatally susceptible (Quach et al. 2013).

3.1.2 Rig-1 Like Receptors

Immune cells can detect virus infections at their membranes: however, viruses do not necessarily infect immune cells, and non-immune cells, while not expressing TLRs, are able to mount effective innate responses against pathogen infections (Diebold et al. 2003).

This is made possible by the presence of Rig-1 Like Receptors (RLRs) in the cytoplasm of all cells, either immune or non-immune, which are able to sense cytosolic viral RNA, as evidenced by the identification of RIG-I, MDA5 and LPG2 (Yoneyama et al. 2004; Yoneyama et al. 2005). Those three receptors share the same structure (Figure 7)

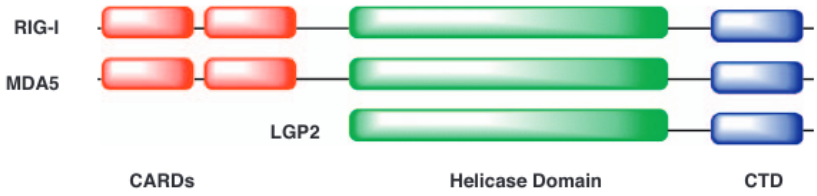


Figure 7: Structure of RIG-I-like receptors.

The CTD (C-Terminal Domain) mediates specific recognition of the RNA ligand, while the CARD domain allows binding to the MAVS adaptator and downstream signaling. From Bowzard et al. 2011

While LPG2 is considered as an RLR and displays the highest RNA binding affinity among them, the absence of a CARD domain prevents it from activating downstream antiviral signal (Wu and Chen 2014). It has been suggested it acts as a regulator of RIG-I and MDA5 signaling, specifically of MDA5 sensing (Bruns and Horvath 2015).

RIG-I senses ligands from negative-strand ssRNA viruses, like influenza A virus, such as 5'-triphosphate-bearing panhandle structures, and from dsRNA viruses such as reovirus, through their 5'-diphosphate pattern (Pichlmair et al. 2006; Goubau et al. 2014). These structures are not found in the host cells, therefore allowing easy discrimination from the host's own RNAs. RIG-I also recognizes 5'-hydroxyl and 3'-monophosphoryl short RNA molecules generated by the OAS/RNase L system, though these can be from both cellular or viral origin (Malathi et al. 2007; Malathi et al. 2010; Malathi et al. 2014). Recent evidence suggests co-sensors may play a role in enhancing RIG-I antiviral response, such as DDX6 (Núñez et al. 2018).

MDA5 binds longer RNAs with high molecular weights, such as synthesized poly I:C and the long replicative form dsRNA of picornaviruses (Pichlmair et al. 2009; Feng et al. 2014). However, its weak RNA-binding activity appears to be rather aspecific, and there is no evidence that specific terminal groups are required for MDA5 ligands.

In the absence of their ligands, RIG-I and MDA5 present sequestered CARD domains to prevent signaling. Linking to their ligand induces a conformation change allowing for the release of their CARD domains, which become available for ubiquitination by

K63 polyubiquitin chains. This binding is essential for RIG-I and MDA5 activation and their signaling pathway down to the IRF3 transcription factor, which in turn triggers the induction of type I interferons and numerous pro-inflammary cytokines (Wu and Chen 2014).

3.1.3 DNA sensors

The role of DNA sensors has been long underestimated, owing to their less extent characterization. DAI (DNA-dependent Activator of IFN-regulatory factors) was the first cytosolic DNA sensor inducting the expression of type I interferon identified (Takaoka et al. 2007), and the major adaptor STING (STimulator of INterferon Genes) was afterwards identified as acting upstream of TBK1 kinase (Ishikawa et al. 2009; Jin et al. 2008; Zhong et al. 2008; Sun et al. 2009) and IRFs, notably IRF3 and IRF7.

Following this discovery, multiple other DNA sensors dependent on STING were characterized, amongst which IFI16 (Unterholzner et al. 2010), DDX41 (Zhang et al. 2011) and cGAS (Sun et al. 2013; Wu et al. 2013). The exact role and importance of such sensors in DNA PAMPs recognition is still unclear however.

IFI16 (InterFeron Induced protein 16) has notably been identified as a DNA sensor able to trigger a STING-TBK1 mediated Interferon response in both the cytoplasm and nucleus of infected cells (Unterholzner et al. 2010; Kerur et al. 2011; Li et al. 2012; Orzalli et al. 2012; Diner et al. 2016). It belongs to the PYHIN gene family and is classified as an ALR (AIM2-Like Receptors), and present two important domains: the Pyrin domain, which provides it with protein-protein interaction abilities, and HIN (Hematopoietic Interferon inducible Nuclear factor) domains at the C-term to interact with DNA. IFI16 binds viral DNA independently of sequence motifs, but through the presence of secondary structures, including products from reverse transcription of RNA viral genomes such as HIV's (Jakobsen et al. 2013; Hotter et al. 2019). On top of its ability to activate the interferon response through STING, IFI16 also has a role in the inflammasome pathway and caspase 1 activation. Infection by herpes viruses leads to activation of caspase 1 but rarely results in cell death (Ansari et al. 2015; Merkl et al. 2018). IFI16 might also provide an antiretroviral checkpoint against Human Endogenous RetroVirus (HERV) induced carcinogenesis by detecting the ssDNA products of HERV retrotranscription (Hurst et al. 2019).

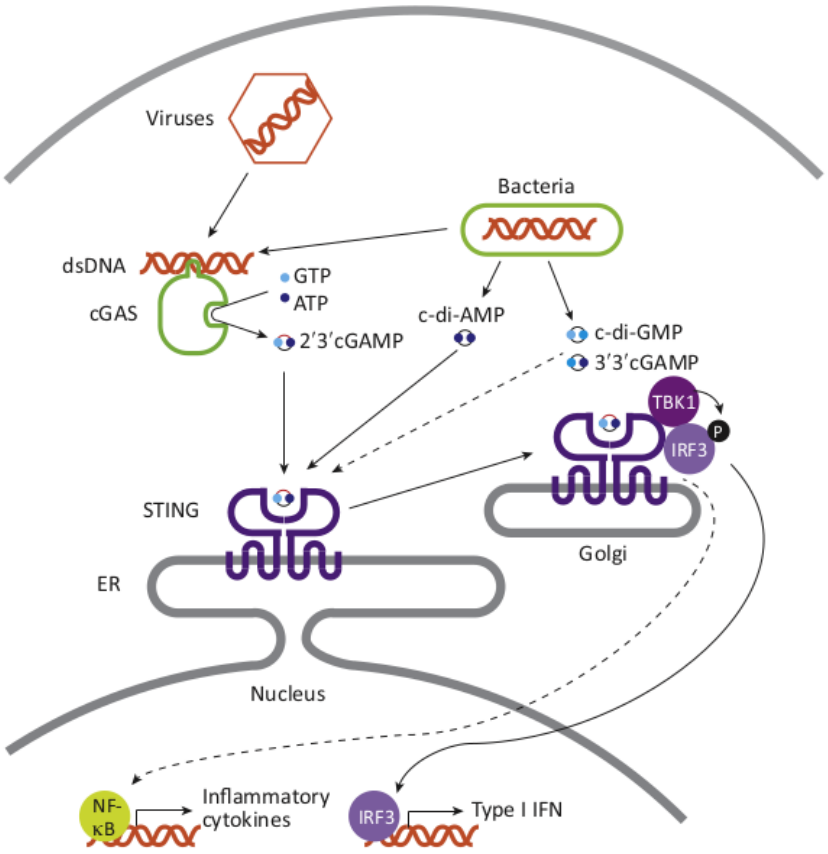


Figure 8: Schematic Representation of the cGAS-STING Pathway in Mammals.

Detection of cytosolic dsDNA by cGAS drives the synthesis of noncanonically linked cyclic dinucleotide 2'3' cGAMP. Its binding activates STING dimers localized in the ER membrane, which translocate to an ER-Golgi intermediate compartment. There, TBK1 phosphorylates the C-terminal tail of STING. This ensure the recruitement of IRF3 for phosphorylation by TBK1, allowing its dimerization, nuclear translocation, and transcription of target genes, including type I interferon. The mechanisms of activation of other downstream signaling pathways (such as NF-kB) are not well understood. Cyclic dinucleotides of bacterial origin can also activate STING; however, they bind to STING with lower affinity than 2'3' cGAMP. From Margolis et al. 2017.

The cytosolic DNA sensor cyclic GMP-AMP synthase cGAS is activated by sensing B-type dsDNA with high affinity, and hybrid DNA:RNA and ssDNA with low affinity (reviewed in Wan et al. 2020). These ligands are not usually found in the cytoplasm of cells, notably dsDNA, which is usually circumscribed to the nucleus or the mitochondria. Its unwarranted presence can either be due to exogenous input, such as the internalization of pathogens, or endogenous, through the inaccurate segregation and release of genomic DNA for example. Binding of cGas to its ligands leads to second messenger cGAMP synthesis. cGAMP has a very high affinity for STING, which dimerization it promotes. It can also be transferred to other cells through diverse mechanisms to pass on the danger signal and pre-notify cells (through gap junctions or by being packaged

into virions, for example). Once the STING dimer has been activated, it is translocated from the Endoplasmic Reticulum Golgi intermediate compartment (ERGIC) then to the Golgi, where it binds and is phosphorylated by TANK binding kinase 1 (TBK1). The phosphorylation of STING by activated TBK1 enables the recruitment of interferon regulatory factor 3 (IRF3). TBK1 then phosphorylates IRF3, leading to its dimerization and subsequent translocation to the nucleus where it will induce the expression of Type I Interferon by binding to ISREs. The cGas-STING pathway is activated by a broad spectrum of pathogens, amongst which various viruses. They include dsDNA viruses like HSV and papillomaviruses, but cGAS can also detect the intermediate dsDNA products of retroviruses like Murine Leukemia Virus (MLV) and HIV (also reviewed in Motwani et al. 2019 and in depth for HIV in Yin et al. 2020). Some of those viruses, notably HIV, may be able to use some host proteins in order to evade cGAS sensing (Yin et al. 2020). This ability of cGAS to detect dsDNA of viral origin, and the ability of some viruses to evade it, may have driven a genetic conflict that can be observed in primates, as cGAS appears to have undergone rapid evolution in this lineage (Margolis et al. 2017).

3.2 The interferon family and signaling pathway

3.2.1 Different families of interferon

Interferons are cytokines which were identified as major role-players in immunity through their ability to inhibit the replication of influenza virus (Isaacs and Lindenmann 1987). Three families of interferons, I to III, were afterwards characterized, with IFN-I and IFN-III shown to be involved in the innate immune response against viruses, while IFN-II is associated with adaptive immunity (reviewed in Negishi et al. 2018; Mesev et al. 2019; Lazear et al. 2019). They all share an α-helical bundle structure and present broad similarities in their mode of action.

The IFN-I family is declined in 4 subtypes: α, β, ϵ and ω. The two best described subptypes are IFNα, which has 13 isoforms, and the single IFNβ. IFNβ can be secreted by most cellular types, while IFNα is mostly produced by plasmacytoid Dendritic Cells (pDCs) (Siegal et al. 1999) as its expression requires IRF7 which is only expressed in pDCs.

The IFN-II family is constituted by the single IFNγ and its production is largely

restricted to hematopoietic cells.

The IFN-III family is comprised of four subtypes of IFNλ (1 to 4) (Egli et al. 2014) and are thought to act predominantly at barrier surfaces, notably in epithelial cells, subsets of myeloid cells, and certain neuronal cells.

The IFN-I family, being the most broadly expressed and the largest, represents the main driver of antiviral innate immunity through its capacity to induce protective cellular states. Indeed, the canonical components of the IFN-I signaling pathway are widely expressed: thus, most cells are competent to mount IFN-I-dependent responses against viral infection.

3.2.2 Interferon I signaling pathway

Signalisation through the PRRs will induce production of IFNβ and small quantities of IFNα in the majority of cells. The signal of IFNβ will then amplify the production of IFNα , as both IFNs use the same signaling pathway (reviewed in Ivashkiv and Donlin 2014; Schneider et al. 2014).

Type I IFN links to specific heterodimeric transmembrane receptors called IFNAR (IFNα Receptor), composed of subunits IFNAR1 and IFNAR2, which are constitutively expressed by most cells. IFNAR engagement then activates JAK1 (JAnus Kinase 1) and TYK2 (TYrosine Kinase 2) by cross-phosphorylation. They will then recruit and phosphorylate proteins STAT1 and STAT2 (Signal Transducer and Activator of Transcription), causing their dimerization and nuclear translocation, where they will bind IRF9 to form the trimolecular ISGF3 complex (IFN-Stimulated Gene Factor 3). ISGF3 then recognizes and binds its response element on the cell genome, called ISREs (Interferon Stimulated Response Elements), thereby directly activating the transcription of the genes presenting these sequences in their promoter region. Alternatively, phosphorylated STAT1 can homodimerize and bind other response elements called Gamma-Activated Sequences (GAS).

Thousands of genes have these response elements and thus see their expression augments when IFN-I is produced: they are called Interferon Stimulated Genes (ISGs) (Rusinova et al. 2013). The expression of many of those ISGs induces an antiviral cellular state, and the patterns of their expression are both cell type-dependent and context-dependent, allowing for modulation of the immune response to viral infection.

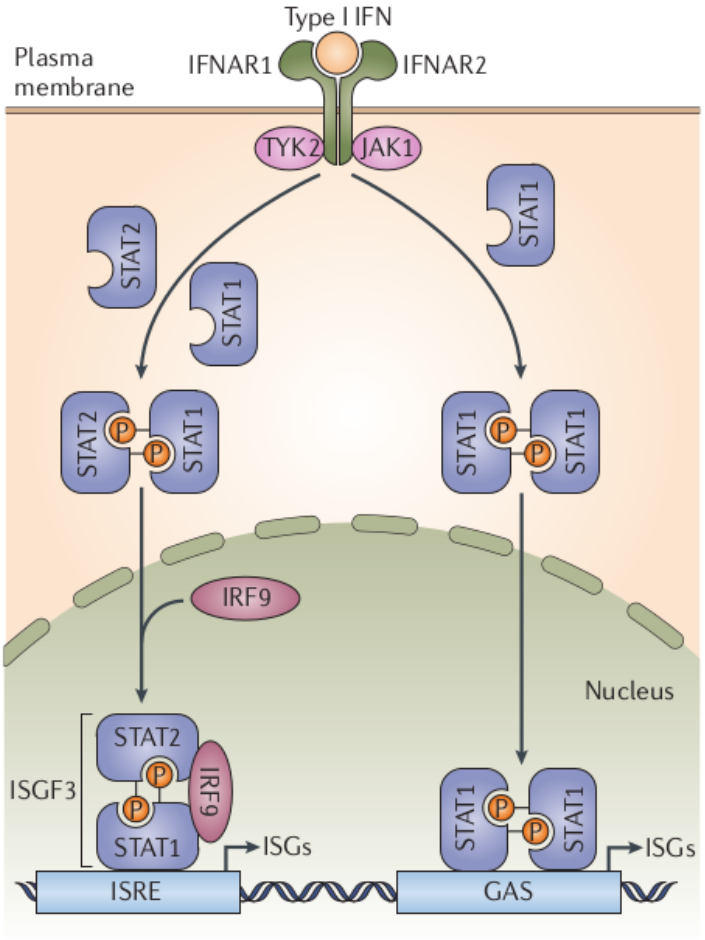


Figure 9: Type I Interferon signaling pathway.

Interferon receptors (IFNAR, composed of subunits IFNAR1 and IFNAR2) recognize and bind type I interferon, leading to the activation of kinases TYK2 and JAK1. Phosphorylation (P) by those kinases of STAT1 allows its homodimerization or heterodimerization with STAT2. Once in the nucleus, the heterodimers bind IRF9 to form the ISGF3 complex, which engages ISREs, whereas homodimers engage GASs. This binding activates transcription of IFN-stimulated genes (ISGs). From Doyle et al. 2015.

3.2.3 Ending the interferon response

The interferon response induces an immune reaction that is stressful to the cells and the organism as a whole. Failing to properly regulate this pathway can lead to numerous diseases and deleterious consequences, such as in the case of autoimmune disorders (Psarras et al. 2017). It is thus of paramount importance that the activation of the IFN pathway remains transitory.

Diverse mechanisms happen concurrently to control the IFN response: from negative regulation of IFNAR expression at the cell surface to limit cell responsiveness, induction of negative regulators, themselves ISGs, and the induction of miRNAs (Ivashkiv and Donlin 2014).

Negative regulators include SOCS (Supressor Of Cytokine Signalling) proteins 1 and

3, which expression is stimulated by type I IFN and which competes with STAT proteins for binding IFNAR, thus suppressing the activity of JAK. Another negative regulator, USP18 (ubiquitin specific peptidase 18), displaces JAK1 away from the receptor. This causes a negative feedback loop to limit both the extent and duration of the interferon response.

MiRNAs control the expression of most of those proteins, adding another level of control: for example, miR155 has been shown to supress expression of various components of the interferon pathway (Hsin et al. 2018; Mahesh and Biswas 2019).

3.3 Interferon Stimulated Genes and their antiviral activities

Type I interferon regulates the expression of a broad array of genes. These genes in turn play multiple roles in regulating interferon signaling and activating the interferon mediated immune response, including an antiviral capacity. These antiviral responses can be varied and targeted against specific viruses or a broad category of them. Table 2 highlights some of those genes, their antiviral mechanism(s), the virus(es) they target and how those viruses might antagonize or escape the antiviral ISG's action. These examples are limited to those ISGs that I later used to validate the tool I developed, and do not aim to be a complete review of antiviral ISGs. Detailed descriptions of some of those genes will be further explored in Introduction – Chapter 4.

However, some ISGs might also facilitate viral infection. For example, TREX1, is involved in the negative regulation of innate immune responses triggered by DNA sensing: its expression mediates the termination of the Interferon response by degrading the immune-stimulating DNA molecules present in the cytosol, so they cannot trigger immune and autoimmune responses mediated by other DNA sensors such as cGAS (Stetson et al. 2008; Ablasser et al. 2014).This in turn positively impacts the infectivity of HIV-1 (Kumar et al. 2018).

Table 2: ISGs involved in antiviral responses

ISG	Full name	Some of the known mechanisms	Examples of viruses targeted	Examples of antagonism and/or evasion mechanisms	References
APOBEC3 family	Apolipoprotein B mRNA editing enzyme catalytic polypeptide-like	Hypermutation by deamination, block reverse transcription	Retro-, Papilloma-, Herpes-, Hepadnaviruses	Viral antagonists: Vif (lentiviruses), Bet (spumaviruses), Gag (gammaretroviruses)	Nakano et al. 2017
GBP5	Guanylate binding protein 5	Suppress the activity of thhe virus-dependency factor furin	Human Immunodeficiency -, Zika -, Measles -, Influenza A virus	Mutation of Vpu, alternative maturation process of glycoproteins	Braun et al. 2019
HERC5	HECT And RLD Domain Containing E3 Ubiquitin Protein Ligase 5	Mediates ISGylation through ISG15 to inhibit retroviral assembly and reduce infectious virus yield for other virusesI Gag	Human Immunodeficiency -, Murine Leukemia -, Human Papilloma-, Marburg -, Influenze A virus	NS1 (?) (Influenza A virus)	Paparisto et al. 2018
IFI16	Interferon Gamma Inducible Protein 16	Viral DNA sensing and activation of innate immune response	DNA -, Retro-, Human Endogenous Retroviruses		Hurst et al. 2019
ISG20	Interferon Stimulated Gene 20	Modulates exo/endogenous mRNA translation differentially, degrades viral RNA	Flavi-, Picorna-, Rhabdo-, Toga-, Orthomyxo-, Retro-, Bunyavirus, Hepadna, Bunyaviruses	Unknown	Espert et al. 2003; Wu et al. 2019
MX1	MyXovirus Dynamin Like GTPase 1	Inhibits vRNP nuclear import, sequesters newly synthesized viral N protein into perinuclear complexes	Influenza A -, Vesicular Stomatis -, Thogoto-, Bunya-, Monkey pox -, African swine fever virus	Nucleoprotein mutations (pandemic IAV)	Mitchell et al. 2013
NT5C3A	5'-Nucleotidase, Cytosolic IIIA	Inhibits reverse transcription	Influenza A -, Classical swine fever virus	Unknown	Wang et al. 2016
RSAD2 (viperin)	Radical S-Adenosyl Methionine Domain Containing 2	Multiple independent mechanisms, including disrupting lipid rafts and viral egress and terminating RNA replication	Human Immunodeficiency -, Human Cytomegalo-, Bunyamwera -, Hepatitis C -, Influenza A virus, Flaviviruses	Unknown	Gizzi et al. 2018
SAMHD1	SAM And HD Domain Containing Deoxynucleoside Triphosphate Triphosphohydrolase 1	Hydrolizes cellular dNTPs and degrades viral RNA	Retro-, Arteri-, Pox-, Herpesviruses	Vpx (HIV-2, some SIV), Vpr (some SIV)	Miyakawa et al. 2019
ZC3HAV1 (ZAP)	Zinc Finger CCCH-Type Containing, Antiviral 1	Recruits RNA exosome complex to degrade viral RNA, may interfere with translation initiation	Retro-, Filo-, Toga-, Alpha-, Hepadnaviruses	Unknown	Luo et al. 2020

Chapter 4

Evolutionary history of primate lentiviruses and the Human Immunodeficiency Virus

The International Committee on Taxonomy of Viruses (Lefkowitz et al. 2018) listed about 5000 viruses, distributed over 143 families, as of 2019 (Siddell et al. 2019). However, this classification mostly includes manually curated taxonomic entries, and metagenomics studies have revealed an abundance and complexity of the virosphere that much exceeds what is presently included by the ICTV. This highlights the complexity of the virus world. The following sections will only detail the impact of lentiviruses on the adaptation of primates, which was the main focus of my work.

4.1 The *retroviridae* family and lentiviruses

The *retroviridae* is one of those 143 families, and is itself composed of two subfamilies: *spumaretrovirinae* and *orthoretrovirinae*. All members of this family are enveloped viruses with an RNA genome, and the name stems from the ability of those viruses to retrotranscribe their genetic material into single-strand (ss) DNA. This intermediate is necessary for the formation of the viral double-strand (ds)DNA form that will integrate into the host's genome. The virus will then replicate at the same time as the cell's genome. To this end, retroviruses encode two specific enzymes: the reverse transcriptase and the integrase.

The *spumaretrovirinae* are commonly referred as foamy viruses, owing to the foamy

aspect large vacuoles give to infected cells. They are highly prevalent in diverse non-primate mammalian and primate species, though often latent, and regroups five genera (Khan et al. 2018).

The *orthoretrovirinae* subfamily is composed of six genera: alpha-retroviruses, beta-retroviruses, delta-retroviruses, epsilon-retroviruses, gamma-retroviruses and lentiviruses. This last genus owes its name to the late onset of the associated pathologies, preceded by a long latent phase, and includes Human Immunodeficiency Viruses (HIV) and Simian Immunodeficiency Viruses (SIV).

Lentiviruses also infect others mammals: horses, small ruminants, bovines and felines (Gifford 2012). However, endogenous lentiviruses were also identified in the lagomorph subgroup (rabbits and hares). Figure 10 highlights the rather extensive range and diversity of this genus, as well as some of their genomic differences and similarities. Most of the differences concern accessory proteins, which will be discussed in further details in the next parts of this introduction. The overall timescale of the lentivirus genus evolution is estimated at a minimum of 12 million years (Gifford 2012).

4.2 Primates lentiviruses: Simian Immunodeficiency Viruses

Non-human primates (NHP) are an important reservoir for lentiviruses. A large number of species are naturally infected by SIVs. For example, out of 73 referenced species of African monkeys, 45 were found serologically SIV-positive (Locatelli and Peeters 2012). However, Asian and New World monkeys do not present evidence of the same (Ayouba et al. 2013). Some SIVs appear to circulate exclusively within the species they are adapted to, as is the case for SIVcol, which infects *Colobus* monkeys. However, numerous cross-species transmission events have been described between species with overlapping habitats or other common points, as highlighted in Figure 11. For example, the SIV from chimpanzees (SIVcpz) results from cross-species transmissions and recombination between SIVs from red-capped mangabeys (SIVrcm) and *Cercopithecus* monkeys (SIVgsn/mon/mus), and an unknown SIV, and then itself crossed the species-barrier to gorillas (SIVgor) (Bailes et al. 2003; Etienne et al. 2013; Bell and Bedford 2017).

The pathogenicity of SIVs is poorly understood, owing in part to the difficulty to

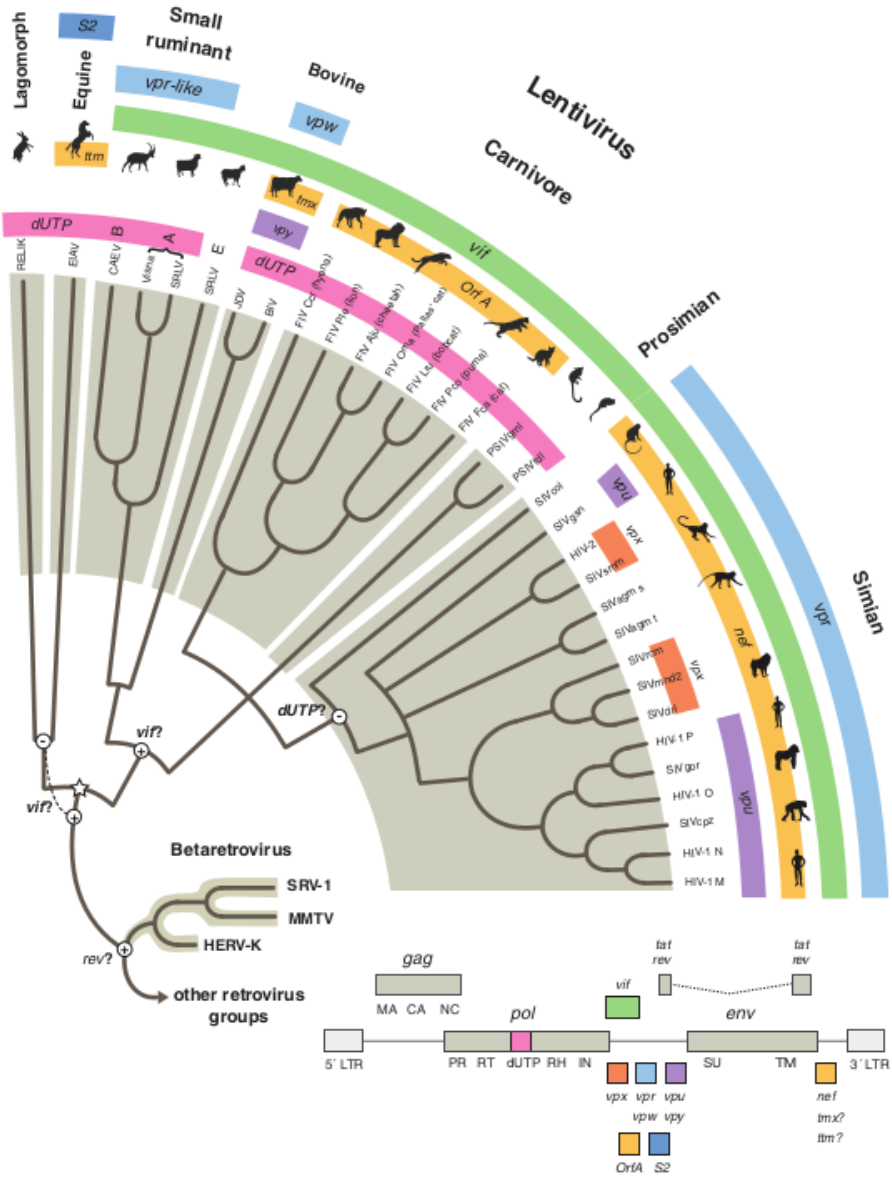


Figure 10: Lentivirus phylogeny, distribution and genetic diversity.

Phylogenetic tree summarizing the genomic organization and evolutionary relationships among known lentiviruses. The star indicates that the root node of the Lentivirus clade is uncertain. The root shown is the phylogenetic midpoint between taxa. A schematic representation of a generalized lentiviral genome is shown at the bottom right of the figure. Lentiviruses encode the fundamental retroviral genes gag, pol, and env, and various accessory genes. The presence of these accessory genes, or genes assumed to be homologs, in various lentiviruses is indicated by corresponding bars along the tips of the tree. Not shown are the lentiviral regulatory genes, rev and tat. White circles on branches show parsimonious interpretations of gene gain (+) and/or loss (–), with two alternatives being shown for vif. Virus names: BIV, bovine immunodeficiency virus; EIAV, equine infectious anemia virus; FIV, feline immunodeficiency virus; HERV-K, human endogenous retrovirus K; MMTV, mouse mammary tumor virus; PSIV, prosimian immunodeficiency virus; RELIK, rabbit endogenous lentivirus K; SRLV, small ruminant lentivirus; SRV-1, simian retrovirus type-1. Genes: dUTP, deoxyuridine triphosphatase; nef, negative factor; rev, regulator of virion expression; tat, trans-activator of transcription; vif, viral infectivity factor. CA, capsid; IN, integrase; MA, matrix; NC, nucleocapsid; PR, protease; RH, RnaseH; RT, reverse transcriptase; SU, surface glycoprotein; TM, transmembrane. From Gifford 2012.

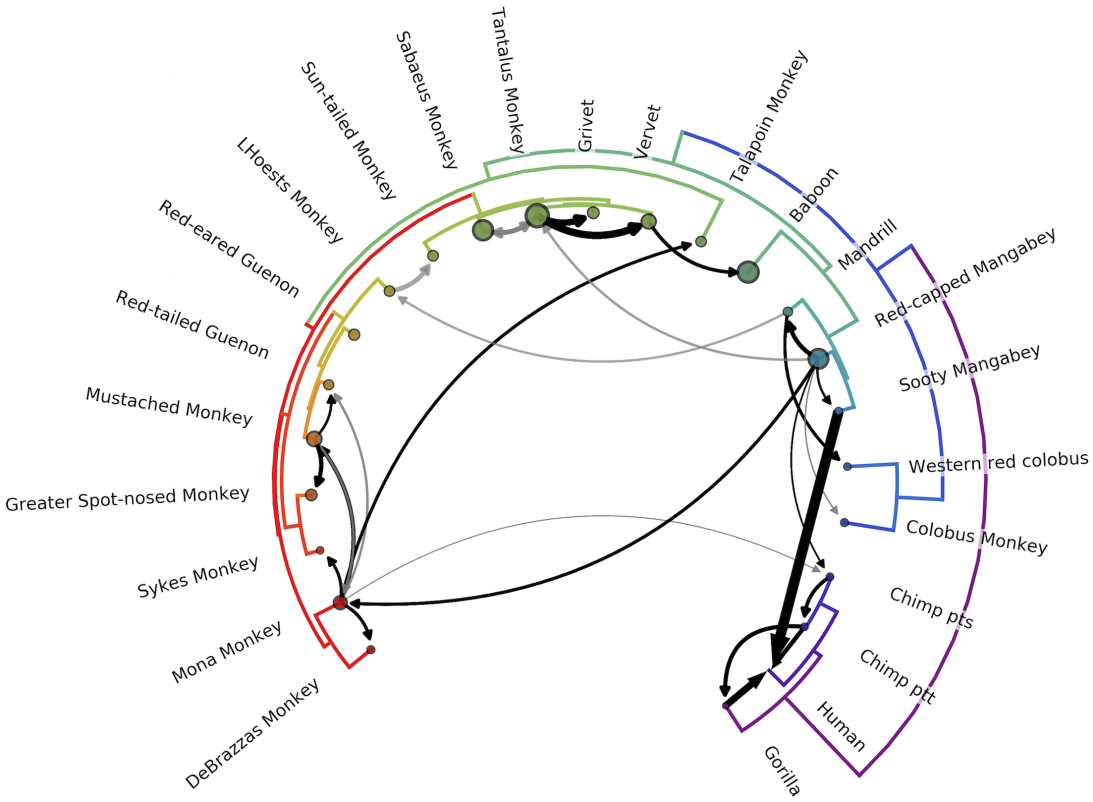


Figure 11: Network of inferred cross species transmissions of primate lentiviruses.

The phylogeny of the host species' mitochondrial genomes forms the outer circle. Cross-species transmission events are represent by arrows between donor and receiver. Width of the arrow indicates the rate of transmission (actual rates = rates indicators). Circle sizes represent network centrality scores for each host. From Bell and Bedford 2017.

assess for the circulation and impact of those viruses in wild primate populations. SIV infections are generally considered as harmless (Pandrea et al. 2008), due to individuals rarely progressing to acquire immunodeficiency. Indeed, the infected species may have gone through long term coadaptation, over millions of years, with their species-specific SIVs. This results in slow progression to immunodeficiency, and deleterious symptoms do not appear before the natural death of the infected individual (Pandrea and Apetrei 2010; Klatt et al. 2012). However, this is not true for all species, as SIVcpz does appear to be pathogenic (Keele et al. 2009; Etienne et al. 2011), though this might be due to it originating from multiple and recent cross-transmission events (Sharp and Hahn 2011). As such, the absence of a co-evolving relationship between chimpanzee and the SIV strain translate to an absence of adaptive mechanisms, leading cross-transmission events to result in increased virulence (Mandell et al. 2014). This highlights the impact that long term interactions with a virus can have on the evolutionary history of its host.

4.3 The evolutionary origin of the Human Immunodeficiency Virus

The emergence of modern HIV in the human population originated from very recent cross-species transmission events from different SIVs, giving rise to both HIV-1 and HIV-2 (reviewed recently in Sauter and Kirchhoff 2019). Those events were probably linked to factors favoring cross-transmissions and spread into the human population: contacts with domesticated monkeys, hunting and/or consumption of bushmeat. Moreover, the spread of these viruses have been highly facilitated by its transmission mode (sexual intercourse, contaminated blood, sharing contaminated needles, etc) (Worobey et al. 2008).

HIV-1 is subdivided in 4 groups: M (Major), O (Outlier), N (non-M, non-O) and P. Group M is responsible for the current pandemic and has, by far, the largest spread, with a subdivision in clades going from A to K and Circulating Recombinant Forms (CRFs). Other groups of HIV-1 are circumscribed to regions of Western Africa. Group N is circumscribed to isolated cases. The spread of HIV-2 has also been limited to a much smaller scale than HIV-1 group M, and cases mostly concentrate in Western Africa (Sharp and Hahn 2011).

Four independent cross-species transmission events gave rise to the different groups of HIV-1. Groups M and N were identified as phylogenetically closer to SIVcpz than any other virus (Keele 2006). Groups O and P originated from a SIV from Western lowland gorillas (SIVgor) with which they cluster phylogenetically (D'arc et al. 2015), and SIVgor itself originated from a cross-species transmission of SIVcpz (Takehisa et al. 2009). As for HIV-2, at least nine independent transmission events from SIVsmm (sooty mangabey) gave rise to groups A to I (Ayouba et al. 2013). These transmission events are recapitulated in Figure 12.

4.4 Epidemiology of AIDS

The first cases of Acquired immunodeficiency syndrome (AIDS) were observed in the early 80s and the identification of HIV-1 as the causal agent followed rapidly (Barre-Sinoussi et al. 1983). The less pathogenic HIV-2 was then discovered in infected patients from Western Africa (Barin et al. 1985).

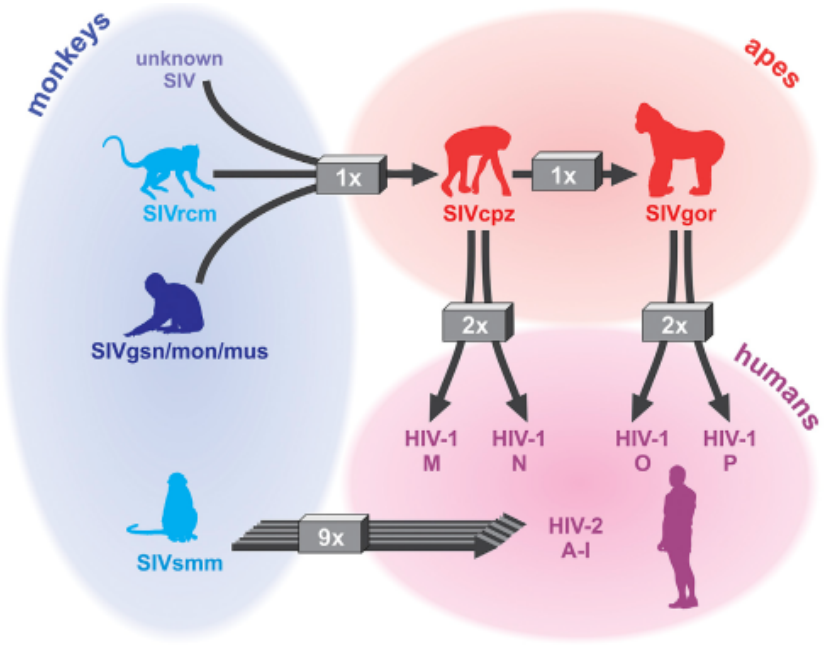


Figure 12: Cross-Species Transmission Events Preceding the Emergence of HIV-1 and HIV-2.

SIVcpz (chimpanzees) arised from cross-transmissions and recombination events between three different SIV strains: an unknown SIV strain, a precursor of today's SIVgsn/mon/mus clade (*Cercopithecus* monkeys) and possibly a precursor of today's SIVrcm from red-capped mangabeys. SIVcpz was subsequently transmitted to gorillas and humans, giving rise to SIVgor and HIV-1 groups M and N, respectively. HIV-1 groups O and P are the result of two zoonotic transmission events of SIVgor, while SIVsmm (sooty mangabeys) was transmitted to humans on at least nine occasions, resulting in the emergence of HIV-2 groups A through I. From Sauter and Kirchhoff 2019.

HIV-1 is responsible for a pandemic that affected an estimated number of 37.9 million people globally in 2018, according to UNAIDS. 24.5 millions of those people were under antiretroviral treatment as of June 2019, lowering their viral load to almost undetectable levels and considerably decreasing their risk of transmission. The number of newly infected people every year is steadily decreasing, from 2.8 million in 2000 to 1.7 million in 2018.

The virus is present in body fluids such as blood, semen, pre-seminal, vaginal and rectal fluids and maternal milk. The main modes of transmission thus occur through blood, sexual transmission, and mother-child transmission.

HIV targets cells from the immune systems such as T lymphocytes (Barre-Sinoussi et al. 1983), macrophages (Gartner et al. 1986) and dendritic cells (Biberfeld et al. 1985). The eventual course of the infection destroys those cells, thus slowly weakening the immune system of the host and leaving them defenseless to other infections. This leads to the phase of AIDS.

4.5 Human Immunodeficiency Virus

4.5.1 Genome

All lentiviruses share a common genomic structure, and diverge from one another essentially through what are called accessory genes, as visible in Figure 10.

Their genome is encoded by two molecules of positive-sense ssRNA of around 9 kb, delimitated by Long Terminal Repeats (LTRs) which are only complete during the dsDNA phase. They are subdivised in three regions: U3, R and U5. U3 serves as a promoter for the expression of the fifteen viral genes.

As all retroviruses do, HIV's genome encodes for three structural proteins: Gag, Pol and Env. Gag (group specific antigen) is a polyprotein encoding multiple smaller proteins that are cleaved during the replication cycle of the virus (Figure 10). It is notably composed of nucleocapsid (NC), capsid (CA), and matrice (MA). They respectively associate with the RNA, protect the core, and form a mesh under the viral membrane. Gag also encodes for peptides Sp1, Sp2 and p6. Env is the envelope protein, encoding for gp41 or TM, the transmembrane protein, and gp120 ou SU, the structural protein, and is inserted within the viral membrane. Pol is cleaved into three viral polymerases: reverse transcriptase (RT), integrase (IN) and protease (PR), indispensable for the proper cleavage of polyproteins (Figure 10).

The genome also encodes for non-structural proteins, called regulatory proteins: Tat and Rev. Finally, the presence of accessory proteins, such as Nef, Vif and Vpr, characterizes complex retroviruses. One of those accessory proteins marks the difference between the genomes of HIV-1 and HIV-2: the first only encodes for vpu, while the second also presents vpx.

These accessory proteins actually play major roles during the viral cycle, as they counteract the action of various host antiviral proteins (reviewed in Faust et al. 2017). For example, the primary function of Vif is to counteract the effect of Apolipoprotein B mRNA editing enzyme catalytic polypeptide-like (APOBEC3) proteins, protecting the viral genome from lethal hypermutations (Etienne et al. 2015; Nakano et al. 2017). This is notably mediated through the ubiquitination and subsequent proteosomal degradation of APOBEC3 proteins. Vpu and Nef mainly target transmembrane proteins hindering the release of the new virions, such as BST-2/Tetherin, and Vpr activates the DNA damage

response. However, all those proteins also possess secondary functions, all geared towards the promotion of an environment favoring infection.

4.5.2 Structure

HIV is an enveloped virus of around 120 nanometers of diameter (Barre-Sinoussi et al. 1983), existing in two different conformations, immature or mature. The immature particle has not undergone cleavage of its polyproteins yet.

The mature particle presents ten to fifteen trimers of envelope proteins gp41 and gp120 at its surface (Zhu et al. 2006). Inside the particle, the cleaved products of Gag reorganize to form a mesh under the membrane (MA), while CA reassociates to form a core by forming hexamers, with the exception of twelve pentamers providing the core with a conical shape (Ganser et al. 1999). This core contains the two viral genomic RNA molecules, covered by the nucleocapsid for protection. The virion also contains the different viral enzymes (PR, RT and IN) and the accessory proteins Vif, Vpr (and Vpx, in the case of HIV-2) and Nef (Figure 13). Those favor the replication of HIV within the host cell as soon as it enters, notably by blocking the activity of restriction factors (Strebel 2013; Sauter and Kirchhoff 2018).

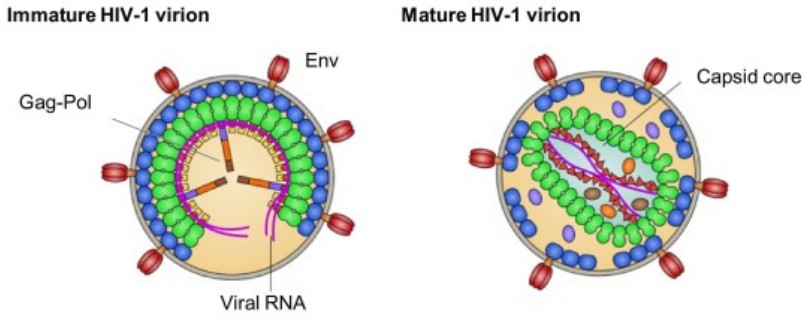


Figure 13: A schematic representation of an HIV virion.

The immature virion present rimers of enveloppe protein (Env) embedded in the viral membrane, and uncleaved Gag-Pol polyproteins. The mature virion is constituted of different elements following the cleavage of Gag-Pol: viral genomic RNA (magenta), matrice MA (blue), capsid CA (green), nucleocapsid NC (red), protease PR (purple), reverse transcriptase RT (orange) and integrase IN (brown). From Novikova et al. 2019.

Viral proteins are not the only ones incorporated in the virions upon budding: host proteins present at the site of budding may also be included, in the viral capsid or within the lipidic membrane. These inclusions may happen through specific and random pro-

cesses, and they have various roles and consequences upon subsequent viral cycles (Ott 2008). For example, restrictions factors such as Interferon-induced transmembrane proteins (IFITMs) may be included at the membrane of the virions and impair their infectivity (Tartour et al. 2014; Compton et al. 2014).

4.6 Viral cycle of HIV

As all viruses, HIV's genome does not encode all proteins necessary for its replication. From the entry step, it usurps the cellular replication machinery until it forms new virions that will then go on to infect new cells: this succession of steps is called the viral cycle (Figure 14).

In retroviruses, the viral cycle can be separated in two main phases. The early phases consist of the attachment of the viral particle to its target cell up to the integration of its retrotranscribed genome within the infected cell's. Late phases start from the transcription of the viral genome to the extracellular maturation of the released particles, i.e. the actual production of new virions (Figure 14).

4.6.1 Early phases

The early phases of the viral cycle of HIV start with the attachment and entry of the virus into the target cell (Figure Chen 2019).

Fusion is followed by the entrance of the viral capsid in the cell, which then loses its integrity to allow access to the viral genome for reverse transcription, becoming the Reverse Transcription Complex/Pre-Integration Complex (RTC/PIC). However, this "uncoating" is the subject of much debate to establish whether it is complete or whether CA remains associated with the RTC/PIC all throughout the early phases (reviewed in Novikova et al. 2019) (Figure 14 step 3).

Lentiviruses infect non dividing cells, contrary to most other retroviruses. As such, they have developed active ways of accessing the nucleus and the cellular genome, and this access is through nuclear pores, nucleoporins NUP, forming a tunnel through the nuclear membrane. Direct interactions between a CA-bearing RTC/PIC with proteins involved in nuclear import such as NUP153, NUP358, TNPO3 and cyclophilin A (CYPA), and

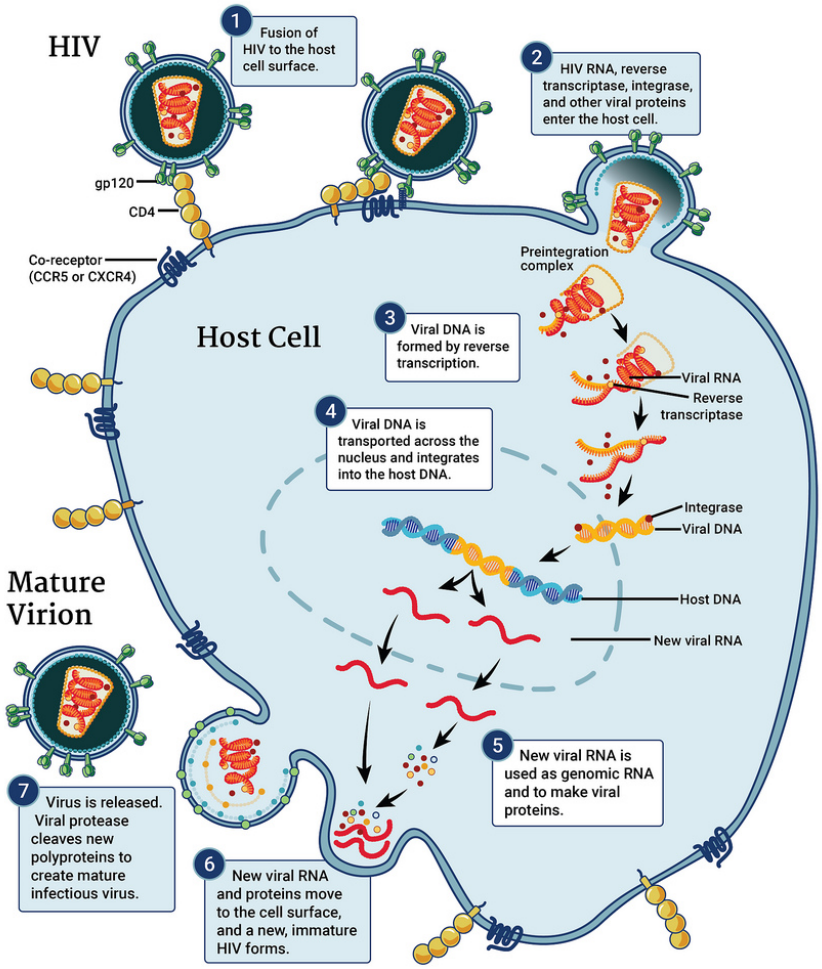


Figure 14: HIV viral cycle.

Complete overview of HIV viral cycle, starting from the recognition of the surface receptor CD4 by its enveloppe protein gp120, up to the release of mature virions in the extracellular environment. Steps 1-4: early phases, steps 5-7: late phases. From the National Institude of Allergy and Infectious Diseases.

the detection of CA in nuclear PICs seem to argue for an incomplete uncoating up to the integration of the viral genome into the host's (Novikova et al. 2019). However, the exact mechanism of transportation of the 50-60 nm CA through the 40 nm nuclear pore remains mysterious and was still actively discussed at the 2020 Cold Spring Harbor Laboratory Conference on Retroviruses (Figure 14 step 4).

CA may well play a role in determining the integration site of the viral genome. Indeed, NUP153, NUP358 and CYPA have been shown to influence the integration site, often favored within the peripheral region of the nucleus (Engelman and Singh 2018). The second main actor of this step is the viral Integrase (IN), which also influences integration targeting and catalyzes the integration reaction itself after forming the intasome

by binding the viral genome (reviewed in Lesbats et al. 2016). Once integrated, the virus is called a provirus and the early phases are finished.

4.6.2 Late phases

The late phases of the viral cycle start with the transcription of the provirus. The proviral genome contains two promoters, a main one and an enhancer. The main promoter, located in region U3 of the 3'LTR, is the one recruiting cellular factors, including RNA polymerase II, to start efficient transcription of the provirus. Proviral transcription generates over 40 transcripts through alternative splicing. The first transcripts out of the nucleus are the ones coding for Tat, Rev, and Nef. Export of the other transcripts from the nucleus is mediated through the Rev protein.

The viral transcripts are then translated by the cellular machinery (Figure 14 step 5). Viral RNAs, like cellular mRNAs, present a 3' poly(A) tail and a 5' cap, which is used to discriminate two mechanisms of translation, cap-dependent or cap-independent (Ohlmann et al. 2014), depending on physiological conditions, infection stage and cellular state (Breyne et al. 2013). Another feature of viral RNA translation is leaky scanning: the ribosome can "jump" to initiate the process at different START codons. This is how the translation of different proteins, such as Vpu and the envelope proteins, are enabled from the same initial RNA. The polyproteic precursors to Gag and Gag-Pol are synthetized from non-spliced genomic RNA (reviewed in Freed 2015).

Gag directs viral assembly through its four subdomains (MA, CA, NC and p6) after its translocation to the main site of assembly at the inner leaflet of the plasma membrane (Freed 2015). Assembly starts through the multimerization of three to four Gag proteins through interactions of their CA domains. This causes the myristoyled basic MA domains to get closer with each other and interact with the phosphatidilinositol-4-5 biphosphates (PIP2) found in specific structures of the plasma membrane, called lipid rafts (Mücksch et al. 2017). The NC domains can interact with genomic RNA through the ψ sequence, thus encapsidating two RNA copies inside the nascent viral particle. These RNA copies interact with each other through their DIS sequences (Dimerization initiation start). Overall, an immature viral particle contains around 5,000 Gag proteins, while the mature ones contain about 1,500 capsid molecules (Briggs et al. 2004).

Meanwhile, the immature envelope protein is translated and glycosylated at the en-

doplasmic reticulum (ER), and then cleaved into subunits gp41 and gp120 at the Golgi apparatus. Its incorporation into the immature virion remains a poorly understood process (Freed 2015).

As this point, the nascent viral particles are found beneath the cellular membrane and the association of proteins are deforming it. The ESCRT (Endosomal Sorting Complexe Required for Transport) cellular machinery is essential for the budding and fission step and is recruited through the late domains of peptide p6 (Bieniasz 2006). Those domains are conserved in other retroviruses, but also in structural proteins of viruses from different families, highlighting the major role of the ESCRT machinery in enveloped virus budding.

Virions acquire their membrane through the budding process. The last step of the viral cycle, maturation, occurs outside of the host cell. The viral protease cleaves Gag to form the different structural proteins, to result in the previously described virion structure (MA at the internal part of the viral membrane, CA to form the viral capsid, and NC associated to the viral genome) (reviewed in Pornillos and Ganser-Pornillos 2019). The produced viral particle is thus competent to infect new cells.

Goals

As exposed in this introduction, HIV interacts with a large number of host proteins, called Viral Interacting Proteins (VIPs), which act to either facilitate (co-factors) or hinder (restriction factors) its replication. Several replication factors have been shown to be under genetic conflict with lentiviruses. In such cases, evolutionary analyses have allowed researchers to identify the hallmarks of this conflict on the host genes encoding for VIPs.

The goal of this study was to reverse the usual approach based on looking for evolutionary hallmarks of a genetic conflict with lentiviruses on genes previously shown to impact HIV replication. We wanted to use the presence of such hallmarks to identify genes with the potential to be antiretroviral effectors, e.g. use evolutionary analyses to select candidate genes for functional characterization.

This approach is not novel and diverse tools have been presented in the Introduction, which aimed to automate the detection of hallmarks of genetic conflict, which can be termed genetic innovations: positive selection, recombination and duplication events. However, they did not satisfy our requirements in the lab, as to the full automation of all steps, starting from the automatic retrieval of homologs in our species of interest to the final detection of the previously cited genetic innovations.

The first goal of my PhD was thus to develop a pipeline for the Detection of Genetic INNovations, called DGINN, and to validate its applicability on genes with known evolutionary histories. This development aimed to entirely automate the workflow routinely used in the lab, based on gold-standard methods, to simplify the analysis of large datasets. We did not seek to establish which software performed the absolute best between different solutions, and thus performed no formal benchmarking, but rather to integrate those which answered our main requirements as to quality and ease of use. The validation of this tool led us to compare our results with manually curated analyses available in the

literature, but also to establish new findings on technical aspects of the pipeline, but also on the evolutionary history of genes already extensively described in the literature. Those findings are exposed in the first chapter of Results.

The development of this tool allowed us to perform evolutionary analyses in a streamlined manner, which was particularly useful during our collaboration with the team of Nicolas Manel at Institut Pasteur, Paris. We performed the evolutionary analyses on both sides of the virus-host interaction they had functionally characterized (second Results chapter).

Once DGINN had been developed, we analyzed datasets of potential VIPs to identify genes of interest in the context of host-lentiviral interactions. These efforts are summarized in the third and fourth chapters of Results, and led to the identification of several targets which are currently undergoing functional characterization in the lab. Such genes are of major interest as their evolutionary history suggest they might have been engaged in long-term genetic conflict with one or several viruses, and may thus have an antiviral role against them. Due to the nature of our datasets, we relied on the hypothesis that lentiviruses could have driven this conflict, and that current lentiviruses such as HIV might be susceptible to restriction by those genes. We thus aimed to characterize the interaction of HIV with our genes of interest.

Results

Chapter 1

DGINN, an automated and highly-flexible pipeline for the Detection of Genetic INNovations on protein-coding genes (full paper)

This paper has been accepted for publication in *Nucleic Acids Research*.

DGINN, an automated and highly-flexible pipeline for the Detection of Genetic INNovations on protein-coding genes

Lea Picard[1,2], Quentin Ganivet[2], Omran Allatif[1], Andrea Cimarelli[1], Laurent Guéguen[2,3,4,*], Lucie Etienne[1,4,*]

[1] CIRI – International Center for Infectiology Research, Inserm U1111, Université Claude Bernard Lyon 1, CNRS UMR5308, Ecole Normale Supérieure de Lyon, Univ Lyon, F-69007, Lyon, France

[2] Laboratoire de Biologie et Biométrie Evolutive, CNRS UMR 5558, Université Claude Bernard Lyon 1, Villeurbanne, France

[3] Swedish Collegium for Advanced Study, Uppsala, Sweden

[4] Contributed equally to this work as senior authors

* Correspondence:

Lucie Etienne, lucie.etienne@ens-lyon.fr

Laurent Guéguen, laurent.gueguen@univ-lyon1.fr

Abstract

Adaptive evolution has shaped major biological processes. Finding the protein-coding genes and the sites that have been subjected to adaptation during evolutionary time is a major endeavor. However, very few methods fully automate the identification of positively selected genes, and widespread sources of genetic innovations as gene duplication and recombination are absent from most pipelines. Here, we developed DGINN, a highly-flexible and public pipeline to Detect Genetic INNovations and adaptive evolution in protein-coding genes. DGINN automates, from a gene's sequence, all steps of the evolutionary analyses necessary to detect the aforementioned innovations, including the search for homologues in databases, assignation of orthology groups, identification of duplication and recombination events, as well as detection of positive selection using five methods to increase precision and ranking of genes when a large panel is analyzed. DGINN was validated on nineteen genes with previously-characterized evolutionary histories in primates, including some engaged in host-pathogen arms-races. Our results confirm and also expand results from the literature, with novel findings on the GBP family. This establishes DGINN as an efficient tool to automatically detect genetic innovations and adaptive evolution in diverse datasets, from the user's gene of interest to a large gene list in any species range.

Running Title: DGINN, an automated pipeline to Detect Genetic INNovations

Key Words: Adaptation, positive selection, duplication, bioinformatics pipeline, recombination, evolution of protein-coding genes, genetic conflict, host-pathogen interaction, primates, HIV, virus

Introduction

Genetic innovation is a major adaptation process that has impacted genome structures and functions over millions of years in response to natural selection. Such changes have shaped key biological functions, such as reproduction, adaptation to a new environment, immunity, sensory-perception, host-pathogen interaction. Adaptation in protein-coding genes can take place through several mechanisms. They include, amongst others, positive selection on coding sequences, duplication events with subsequent divergence of the copies, as well as recombination (1). The first is caused by natural selection that increases the frequency of advantageous mutations, leading to an apparent excess of non-synonymous substitution rates over synonymous ones over evolutionary times. This notably leads to the accumulation of beneficial amino-acid changes at the location of functionally important residues, such as the interface of proteins involved in host-virus interactions. Gene duplication is another important source of genetic novelty, which notably allows to increase the general evolvability (2, 3). The fixation of multiple copies enables diversification of gene function through subfunctionalization or neofunctionalization. Moreover, gene conversion, by recombination between alleles, allows for rapid divergence of the copies. Gene duplication and loss may further be a dynamic and rapid adaptation process (2–4).

These mechanisms fueling genetic novelty are all parts of the response of organisms to selective pressures and must therefore be analyzed as much as possible together to wholly apprehend the evolutionary history of genes. However, despite their frequency and their biological importance and relevance, these diverse evolutionary innovations are not accounted for in most tools and studies analyzing genes under adaptive evolution (5–7). Lastly, performing gold-standard and complete phylogenetic analyses is usually highly hand-curated. Our goal was therefore to design a tool that would incorporate all these mechanisms at the origin of genetic innovation in a robust end-to-end pipeline to identify and characterize new protein-coding genes with signatures of adaptive evolution.

Such a pipeline requires the automation of essential steps. Primarily, searching for homologous gene sequences and identifying orthologous relationships represent a time-consuming and difficult process. No existing tool include these steps, because they either remain essentially hand-curated (Hyphy suite (8), Selecton (9), IDEA (10), JcoDa (11), PoSeiDon (12) and POTION (13)), are restricted to specific vertebrate and prokaryotic species (PhyleasProg (14) and PSP (15)), or rely on published orthologous annotations (essentially from the NCBI HomoloGene) which may become imprecise on non-model species.

Secondly, correct codon alignments are necessary for the accurate detection of residues under positive selection. However, current pipelines rely on protein or nucleotide alignment softwares like ClustalW (16) or Muscle (17), although more recent ones such as PRANK (18) have been repeatedly shown to provide high-quality codon alignments, thereby diminishing false positives during the detection of positive selection (19–22).

Thirdly, we identified the need to include within a single analysis the detection of positive selection signatures by different methods and models, to allow for more specificity and sensitivity of the results, as well as to help "ranking" genes in an evolutionary screening approach (23–28). Moreover, the inclusion of methods in which the experienced user has access to the parameterization of the maximum likelihood models is needed (29). Existing tools rely almost exclusively on PAML codeml (30), which has allowed the

identification of numerous genes under positive selection, but offers limited options for parameterization.

Overall, there seemed to exist a void when it comes to pipelines which fully automate the search for adaptive evolution in protein-coding genes, from retrieving homologous sequences of a gene of interest in any species range, establishing orthologous relationships, reconstructing codon alignments and the corresponding phylogenies, to detecting different genetic innovations using gold-standard and diverse methods to ensure high-degree of confidence in the results. We thus developed an integrative pipeline, that we named DGINN (for Detection of Genetic INNovations) to satisfy those requirements. All scripts are freely available on Github and as a docker on DockerHub. We also focused on user-friendliness and flexibility, so that biologists can use with ease and use only parts of the workflow for various purposes. DGINN was developed as a one-gene workflow and can easily be up-scaled to screen large datasets of dozens or hundreds of genes. Finally, we performed an extensive validation of our pipeline, using published and highly hand-curated phylogenetic data on a set of nineteen primate genes with various evolutionary histories including genes involved in virus-host evolutionary arms-races (1, 31). Through DGINN, we further identified previously uncharacterized signatures of genetic conflict in the primate Guanylate-binding protein (GBP) family, which plays important roles in cell-autonomous immunity against pathogens (32, 33).

Materials and Methods

Pipeline structure

The overall goal of the DGINN pipeline (overviewed in Figure 1) is to provide an easy, integrated, and robust way of detecting genetic innovations from a gene sequence provided by the user on two scales, either on one specific gene for fine-tuned analyses or on large sets of genes of interest for screening purposes.

DGINN is implemented in Python and uses numerous modules, including some from Biopython, as well as several independent softwares. The list of modules and external softwares is provided in the pipeline documentation. All scripts and documentation can be downloaded from Github. To enhance user-friendliness, options are handled through a parameter file, minimizing the complexity of the command line. Importantly, a Docker image is also available for local use without manual installation of the external required softwares. The Docker may also be used to screen large datasets using AWS Batch for example (link). A specific script for the extraction of batch results, parseResults.py, and a graphical interface to produce basic figures with them, have also been developed (see Availability).

The overall workflow of the DGINN pipeline is a succession of eight steps, described hereafter. Of note, DGINN is designed to be extremely flexible as to its uses. The user can enter the workflow at any step with the files resulting from their own analyses, as indicated in Table 1 and Figure 1. The name of the step reflects the very first step performed with the option. For example, starting DGINN at the 'blast' step will make it begin with the blast search, and then execute the whole pipeline. The duplication, recombination and positive selection steps will not be performed if the user has not specifically opted in for them, allowing for maximum flexibility.

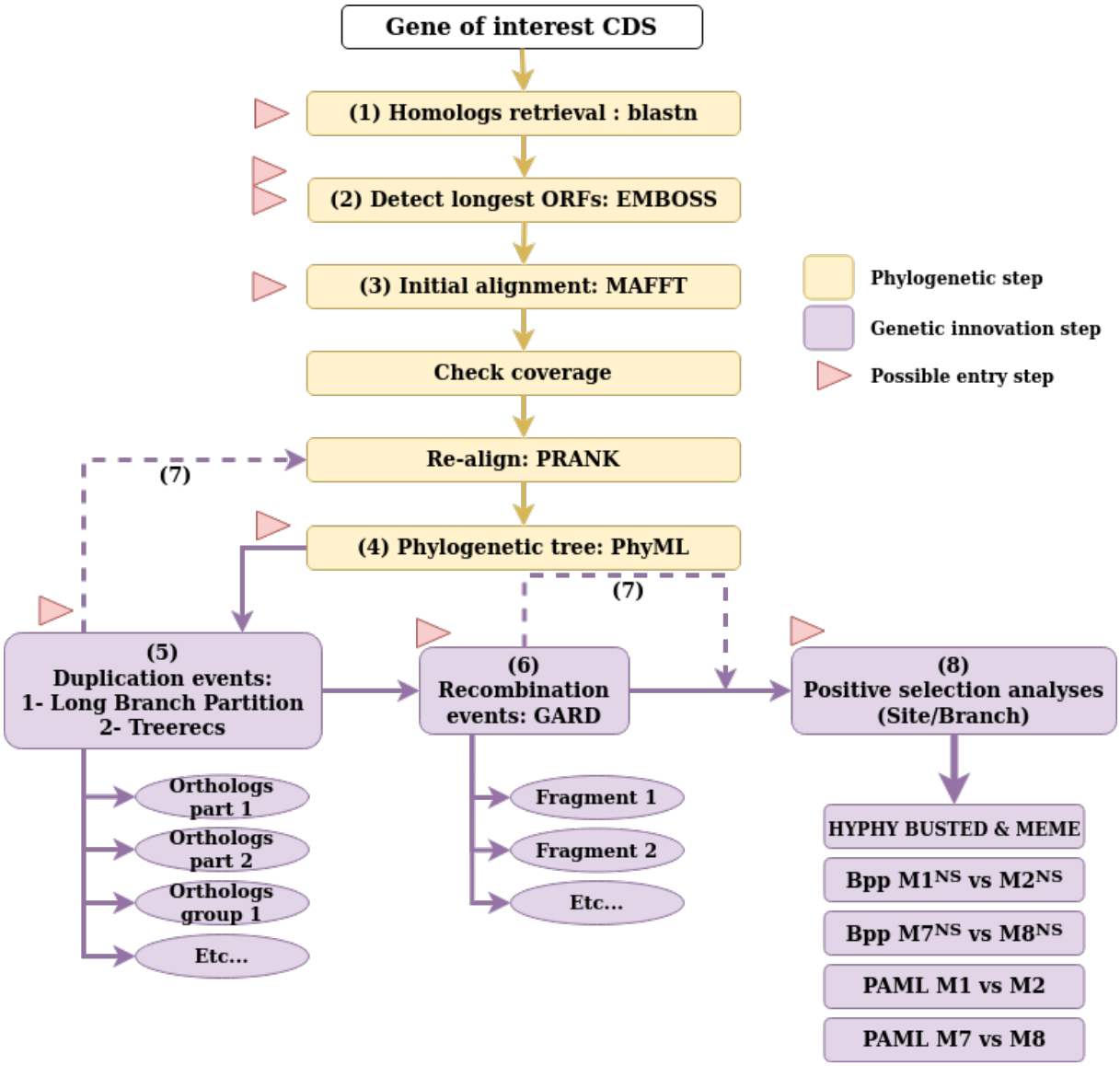


Figure 1: Workflow diagram of DGINN.

Phylogenetic steps (yellow) happen sequentially from the entry point of the pipeline (Steps 1-4). Each genetic innovation step (purple, Step 5, 6, and 7) is optional. All red arrowheads denote possible entry points into the pipeline following file formats from Table 1.

(Step 1) Automated retrieval of homologous genes in species of interest

DGINN uses BLAST+ search (34) against the NCBI databases. The BLAST search can be done against a local database constructed by the user, or online against specific NCBI databases, which allows the user to limit the search to certain sequences, such as ESTs, or certain species, by providing the proper Entry Query, following the syntax used on the NCBI website, as described in their documentation (Entrez Searching Options). BLAST+ is used by providing the coding sequence of the gene of interest against a nucleotide databank (blastn). We decided not to use blastp (protein query against protein database) as it significantly complicated the recuperation of the nucleotide sequences afterwards, which are indispensable to the rest of the pipeline. Moreover, nucleotide databases include more sequences and thus allow for a more exhaustive search. The number and speed of requests against NCBI databases can be increased

Table 1: Overview of the possible entry steps into DGINN.

DGINN can be entered at different steps to enhance flexibility. If the user introduces the name of the proper entry step option and inputs the appropriate files for this option in the parameter file, DGINN will start at that step, ignoring the upstream steps. If the user wants to perform the detection of duplication and orthologous groups, the user has to provide a species tree through the parameter file (see methods and GitHub readme for details).

Step	Name of entry step option	Input files	Format
0	blast	CDS of the gene of interest	Fasta
1	accession	List of blast results	NCBI tabulated format
2	fasta	List of accession identifiers (one per line)	Text file
2	orf	mRNA sequences of homologs	Fasta
3	alignment	CDS sequences of homologs/orthologs	Fasta
4	tree	(codon) alignment of homologs/orthologs	Fasta
5	duplication	(codon) alignment, gene tree	Fasta, newick
6	recombination	(codon) alignment	Fasta
8	positiveSelection	codon alignment, gene tree	Fasta, newick

through the acquisition of an NCBI API key, available online. This ensures access to the largest possible number of sequences, including those not annotated as orthologous or paralogous sequences. The user may modify minimum e-value, coverage, and identity values to reflect the specificities of the database and the species set against which they are using BLAST+. Because we validated our pipeline on primate evolution, we set those with default values of 10-4, 50%, and 70%, respectively, to retrieve a maximum of homologous sequences without too many unrelated sequences.

(Step 2) Elimination of overly long sequences and isolation of Open Reading Frames (ORFs)

Because the user may want to cast a wide net in terms of homologue retrieval, and thus use low coverage and identity for the blastn search (Step 1), a variety of resulting hits are retrieved, including overly long sequences from whole contigs or chromosomes originating from whole genomes where annotations are still an ongoing process. Those sequences considerably increase the analysis time if not properly curated, and the process of automatically detecting in any species the corresponding ORF in a contig is a highly complex task that we did not be include in this pipeline. In DGINN, we identify and remove such sequences based on the median length of all the retrieved sequences: if the median is longer than 10,000 nucleotides, any sequence longer than twice the median are taken out, otherwise sequences are deleted if they exceed three times the median length as the default method. Alternatively, the user can chose to eliminate sequences based on another factor of the median length, or to eliminate outliers based on the InterQuartile Range (IQR) approach. The remaining sequences are searched for ORFs using ORFinder from the EMBOSS package (35) to keep only the coding sequence of each gene. The longest detected ORF of each sequence is selected for further analysis.

(Step 3) Initial codon alignment

Positive selection analyses rely on identifying substitutions leading to amino-acid changes over those being silent. Therefore, a codon alignment of good quality is essential. However, very few softwares propose true codon-alignment modes. To date, the best codon aligners are PRANK (18) and MACSE (36). PRANK has been shown to produce the best alignments for positive selection analyses (19–22, 37).

From our observations, MACSE also produced high-quality codon alignments, but it was significantly slower than PRANK. We therefore selected the latter as the best solution for both quality alignments and lower computational time. To gain rapidity, we first perform a initial nucleotide alignment by MAFFT (38) with automactic settings (mafft –auto, v7.3) after which we added a quality control step to eliminate sequences that did not align properly, using Python homemade scripts, based on alignment coverage against the query (either the user-provided value or default of 50%). PRANK alignments are performed with the codon model and without forcing insertions to be skipped, and otherwise default settings (prank -F -codon; version 150803). In this validation, both the initial and the second alignment were performed using PRANK.

(Step 4) Construction of the initial phylogenetic gene tree

The gene's phylogenetic reconstruction is performed with PhyML v3.2 (39). We opted for a HKY+G+I model as default, because it offers the best combination of realistic phylogenies without being too time-consuming. As the produced trees are only intended for screening purposes at this step, we also opted to use approximate Likelihood Ratio Test (aLRT) for the statistical support of the branches (40). Users can provide their own options for PhyML through the parameter file should they wish to use other models and statistics.

(Step 5) Identification of duplication events and orthologous groups

As previous steps retrieved homologues without relying on synteny or gene annotation, we implemented two strategies to identify duplicated genes and to constitute orthologous groups necessary for the positive selection analyses. DGINN first identifies the overly "long branches" within the gene tree. By default, we define a "long branch" as a branch which length is at least 50 times longer than the mean of all branch lengths in the tree (i.e. the estimated number of substitutions per position is at least 50 times superior in the "long branch" compared to the mean). Alternatively, the user can chose to cut branches based on another factor of the overall mean of the substitution rate, or to eliminate outliers based on the InterQuartile Range (IQR) approach. When "long branches" are identified, the tree is cut along those "long branches" and the groups of sequences subsequently constituted are re-aligned (back to step 3) and their trees recomputed separately (step 4). This constitutes a first method of separating highly divergent groups of genes, between which detection of positive selection may be ambiguous because of suspicion of paralogy and branch length saturation. However, for multigenic families that include paralogues that have recently diverged, the gene members cannot be separated solely based on the relative lengths of the tree branches. We therefore included a phylogenetic reconciliation method, TreeRecs (41), to identify genes sharing a common evolutionary history in our species of interest. To identify duplication events, TreeRecs reconciles a user-provided species tree or cladogram to each gene tree. From the reconciled tree, DGINN establishes groups of orthologues based on ancestral duplication events annotated on the reconciled tree. Since interspecific positive selection analyses rely on the comparison of several orthologous sequences, orthologous groups resulting from very recent duplications may have too few sequences to be informative for those analyses. So we chose to ignore duplication events that were not ancestral enough, by taking

into account the minimal number of species represented downstream of the event. This number is user-determined. We decided on a default setting of a minimum of eight species to extract a duplication group from the original alignment, based on the results obtained by Anisimova et al. (42), and in primates specifically by McBee et al. (27). Duplication events on nodes that do not have at least two species in common in the groups formed on either side of the node are considered dubious: the corresponding annotated events are then ignored by DGINN. After extraction based on ancestral duplication events, the orthologous groups are realigned using PRANK as in Step 3.

To run this step, the user has to provide, through the parameter file, a valid species tree (cladogram) of the species of interest. The format is a newick file with the species names following DGINN's nomenclature (speSpe). If this file is absent, DGINN will not separate the sequences into orthologous groups.

(Step 6) Identification of recombination events and splitting of alignments along the significant breakpoints

To account for recombination, DGINN includes GARD from HYPHY (43) with standard parameters. The breakpoints are then assessed for statistical significance using a likelihood ratio test (LRT) with $p < 0.05$ against a null hypothesis that there is no breakpoint at that position. If any breakpoint is found significant, the sequence alignment is then cut longitudinally at the breakpoint(s) to produce non-recombinant sequence alignments (preserving the codon units). These non-recombinant codon cut alignments, as well as the original one, will become the input in the following steps (and named fragPos1-Pos2).

(Step 7) Construction of the final phylogenetic trees

Following the analyses of duplication and recombination events (steps 5-6), new codon-wise alignments using PRANK (same parameters as in step 3) and new phylogenies using PhyML (same parameters as in step 4) are built for groups of non-recombinant fragments (see step 6) of orthologous genes (see step 5). These final codon alignments and gene trees will further provide the input for the positive selection analyses.

(Step 8) Positive selection analyses

Numerous softwares exist to identify positive selection on coding sequences. DGINN includes several methods of positive selection analyses, which the user can chose to turn on or off independently. Those analyses make extensive use of three packages: HYPHY (8), PAML codeml (30) through the ETE toolkit (link), and Bio++ (44).

From the HYPHY package, we included two methods. First, we included BUSTED (Branch-Site Unrestricted Statistical Test for Episodic Diversification), a random effect model which allows for gene-wide detection of episodic positive selection (45). Results are considered positive in the DGINN pipeline for a p-value < 0.05 for the LRT of the models admitting vs not admitting positive selection. Second, we included MEME (Mixed Effects Model of Evolution), which detects individual sites subjected to episodic positive selection based on a mixed effects model (46). These models are complementary, as BUSTED

evaluates positive selection at the gene level and MEME at the site level. Contrary to BUSTED and MEME, the codon substitution models used in PAML codeml focus on pervasive positive selection and not episodic events. Briefly, the codon alignments are fitted to models that do not allow for positive selection, M1 (with two classes $\omega < 1$ and $\omega = 1$) or M7 (where the $\omega < 1$ class is modeled as a gamma law of n classes, n=5 as default in DGINN), and the corresponding models allowing for positive selection with one class of $\omega > 1$ (M2 or M8, respectively). Statistical significance of positive selection is determined through a chi-squared test of the LRT of both associated models (M1 vs M2, and M7 vs M8) to derive p-values. Results are considered positive in the DGINN pipeline for a p-value < 0.05.

However, PAML codeml relies on the assumption of stationarity (i.e. that the base composition of sequences is at the equilibrium of the evolutionary process), which may impact the detection of selection (47). It is also limited with regards to its parameterization. Therefore, we also integrated the parameterizable Bio++ library to propose similar models but without stationarity assumption (Bio++ models $M1^{NS}$ vs $M2^{NS}$, and $M7^{NS}$ vs $M8^{NS}$). Similarly, DGINN considers significant positive selection if p-value < 0.05 of each model comparison.

If positive selection is determined with PAML or Bio++, the pipeline will proceed to the identification of the sites under positive selection, using the Bayes Empirical Bayes statistics (BEB) from the M2 and M8 in PAML codeml and the Bayesian Posterior Probabilities (PP) from the $M2^{NS}$ and $M8^{NS}$ models in Bio++. Sites are considered as under significant positive selection if BEB or PP > 0.95.

To detect specific branches/lineages under positive selection, DGINN uses Bio++ to include a method similar to the Free-Ratio test available in PAML codeml, called One Per Branch in DGINN (OPB). The ω ratio is calculated along the branches of the phylogenetic tree by using a M0 model where all parameters but ω are homogeneous. As this step is independent and the Bio++ parameter file is fully accessible, an experienced user can choose any model they wish, allowing for maximum flexibility.

Each of those methods can be opted in or out through the parameter file, so that users can run any subset they want.

Pipeline parallelization

DGINN has been developed with the intention to analyze each gene independently, with parallelization over large datasets being handled in a cluster environment. This is done through user-made scripts (such as job arrays) and facilitated through configuration parameters that are specific to this use. -i/–infile allows for easier parallelization by eliminating the need to create parameter files for each analyzed gene. -host/–hostfile allows the user to indicate the cluster hostfile to avoid conflicts when starting mpi processes.

Also, if the query genes are from human, a separate script is provided for downloading their CCDS sequences prior to using DGINN itself. This script, called CCDSquery.py and available on the Github, only requires a table as its entry, with HUGO Gene Nomenclature Committee (HGNC) approved symbols in one column and the corresponding CCDS accessions in another. This table can be obtained through the HGNC biomart (link).

Results extraction

An independent script, parseResults.py, is provided to extract the essential results after running the pipeline. This script outputs a table (described in DGINN's documentation) which compiles, for each analyzed gene, the results regarding duplication and recombination events, and the different methods of positive selection detection used (including significance of each method and sites identified). This script only requires the path to the directory containing DGINN's results as input.

An R Shiny App (see Availability) has been further designed to help the user visualize the results quickly, which only necessitates the file produced by parseResults.py. This app will output the figures in the same format as those shown in Figures 3-4.

Validation dataset and method

To test our pipeline, we used a dataset of nineteen primate genes, for which evolutionary histories and positive selection profiles are either known and described in the literature or have been established within our laboratory in the past years (Table 2). We grouped those genes in three categories based on the clusters described by Murrell et al. (48): "canonical arms-race genes" such as APOBEC3G and SAMHD1 (Table 2, red column), "genes described as presenting various selection profiles" (Table 2, green column), such as HERC5 or SERINC3, either regarding the methods employed to detect positive selection or the strength of the detected signal, and "genes under no positive selection pressure" such as GADD45A and RHO/rhodopsin (Table 2, blue column). The goal was to validate our automatic DGINN method using data and findings from highly hand-curated phylogenetic and evolutionary analyses, and if possible enrich them. To assess the pertinence of our detection of duplication events, we included nine genes belonging to multigene families (annotated with an asterisk in Table 2). A gene was considered as part of a multigene family if it had at least one paralogue with over 50% reciprocal identity amongst primates (according to Ensembl). A member of the APOBEC3 gene family was also included as an extreme example of genes involved in virus-host evolutionary arms-races and that have undergone numerous genetic innovations (49–52). Another example of multigene family member included is HERC5, which exhibits antiviral activity (53) and described in the literature as evolving under positive selection (54). Given that in this latter case the analyses were performed on a limited number of primate species (seven species) and that this may conduct to a bias in the signature of positive selection, HERC5 was included in the "various" category rather than in the "canonical" one.

The primate species tree used to assess for duplication events is based on the one established by Perelman et al. (55) and updated by Pecon-Slattery (56), with minor modifications: species' names according to the six-letter naming system nomenclature that is used in DGINN (and is similar to UCSC genome's nomenclature: the first three characters of the organism's genus and species classification in the format gggSss; e.g. Homo sapiens becomes homSap), species names were updated (e.g. Tarsius syrichta was replaced by carSyr for Carlito syrichta), Rhinopithecus bieti (rhiBie) and Rhinopithecus roxellana (rhiRox) were added as the closest relatives of Rhinopithecus brelichi (rhiBre). This modified tree is available on DGINN's Github (link).

Table 2: Validation dataset of nineteen primate genes with various evolutionary histories.

Genes are categorized according to their selection profiles as reported in the literature. Asterisks (*) denote a gene belonging to a gene family with at least one paralogue in primates presenting over 50% reciprocal identity with the query gene according to Ensembl. The corresponding literature reference for each gene of the validation dataset is indicated in the second column of each category (23, 48, 54, 70, 72–77). (Of note: Although there have been some contradictory reports on FOXP2 recent evolution in humans, this gene has been described under negative selection at the primate evolution scale (78)).

Canonical arms race genes		Variable signs of positive selection		No positive selection	
APOBEC3F *	Murrel *et al.*, 2016 (47)	HERC5	Woods *et al.*, 2014 (53)	FOXP2 *	Murrel *et al.*, 2016 (47)
IFI16 *	Cagliani *et al.*, 2014 (75)	NT5C3A	In house analysis	GADD45A *	In house analysis
GBP5 *	McLaren *et al.*, 2015 (74)	RB1	Murrel *et al.*, 2016 (47)	GMPR *	In house analysis
MX1 *	Mitchell *et al.*, 2012 (72)	SERINC3 *	Murrel *et al.*, 2016 (47)	ISG20	In house analysis
RSAD2/Viperin	Lim *et al.*, 2012 (76)	SHH *	Murrel *et al.*, 2016 (47)	RHO	Murrel *et al.*, 2016 (47)
SAMHD1	Laguette et al., 2012 (69) Lim et al., 2012 (70)	SMC6	Abdul *et al.*, 2018 (23)	TREX1	In house analysis
ZC3HAV1/ZAP	Kerns *et al.*, 2008 (71)				

Reconstruction of the evolutionary history of primate Guanylate-binding protein (GBP) family

Homologs for human GBP4 and GBP6 were retrieved online through Blastn (link) against the nr database limited to primates (taxid:9443). Sequences were manually selected to span as many primate species as available. Their accession numbers were added to the list of accession numbers previously obtained from the DGINN run from the human GBP5 query, then DGINN was run from the accession step to the duplication step (steps 2-5) to determine the new orthologous relationships and reconstruct the different gene trees.

Resources

DGINN was run on the nineteen genes in a cluster environment (PSMN) in two stages. The first one ran from blast step against the NCBI non-redundant nucleotide nr/nt database circumscribed to primate species, with default settings (2 CPUs for each gene) until the identification of recombination events (steps 1-7, Figure 1). The second stage focused solely on positive selection analyses (step 8, 1 CPU for each alignment). Running times are summarized in Table 3.

Availability

All scripts and documentation are freely available on Github (link) and as a Docker on DockerHub (link). Example files to test DGINN are available to the users on GitHub. A specific script for the extraction of batch results, parseResults.py, is also available on the same Github. A graphical interface, which uses the file produced by parseResults.py as input and produces basic figures from the results (as in Figures 3-4), has also been developed and is available under the following link.

Results and Discussion

1- Presentation and novelties of the DGINN pipeline

The DGINN pipeline presents an end-to-end solution for the phylogenetic and automated detection of genetic innovations on protein-coding genes that are suspected to have undergone adaptive evolution. It

automates the search for homologous sequences, their codon alignment and the reconstruction of phylogenetic histories. This is followed by the identification of marks of genetic innovations: (i) duplication events (also allowing for the identification of orthologous groups), (ii) recombination events (also limiting bias in subsequent positive selection analyses), (iii) positive selection through different methods.

The detailed presentation of the steps is found in the Method section. Key novelties of the DGINN pipeline include a major focus on its flexibility of use: as such, it is possible to enter at any step in the pipeline without deep knowledge of the command line. The possibility to search within a single pipeline for diverse mechanisms of genetic innovations and using different methods for positive selection analyses translates to saved time compared to independent performance of each analysis. Moreover, though DGINN is designed to screen large datasets, it can also be used to perform gold-standard analyses on a single gene of interest with ease. For example, in the analyses of Lahaye et al. (57), positive selection analyses on the NONO gene were performed through the use of DGINN to determine the evolutionary history of this newly discovered sensor of the Human Immunodeficiency Virus (HIV) capsid. Finally, DGINN includes key features detailed hereafter which are novel in such pipelines and allow for a more versatile use than just the detection of positive selection.

Automatic retrieval of homologous sequences and constitution of orthologous groups by tree reconciliation

The first important step for the identification of genetic innovations in a protein-coding gene is the retrieval of orthologous sequences of this gene, in as many species as possible in a given range, clade or family of interest to the user. Automating this step is a challenge as the evolutionary characteristics of orthologous genes vary a lot (between organisms, between copies in different species, according to different molecular clocks or environmental constraints). Usually, this step is time consuming and demands high manual curation. This is even more true for genes that have rapidly evolved. Most available tools for the detection of positive selection rely on user-provided alignments or are limited to fixed input species as PosiGene (7). To circumvent these limits, DGINN uses BLAST against the NCBI online databases (see Methods – Steps 1-2). This approach makes the search for homologues simpler and relies on a widely-used and well-known tool, BLAST, which can be parameterized by the user. As true orthologous genes are identified through a subsequent reconciliation step, the user can cast a wide net by tuning parameters in terms of minimum coverage, e-value, identity, and species concerned.

From a set of homologous sequences, true orthologous groups are identified through a reconciliation software, Treerecs (41) and additional homemade scripts (Steps 3-5). Using tree reconciliation instead of annotations or tools such as OMA or Eggnogg (58, 59) may be particularly advantageous when working with non-model species, unknown genes, and recent duplication events. By separating the two phases of homologs retrieval and orthologs identification, we ensure that the user can change BLAST parameters without compromising the validity of the subsequent positive selection analyses.

DGINN detects gene duplication events, which may themselves be hallmarks of genetic innovation

While tools for the detection of positive selection abound, they often leave aside the detection of other

hallmarks of genetic innovations, such as duplications (2). Very often, duplicated genes are even taken out of the analysis entirely to avoid bias during the detection of positive selection (5). However, this may lead to missing potential genes of interest and dismissing the gene copies that have been under adaptive evolution. On the contrary, DGINN looks for duplication events, as signals of potential genetic innovation as well as to identify relevant groups of orthology for further analyses. Similarly, tools which perform orthologous assignments from annotations cannot be trusted to detect either recent duplications or ancient ones on non-model species. To our knowledge this is the first time this feature is included in an automated pipeline searching for genetic innovation. The importance of accounting for those events is shown through the numerous genes involved in genetic conflicts which have undergone duplications and subsequent diversification (2). For example, many antiviral effectors, also called restriction factors, belong to multigene families, where duplicated copies have evolved varied antiviral functions and/or virus-host interfaces/determinants, such as the Mx (Myxovirus resistance) Dynamin Like GTPases Mx1 and Mx2 (60), the guanylate-binding proteins GBPs (61, 62), the primate APOBEC3 gene family (49–51, 63) or the genes from the TRIM family (26).

Accounting for recombination allows for the detection of an important source of genetic innovation, while also avoiding bias in subsequent positive selection analyses

DGINN uses GARD to detect significant recombination breakpoints along the aligned sequences. As previously mentioned, recombination and gene conversion may be major sources of genetic innovations (in particular in the context of large gene families), and are widely ignored in existing pipelines. One example is the TRIMcyp gene present in certain primate species which results from the recombination and fusion of a cypA gene with the antiviral TRIM5 gene leading to a change of antiviral specificity (26). Moreover, recombination may also itself bias phylogenetic reconstruction and positive selection analyses (64, 65), as exemplified by the multiple recombination and gene conversion events that occurred in the Mx gene family during mammalian evolution (66). To date, only the PSP (15) and PoSeiDon (12) pipelines account for such events in their workflow. In DGINN, detecting recombination events thus serves two purposes: identifying one possible hallmark of genetic innovation and avoiding bias in positive selection analyses.

DGINN integrates numerous methods for the detection of positive selection

The detection of signatures of positive selection is a key part of the pipeline. Indeed, very few pipelines include different models than the ones from PAML (9, 15). In DGINN, we decided to implement various methods with different underlying models, so the results obtained are more robust and can be balanced between methods. It also helps to "rank" the importance of signatures on genes when a large dataset is screened. The methods and models are described in the Method section, Step 8. In addition to the most used PAML codeml, we included Bio++ bppml with similar but non-stationary models. Of note, on our validation dataset, Bio++ bppml consistently performed better than PAML codeml when it comes to calculating likelihoods (Supplementary Table 1). Moreover, because of its versatility, Bio++ allows for more parameterization and the easy declaration of many modelings that would permit to detect positive selection under user-defined scenarios (e.g. using non-homogeneous mixture models, or other kinds of

models such as allowing amino-acid specificity or simultaneous substitutions (67, 68)).

Lastly, HYPHY is a good complement in those analyses, as shown in various studies (23, 25–28). We thus decided to include two methods from the HYPHY package: one that considers the impact of positive selection at the level of the gene itself, using a branch-site model (BUSTED, (45)), and another one which detects episodic positive selection at the site level (MEME, (46)). However, codon models have long running times, and users may not want to run all of these methods in one go if their goal rests on fast answers. Running times of Bio++ non-stationary models outperformed PAML codeml models in almost every instance in the validation dataset presented hereafter: 17 out of 19 analyses were faster in either $M1^{NS}$ vs $M2^{NS}$ and $M7^{NS}$ vs $M8^{NS}$, compared with codeml M1 vs M2 and M7 vs M8 (Table 3). Moreover, Bio++ parameter files can be easily modified to accelerate the modeling even further. As such, we would suggest the use of Bio++ only for such users for whom time is of the essence.

Table 3: DGINN validation running times.

For each gene, the running time of "Steps 1-7" and "Step 8" (Figure 1) is shown, as well as a break-down of the running times of each of the methods run during Step 8. Times for Step 8 (positive selection analyses) are only shown for the query genes of the validation dataset following attribution of orthologous groups (Table 4).

	Steps 1 – 7	Step 8							Balance speed and results (Steps 1-7, Bio++ only)
		TOTAL	BUSTED	PAML M1 vs M2	PAML M7 vs M8	Bio++ M1 vs M2	Bio++ M7 vs M8	MEME	
APOBEC3F	12:18:39	04:11:18	00:02:06	00:09:09	00:55:26	01:37:42	00:59:40	00:18:18	14:56:01
FOXP2	05:26:33	5 days, 23:28:04	00:10:41	22:26:10	3 days, 2:56:09	00:22:23	02:49:05	00:14:29	08:38:01
GADD45A	01:24:26	02:17:59	00:01:00	00:17:43	01:03:46	00:16:32	00:16:42	00:09:15	01:57:40
GBP5	14:04:30	06:43:06	00:07:57	00:41:13	02:17:19	00:24:26	00:54:38	01:12:38	15:23:34
GMPR	03:51:52	07:50:42	00:06:22	01:19:12	04:13:22	00:23:07	00:28:42	00:46:10	04:43:41
HERC5	04:03:01	15:36:40	00:07:03	01:43:54	06:02:21	01:42:03	01:39:28	02:24:10	07:24:32
IFI16	05:45:16	5 days, 8:01:45	01:09:12	18:32:28	4 days, 1:14:47	01:28:31	01:22:09	03:19:20	08:35:56
ISG20	01:34:50	08:01:08	00:04:04	01:23:59	04:51:58	00:20:49	00:18:46	00:29:56	02:14:25
MX1	02:34:42	1 day, 18:16:17	00:12:33	09:14:12	1 day, 0:17:30	01:14:08	03:18:13	01:56:16	07:07:03
NT5C3A	00:53:48	4 days, 20:17:21	00:17:33	13:42:28	1 day, 15:07:57	00:42:06	2 days, 0:47:25	10:08:30	2 days, 02:23:19
RB1	00:14:09	14:16:03	00:20:44	01:59:13	07:46:37	00:27:46	01:40:49	01:25:10	02:22:44
RHO	00:06:30	02:05:53	00:02:46	00:12:45	00:34:20	00:15:19	00:24:46	00:24:53	00:46:35
RSAD2	01:11:31	18:40:09	00:09:17	03:40:53	10:40:19	00:29:09	01:13:49	01:09:52	02:54:29
SAMHD1	00:51:44	3 days, 13:26:08	00:10:46	19:30:38	2 days, 6:50:22	01:23:20	02:07:07	02:32:58	04:22:11
SERINC3	01:21:46	11:55:16	00:05:58	01:37:51	05:24:02	00:56:57	02:17:20	00:59:16	04:36:03
SHH	01:54:44	06:12:50	00:04:46	00:56:20	03:15:44	00:13:25	00:54:23	00:34:22	03:02:32
SMC6	02:48:41	2 days, 20:34:48	00:16:53	14:16:56	1 day, 21:28:49	00:40:12	03:00:34	02:08:19	06:29:27
TREX1	01:01:04	09:43:10	00:04:11	01:18:01	04:00:57	00:22:24	01:46:02	00:04:34	03:09:30
ZC3HAV1	02:38:52	2 days, 13:27:12	00:16:12	11:03:56	1 day, 10:43:01	01:36:41	04:24:26	01:54:47	08:39:59

2- Validation

We tested our pipeline on nineteen primate genes selected for their various evolutionary histories and positive selection profiles (Table 2). These genes were grouped in three categories based on the clusters described in Murrell et al., 2016: "canonical arms-race genes" such as MX1 and SAMHD1, "genes described as presenting various selection profiles", such as HERC5 or SERINC3 "genes under no positive selection pressure" such as GADD45A and RHO/rhodopsin (Table 2). The intermediate category was attributed on the basis of the methods employed to detect positive selection or the strength of the detected signal (see Method section).

An overview of the complete execution of DGINN on a protein-coding gene

A brief overview of DGINN's workflow on a specific gene, HERC5, is presented in Figure 2. The Blast search returned 71 primate homologous sequences, of which twelve were eliminated by the subsequent filters, yielding to a total of 59 sequences. As a duplication event was detected by Treerecs, these 59 sequences were then automatically (and correctly) split into two groups: one with 32 sequences

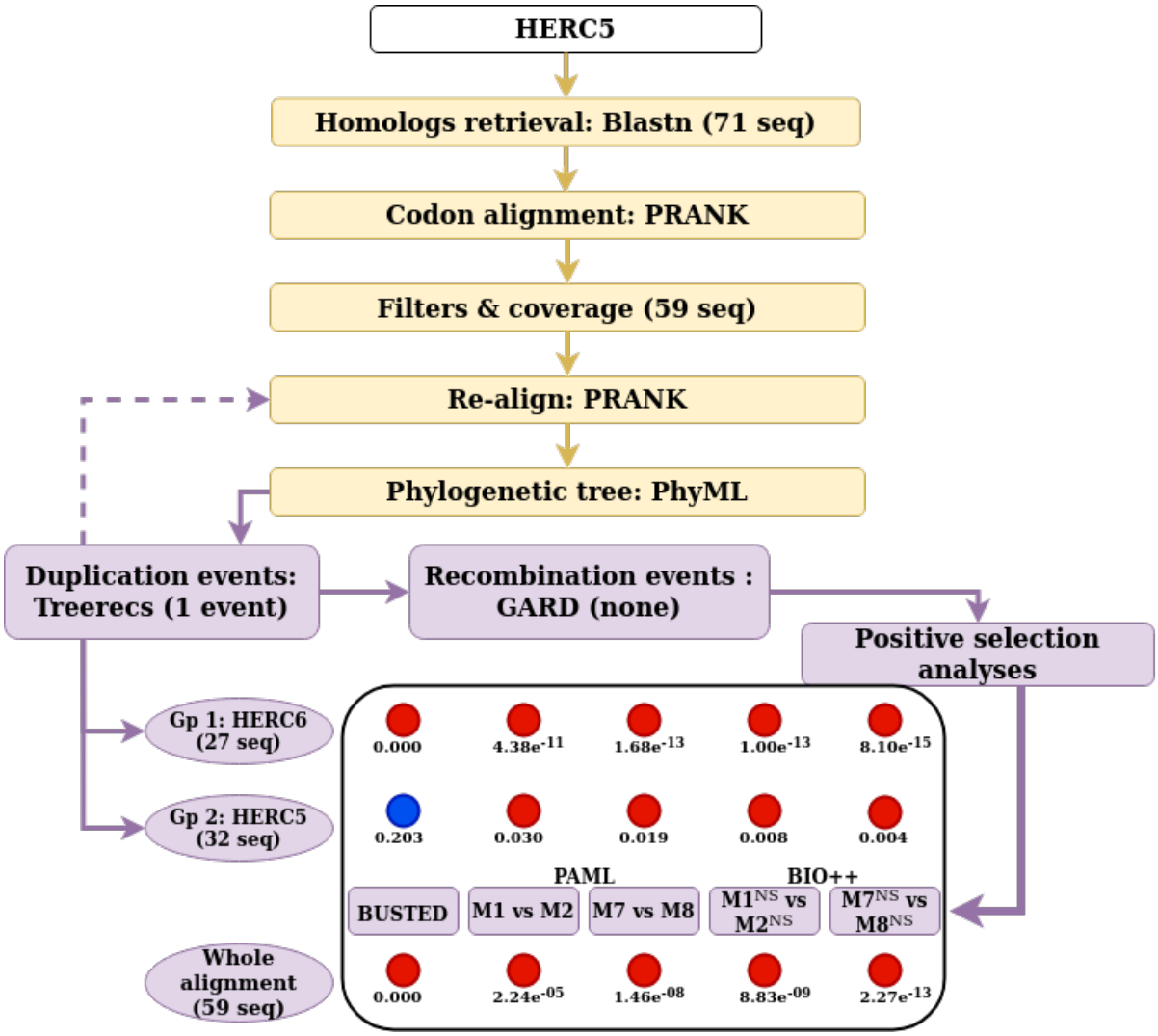


Figure 2: Example of workflow on the HERC5 primate gene.

The workflow follows the diagram from Figure 1. Using human HERC5 CDS as the starting point in DGINN gave results for both HERC5 and HERC6. The number of sequences (seq) retrieved or left after each step is indicated. In the bottom panel, each colored circle represents the results from one of the five methods to detect positive selection at the gene level, with red representing significant evidence of positive selection and blue no significant evidence. P-values are indicated below the colored circles. Gp, orthologous group.

corresponding to HERC5 and one with 27 sequences corresponding to HERC6. No recombination event was identified and the positive selection analyses then followed. All methods found highly significant evidence of positive selection on the complete alignment of 59 mixed HERC5-HERC6 sequences, with p-values ranging from 2.24e-05 to 2.27e-13 for PAML and Bio++ models. However, after separating the two paralogues into orthologous groups, it appeared that most of this signal was driven by the positive selection on HERC6 (p-values of 4.38e-11 to 8.10e-15 for PAML and Bio++ models), while the signal on HERC5 sequences was present but much more modest (p-values, 0.030 to 0.004), with BUSTED even returning a non-significant p-value for positive selection on that alignment. The positive selection results therefore highlight the necessity to properly separate paralogues from each other prior to performing the analyses. For a query on the HERC5 gene, keeping the initial mixed alignment could have caused a mistaken conclusion that primate HERC5 has been under very strong positive selection, though the

signal was mostly driven by HERC6. Moreover, the sites identified as under positive selection on that alignment would also be erroneous. Overall, the complete DGINN analyses with HERC5 as query took less than 20 hours (Table 3, 4h03 for the data mining and phylogenetics, and 15h36 for the detection of genetic innovations per se).

Detection of ancestral duplications allows for proper assignation of orthologous groups

We identified genes as belonging to multigene families if at least one member had over 50% reciprocal identity with our gene query according to ENSEMBL annotations (Table 2). Given this definition, we were able to retrieve multiple family members for the majority of the genes belonging to such families, when performing BLAST with the minimum coverage (50%) and identity (70%) values. The sole exception was SERINC3, for which no paralogue was returned through our Blast search. Two additional exceptions were observed, first with HERC5, for which the Blast search also returned HERC6 sequences, though reciprocal identity between the two paralogues was below our threshold. The second case concerned TREX1, for which the Blast search also returned sequences annotated as ATRIP, an adjacent gene. Given that read-through transcription of TREX1-ATRIP occurs naturally and yields a non-coding transcript, it is probable that those sequences annotated ATRIP actually represents the non-coding transcript and not the mRNA of the ATRIP gene. This explains the retrieval of ATRIP-annotated genes through Blast despite the two genes not being strictly homologous.

DGINN efficiently reconstructed orthologous groups (Table 4). Indeed, in the case of multigene families (from two to five paralogues retrieved here), we were able to properly reconstruct orthologous groups for our genes of interest, without mixture with other paralogues. Our approach allowed us to separate the different family members retrieved through BLAST in groups which did not mix paralogous sequences through long branch partition (LB) and/or through reconciliation (Treerecs). For example, using the human CCDS sequence of FOXP2 as input in DGINN, we retrieved sequences from both FOXP2 and its paralogue FOXP1. The tree reconstructed from their alignment featured a branch over 50 times longer than the mean length of the tree's branches, and by automatically splitting the sequences separated by that branch, we were able to reconstitute two groups corresponding to the paralogues. However, paralogues from other families may not have diverged enough for long branch partition to be able to properly discriminate them into different groups. We resolved those through TreeRecs, reconciling the tree obtained from the Blast-retrieved sequences with the primate species tree. This is the case, for example, of the immune sensor IFI16, which was properly assigned to a different group than MNDA through our Treerecs-based approach. Non-annotated sequences (such as those referred as LOCXXX in databases) were also assigned to groups through this process, showing that this method of attributing orthologous relationships might help with non-annotated sequences in the databases.

Of our nineteen genes of interest, only one presented some inaccuracies in the distribution of sequences to orthologue groups. With an APOBEC3F query, DGINN erroneously divided APOBEC3F itself in two different groups (group 3 and 5, Table 4). By further analyzing all the retrieved paralogues, we observed two mixes: in the APOBEC3F query, group 2 contained APOBEC3D and APOBEC3B sequences and APOBEC3B was split in two groups, and a similar pattern occurred in the GBP5 query, with GBP1 in

Table 4: Groups of orthologues reconstructed by DGINN, using long-branch partition and TreeRecs for identification of duplication events.

For each gene of the validation dataset, are represented the orthologous groups that were identified, the number of sequences per group, the orthologues present in the group and the method used to separate the groups (long branch (LB) partition or TreeRecs-based). Groups kept for subsequent analyses are highlighted in yellow.

Query	Group	Number of sequences	Gene	Type
APOBEC3F	1	11	APOBEC3B	Treerecs-based
	2	56	APOBEC3D + APOBEC3B	
	3	16	APOBEC3F	
	4	94	APOBEC3G	
	5	10	APOBEC3F	
FOXP2	1	77	FOXP2	LB
	2	59	FOXP1	
GADD45A	1	30	GADD45A	LB
	2	4	GADD45B	
GBP5	1	48	GBP3	Treerecs-based
	2	28	GBP1	
	3	32	GBP7 + GBP1	
	4	36	GBP2	
	5	24	GBP5	
GMPR	1	104	GMPR2	LB
	2	34	GMPR	
HERC5	1	27	HERC6	Treerecs-based
	2	32	HERC5	
IFI16	1	60	IFI16	Treerecs-based
	2	26	MNDA	
MX1	1	60	MX2	LB
	2	55	MX1	
SHH	1	25	IHH	LB
	2	22	SHH	
TREX1	1	35	TREX1	LB
	2	14	ATRIP	

groups 2 and 3 (Table 4). These errors could be explained by the particularly complicated evolutionary histories of those two expanded gene families during primate evolution (49, 51, 63). This highlights a need to improve the management of the detection of duplication events in further versions of DGINN. Importantly, because such genes would be tagged by DGINN with "detected duplication events", these cases would anyway not be missed by the user and the gene of interest could be reanalyzed through DGINN after curation.

Using several positive selection methods together allows for more sensitivity and specificity and a "ranking" of genes' positive selection status during screening

Positive selection results were analyzed according to two different aspects. The first aspect focused on how many methods found a gene with significant evidence of positive selection (Figure 3, left panel – produced using the Shiny app openly available). The methods considered at this point were those on which a LRT could be performed: HYPHY BUSTED, the M1 vs M2 and M7 vs M8 models of PAML

Codeml, and the $M1^{NS}$ vs $M2^{NS}$ and $M7^{NS}$ vs $M8^{NS}$ models of Bio++ bppml. Genes were ranked according to the number of positive results. This allowed us to compare the results obtained for the three categories of genes (Table 2). The canonical arms-race genes were all detected under positive selection by all five methods, with the exception of RSAD2 which was detected by four methods (Figure 3). Genes which presented variable signs of positive selection in the literature (green category, Table 2) also fell into a middle category in the DGINN screen. Genes without signs of positive selection in previous studies (blue category, Table 2) displayed low signs of positive selection: detected by less than two methods in DGINN. Two genes were detected by two methods: FOXP2 and RHO. FOXP2 was detected by both PAML M7 vs M8 and Bio++ $M7^{NS}$ vs $M8^{NS}$, but both the mean omega and the very low number of sites detected under positive selection (n=1) suggested artefactual results. Similarly, RHO was detected by BUSTED and Bio++ $M7^{NS}$ vs $M8^{NS}$, but only two sites were detected. Therefore, our DGINN screen efficiently recapitulated results from published studies.

These results further highlight the advantage of using different methods within a single analysis to confirm results and discriminate for false positives. Doing this validation also showed that amongst those methods, BUSTED and PAML Codeml M7 vs M8 appeared the least conservative methods to detect positive selection.

If one would run less methods because of time constraint, our validation results indicate that running Bio++ methods would best balance running times, sensitivity and specificity. The second aspect taken into account focused on the percentage of positively-selected sites. Overall, the arms-race genes displayed higher proportions of positively selected sites (2.4%-8.8%) compared to other genes (Figure 3, right side). However, this does not represent a hard rule, as some of those arms-race genes show rather low percentages, such as MX1 (around 3.2%). This suggests that ranking genes by the number of significant methods rather than the proportion of positive selection sites, as in Figure 3, is a better proxy for positive selection status.

DGINN recapitulates and expands the findings from previously published profiles of positively selected sites along genes.

To identify the domains that have evolved under positive selection, we mapped every positively selected site detected by DGINN by a peak along the alignment (Figure 4, using the Shiny app). We further represented the height of the peak proportional to the number of methods detecting that site under significant positive selection, amongst five methods, M2 and M8 results of PAML codeml, $M2^{NS}$ and $M8^{NS}$ results of Bio++ bppml and HYPHY MEME (Figure 4). Overall, we observed similar patterns as described in the literature, especially on the canonical arms-race genes. For example, in the case of SAMHD1, we found most positively selected sites at the N- and the C-termini (Figure 4). This is in accordance with the findings that the N-ter and C-ter domains both play a role in the antiviral/escape determinants of primate SAMHD1 and that rapid evolutions at these sites are certainly adaptive as a result of lentiviral selective pressure (69–71). In the case of ZC3HAV1/ZAP, we found the positively selected sites cluster at both extremities of the alignment (Figure 4). However, the middle portion without positively selected sites corresponds to a gap-enriched region in the alignment linked to the different possible isoforms of the

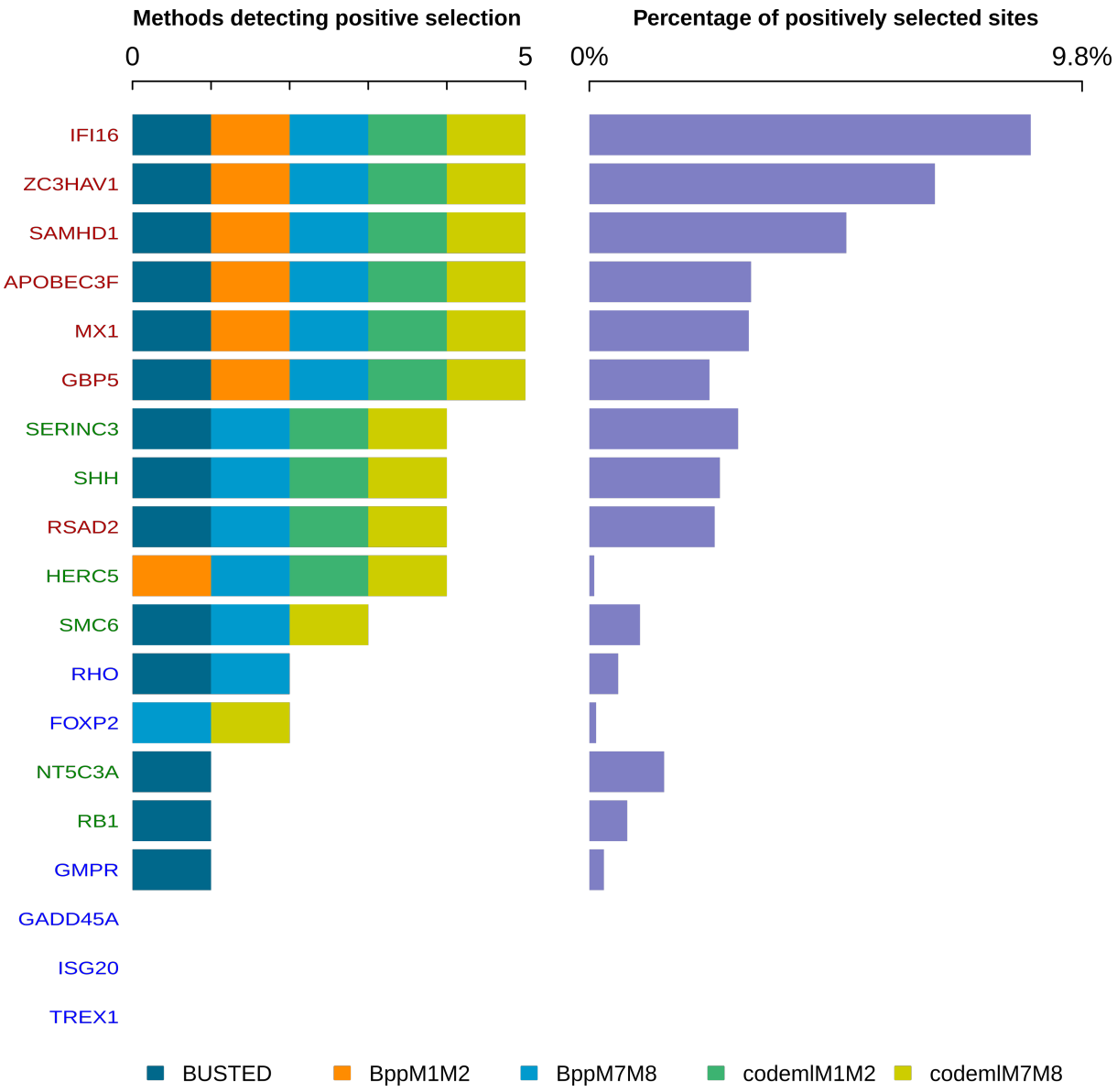


Figure 3: DGINN results on the validation dataset.

The nineteen primate genes studied are color-coded according to their selection profile category (Table 2). Left panel, number of methods detecting significant positive selection for each alignment; each method is color-coded (embedded legend). Right panel, percentage of positively selected sites (by at least one method) over the length of the query coding sequence. Genes are ordered by descending number of methods detecting positive selection then descending percentage of positively selected sites.

gene. Interestingly, this shows that the maintenance of these gap regions in the alignment did not lead to an excess of false positive detection in DGINN. If we now consider the main ORF (with the gap-enriched region ignored), it appears that the positively selected sites are spread over the whole length of the gene (Figure 4). Previous results established that the C-ter domain in particular was under significant positive selection (72). In contrast, the N-ter domain was not detected, probably because we used more methods and had more species/sequences available for analyses.The differences between our results and the published ones for APOBEC3F (48) were mainly due to the sequences used for the positive selection analyses. Indeed, our analyses excluded four species that were correctly retrieved in the early steps of DGINN but were erroneously assigned by TreeRecs to another group, so the detection of positive selection was only

performed on a subset of primate sequences, spanning solely Old World monkeys. However, we have included the solution to such problems in DGINN thanks to its high flexibility. The user may retrieve the gene sequences (here APOBEC3F) from the different groups and re-enter DGINN at step 3/alignment (Figure 1 and Table 2) to obtain the complete evolutionary history and positive selection analyses.

For MX1, we were first surprised that we did not detect such a high signal of positive selection in the L4 loop as described previously (Figure 4) (73). However, we found that this was mainly due to differences in the alignments, because PRANK (as opposed to ClustalX used in the previous study) introduced many gaps in the L4 loop region due to the extremely-high divergence of the region. Whether MX1 adaptation to viral countermeasures has occurred by accumulation of non-synonymous changes and/or by indels in the L4 loop remains to be determined.

In the case of HERC5, four methods detected the gene as under positive selection during primate evolution (Figures 2-3), but only one site was identified as positively selected (Figure 4). These results differ from the ones reported previously (54), who found a much larger number of residues under positive selection (n=50). This discrepancy, however, can be explained by the fact that the previous study identified positive selection on an alignment that included six non-primate species and only seven primate species, while ours focused exclusively on primates and included twenty species. It is therefore possible that a stronger selective pressure has occurred in placental mammals outside of primate evolution. Interestingly, in DGINN, our Blast search with HERC5 as query also automatically retrieved HERC6 sequences (Figure 2). The latter were then correctly assigned to a different orthologous group than HERC5. As previously reported (Paparisto et al., 2018), we identified strong evidence of positive selection on HERC6 (with five methods, Figure 2). This could mean that while both HERC5 and HERC6 have been evolving under positive selection in mammals, they have been subjected to different evolutionary constraints in primates, with a lower selective pressure on primate HERC5 vs HERC6. It further shows that DGINN is an efficient tool to screen not only the query genes but also the evolutionary history of their closest gene relatives that may have themselves be subjected to positive selection and would otherwise be missed by most analyses.

Identification of the loss of GBP5 during primate evolution using DGINN

The positive selection results obtained through DGINN screening for GBP5 showed strong positive selection (identified by five methods). This was in accordance with previous results from McLaren et al., 2015. By analyzing the phylogenetic tree generated by DGINN for all the homologs retrieved with the GBP5 query (after step 4, Figure 5A), we found that no sequence from Old World monkeys were retrieved for GBP5 through our Blast search. This absence was confirmed in the tree reconstructed with only GBP5 sequences after orthologue group attribution (step 5, Figure 5B). However (and as expected), the entire GBP gene family was not retrieved by DGINN using human GBP5 as query (with blastn 70% identity and 50% coverage); in particular, GBP4 and GBP6 were too divergent to be retrieved by DGINN. To reconstruct to GBP family evolutionary history, we independently retrieved primate sequences of GBP4 and GBP6 by blastn and added the new sequences to a large GBP family sequence file. This served as input to DGINN steps 2-5 to automatically perform alignments, phylogenies, and duplication/orthologous

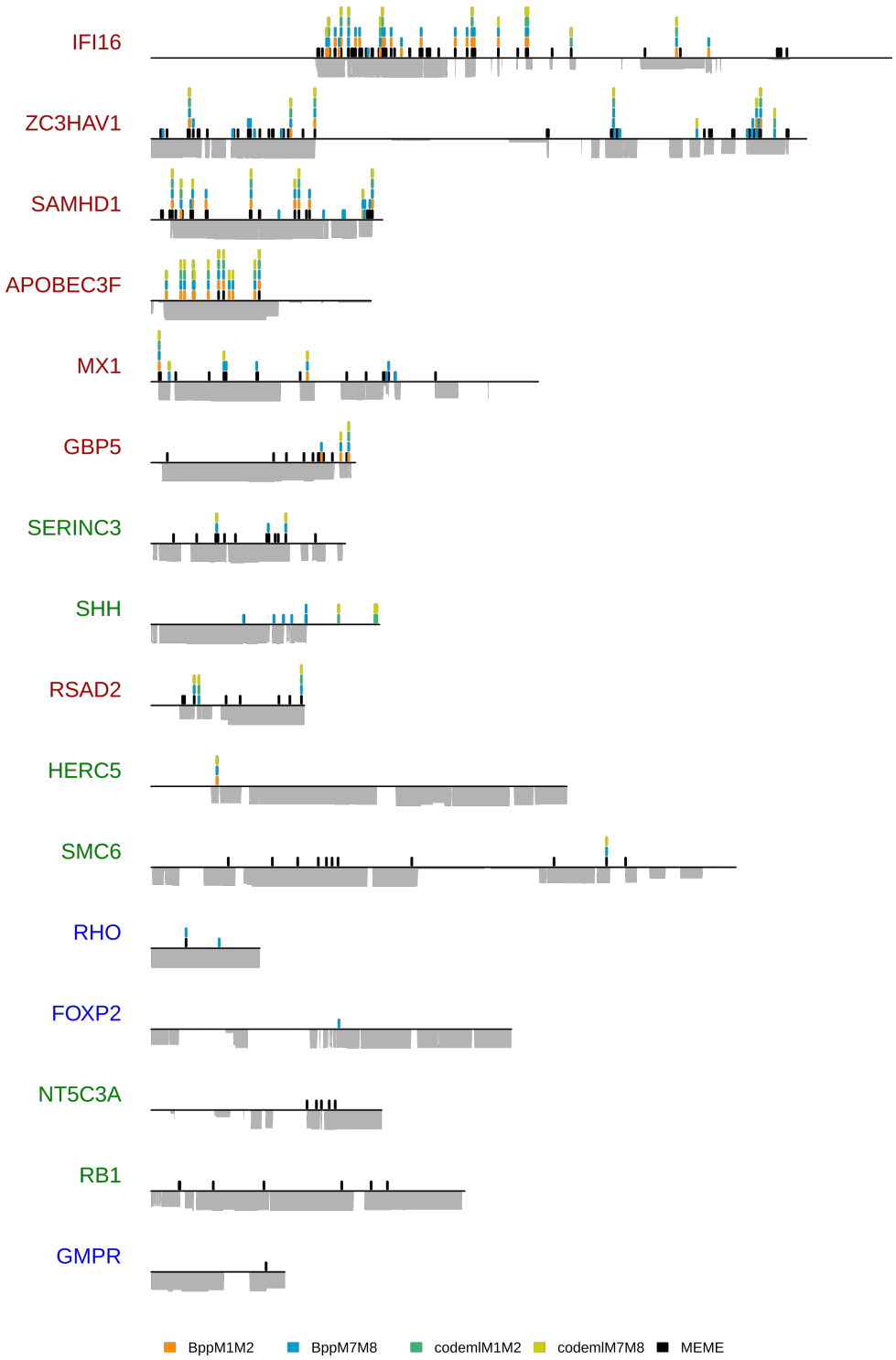


Figure 4: Positive selection patterns on nineteen primate genes.

The nineteen genes studied are color-coded according to their selection profile category (Table 2) and follow the same order as in Figure 3. Genes without positively selected sites were excluded from this representation. Positively selected sites are represented as a spike at their position on the alignment. Height of the peak is proportional to the number of methods that have identified the site as being under positive selection (posterior probabilities > 0.95 for Bio++ and PAML codeml, and p-value < 0.10 for MEME), with each method being represented by a different color (embedded legend). HYPHY MEME sites were only mapped if the gene was detected as under positive selection by BUSTED ($p < 0.05$). For each gene, alignement coverage is represented under the line representing the length of the alignment in light grey.

group detection. The final tree confirmed that GBP5 is absent in Old World Monkeys (Figure 5C). This might also be the case for GBP4, for which we did not retrieve sequences from Old World Monkeys;

with the exception of two sequences from Papio anubis and Mandrillus leucophoeus that were annotated as "GBP4" but did not follow a typical orthologous phylogeny and branched more closely with GBP7 in our phylogeny (Figure 5C). Genomic analyses of the GBP locus in several primates confirmed that GBP5 has been lost in the ancestor of Old World Monkeys during primate evolution, and that it may also be the case for GBP4 (Figure 5D). To explain our retrieval of the two sole Old Word Monkeys sequences, and their position in the phylogeny, one hypothesis could be that GBP4 has indeed been lost at a similar point in primate evolution than GBP5, and was then regained in some Old World monkey species through a duplication of GBP7. Overall, these results show that GBP5 has been subjected to strong positive selection during primate evolution but has also entirely been lost in the Cercopithecinae. Whether part of this has been driven by pathogens such as lentiviruses (33) or bacteria (32) should be investigated.

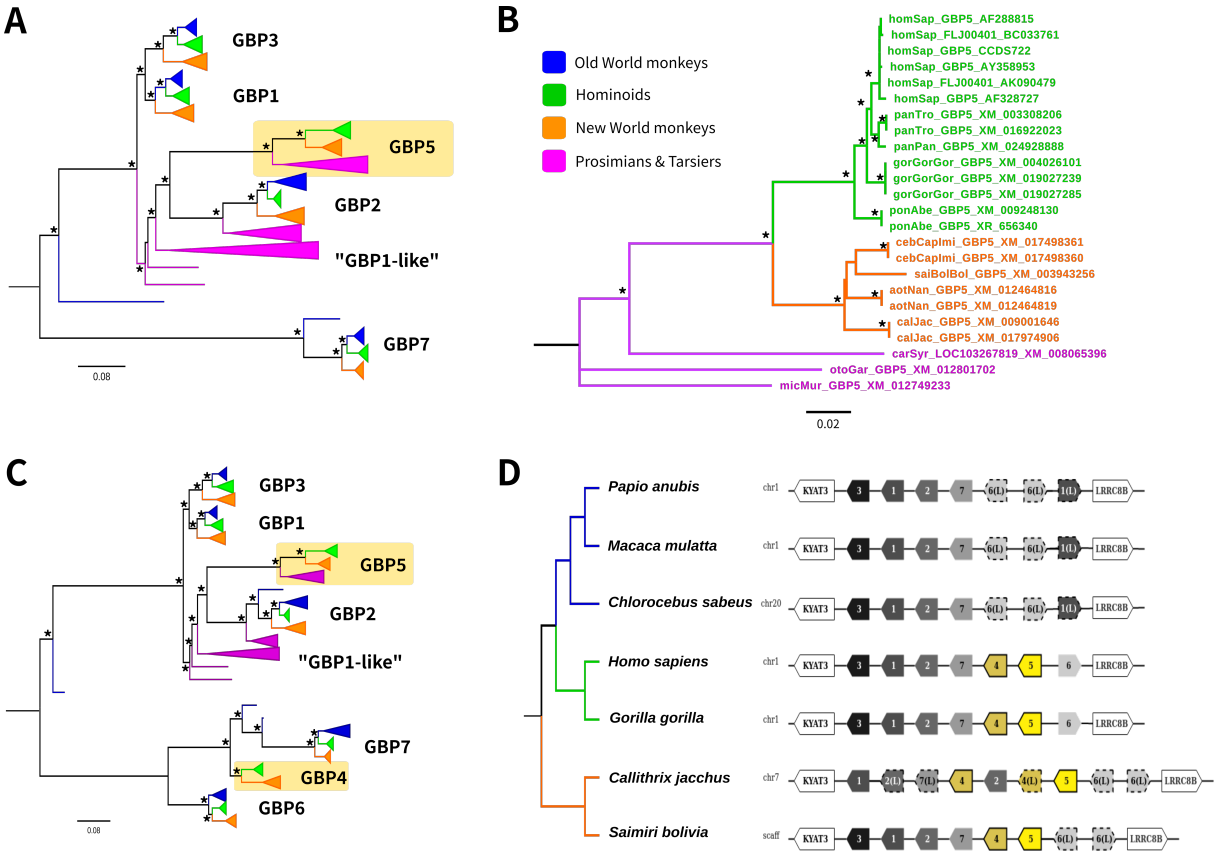


Figure 5: Evolutionary history of the primate GBP family.

(A) Maximum-likelihood phylogeny established through DGINN based on a run on the GBP5 query (step 4). The four main primate lineages are identified by color-coding: Old World monkeys, blue; Hominoids, green; New World monkeys, orange; prosimians, purple/pink. Asterisks (*) denote nodes that are statistically supported by aLRT > 0.90. The GBP5 group, which lacks Old World monkey sequences, is boxed in yellow. The scale bar represents the number of nucleotide substitutions per site and the tree was midpoint rooted. **(B)** Maximum-likelihood phylogeny of the GBP5 group of primate orthologues established through DGINN screen (step 7). **(C)** Maximum-likelihood phylogeny of the whole GBP family performed in DGINN after manual addition of primate GBP4 and GBP6 sequences. **(D)** Diagram of the genomic locus of the GBP gene family in seven simian primate species. The reference genomes from the NCBI used were: papAnu (*Papio anubis*): Panu_3.0, macMul (*Macaca mulatta*): Mmul10, chlSab (*Chlorocebus sabaeus*): Chlorocebus_sabeus 1.1, homSap (*Homo sapiens*): GRCh38.p13, gorGor (*Gorilla gorilla*): gorGor4, calJac (*Callithrix jacchus*): Callithrix jacchus-3.2, saiBol (*Saimiri boliviensis*): saiBol1.0. X(L) annotations with dotted outlines represent genes for which the orthology and paralogy relationships could not be completely ascertained.
All alignments and phylogenies for panel A, B and C (referred as 5A_aln, 5A_tree etc...) can be found on the Github repository.

Conclusion

We have developed DGINN, an integrative pipeline for the automatic detection of genetic innovations, and made it freely available through both GitHub and Docker. DGINN was validated for screening usage against nineteen primate genes (all results are available on GitHub in the corresponding repository). It automates and streamlines those analyses, allowing the user to simply provide the coding sequence of their gene of interest and a parameter file to complete the whole workflow, from retrieval of homologous sequences to the detection of orthology relationships, recombination events and positive selection.

Through our validation, we confirmed and expanded on results previously established in the literature. Genes described as engaged in arms-races with viruses were found under strong positive selection by all five methods included in DGINN. Our analyses allowed us to establish clearer profiles for the genes belonging to the "varied" category, owing to our inclusion of different methods for positive selection: this way, we were able to establish that some genes previously thought to present moderate signs of positive selection presented stronger signs than suspected. Little evidence of positive selection was found on the genes belonging to "no positive selection" category, in accordance to the literature.

An important feature of DGINN is its flexibility, which allows usage beyond its screening capacity. Indeed, in cases of dubious results, the possibility remains for the user to curate their input files and perform the appropriate analyses by entering DGINN at any of the downstream steps. This also means that the "positive selection" part might be of primary interest to scientists wishing to perform gold-standard positive selection analyses on their favorite gene, because they could enter their curated alignment and phylogeny and obtain results of positive selection analyses from five methods in a single query.

Using DGINN to analyze nineteen primate genes also allowed us to enrich some findings, notably on the importance of detecting duplications and properly ascribing orthologue groups, as exemplified by the case of HERC5 and its paralogue HERC6 in primates. The ability to check multiple members of a query's gene family is a major advantage of DGINN, as it may allow the user to identify genes bearing signs of genetic innovations that they would not have analyzed otherwise. Improving the constitution of orthologue groups will remain an objective in future versions of DGINN.

Acknowledgements

We thank Stéphanie Jacquet for her comments on the manuscript. We also thank Bastien Bousseau, Marie Cariou, Hélène Dutartre, Laurent Modolo, Xavier Morelli, and Guy Perrière for helpful discussions on this project. We gratefully acknowledge support from the PSMN (Pôle Scientifique de Modélisation Numérique) of the ENS de Lyon for the computing resources, and the PRABI (Pôle Rhône-Alpes de BioInformatique) for further bioinformatics support. We thank all the contributors of publically available genome sequences, as well as the scientists who developed the methods included in DGINN. This work was funded by the ANR LABEX ECOFECT (ANR-11-LABX-0048 of Université de Lyon, within the program "Investissements d'Avenir" (ANR-11-IDEX-0007) operated by the French National Research Agency) to LE and LG. LE is supported by the CNRS and by grants from the amfAR (Mathilde Krim Phase II Fellowship #109140-58-RKHF), the "Fondation pour la Recherche Médicale" (FRM "Projet Innovant" #ING20160435028), the FINOVI ("recently settled scientist" grant), the ANRS (#ECTZ19143, #ECTZ118944), and a JORISS incubating grant. LG is supported by the Université Claude Bernard Lyon 1 and the Swedish Center of Advanced Study. AC is supported by the CNRS and by grants from the ANRS, Sidaction and the ENS-L.

Author Contributions

Conceptualization and Supervision: LE, LG
Formal analysis: LP, LE, LG
Pipeline development: LP, LG, QG
Funding acquisition: LE, LG
Investigation: LP, QG, OA, AC, LG, LE
Methodology: LP, LG, LE
Project administration: LE, LG
Resources: AC, LE, LG
Writing – original draft: LP, LE, LG
Writing – review and editing: All the authors

References

1. Daugherty,M.D. and Malik,H.S. (2012) Rules of Engagement: Molecular Insights from Host-Virus Arms Races. Annual Review of Genetics, 46, 677–700.

2. Daugherty,M.D. and Zanders,S.E. (2019) Gene conversion generates evolutionary novelty that fuels genetic conflicts. Current Opinion in Genetics & Development, 58–59, 49–54.

3. Kondrashov,F.A. (2012) Gene duplication as a mechanism of genomic adaptation to a changing environment. Proc. R. Soc. B, 279, 5048–5057.

4. McLaughlin,R.N. and Malik,H.S. (2017) Genetic conflicts: the usual suspects and beyond. The Journal of Experimental Biology, 220, 6–17.

5. Kosiol,C., Vinař,T., da Fonseca,R.R., Hubisz,M.J., Bustamante,C.D., Nielsen,R. and Siepel,A. (2008) Patterns of Positive Selection in Six Mammalian Genomes. PLoS Genet, 4, e1000144.

6. Hawkins,J.A., Kaczmarek,M.E., Müller,M.A., Drosten,C., Press,W.H. and Sawyer,S.L. (2019) A metaanalysis of bat phylogenetics and positive selection based on genomes and transcriptomes from 18 species. Proc Natl Acad Sci USA, 116, 11351–11360.

7. Sahm,A., Bens,M., Platzer,M. and Szafranski,K. (2017) PosiGene: automated and easy-to-use pipeline for genome-wide detection of positively selected genes. Nucleic Acids Research, 45, e100–e100.

8. Pond,S.L.K., Frost,S.D.W. and Muse,S.V. (2005) HyPhy: hypothesis testing using phylogenies. Bioinformatics, 21, 676–679.

9. Stern,A., Doron-Faigenboim,A., Erez,E., Martz,E., Bacharach,E. and Pupko,T. (2007) Selecton 2007: advanced models for detecting positive and purifying selection using a Bayesian inference approach. Nucleic Acids Research, 35, W506–W511.

10. Egan,A., Mahurkar,A., Crabtree,J., Badger,J.H., Carlton,J.M. and Silva,J.C. (2008) IDEA: Interactive Display for Evolutionary Analyses. BMC Bioinformatics, 9, 524.

11. Steinway,S.N., Dannenfelser,R., Laucius,C.D., Hayes,J.E. and Nayak,S. (2010) JCoDA: a tool for detecting evolutionary selection. BMC Bioinformatics, 11, 284.

12. Fuchs,J., Hölzer,M., Schilling,M., Patzina,C., Schoen,A., Hoenen,T., Zimmer,G., Marz,M., Weber,F., Müller,M.A., et al. (2017) Evolution and Antiviral Specificities of Interferon-Induced Mx Proteins of Bats against Ebola, Influenza, and Other RNA Viruses. Journal of Virology, 91.

13. Hongo,J.A., de Castro,G.M., Cintra,L.C., Zerlotini,A. and Lobo,F.P. (2015) POTION: an end-to-end pipeline for positive Darwinian selection detection in genome-scale data through phylogenetic comparison of protein-coding genes. BMC Genomics, 16.

14. Busset,J., Cabau,C., Meslin,C. and Pascal,G. (2011) PhyleasProg: a user-oriented web server for wide evolutionary analyses. Nucleic Acids Research, 39, W479–W485.

15. Su,F., Ou,H.-Y., Tao,F., Tang,H. and Xu,P. (2013) PSP: rapid identification of orthologous coding genes under positive selection across multiple closely related prokaryotic genomes. BMC Genomics, 14, 924.

16. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research, 22, 4673–4680.

17. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research, 32, 1792–1797.

18. Loytynoja,A. and Goldman,N. (2008) Phylogeny-Aware Gap Placement Prevents Errors in Sequence Alignment and Evolutionary Analysis. Science, 320, 1632–1635.

19. Fletcher,W. and Yang,Z. (2009) INDELible: A Flexible Simulator of Biological Sequence Evolution. Molecular Biology and Evolution, 26, 1879–1888.

20. Privman,E., Penn,O. and Pupko,T. (2012) Improving the Performance of Positive Selection Inference by Filtering Unreliable Alignment Regions. Molecular Biology and Evolution, 29, 1–5.

21. Jordan,G. and Goldman,N. (2012) The Effects of Alignment Error and Alignment Filtering on the Sitewise Detection of Positive Selection. Molecular Biology and Evolution, 29, 1125–1139.

22. Markova-Raina,P. and Petrov,D. (2011) High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 Drosophila genomes. Genome Research, 21, 863–874.

23. Abdul,F., Filleton,F., Gerossier,L., Paturel,A., Hall,J., Strubin,M. and Etienne,L. (2018) Smc5/6 Antagonism by HBx Is an Evolutionarily Conserved Function of Hepatitis B Virus Infection in Mammals. Journal of Virology, 10.1128/JVI.00769-18.

24. Elde,N.C., Child,S.J., Geballe,A.P. and Malik,H.S. (2009) Protein kinase R reveals an evolutionary model for defeating viral mimicry. Nature, 457, 485–489.

25. Shultz,A.J. and Sackton,T.B. (2019) Immune genes are hotspots of shared positive selection across birds and mammals. eLife, 8, e41815.

26. Malfavon-Borja,R., Sawyer,S.L., Wu,L.I., Emerman,M. and Malik,H.S. (2013) An Evolutionary Screen Highlights Canonical and Noncanonical Candidate Antiviral Genes within the Primate TRIM Gene Family. Genome Biology and Evolution, 5, 2141–2154.

27. McBee,R.M., Rozmiarek,S.A., Meyerson,N.R., Rowley,P.A. and Sawyer,S.L. (2015) The Effect of Species Representation on the Detection of Positive Selection in Primate Gene Data Sets. Molecular Biology and Evolution, 32, 1091–1096.

28. Rowley,P.A., Ho,B., Bushong,S., Johnson,A. and Sawyer,S.L. (2016) XRN1 Is a Species-Specific Virus Restriction Factor in Yeasts. PLoS Pathog, 12, e1005890.

29. van der Lee,R., Wiel,L., van Dam,T.J.P. and Huynen,M.A. (2017) Genome-scale detection of positive selection in nine primates predicts human-virus evolutionary conflicts. Nucleic Acids Research, 45, 10634–10648.

30. Yang,Z. (2007) PAML 4: Phylogenetic Analysis by Maximum Likelihood. Molecular Biology and Evolution, 24, 1586–1591.

31. Duggal,N.K. and Emerman,M. (2012) Evolutionary conflicts between viruses and restriction factors shape immunity. Nat Rev Immunol, 12, 687–695.

32. Kim,B.-H., Shenoy,A.R., Kumar,P., Bradfield,C.J. and MacMicking,J.D. (2012) IFN-Inducible GTPases in Host Cell Defense. Cell Host & Microbe, 12, 432–444.

33. Krapp,C., Hotter,D., Gawanbacht,A., McLaren,P.J., Kluge,S.F., Stürzel,C.M., Mack,K., Reith,E., Engelhart,S., Ciuffi,A., et al. (2016) Guanylate Binding Protein (GBP) 5 Is an Interferon-Inducible Inhibitor of HIV-1 Infectivity. Cell Host & Microbe, 19, 504–514.

34. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. BMC Bioinformatics, 10, 421.

35. Rice,P. (2000) EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics, 16, 276–277.

36. Ranwez,V., Harispe,S., Delsuc,F. and Douzery,E.J.P. (2011) MACSE: Multiple Alignment of Coding SEquences Accounting for Frameshifts and Stop Codons. PLoS ONE, 6, e22594.

37. Schneider,A., Souvorov,A., Sabath,N., Landan,G., Gonnet,G.H. and Graur,D. (2009) Estimates of Positive Darwinian Selection Are Inflated by Errors in Sequencing, Annotation, and Alignment. Genome Biology and Evolution, 1, 114–118.

38. Katoh,K., Misawa,K., Kuma,K. and Miyata,T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res., 30, 3059–3066. 39. Guindon,S., Dufayard,J.-F., Lefort,V., Anisimova,M., Hordijk,W. and Gascuel,O. (2010) New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Systematic Biology, 59, 307–321.

40. Anisimova,M. and Gascuel,O. (2006) Approximate Likelihood-Ratio Test for Branches: A Fast, Accurate, and Powerful Alternative. Systematic Biology, 55, 539–552.

41. Comte,N., Morel,B., Hasic,D., Guéguen,L., Boussau,B., Daubin,V., Penel,S., Scornavacca,C., Gouy,M., Stamatakis,A., et al. (2019) Treerecs: an integrated phylogenetic tool, from sequences to reconciliations Bioinformatics.

42. Anisimova,M., Bielawski,J.P. and Yang,Z. (2002) Accuracy and Power of Bayes Prediction of Amino Acid Sites Under Positive Selection. Molecular Biology and Evolution, 19, 950–958.

43. Kosakovsky Pond,S.L., Posada,D., Gravenor,M.B., Woelk,C.H. and Frost,S.D.W. (2006) GARD: a genetic algorithm for recombination detection. Bioinformatics, 22, 3096–3098.

44. Guéguen,L., Gaillard,S., Boussau,B., Gouy,M., Groussin,M., Rochette,N.C., Bigot,T., Fournier,D., Pouyet,F., Cahais,V., et al. (2013) Bio++: Efficient Extensible Libraries and Tools for Computational Molecular Evolution. Molecular Biology and Evolution, 30, 1745–1750.

45. Murrell,B., Weaver,S., Smith,M.D., Wertheim,J.O., Murrell,S., Aylward,A., Eren,K., Pollner,T., Martin,D.P., Smith,D.M., et al. (2015) Gene-Wide Identification of Episodic Selection. Molecular Biology and Evolution, 32, 1365–1371.

46. Murrell,B., Wertheim,J.O., Moola,S., Weighill,T., Scheffler,K. and Kosakovsky Pond,S.L. (2012) Detecting Individual Sites Subject to Episodic Diversifying Selection. PLoS Genetics, 8, e1002764.

47. Guéguen,L. and Duret,L. (2018) Unbiased Estimate of Synonymous and Nonsynonymous Substitution Rates with Nonstationary Base Composition. Molecular Biology and Evolution, 35, 734–742.

48. Murrell,B., Vollbrecht,T., Guatelli,J. and Wertheim,J.O. (2016) The Evolutionary Histories of Antiretroviral Proteins SERINC3 and SERINC5 Do Not Support an Evolutionary Arms Race in Primates. Journal of Virology, 90, 8085–8089.

49. Nakano,Y., Aso,H., Soper,A., Yamada,E., Moriwaki,M., Juarez-Fernandez,G., Koyanagi,Y. and Sato,K. (2017) A conflict of interest: the evolutionary arms race between mammalian APOBEC3 and lentiviral Vif. Retrovirology, 14, 31.

50. Etienne,L., Bibollet-Ruche,F., Sudmant,P.H., Wu,L.I., Hahn,B.H. and Emerman,M. (2015) The Role of the Antiviral APOBEC3 Gene Family in Protecting Chimpanzees against Lentiviruses from Monkeys. PLOS Pathogens, 11, e1005149.

51. Desimmie,B.A., Delviks-Frankenberrry,K.A., Burdick,R.C., Qi,D., Izumi,T. and Pathak,V.K. (2014) Multiple APOBEC3 Restriction Factors for HIV-1 and One Vif to Rule Them All. Journal of Molecular Biology, 426, 1220–1245.

52. Sawyer,S.L., Emerman,M. and Malik,H.S. (2004) Ancient Adaptive Evolution of the Primate Antiviral DNA-Editing Enzyme APOBEC3G. PLoS Biol, 2, e275.

53. Kluge,S.F., Sauter,D. and Kirchhoff,F. (2015) SnapShot: antiviral restriction factors. Cell, 163, 774–e1.

54. Woods,M., Tong,J., Tom,S., Szabo,P., Cavanagh,P., Dikeakos,J., Haeryfar,S. and Barr,S. (2014) Interferon-induced HERC5 is evolving under positive selection and inhibits HIV-1 particle production by a novel mechanism targeting Rev/RRE-dependent RNA nuclear export. Retrovirology, 11, 27.

55. Perelman,P., Johnson,W.E., Roos,C., Seuánez,H.N., Horvath,J.E., Moreira,M.A.M., Kessing,B., Pontius,J., Roelke,M., Rumpler,Y., et al. (2011) A Molecular Phylogeny of Living Primates. PLoS Genetics, 7, e1001342.

56. Pecon-Slattery,J. (2014) Recent Advances in Primate Phylogenomics. Annu. Rev. Anim. Biosci., 2, 41–63.

57. Lahaye,X., Gentili,M., Silvin,A., Conrad,C., Picard,L., Jouve,M., Zueva,E., Maurin,M., Nadalin,F., Knott,G.J., et al. (2018) NONO Detects the Nuclear HIV Capsid to Promote cGAS-Mediated Innate Immune Activation. Cell, 175, 488-501.e22.

58. Altenhoff,A.M., Glover,N.M., Train,C.-M., Kaleb,K., Warwick Vesztrocy,A., Dylus,D., de Farias,T.M., Zile,K., Stevenson,C., Long,J., et al. (2018) The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. Nucleic Acids Research, 46, D477–D485.

59. Huerta-Cepas,J., Szklarczyk,D., Forslund,K., Cook,H., Heller,D., Walter,M.C., Rattei,T., Mende,D.R., Sunagawa,S., Kuhn,M., et al. (2016) eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res, 44, D286–D293.

60. Haller,O., Staeheli,P., Schwemmle,M. and Kochs,G. (2015) Mx GTPases: dynamin-like antiviral machines of innate immunity. Trends in Microbiology, 23, 154–163.

61. Tretina,K., Park,E.-S., Maminska,A. and MacMicking,J.D. (2019) Interferon-induced guanylate-binding proteins: Guardians of host defense in health and disease. Journal of Experimental Medicine, 216, 482–500.

62. Huang,S., Meng,Q., Maminska,A. and MacMicking,J.D. (2019) Cell-autonomous immunity by IFN-induced GBPs in animals and plants. Current Opinion in Immunology, 60, 71–80.

63. Münk,C., Willemsen,A. and Bravo,I.G. (2012) An ancient history of gene duplications, fusions and losses in the evolution of APOBEC3 mutators in mammals. BMC Evol Biol, 12, 71.

64. Anisimova,M., Nielsen,R. and Yang,Z. (2003) Effect of Recombination on the Accuracy of the Likelihood Method for Detecting Positive Selection at Amino Acid Sites. Genetics, 164, 1229–1236.

65. Posada,D. and Crandall,K.A. (2002) The Effect of Recombination on the Accuracy of Phylogeny Estimation. Journal of Molecular Evolution, 54, 396–402.

66. Mitchell,P.S., Young,J.M., Emerman,M. and Malik,H.S. (2015) Evolutionary Analyses Suggest a Function of MxB Immunity Proteins Beyond Lentivirus Restriction. PLoS Pathog, 11, e1005304.

67. Weber,C.C. and Whelan,S. (2019) Physicochemical Amino Acid Properties Better Describe Substitution Rates in Large Populations. Molecular Biology and Evolution, 36, 679–690.

68. Zaheri,M., Dib,L. and Salamin,N. (2014) A Generalized Mechanistic Codon Model. Molecular Biology and Evolution, 31, 2528–2541.

69. Fregoso,O.I., Ahn,J., Wang,C., Mehrens,J., Skowronski,J. and Emerman,M. (2013) Evolutionary Toggling of Vpx/Vpr Specificity Results in Divergent Recognition of the Restriction Factor SAMHD1. PLoS Pathogens, 9, e1003496.

70. Laguette,N., Rahm,N., Sobhian,B., Chable-Bessia,C., Münch,J., Snoeck,J., Sauter,D., Switzer,W.M., Heneine,W., Kirchhoff,F., et al. (2012) Evolutionary and Functional Analyses of the Interaction between the Myeloid Restriction Factor SAMHD1 and the Lentiviral Vpx Protein. Cell Host & Microbe, 11, 205–217.

71. Lim,E.S., Fregoso,O.I., McCoy,C.O., Matsen,F.A., Malik,H.S. and Emerman,M. (2012) The Ability of Primate Lentiviruses to Degrade the Monocyte Restriction Factor SAMHD1 Preceded the Birth of the Viral Accessory Protein Vpx. Cell Host & Microbe, 11, 194–204.

72. Kerns,J.A., Emerman,M. and Malik,H.S. (2008) Positive Selection and Increased Antiviral Activity Associated with the PARP-Containing Isoform of Human Zinc-Finger Antiviral Protein. PLoS Genetics, 4, e21.

73. Mitchell,P.S., Patzina,C., Emerman,M., Haller,O., Malik,H.S. and Kochs,G. (2012) Evolution-Guided Identification of Antiviral Specificity Determinants in the Broadly Acting Interferon-Induced Innate Immunity Factor MxA. Cell Host & Microbe, 12, 598–604.

74. McLaren,P.J., Gawanbacht,A., Pyndiah,N., Krapp,C., Hotter,D., Kluge,S.F., Götz,N., Heilmann,J., Mack,K., Sauter,D., et al. (2015) Identification of potential HIV restriction factors by combining evolutionary genomic signatures with functional analyses. Retrovirology, 12.

75. Cagliani,R., Forni,D., Biasin,M., Comabella,M., Guerini,F.R., Riva,S., Pozzoli,U., Agliardi,C., Caputo,D., Malhotra,S., et al. (2014) Ancient and Recent Selective Pressures Shaped Genetic Diversity at AIM2-Like Nucleic Acid Sensors. Genome Biology and Evolution, 6, 830–845.

76. Lim,E.S., Wu,L.I., Malik,H.S. and Emerman,M. (2012) The function and evolution of the restriction factor viperin in primates was not driven by lentiviruses. Retrovirology, 9, 55.

77. Lim,E.S., Fregoso,O.I., McCoy,C.O., Matsen,F.A., Malik,H.S. and Emerman,M. (2012) The Ability of Primate Lentiviruses to Degrade the Monocyte Restriction Factor SAMHD1 Preceded the Birth of the Viral Accessory Protein Vpx. Cell Host & Microbe, 11, 194–204.

78. Atkinson,E.G., Audesse,A.J., Palacios,J.A., Bobo,D.M., Webb,A.E., Ramachandran,S. and Henn,B.M. (2018) No Evidence for Recent Selection at FOXP2 among Diverse Human Populations. Cell, 174, 1424-1435.e15.

Supplementary Information

Supplementary Table 1. Log likelihoods calculated by BIO++ and PAML codeml for each of the different models.

File	Method	M1	M2	M7	M8
APOBEC3F	BPP	-2507,37	-2491,43	-2511,01	-2491,47
	PAML	-3945,63	-3930,68	-3945,88	-3930,68
FOXP2	BPP	-3997,30	-3995,85	-3999,96	-3995,98
	PAML	-5860,15	-5858,43	-5861,65	-5858,44
GADD45A	BPP	-1239,99	-1239,99	-1241,28	-1238,69
	PAML	-1291,75	-1291,75	-1292,06	-1289,64
GBP5	BPP	-6138,60	-6131,37	-6139,83	-6131,78
	PAML	-6516,42	-6504,84	-6518,34	-6505,90
GMPR	BPP	-3027,30	-3027,30	-3027,12	-3025,84
	PAML	-3425,82	-3425,82	-3423,78	-3423,53
HERC5	BPP	-10433,56	-10430,06	-10434,33	-10430,37
	PAML	-11950,25	-11945,50	-11951,66	-11946,35
IFI16	BPP	-11099,74	-11075,12	-11103,87	-11075,78
	PAML	-18264,89	-18223,86	-18270,40	-18225,35
ISG20	BPP	-1837,14	-1837,14	-1836,69	-1836,69
	PAML	-3048,57	-3048,57	-3047,88	-3047,42
MX1	BPP	-8537,86	-8532,47	-8543,10	-8531,57
	PAML	-11329,45	-11316,93	-11335,61	-11317,74
NT5C3A	BPP	-2324,85	-2324,85	-2324,19	-2323,64
	PAML	-4483,43	-4483,43	-4482,88	-4482,16
RB1	BPP	-6873,89	-6873,89	-6874,36	-6872,91
	PAML	-7244,38	-7244,38	-7242,93	-7242,93
RHO	BPP	-2910,93	-2910,93	-2913,24	-2908,65
	PAML	-2967,07	-2967,07	-2953,69	-2953,12
RSAD2	BPP	-4248,12	-4248,12	-4248,75	-4243,00
	PAML	-4875,51	-4868,59	-4874,28	-4865,51
SAMHD1	BPP	-7522,22	-7495,71	-7534,66	-7497,30
	PAML	-8219,63	-8186,06	-8225,38	-8187,80
SERINC3	BPP	-4751,87	-4749,23	-4755,82	-4749,69
	PAML	-5373,70	-5370,57	-5376,40	-5371,08
SHH	BPP	-3576,79	-3576,79	-3580,07	-3575,89
	PAML	-4855,15	-4841,25	-4853,34	-4837,51
SMC6	BPP	-8971,83	-8971,83	-8974,05	-8970,65
	PAML	-12342,77	-12342,77	-12341,63	-12336,91
TREX1	BPP	-3421,42	-3421,42	-3421,35	-3421,15
	PAML	-5080,28	-5080,27	-5080,76	-5080,15
ZC3HAV1	BIO++	-12413,91	-12384,63	-12418,39	-12392,87
	PAML	-17617,28	-17585,61	-17619,80	-17585,77

Chapter 2

Host and virus evolutionary analyses of NONO, a sensor of HIV capsid

2.1 Introduction to the paper

During my PhD, we collaborated with the Manel Lab from Institut Curie in Paris to decipher the evolutionary determinants of the NONO-Capsid interactions. They had identified their gene of interest, NONO, as interacting with the HIV capsid and promoting the innate immune response by participating in the activation of the cGAS pathway. I performed the evolutionary analyses on both host side (NONO) and virus side (capsid) in the manner described hereafter.

2.2 Material and methods

Host phylogenetic and evolutionary analyses

Orthologous sequences of the primate NONO gene were retrieved from publically available databases using UCSC Blat and NCBI Blastn with the human sequence as the query. Sequences annotated as non-coding (XR/NR identification in the NCBI databases) were discarded. We confirmed the conserved synteny of the selected orthologous genes using UCSC and NCBI. In total, the orthologous sequences of 20 primate species were included and codon-aligned using PRANK v150803 with the default parameter -F (Löytynoja, 2014). We used GARD from HYPHY to assess for recombination with a cut-off at $p<0.05$ (Pond et al., 2005, Kosakovsky-Pond et al., 2006). PhyML v3.2 was used for

the phylogenetic reconstructions with a HKY+G+I model and 1,000 bootstrap replicates for statistical support of the branches (Guindon et al., 2010).

Marks of positive selection were assessed using maximum-likelihood tests performed by three softwares: HYPHY (Pond et al., 2005), PAML Codeml (Yang et al., 2010), and Bio++ (Guéguen et al., 2013, Guéguen and Duret, 2018). In HYPHY, we used the BUSTED method to detect gene-wide evidence of positive selection within a codon alignment (Murrel et al., 2015). In PAML Codeml and Bio++, we used the NONO gene tree inferred with PhyML as input. The gene-coding sequence alignments were fit to models that disallow (M1, two classes of ω (ω = dN/dS): $\omega < 1$ and $\omega = 1$) or allow (M2, three classes: $\omega < 1$, $\omega = 1$, and $\omega > 1$) for positive selection (Yang et al., 2007). The likelihood of the models was compared using a chi-squared test to derive p-values. To detect (episodic) site-specific positive selection, we used MEME and FUBAR from HYPHY (Murrell et al., 2012, Murrell et al., 2013), the BEB (Bayes Empirical Bayes) analysis from the M2 model in PAML Codeml, and the Bayesian Posterior Probabilities (PP) from the M2 model in Bio++.

Virus phylogenetic analyses

The nucleotide sequences of the gag gene from primate lentiviruses were retrieved from the Los Alamos HIV sequence database curated alignments (link) on June, 24th 2018. Three alignments were recovered: the first one with "HIV-1/SIVcpz" sequences, the second one with "HIV-2/SIVsmm" sequences, and the third one with all SIV and some HIV sequences (named "Other SIV (includes HIV1 and HIV2 sequences)" in the database).

Each alignment was then ungapped, translated and aligned using MAFFT (Katoh et al., 2002). In the obtained amino-acid alignments, positions which had gaps in more than 10% of the sequences were trimmed using SeqMagick (from Frederick Matsen group). The NONO-binding motifs (Figure 16) were subsequently isolated and separated according to the lentiviral lineage/group. A sequence logo was then produced for each motif and each lineage/group of viruses using WebLogo3.

To reconstruct the lentiviral phylogeny of gag, we used the nucleotide sequences from the third primate lentiviral alignment (total, 202 HIV and SIV sequences) and we re-aligned them using PRANK with the default parameter -F (Löytynoja, 2014). We discarded the following sequences that were poorly aligned: SMM.US.04.M934.JX860421, H2U.FR.96.12034.AY530889 and SAB.SN.x.SAB1.U04005. We used GARD from HY-

PHY to assess for recombination with a cut-off at p¡0.05 and found no evidence of significant recombination in this alignment. We reconstructed the phylogeny using PhyML v3.2 with a GTR+G+I model (best model according to Smart Model Selection, SMS – Lefort et al., 2017) and aLRT for branch support (Guindon et al., 2010). The clade of SIVcol sequences was used as outgroup to root the tree, in accordance with accepted lentiviral phylogenies (Bell and Bedford 2017, Gifford et al 2008). Branches were collapsed using FigTree v1.4.3.

2.3 Figures

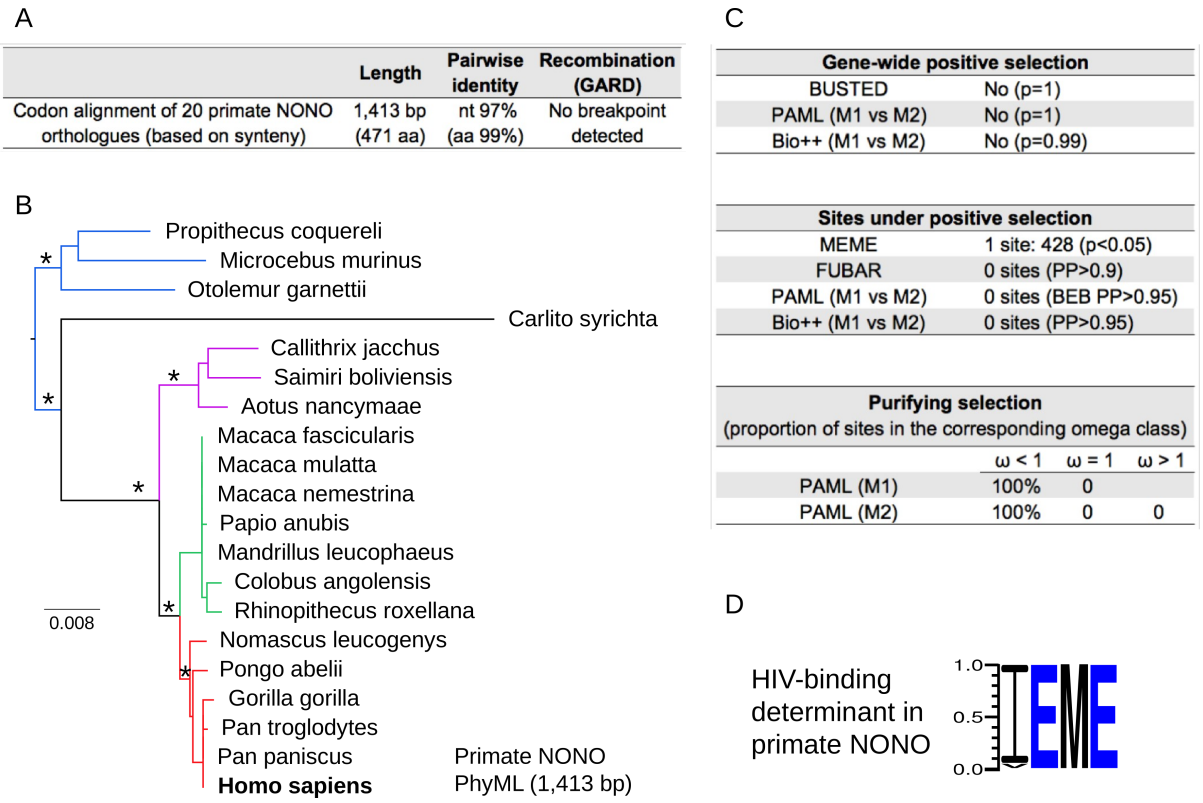


Figure 15: NONO and the "IEME" CA-binding determinant have been highly conserved during primate evolution.

Phylogenetic analyses of primate NONO were performed on twenty orthologous nucleotide sequences of primate NONO that were aligned with PRANK. **A**, Pairwise identity was computed in Geneious, Biomatters. Recombination analysis was performed with GARD. **B**, Phylogeny was performed with PhyML with an HKY+G+I model and 1,000 bootstrap replicates as statistical support. Only bootstrap values above 900/1,000 are shown here by an asterisk above the branches. The scale bar indicates the number of nucleotide substitutions per site. **C**, Positive selection analyses show that NONO has been under purifying selection during primate evolution. Top panel shows the results of three "gene-wide" positive selection analyses (i.e. HYPHY BUSTED, PAML Codeml, Bio++) with the p-values from the maximum-likelihood ratio tests (LRT) indicating whether the model that allows positive selection better fits the data. Middle panel shows results from "site-specific" positive selection analyses (i.e. HYPHY MEME and FUBAR, PAML Codeml, Bio++). Statistical thresholds for significance are indicated (PP, posterior probabilities; p, p-value). BEB, Bayes Empirical Bayes. See Methods for details. Lower panel shows the proportion of sites falling into each omega class (omega = dN/dS) computed in PAML Codeml. Both M1 and M2 models show that all sites of primate NONO have a dN/dS < 1, reflecting purifying selection. **D**, Sequence logo (WebLogo3) of the lentivirus capsid-binding region on NONO showing that the motif is highly conserved in primates, with only tarsier (Carlito syrichta) harboring an I to V amino acid change at the first position. **[Corresponds to Figure S6D-G in the paper]**

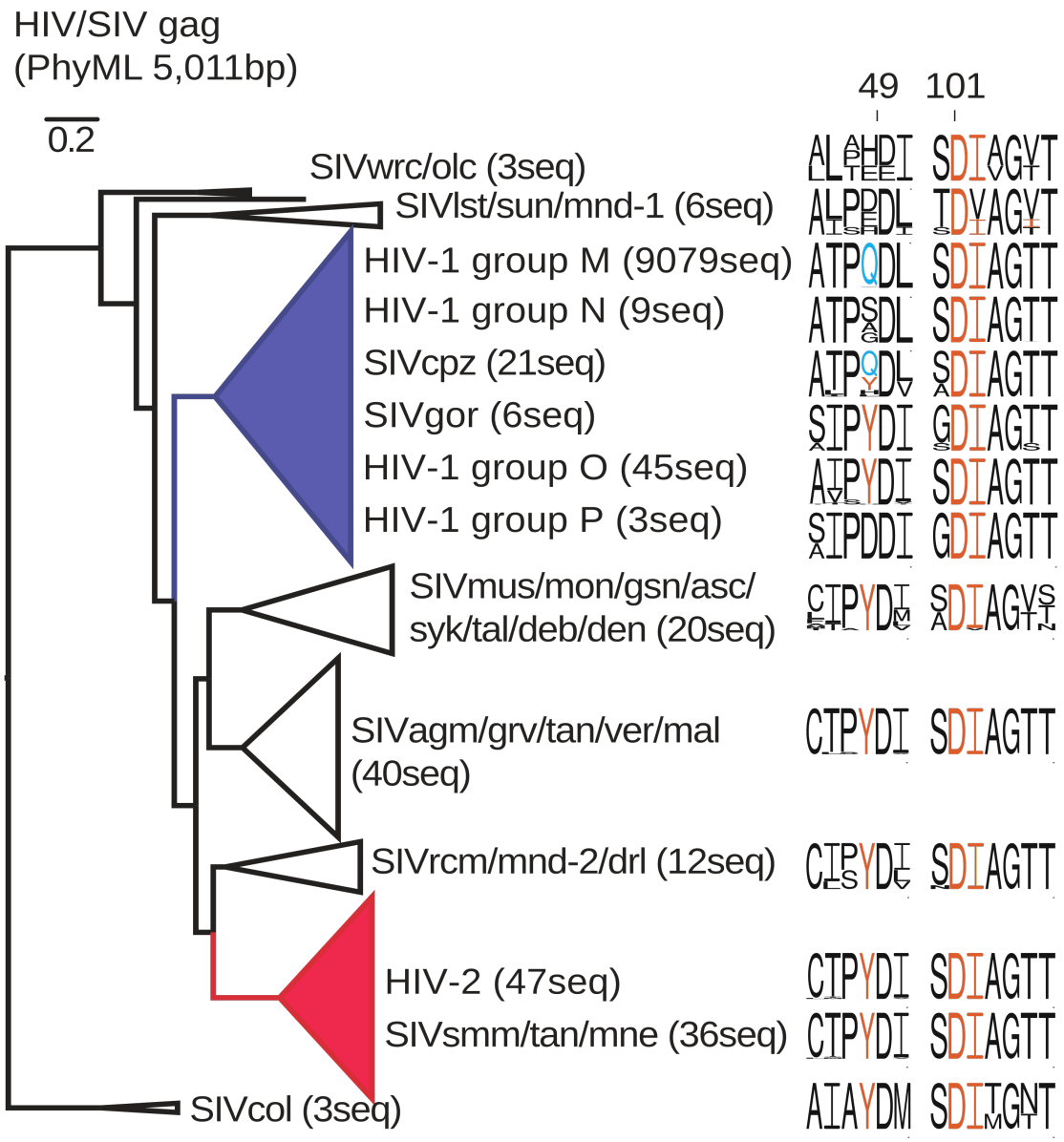


Figure 16: NONO-binding determinants in the viral capsid across primate lentiviruses.

Phylogenetic analysis of primate lentiviruses was performed on 202 nucleotide sequences that were aligned with PRANK and tree reconstruction (on the left) was performed with PhyML with an GTR+G+I model and aLRT as statistical support. The composition of the amino-acids corresponding to those involved in the NONO-HIV CA interaction is represented as sequence logos (WebLogo3), for each primate lentiviral lineage/group. The number of sequences in each lineage/group used to compute the sequence logos are indicated (seq, sequences). The clade that includes HIV-1 and HIV-2 strains are highlighted in blue and red, respectively. The color-coding in the logos are as default in WebLogo3, except for the Q50, which is solely found in HIV-1 group M and SIVcpz and is highlighted in orange. [**Corresponds to Figure 6K in the paper**]

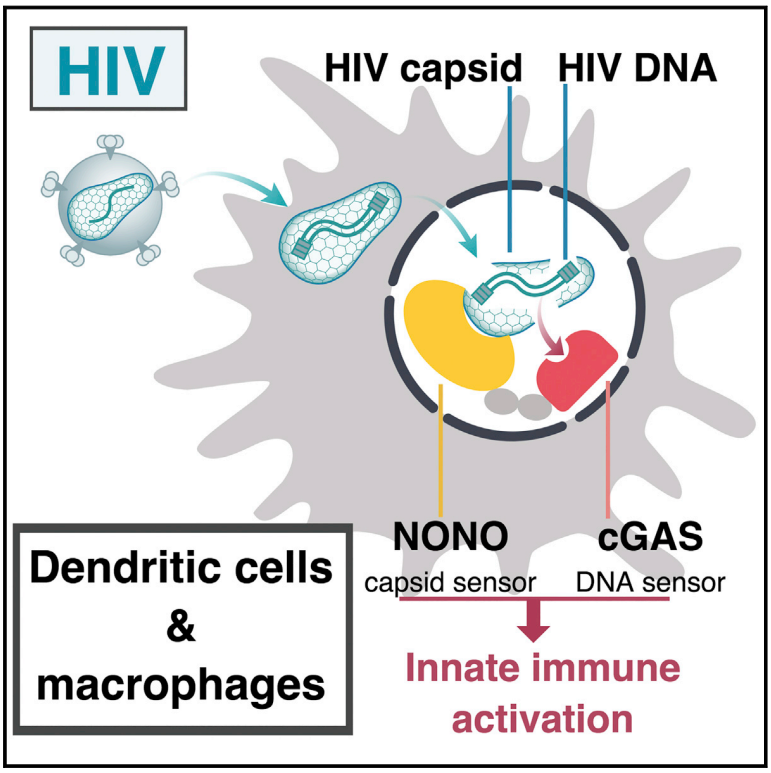
2.4 Paper

Article

Cell

NONO Detects the Nuclear HIV Capsid to Promote cGAS-Mediated Innate Immune Activation

Graphical Abstract



Authors

Xavier Lahaye, Matteo Gentili, Aymeric Silvin, ..., Charles S. Bond, Laurence Colleaux, Nicolas Manel

Correspondence

nicolas.manel@curie.fr

In Brief

The cellular factor NONO activates cGAS-mediated innate immune defenses against HIV-2 infection via viral capsid binding.

Highlights

- HIV-2, not HIV-1, activates innate immunity in macrophages and dendritic cells
- NONO protein binds to the HIV-2 capsid protein with more affinity than HIV-1
- NONO is an innate immune sensor of the HIV capsid in the nucleus
- NONO associates with the sensor cGAS in the nucleus and enables sensing of HIV DNA

Lahaye et al., 2018, Cell *175*, 488–501
October 4, 2018 © 2018 Elsevier Inc.
https://doi.org/10.1016/j.cell.2018.08.062

CellPress

Cell

Article

NONO Detects the Nuclear HIV Capsid to Promote cGAS-Mediated Innate Immune Activation

Xavier Lahaye,[1] Matteo Gentili,[1] Aymeric Silvin,[1] Cécile Conrad,[1] Léa Picard,[2,3] Mabel Jouve,[1] Elina Zueva,[1] Mathieu Maurin,[1] Francesca Nadalin,[1] Gavin J. Knott,[4] Baoyu Zhao,[5] Fenglei Du,[5] Marlène Rio,[6,7] Jeanne Amiel,[6,7] Archa H. Fox,[4,8,9] Pingwei Li,[5] Lucie Etienne,[2] Charles S. Bond,[4] Laurence Colleaux,[6,7] and Nicolas Manel[1,10,*]

[1]Immunity and Cancer Department, Institut Curie, PSL Research University, INSERM U932, 75005 Paris, France
[2]CIRI-International Center for Infectiology Research, Inserm U1111, CNRS UMR5308, Ecole Normale Supérieure de Lyon, Université Claude Bernard Lyon 1, Univ Lyon, 69007 Lyon, France
[3]LBBE-Laboratoire de Biométrie et Biologie Evolutive CNRS UMR 5558, Universite Lyon 1, Univ Lyon, 69622 Villeurbanne, France
[4]School of Molecular Sciences, The University of Western Australia, Crawley, WA 6009, Australia
[5]Department of Biochemistry and Biophysics, Texas A&M University, College Station, TX 77843, USA
[6]INSERM UMR 1163, Paris-Descartes-Sorbonne Paris Cité University, Institut IMAGINE, Necker-Enfants Malades Hospital, 75015 Paris, France
[7]Service de Génétique, Hôpital Necker-Enfants Malades, AP-HP, Paris, France
[8]School of Human Sciences, The University of Western Australia, Crawley, WA 6009, Australia
[9]The Harry Perkins Institute of Medical Research, QEII Medical Centre, Nedlands, WA 6009, Australia
[10]Lead Contact
*Correspondence: nicolas.manel@curie.fr
https://doi.org/10.1016/j.cell.2018.08.062

SUMMARY

Detection of viruses by innate immune sensors induces protective antiviral immunity. The viral DNA sensor cyclic GMP-AMP synthase (cGAS) is necessary for detection of HIV by human dendritic cells and macrophages. However, synthesis of HIV DNA during infection is not sufficient for immune activation. The capsid protein, which associates with viral DNA, has a pivotal role in enabling cGAS-mediated immune activation. We now find that NONO is an essential sensor of the HIV capsid in the nucleus. NONO protein directly binds capsid with higher affinity for weakly pathogenic HIV-2 than highly pathogenic HIV-1. Upon infection, NONO is essential for cGAS activation by HIV and cGAS association with HIV DNA in the nucleus. NONO recognizes a conserved region in HIV capsid with limited tolerance for escape mutations. Detection of nuclear viral capsid by NONO to promote DNA sensing by cGAS reveals an innate strategy to achieve distinction of viruses from self in the nucleus.

INTRODUCTION

In vertebrates, recognition of viruses by dendritic cells (DCs) and macrophages is critical to induce an activated state that stimulates innate and adaptive immune responses. Sensing of viruses by DCs and macrophages relies on the ability to recognize specific elements that are associated with the virus and absent from cells. A limited number of nucleic acid sensors have been identified (Vance, 2016). The cytosolic DNA sensor cyclic GMP-AMP synthase (cGAS) is critical for the activation of DCs and macrophages via IRF3 in response to several DNA viruses and the lentivirus HIV (Gao et al., 2013; Lahaye et al., 2013). In addition, cGAS appears to play a much broader role in response to a large number of microbes and self DNA (Chen et al., 2016). The universal nature of this nucleic acid-based recognition raises the question of how DCs and macrophages achieve their high level of sensitivity while maintaining sufficient specificity in the recognition of viral infection. This suggests that additional mechanisms may exist to control cGAS-mediated sensing in the case of viral infection.

HIV-2 infects 1–2 million individuals (Visseaux et al., 2016). The majority of HIV-2 infected individuals do not progress to AIDS, control viral replication, induce a potent immune response against the virus, and exhibit partial cross-protection against HIV-1 (Esbjörnsson et al., 2012; Rowland-Jones and Whittle, 2007). HIV-2 degrades the restriction factor SAMHD1 through its Vpx protein, leading to efficient viral DNA synthesis and innate immune activation in DCs through cGAS. In contrast, DCs neglect sensing of HIV-1 through cGAS because the virus does not degrade SAMHD1 (Sáez-Cirión and Manel, 2018). Depletion of SAMHD1 using HIV-2/SIVmac Vpx protein sensitizes DCs for HIV-1 infection and restores innate immune activation in response to the virus (Manel et al., 2010). Viral infection and recognition in DCs is thus likely one of several key processes that triggers protective immune responses during the course of HIV-2 infection. However, Vpx is simultaneously required for HIV-2/SIVmac viral replication in T cells (Shingai et al., 2015; Yu et al., 2013), rendering the identification of other factors of innate sensing desirable.

HIV double-stranded DNA (dsDNA) plays an essential role in cGAS-mediated recognition (Herzner et al., 2015; Yoh et al., 2015) but it is not sufficient in DCs (Lahaye et al., 2013). The viral capsid, which associates with viral dsDNA up to the nucleus interior (Chin et al., 2015; Peng et al., 2014), has a critical role in enabling cGAS-mediated sensing of HIV dsDNA. The HIV-2

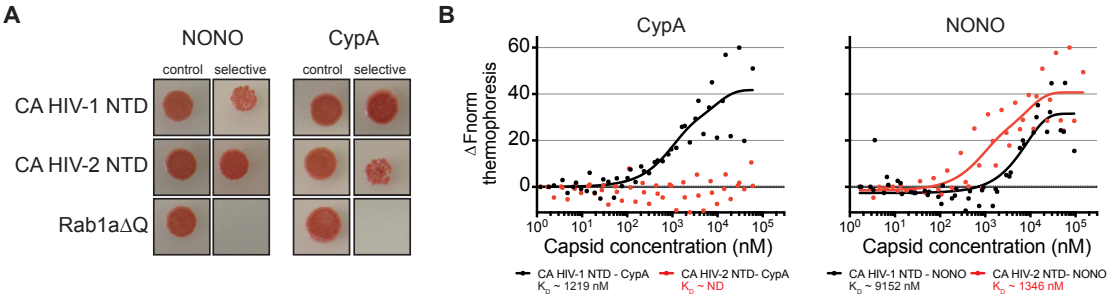


Figure 1. NONO Directly Binds to HIV-1 and HIV-2 Capsid Proteins

(A) Interactions between the N-terminal domains (NTD) of the capsid (CA) proteins of HIV-1 or HIV-2 or a negative control protein Rab1aΔQ with NONO and CypA, measured by Y2H (n = 6, one representative experiment is shown).

(B) Interaction between CypA, NONO, and recombinant NTD of HIV-1 and HIV-2 capsids, measured by MST (n = 3).

See also Figure S1.

capsid is permissive to cGAS-mediated sensing during the early steps of infection, before the incoming viral DNA integrates. In contrast, the HIV-1 capsid evades sensing of the dsDNA before integration, even if Vpx is provided, and viral integration and expression of newly synthesized Gag protein is required to activate innate immunity. Capsid mutations alter HIV-1/2 DNA sensing by cGAS, implicating interactions with host factors (Lahaye et al., 2013; Manel et al., 2010; Rasaiyaah et al., 2013). Cyclophilin A (CypA) binds to the HIV-1/2 capsid and modulates recognition of HIV-1/2 by DCs and macrophages, but the outcome of this interaction is virus strain- and cell-type-specific. These observations suggested the existence of a capsid-binding factor that would be essential for cGAS-mediated recognition of HIV-1/2.

RESULTS

NONO Directly Binds to HIV Capsid Protein

Given the physio-pathological differences between HIV-1 and HIV-2 infections, we reasoned that host factor(s) implicated in viral capsid recognition would preferentially bind to the HIV-2 capsid. We performed a yeast two-hybrid (Y2H) screen for protein fragments interacting with the HIV-2 capsid. NONO was identified with the best confidence score (Figure S1A). The NONO clones coded for fragments that all contained domain 256–310, representing the HIV-2 capsid-binding domain. Full-length NONO interacted with full-length and N-terminal domain (NTD) of HIV-2 capsid and the NTD of HIV-1 capsid (Figures 1A, S1B, and S1C). We compared the strengths of the interaction with CypA as control. As predicted, the two capsids interacted with CypA, and the interaction with HIV-1 capsid was stronger than observed with HIV-2 capsid (Figures 1A and S1C). In contrast, the HIV-2 capsid interacted more strongly with NONO than the HIV-1 capsid. To determine direct protein-protein interactions, we expressed and purified recombinant CypA, a NONO fragment that produces a soluble protein (domain 35–312) (Knott et al., 2016b) and capsid NTDs (Figure S1D). Microscale thermophoresis (MST) confirmed a direct binding between NONO and the capsids NTD of HIV-1 or HIV-2. As expected in this assay (Lahaye et al., 2013), the capsid NTD of HIV-1 binds to CypA while HIV-2 binding is below the detection limit (Figure 1B). In contrast, capsid NTD affinity of HIV-2 for NONO was higher than HIV-1 capsid NTD. Using MST, the full-length HIV-1 capsid bound NONO with the same affinity as the NTD (Figure S1E), suggesting that the lack of detectable interaction of full-length capsid in yeast is not due to a reduced affinity (Figure S1B).

NONO Is Not Essential for HIV Infection

The conserved interaction of NONO with the HIV-1 and HIV-2 capsids suggested that it could be a host factor required for HIV infection. Depletion of NONO with CRISPR/Cas9 in THP-1 cells (Figures S2A and S2B) and U87 cells (Figures S2C and S2D) or with RNAi in THP-1 cells (Figures S2E and S2F) and HeLa cells (Figures S2G and S2H) had no major impact on HIV-1 and HIV-2 early phases of replication. To address late phases, we generated NONO-deficient HEK293FT cells, and this had no impact on viral production (Figure S2I) and titer of viral progeny (Figures S2J and S2K). While Vpx was encapsidated in HIV-2 particles as expected, NONO was not packaged in HIV-1 or HIV-2 particles (Figure S2I). Thus, NONO is not a general host-dependency factor of HIV infection.

NONO Is Essential in Dendritic Cells and Macrophages for Immune Activation after HIV Infection

These results raised the alternative possibility that NONO could be implicated in the recognition of HIV by the innate immune system. To determine if NONO was implicated in HIV recognition, we depleted NONO in monocyte-derived DCs (MDDCs), using Vpx and small hairpin RNA (shRNA) that induced intermediate (shRNA#1) or near complete (shRNA#5) depletion, and utilized a previously validated cGAS shRNA as a control (Figure 2A). Similar to cGAS depletion, NONO depletion inhibited the expression of type I interferon (IFN), co-stimulatory molecule CD86, and inflammatory cytokine IP-10 (CXCL10) after infection by HIV-1 or HIV-2 GFP-reporter viruses (Figures 2B–2D and S3A). In contrast, NONO depletion did not impair DC activation by TLR3 agonist poly(I:C) or TLR7/8 agonist R848 (Figures 2D and S3C). Activation by HIV-1 and HIV-2 capsid mutants that favor DC activation over infection was also dependent on NONO

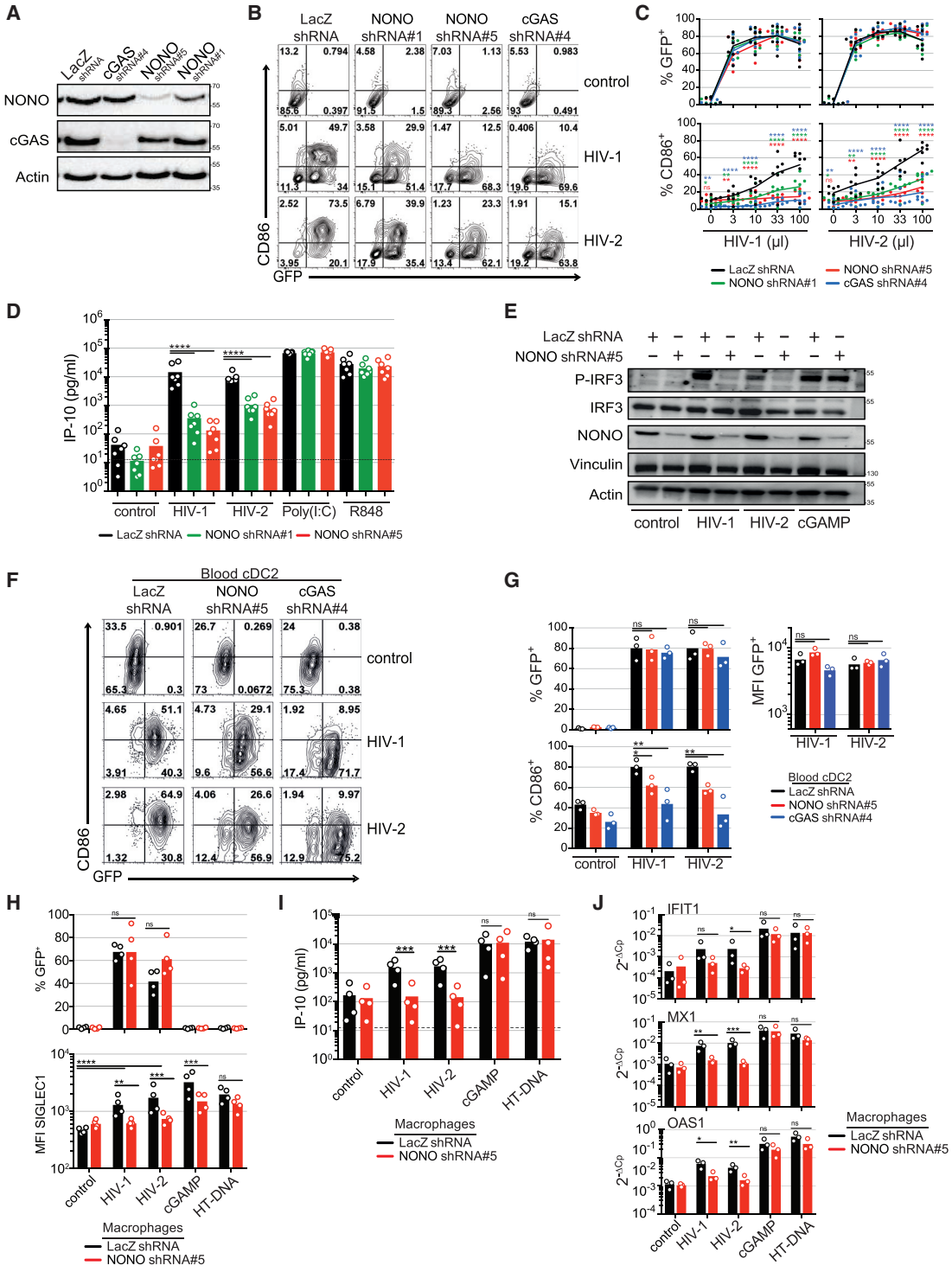


Figure 2. NONO Is Essential for Immune Activation after HIV Infection in Dendritic Cells and Macrophages

(A) Expression of NONO, cGAS, and actin in MDDCs at day 4, transduced at day 0 with a control shRNA against LacZ or individual shRNA against NONO or cGAS, combined with Vpx-containing VLPs (n = 7, one representative experiment is shown).

(B) GFP and CD86 expression in MDDCs as in (A) and infected at day 4 for 48 hr with HIV-1 or HIV-2 GFP-reporter viruses (n = 7, one representative experiment is shown). The shRNA transduction at day 0 includes Vpx, which abrogates the SAMHD1 restriction for HIV-1 at day 4.

(C) Dose-response expression of GFP and CD86 as in (B) (n = 7, paired repeated measure [RM] ANOVA). Virus inoculum volume is indicated.

(legend continued on next page)

expression, while activation by the cGAS agonist HT-DNA, or the cGAS product and stimulator of interferon genes (STING) agonist cyclic guanosine monophosphate-adenosine monophosphate (cGAMP), was not affected by NONO depletion (Figure S3D). Dose-titrations on multiple donors confirmed that the response to HIV required NONO expression, but not the response to cGAMP or HT-DNA (Figure S3D). Accordingly, NONO depletion did not inhibit cGAS messenger or protein expression (Figures 2A and S3E). Viral infection, monitored by GFP reporter expression, was not affected by NONO or cGAS depletion in MDDCs (Figures 2B, 2C, and S3B). The amount of reverse-transcribed DNA, nuclear HIV DNA, and integrated HIV were also not affected by NONO depletion, indicating that the lack of DC activation was not due to a limiting dose of DNA (Figure S3F). To confirm production of antiviral interferons, we measured the production of type I and type III IFN. IFN-β was induced by HIV-2 infection in a NONO-dependent manner (Figure S3G), while it was below the detection limit for HIV-1. We found that MDDCs secreted high levels of IFN-λ1, a type III IFN previously reported to be co-expressed with type I IFN in response to viral infection (Odendall et al., 2014) (Figure S3H). IFN-λ1 induction by HIV-1 and HIV-2 infection required NONO, but not its induction by cGAMP (Figure S3H). The induction of the IFN-stimulated gene (ISG) SIGLEC1 was also inhibited by NONO depletion after infection of DCs with a replication-competent HIV-1 (Figure S3I).

Type I and III interferons are IRF3 target genes and HIV recognition in DCs through cGAS leads to IRF3 phosphorylation and requires IRF3. IRF3 phosphorylation in response to HIV-1 or HIV-2 infection was abrogated in cells depleted for NONO (Figure 2E). The total level of IRF3 was not altered. Phosphorylation of IRF3 after transfection of cGAMP, the product of cGAS, was not altered in NONO-depleted DCs. Thus, NONO is required for IRF3 activation in DCs in response to HIV, and does not appear to play a general role in IRF3 activation through the cGAMP-STING pathway. To validate the role of NONO in primary cells, we purified primary cDC2 ($CD1c^+$ DCs) from human blood and depleted NONO and cGAS using lentiviral vectors (Figure S3J) (Silvin et al., 2017). HIV-1 or HIV-2 infection induced CD86 expression in cDC2, and this was inhibited in the absence of cGAS or NONO (Figures 2F, 2G, and S3K). The rate of viral infection was not significantly altered by the absence of cGAS or NONO (Figures 2F, 2G, and S3K).

Similar to DCs, HIV-1 can activate cGAS in monocyte-derived macrophages in the presence of Vpx (Gao et al., 2013). NONO depletion in macrophages did not significantly alter the rate of infection with HIV-1 or HIV-2 (Figure 2H). However, the induction of IP-10 protein was inhibited by NONO depletion in response to HIV-1 or HIV-2, but not in response to cGAMP or HT-DNA (Figure 2I). The expression of the ISGs IFIT1, MX1, and OAS1 was also inhibited by NONO depletion in response to infection, but not to stimulation with synthetic agonists (Figure 2J).

HIV Capsid Reaches the Nucleus of Dendritic Cells and Interacts with NONO

NONO is a pan-nuclear protein in DCs (Figure S4A). We thus tested if the incoming HIV-2 capsid protein localized to the nucleus of DCs. Using a staining protocol optimized for nuclear capsid detection (Chin et al., 2015) and integrase inhibitor to restrict detection to incoming viral capsids, we found that the incoming HIV-2 capsid reaches the interior of the nucleus upon infection at MOI of 2 (Figures 3A, 3B, and S4B). Nuclear/cytoplasmic fractionation analyses confirmed this finding and further showed that newly synthesized GAG can also be found in the DC nuclear fraction (Figures 3C and S4C). Immuno-electron microscopy also identified incoming nuclear capsid as intranuclear structures stained by anti-capsid antibody (Figures 3D and S4D). Their dimensions were consistent with the shape of viral cores (Figures 3D and S4D) but reduced compared to mature core of extracellular particles (Figure S4E) (Briggs et al., 2003). Proximity-ligation assay (PLA) also identified nuclear capsid, with an average of 11 foci per cell (Figures 3E and 3F). To exclude that nuclear capsid was newly expressed Gag due to leakiness of the integrase inhibitor, we infected DCs with HIV-2 deleted for Gag (ΔGag) complemented with encapsidation-signal deficient HIV-2 ($\Delta\Psi$). As expected, Gag was not expressed in DCs infected with HIV-2 ΔGag (Figure 3G, $\Delta\Psi$+ΔGag). Similar to HIV-2, HIV-2 ΔGag induced CD86, IP-10, and IFN in dendritic cells in the absence of integration and in a cGAS- and NONO-dependent manner (Figure 3H). The incoming capsid for HIV-2 ΔGag viruses was also found in the nucleus (Figure 3I). To test if the incoming viral capsid could interact with NONO in the nucleus, we performed double-labeled immuno-electron microscopy. Nuclear structures positive for capsid were also positive for NONO (Figure 3J). We analyzed 157 CA-positive nuclear structure and 60% contained at least one NONO gold particle (Figure 3K). The double-labeled structures were also observed on two sequential sections (Figure S4F). In contrast, NONO was not detected on the capsid from enveloped viral particles released by infected macrophages (Figure S4G). Altogether,

(D) IP-10 production by MDDCs transduced as in (A) and infected with HIV-1 or HIV-2 GFP-reporter viruses (33 μL) or treated with poly(I:C) (1.33 μg/mL) or R848 (0.33 μg/mL) for 48 hr (n = 7, paired RM ANOVA on log-transformed data).
(E) Western blot of phospho-Ser396-IRF3 in MDDCs transduced as in (A) and infected 16 hr with HIV-1 or HIV-2 GFP-reporter viruses or transfected 12 hr with cGAMP (1.33 μg/mL) (n = 3, one representative experiment is shown).
(F) GFP and CD86 expression in cDC2 transduced and infected as in (B) (n = 5, one representative experiment is shown).
(G) Expression of GFP and CD86 as in (F) (n = 3, paired RM ANOVA, virus inoculum: 33 μL).
(H) GFP and SIGLEC1 expression in macrophages transduced and infected at day 9 as in (B) or treated with cGAMP (1.33 μg/mL) or HT-DNA (1.66 μg/mL) (n = 4, paired RM ANOVA).
(I) IP-10 production by macrophages as in (H) (n = 4, paired RM ANOVA on log-transformed data).
(J) Relative expression of IFIT1, MX1, OAS1 by real-time qPCR by macrophages as in (H) (n = 4, paired RM ANOVA on log-transformed data).
*p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001; lines and bars indicate mean; ns, not statistically significant; MFI, mean fluorescence intensity.
See also Figures S2 and S3.

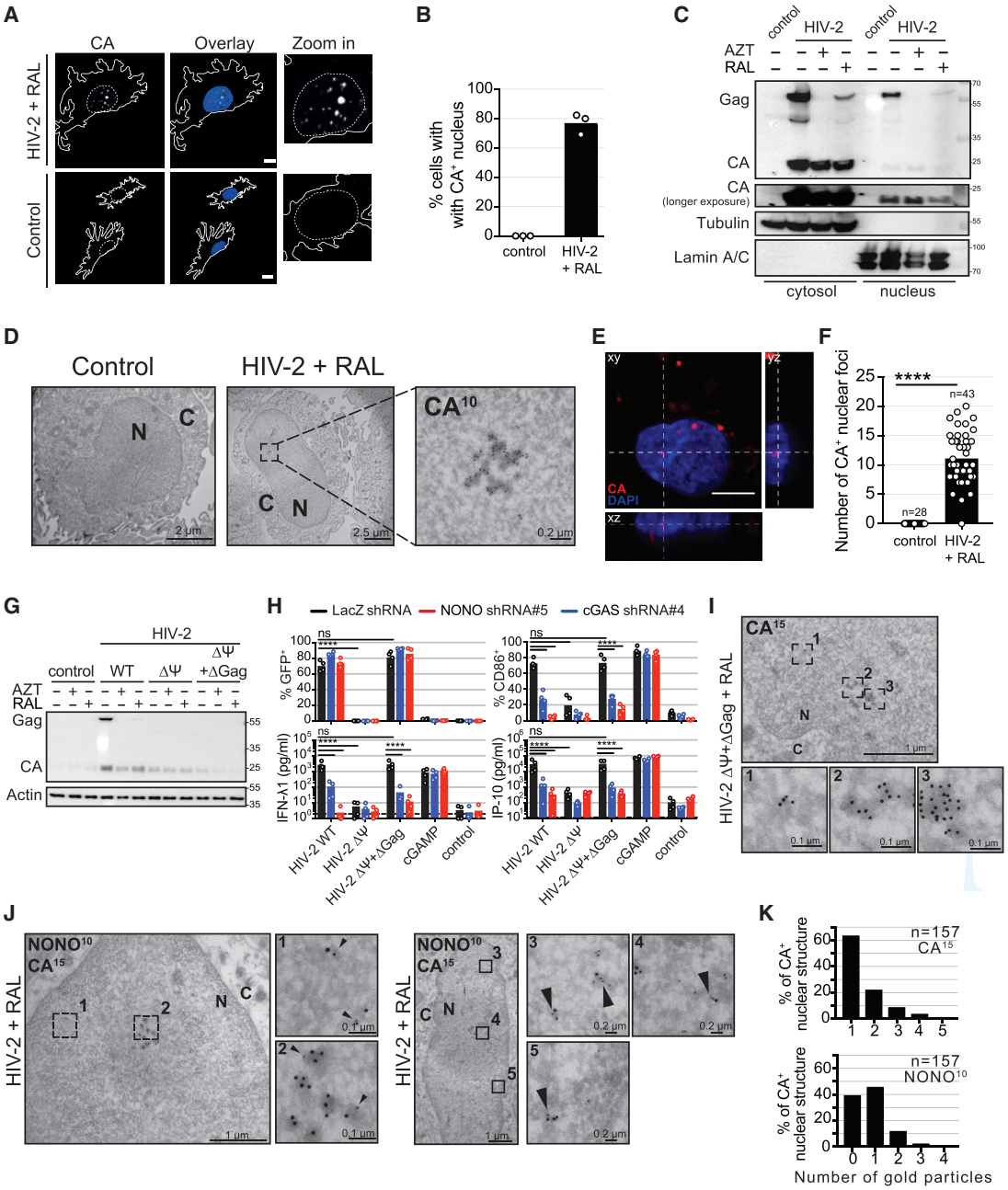


Figure 3. The HIV-2 Capsid Reaches the Nucleus of Dendritic Cells

(A) Detection of capsid from incoming virus in the nucleus of DCs by immunofluorescence. Control MDDCs or MDDCs infected with HIV-2 GFP-reporter virus and Raltegravir (HIV-2 + Raltegravir [RAL]; 20 μM) for 16 hr and processed with proteinase K treatment to detect nuclear capsid (CA) (green). Dash lines, nucleus contour; plain lines, cell contour; blue, DAPI. Scale bar, 10 μm (n = 3, one representative experiment is shown).

(B) Quantification of capsid-containing nuclei as in (A) (MOI = 2; n = 3; 15 cells analyzed on average per condition and per donor).

(C) Detection of GAG, CA, tubulin, and lamin A/C after cytoplasmic and nuclear fractionation. MDDCs were infected 24 hr in presence or not of RAL or azidothymidine (AZT, 25 μM), with HIV-2 (n = 4, one representative experiment is shown).

(D) Detection of capsid (CA^{10}) from incoming virus in the nucleus of DCs infected as in (A) by immuno-electron microscopy (N, nucleus; C, cytosol; n = 2 independent experiments, one representative experiment is shown).

(E) Orthogonal projection of intranuclear capsid (CA, red) from DCs infected as in (A), after PLA for capsid. Blue, DAPI. Scale bar, 5 μm (n = 2, one representative experiment is shown).

(F) Quantification of capsid-containing nuclei as in (E) (n = 2, one representative experiment is shown, numbers of cells analyzed is indicated).

(legend continued on next page)

we conclude that the incoming HIV-2 capsid reaches the nucleus and directly associates with NONO.

NONO Is Required for the Presence of cGAS in the Nucleus

These results raised the possibility that NONO could regulate cGAS localization in cells. While cGAS was initially reported to localize mainly in the cytosol (Sun et al., 2013), cGAS expressed in interphase can access the nuclear compartment as a result of nuclear envelope rupture or mitosis (Denais et al., 2016; Mackenzie et al., 2017; Raab et al., 2016; Yang et al., 2017), and cGAS was detected in the nucleus of human primary fibroblasts (Orzalli et al., 2015). Using nuclear/cytoplasmic fractionation, we found that endogenous cGAS was present in both the nucleus and the cytosol in monocyte-derived DCs (MDDCs), macrophages (MDMs), THP-1, and HeLa cells (Figure 4A). NONO was found in the nucleus, and STING was detected only in the cytoplasm. Calnexin was detected only in the cytoplasmic fraction of DCs and macrophages, thus excluding a contamination of nuclear fractions with endoplasmic reticulum. The presence of cGAS in the nucleus raised the question of how cGAS is prevented from massive activation by nuclear DNA. Using an *in vitro* enzymatic assay for cGAS activity, we found that cGAS is not activated by nucleosomes (Figures 4B and 4C). Extracting DNA from nucleosomes rescued cGAS enzymatic activity similarly to naked DNA. Spiking naked DNA into nucleosomes also rescued cGAS enzymatic activity, excluding a dominant-negative effect (Figure S4H). Thus, the nucleosomal state of nuclear DNA limits cGAS activation. We next tested the role of NONO on cGAS localization in DCs. Upon NONO depletion, the baseline level of nuclear cGAS was reduced, while the cytoplasmic cGAS level was not affected (Figures 4D, 4E, S4I, and S4J). The expression of cGAS was increased in control cells treated with HIV-1 or cGAMP, in agreement with the production of IFN, because cGAS is an ISG (Schoggins et al., 2011). Of note, we also detected the presence of unprocessed Gag protein of HIV-1 and HIV-2 in the nucleus of DCs (Figures 4D and S4I). These results suggested that NONO could interact with cGAS. Using co-immunoprecipitation, both NONO and cGAS pulled down each other (Figures 4F and 4G). The cGAS-NONO interaction was resistant to benzonase, supporting putative protein-protein interactions (Figures 4H and S4K). To visualize the localization of cGAS as a function of NONO, we quantified cGAS intensity by immunofluorescence microscopy in control or NONO-depleted DCs (Figures 4I and 4J). The level of nuclear cGAS was linearly correlated with the level of nuclear NONO in DCs treated with a NONO shRNA. In control cells, the level of nuclear cGAS was correlated in the fraction of cells expressing lower levels of NONO (Figure S4L). In contrast, the level of cytosolic cGAS was not correlated with NONO. We conclude that NONO forms a complex with cGAS in the nucleus and that it is required for the presence of cGAS in the nucleus, but has no impact on the cytosolic pool of cGAS.

NONO Is Required for cGAS-Mediated Sensing of HIV DNA

These results suggested that NONO could have a direct role in cGAS-mediated sensing of HIV DNA. cGAMP production by cGAS results in phosphorylation of STING at Serine 366. HIV-2 infection induced phosphorylation of STING in DCs, and this was inhibited by depletion of cGAS or NONO (Figure 5A). Depletion of NONO or cGAS had no impact on STING phosphorylation in response to cGAMP. Next, we immunoprecipitated cGAS with a chromatin immunoprecipitation (ChIP) protocol to determine the role of NONO in the association of HIV-2 DNA with cGAS in the nucleus (Figure 5B). HIV-2 DNA was associated with cGAS in control infected cells, and this was reduced by 21-fold in NONO-depleted cells (Figure 5C). Altogether, these data show that NONO is required for cGAS recognition of HIV-2 DNA and STING activation.

NONO Recognizes a Site of Genetic Fragility in HIV Capsids

These data suggested that NONO is a HIV capsid recognition receptor of innate immunity. Considering that virus recognition by the innate immune system is critical for inducing antiviral responses, it was intriguing that HIV capsid recognition by NONO would be based on protein-protein interactions, which are presumably sensitive to escape mutations. However, the HIV capsid is considered a genetically fragile structure, which poorly tolerates mutational changes (Rihn et al., 2013). To test if NONO recognized a site of genetic fragility in the capsid protein, we mapped single amino acids essential for NONO binding to capsid (Figure S5A). We found that D101 and I102 were essential for NONO binding to HIV-2 capsid, while Y49 had a partial role (Figures 6A and S5B). D101 and I102 are conserved in HIV-1 capsid, while Y49 is Q50 in HIV-1 (Figure 6B). These residues are surface-exposed and closely located in space,

(G) Detection of GAG, CA, and actin. MDDCs were infected 24 hr in presence or not of RAL or AZT, with HIV-2 (HIV-2 WT), HIV-2 ΔΨ, or HIV-2 ΔΨ+ΔGag (n = 2, one representative experiment is shown).

(H) GFP, CD86 expression, and IP-10, IFN-λ1 production by MDDCs transduced at day 0 with a control shRNA against LacZ or shRNAs against cGAS or NONO and subsequently infected 48 hr at day 4 as in (G) or treated with cGAMP (1.33 μg/mL) for 48 hr (n = 4, paired RM ANOVA, ****$p < 0.0001$, ns, not statistically significant, on log-transformed data for IP-10 and IFN-λ1 production).

(I) Detection of capsid (CA^{15}) from incoming virus in the nucleus of DCs by immuno-electron microscopy infected 16 hr with HIV-2 ΔΨ+ΔGag GFP-reporter virus in presence of RAL.

(J) Detection of endogenous NONO ($NONO^{10}$) and capsid (CA^{15}) from incoming virus in the nucleus of DCs by immune-electron microscopy. MDDCs were infected as in (A) (N, nucleus; C, cytosol; n = 2 independent experiments, one representative experiment is shown).

(K) Quantification of gold particles ($NONO^{10}$ or CA^{15}) associated with capsid-positive nuclear structures as in (J) (n = 2 independent experiments, numbers of CA^{+} nuclear structures analyzed are indicated).

Bars indicate mean.

See also Figures S4A–S4G.

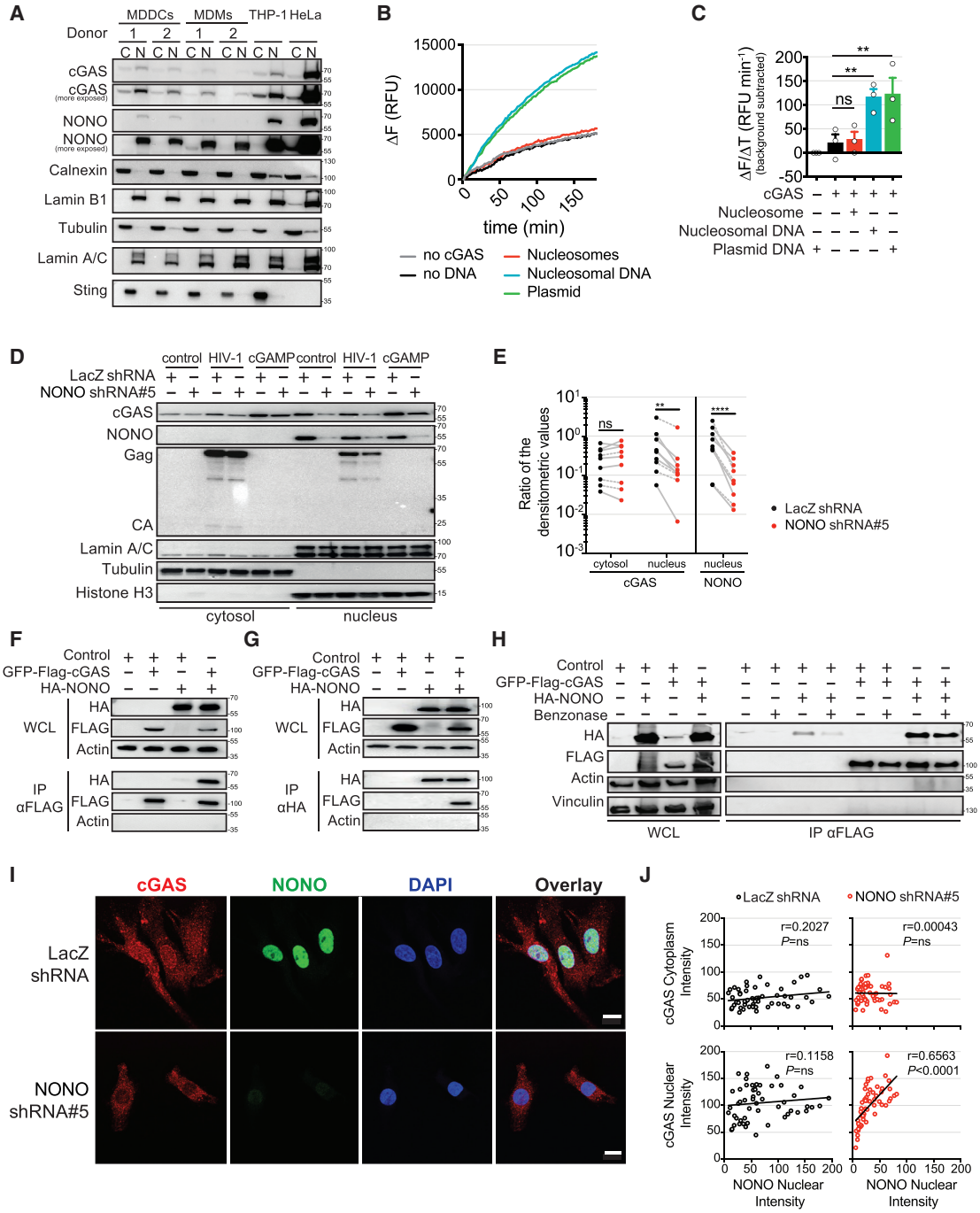


Figure 4. NONO Interacts with cGAS in the Nucleus

(A) Detection of cGAS, NONO, calnexin, lamin B1, tubulin, lamin A/C, and STING after cytoplasmic and nuclear fractionation of MDDCs, MDMs, THP-1, and HeLa cells.

(B) *In vitro* activity of cGAS in the presence of nucleosomes, nucleosomal DNA, or plasmid DNA. Mean values of initial cGAS reaction rates was subtracted to time 0 and inverted (ΔF, RFU, relative fluorescence units; n = 3, one representative experiment is shown).

(C) *In vitro* activity of hcGAS. From (B), background fluorescence was subtracted, initial rates were calculated as a slope of the linear intervals and defined as ΔF/Δt (relative fluorescence units per minute) (n = 3; paired RM ANOVA).

(D) Detection of cGAS, NONO, GAG, CA, tubulin, lamin A/C, and histone H3 after cytoplasmic and nuclear fractionation. MDDCs transduced with shRNA against NONO or control (LacZ) and infected 24 hr at day 4 with HIV-1 or HIV-2 or transfected 16 hr with cGAMP (1.33 μg/mL) (n = 6, one representative experiment is shown).

(legend continued on next page)

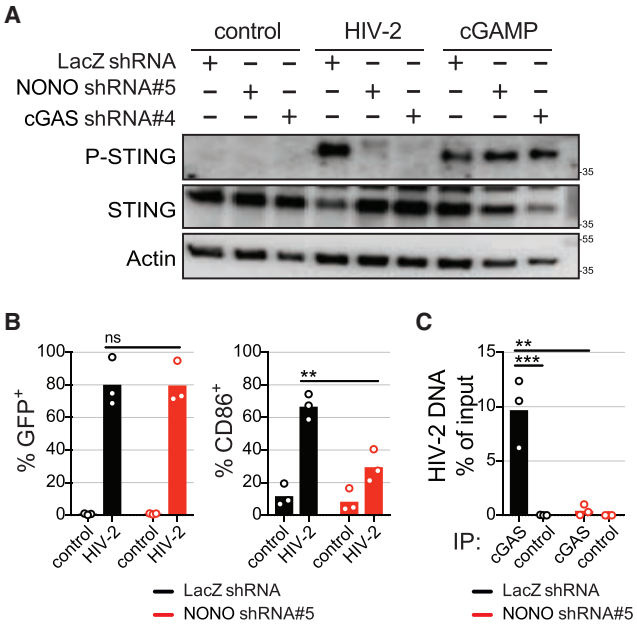


Figure 5. NONO Is Required for cGAS-Mediated Sensing of the HIV-2 DNA

(A) Western blot of phospho-Ser366-STING expression in MDDCs transduced at day 0 with a control shRNA against LacZ or shRNAs against cGAS or NONO and infected at day 4 with HIV-2 GFP-reporter viruses (16 hr) or transfected with cGAMP (12 hr, 1.33 μg/mL) (n = 4, one representative experiment is shown).

(B) GFP and CD86 expression in MDDCs transduced at day 0 with a control shRNA against LacZ or a shRNA against NONO and infected at day 4 with HIV-2 GFP-reporter virus for 48 hr (n = 3, paired RM ANOVA).

(C) Immunoprecipitation of HIV-2 DNA in infected cells (16 hr) as in (B) with an antibody against cGAS or IgG control (n = 3, paired RM ANOVA).

Bars indicate mean; $^{**}p < 0.01$, $^{***}p < 0.001$.

consistent with a protein-binding interface (Figure 6C). We produced viruses with mutations D101A and I102A in HIV-2, and the corresponding D103A and I104A in HIV-1. The mutant viruses showed a 3- and 2-log reductions in viral titers, respectively (Figure 6D). In producer cells, HIV-2 D101A and I102A showed normal levels of cell-associated Gag but reduced levels of cell-associated CA, and HIV-1 D103A and I104A showed reduced levels of cell-associated Gag and CA (Figure S5C). Extracellular viral proteins were reduced for HIV-1 and HIV-2 mutants (Figure S5C). HIV-1 and HIV-2 do not normally package NONO in viral particles during viral production (Figure S5D), indicating that the reduced viral titer was not due to a loss of NONO binding during viral production. The reduction in viral titers was not due either to gaining a dependency on cyclophilin A inhibition (Figure 6D). Electron microscopy showed that one defect of HIV-1 I104A was the accumulation beneath the plasma membrane of virus-producing cells and a failure to bud (Figure 6E). These results indicate that HIV-1 D103A, I104A, and HIV-2 D101A and I102A are severely defective. We next tested if a moderate polymorphic exchange could be tolerated. We generated HIV-2 Y49Q and its counterpart HIV-1 Q50Y. While Y49Q diminished the interaction with NONO for HIV-2 capsid, Q50Y increased it for HIV-1 capsid (Figure 6F). The interactions with CypA were not altered by the mutations (Figure S5E). Gag expression, budding, and capsid maturation in the particles was normal (Figure S5F). However, HIV-1 Q50Y and HIV-2 Y49Q showed 1- and 2-log decreases in viral titers, respectively (Figure 6G). To establish a causal link between NONO recognition and sensing, we examined the activation of DCs in response to infection by capsid mutants. HIV-1 D103A, I104A and HIV-2 D101A, I102A were too defective to be tested in this assay. We infected DCs with capsid-normalized amount of HIV-1 Q50Y, HIV-2 Y49Q, and wild-type (WT) viruses. HIV-1 Q50Y showed a profound defect in viral integration and GFP expression (Figures S5G and S5H), which prevented further analysis because immune activation by HIV-1 in DCs with Vpx requires integration and Gag expression (Manel et al., 2010). HIV-2 Y49Q was also profoundly defective for integration and GFP expression, independently of NONO (Figures 6H and 6I). HIV-2 Y49Q produced 10-fold less viral DNA than its WT counterpart per unit of capsid (Figure 6I). HIV-2 WT readily induced IFN-λ1 in a NONO-dependent manner (Figures 6H and 6J). At comparable levels of input capsid, induction of IFN-λ1 by HIV-2 Y49Q was 26-fold less than WT (Figures 6H and 6J). At comparable levels of viral DNA, residual induction by HIV-2 Y49Q was also independent of NONO, and 2-fold less than WT (Figures 6H and 6J). These results raised the question of the conservation of the NONO-capsid interaction. Using Y2H, we find that NONO binding is conserved in SIVmac and SIVsmm capsids (Figures S6A–S6C). D101 and I102 are conserved throughout the primate lentiviruses (Figure 6K). Residue Y49 is highly conserved throughout the primate lentiviruses, while the HIV-1 group M and SIVcpz lineages

(E) Quantification of the densitometric values from nuclear NONO and cytosolic or nuclear cGAS after cytoplasmic and nuclear fractionation in MDDCs as in (D). Cytosol ratio was calculated over tubulin, nuclear ratios were calculated by over histone H3 (solid lines) or lamin B1 (dash lines) (n = 9; paired RM ANOVA).

(F) Pull-down of NONO by cGAS. 293FT cells were co-transfected with indicated plasmids and processed for anti-FLAG immunoprecipitation (IP) and western blot (n = 3, one representative experiment is shown).

(G) Pull-down of cGAS by NONO. 293FT cells were co-transfected with indicated plasmids and processed for anti-HA immunoprecipitation and western blot (n = 3, one representative experiment is shown).

(H) Pull-down of NONO by cGAS is resistant to nucleic acid digestion. 293FT cells were co-transfected with indicated plasmids and processed for anti-FLAG immunoprecipitation with or not benzonase treatment and western blot (n = 3, one representative experiment is shown).

(I) Detection of endogenous cGAS, NONO, and DNA (DAPI) by confocal microscopy in MDDCs transduced with shRNA against NONO or control (LacZ) at day 5. Scale bar, 5 μm (n = 2, one representative experiment is shown).

(J) Quantification of the density of cGAS fluorescence in the cytoplasm and in the nucleus, relative to the density of NONO fluorescence in the nucleus, in cells as in (I) (n = 2, the line represents the linear regression; r, Pearson r correlation; p, p value; one representative experiment is shown).WCL, whole cell lysate; ns, not statistically significant; bars indicate mean; $^{**}p < 0.01$, $^{****}p < 0.0001$.

See also Figures S4H–S4L.

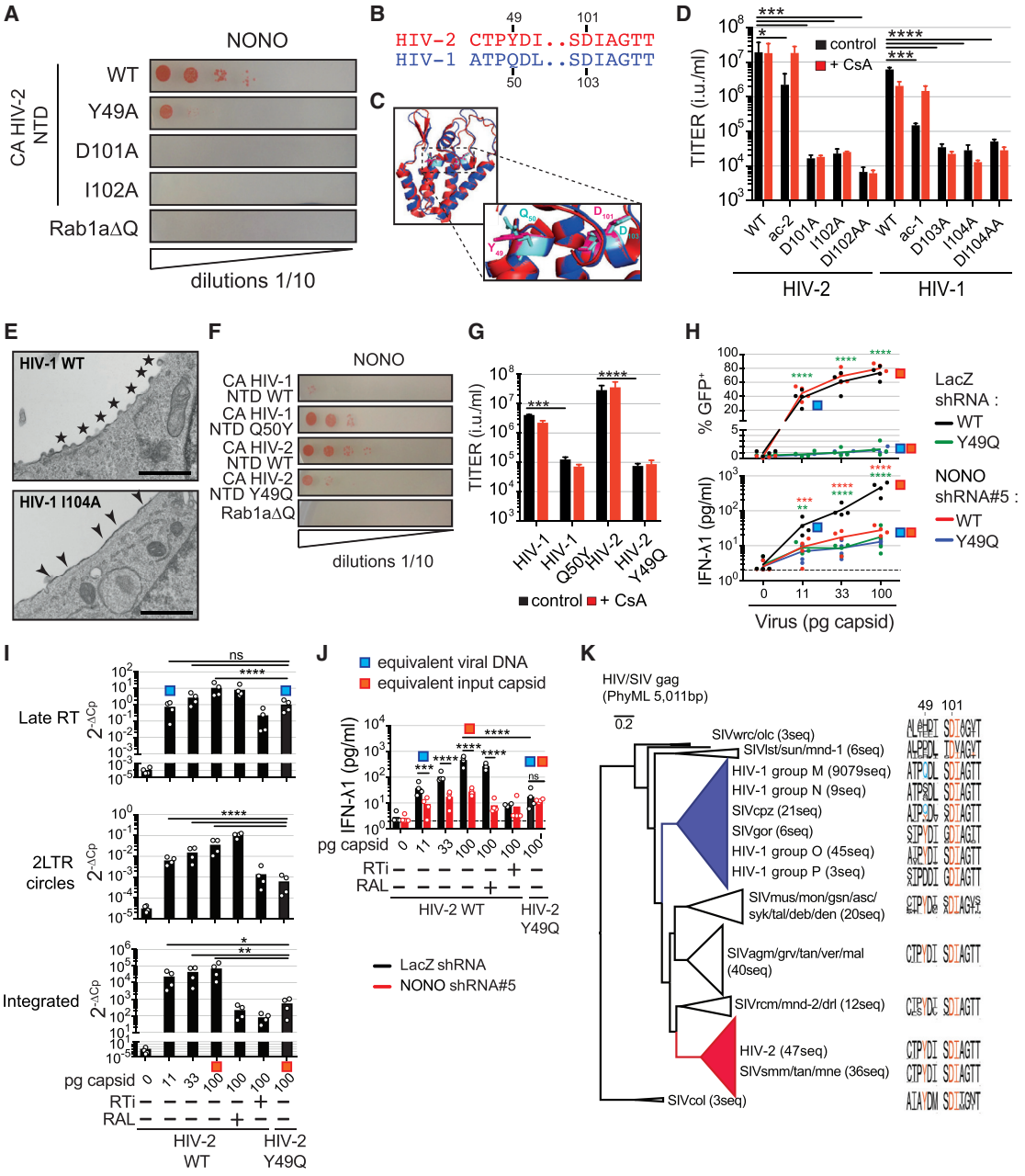


Figure 6. NONO Recognizes a Site of Genetic Fragility in HIV-1 and HIV-2 Capsids

(A) Interaction of HIV-2 CA NTD WT and mutants Y49A, D101A, and I102A, or control Rab1aΔQ with NONO, measured by Y2H and dilutions to estimate the strength of the interaction (n = 4, one representative experiment is shown).

(B) HIV-1 and HIV-2 capsid residues implicated into the NONO-HIV-2 CA interaction.

(C) Capsid residues (cyan, HIV-1; pink, HIV-2) implicated in the interaction with NONO in space. HIV-2 CA NTD is shown in red (PDB: 2WLV), HIV-1 CA NTD in blue (PDB: 1AK4). Bottom, magnification.

(D) Infectious titer of the indicated mutated GFP-reporter viruses on GHOST cells in the presence or absence of CsA (2 μM) (n = 3, paired RM ANOVA).

(E) Plasma membrane of HIV-1 WT and I104A from transfected 293FT, observed by transmission electron microscopy (scale bar, 1 μm). Stars show budding events in HIV-1 WT, arrows show a thickened plasma membrane in HIV-1 I104A (n = 2, one representative experiment is shown).

(F) Interaction of HIV-1 WT, HIV-2 WT, HIV-1 Q50Y and HIV-2 Y49Q CA NTD, or control Rab1aΔQ with NONO, measured by Y2H and dilutions (n = 3, one representative experiment is shown).

(G) Infectious titer of the indicated mutated GFP-reporter viruses as in (D) (n = 2, paired RM ANOVA).

(H) GFP expression and IFN-λ1 production in DCs transduced at day 0 with a control shRNA against LacZ or shRNAs against NONO and subsequently infected at day 4 for 48 hr with normalized input of CA with HIV-2 WT or HIV-2 Y49Q GFP-reporter viruses (n = 4, paired RM ANOVA, on log-transformed data for IFN-λ1 production).

(legend continued on next page)

mainly encode Q50 (Figure 6K). NONO is also highly conserved in vertebrates and primates and has evolved under purifying selection (Figures S6D–S6F). Using Y2H mapping, we found that I275 and E278 residues in NONO, which are aligned in space (Passon et al., 2012), are essential for interaction with HIV-2 capsid and conserved in primate NONO proteins, except in tarsier (Figures S6G–S6I), in agreement with recognition of a conserved molecular pattern by NONO and the evolution of several innate immune sensors (Quintana-Murci and Clark, 2013). We conclude that NONO recognizes a conserved interface in HIV capsid, which cannot tolerate the escape mutations tested as these produce a profound fitness cost, consistent with genetic fragility.

Genetic Validation Using Dendritic Cells of NONO-Deficient Patients

Genetic validation is critical for the identification of pattern recognition receptors (Vance, 2016). Loss-of-functions mutations in NONO have been reported in patients with cognitive disabilities (Mircsof et al., 2015). We obtained PBMCs from 2 independent patients and generated monocyte-derived DCs. We confirmed the lack of expression of NONO in mononuclear cells (Figure 7A). DCs deficient for NONO showed a profound reduction in their ability to produce IP-10 and IFN-λ1 after infection by HIV-1 with Vpx or HIV-2, but not after HT-DNA or cGAMP treatments (Figures 7B, S7A, and S7B). A dose-response analysis confirmed that the unaltered response to cGAMP was not due to saturation (Figure S7B). The rate of HIV-1 and HIV-2 infection was also not affected by NONO deficiency. Expression of the ISGs MX1, IFIT-1, OAS1, and CXCL10 was also reduced by NONO deficiency in response to infection (Figure 7C). Using whole-genome gene expression analysis, 123 genes were significantly upregulated by HIV-2, and 52 were previously recognized ISGs (Figures 7D and 7E). 93% of the upregulated genes were dampened in NONO-deficient cells after HIV-2 infection. A similar pattern was observed in DCs infected by HIV-1 with Vpx, but gene induction in WT cells was reduced for HIV-1 with Vpx as compared to HIV-2. The genes were not expressed at a lower baseline in the uninfected NONO-deficient sample. For one NONO-deficient patient, we also stimulated cells with cGAMP (Figures 7F and 7G). Unlike HIV-2, there was no difference between WT and knockout (KO) for cGAMP stimulation, thus excluding global dysregulations in NONO KO cells.

DISCUSSION

NONO fits the criteria for an HIV capsid sensor essential for innate immune activation of DCs and macrophages. We propose that NONO is required for the presence of cGAS in the nucleus, and that the chromatin state limits cGAS activation by self DNA; upon nuclear entry of HIV-2, the viral capsid is recognized by NONO, leading to recruitment of HIV-2 DNA in the vicinity of cGAS (Figure 7H). NONO was previously detected within protein complexes associated with HIV complexes or HIV-related host factors, but the nature and significance of these associations remained unresolved (Milev et al., 2012; St Gelais et al., 2015). NONO is a multifunctional RNA- and DNA-binding protein scaffold implicated in transcription, splicing, DNA damage response, circadian rhythm, and neuronal development (Knott et al., 2016a). NONO localizes to the nucleoplasm and can also participate to the formation of nuclear paraspeckles. Paraspeckles are intranuclear bodies that require the lncRNA NEAT1 for assembly (Fox et al., 2018). At least three findings argue against a role for paraspeckles in NONO recognition of HIV. First, NONO staining is pan-nuclear in DCs with no obvious nuclear bodies. Second, the NONO-cGAS complex was resistant to nuclease treatment. Third, NEAT1 was expressed at very low level in DCs in our gene expression analysis and this was unaltered by the stimulations. Although we did not observe a general dysregulation in immune activation in NONO-deficient cells, we do not discard the possibility that NONO or paraspeckles could play immuno-regulatory roles beyond HIV recognition, perhaps in relation to other functions of nuclear cGAS to be uncovered. STING and NONO localizations were mutually exclusive in DCs, macrophages, and THP-1, and STING expression was comparatively undetected in HeLa cells, in agreement with studies that did not detect a biologically active IFN response in response to transfected DNA in these cells (Gentili et al., 2015; Lau et al., 2015; Rasaiyaah et al., 2013). In contrast, NONO was found in HeLa cells to co-sediment with a ribonuclear complex that contained cGAS and STING and that could regulate the expression of IFN in response to transfection of a synthetic dsDNA oligonucleotide (Morchikh et al., 2017). These discrepancies warrant further studies.

We find cGAS to be present in the nucleus of DCs, macrophages and two cell lines at steady state. DCs arise from precursors with high proliferative potential (Lee et al., 2015). cGAS can enter the nucleus during each of the previous mitoses that took place in the DC precursors. Transient nuclear envelope ruptures also occur physiologically during migration of non-cycling DCs in tissues (Raab et al., 2016). Other factors could play an active role in the nuclear import of cGAS. In DCs, we find that NONO is essential for the presence of cGAS in the nucleus. This may be leveraged in future work to identify nuclear functions of cGAS beside HIV recognition.

(I) Quantification of late RT, 2LTR circles, and integrated viral cDNA products 24 hr after infection of LacZ cells as in (H), in presence RAL or AZT with 10 μM nevirapine (NVP) (= RTi) when indicated (n = 4; paired RM ANOVA on log-transformed data).
(J) IFN-λ1 production in MDDCs transduced and infected as in (H) and treated as in (I). Comparable levels of input capsid and viral DNA between HIV-2 WT and Y49Q are indicated by a red or blue square, respectively (n = 4; paired RM ANOVA on log-transformed data).
(K) Phylogenetic analysis of primate lentiviruses (*gag*, n = 202 sequences). The clades that include HIV-1 and HIV-2 strains are highlighted in dark blue and red, respectively. Logos (from 9,290 HIV/SIV sequences) show the amino acids corresponding to those involved in the NONO-HIV CA interaction (seq, sequences; positions 49, 101, and 102 are highlighted: residues identical to HIV-2 are in orange, while Q50 in HIV-1 group M and SIVcpz is highlighted in light blue).
*p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001; ns, not statistically significant; bars and lines indicate mean ± SEM.
See also Figures S5 and S6.

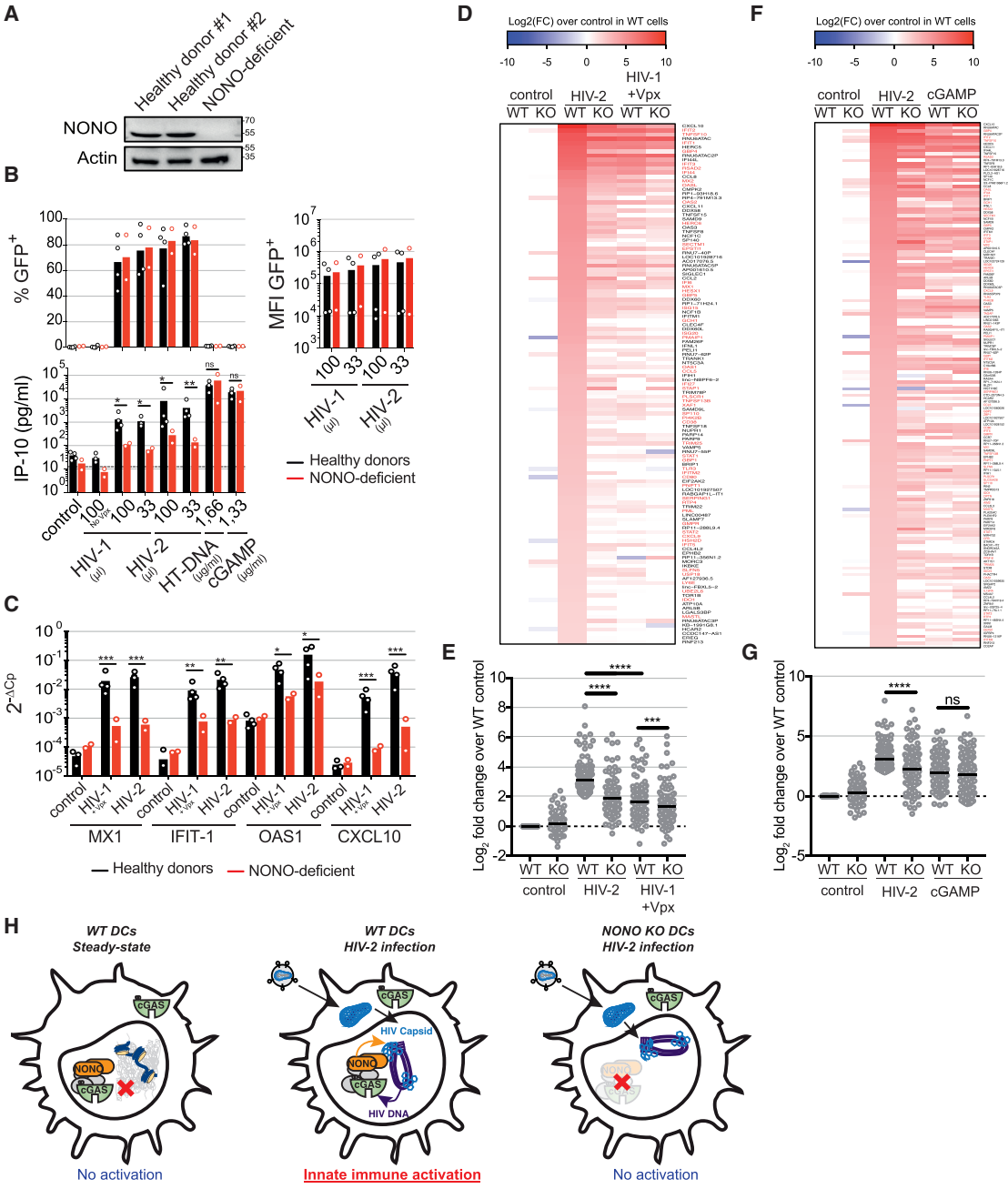


Figure 7. Compromised Immune Response to HIV-1 and HIV-2 Infection in Dendritic Cells of NONO-Deficient Patients

(A) NONO and actin protein expression in mononuclear PBLs (CD14-CD4-fraction) of NONO-deficient patient compare to healthy donors (n = 2, one representative experiment is shown).

(B) GFP expression and IP-10 production in MDDCs from healthy donor or NONO-deficient patients 48 hr after infection with HIV-1 + Vpx or HIV-2 GFP-reporter viruses or transfected with HT-DNA or cGAMP (n = 4 for healthy donors and n = 2 for NONO-deficient patients, unpaired RM ANOVA).

(C) Expression of MX1, IFIT-1, OAS1, and CXCL10 after 24 hr of infection with HIV-1 + Vpx or HIV-2 GFP-reporter viruses (33 µL) (n = 4 for healthy donors and n = 2 for NONO-deficient patients, unpaired RM ANOVA).

(D) Expression of genes induced in DCs from four healthy donors (WT) and two NONO-deficient patient (KO) after 24 hr of infection with HIV-1 + Vpx or HIV-2 GFP-reporter viruses (33 µl). Genes induced by HIV-2 in DCs from healthy donors were selected. ISG previously recognized in another study (Schoggins et al., 2011) are highlighted in red.

(E) Log_2 fold change over WT control of genes induced in DCs as in (D) (n = 4 for healthy donors and n = 2 for NONO-deficient patients, Friedman test with Dunn's multiple comparisons).

(legend continued on next page)

We find that the HIV-2 capsid has increased affinity for NONO as compared to HIV-1, which is consistent with the reduced pathogenicity and better immune control of HIV-2 over HIV-1. HIV-2 contains Vpx and its capsid efficiently binds NONO, enabling recognition of the incoming virus. HIV-1 recognition with Vpx requires viral integration and expression, and we find accumulation of newly synthesized HIV-1 Gag in the nucleus of DCs. The existence of small nuclear pool of HIV-1 Gag has previously been described, but its function and conformation is not known (Grewe et al., 2012). New viral DNA synthesis after HIV-1 Gag expression and cleavage of newly expressed Gag is not required for sensing in dendritic cells (Gao et al., 2013; Manel et al., 2010; Sunseri et al., 2011). Thus, while newly expressed Gag is not expected to interact with viral DNA in *cis*, it may interact in *trans* with incoming capsid-DNA complexes through NONO in the nucleus. NONO forms multimers in the nucleus (Knott et al., 2016a), which could connect newly synthesized Gag to incoming capsid-DNA complexes. Furthermore, we find that the gene signature induced by HIV-1 with Vpx is dampened as compared to HIV-2, and that HIV-2 induces higher levels of IFN-I, IFN-III, and CD86 than HIV-1, even if Vpx is provided. Consistent with this, in macrophages, the level of ISG induction was similar for HIV-2 and HIV-1 with Vpx, despite a lower rate of infection for HIV-2. Levels of IP-10 protein were comparable between HIV-2 and HIV-1 with Vpx, possibly related to post-transcriptional regulation (Casrouge et al., 2011). There was also lower level of phospho-IRF3 for HIV-2 than HIV-1 with Vpx at the time point tested. This experiment was not designed to compare the magnitude of phospho-IRF3 between stimuli, but it suggests that the dynamic of IRF3 phosphorylation may be different between the two viruses. We conclude that the decreased affinity of HIV-1 capsid for NONO as compared to HIV-2 contributes to a reduced magnitude of innate immune activation, even if Vpx is provided to HIV-1.

We show that NONO directly associates with HIV-2 capsids in the nucleus. It is currently thought capsid is co-imported with the viral DNA in the nucleus. We speculate that other active import mechanisms of capsid and/or Gag in the nucleus could also take place.

NONO recognizes a surface on HIV capsids that appears to have a limited ability to tolerate mutations. Supporting this notion, this region was recently implicated in the regulation of dynamic pores in the HIV capsid that are critical for virus viability and thus likely highly sensitive to mutations (Jacques et al., 2016). Furthermore, HIV-1 D103, I104, and Q50 are located within CTL epitopes that have been associated with control of viral replication, consistent with a fitness cost of mutating them (Honeyborne et al., 2007; Migueles et al., 2014).

The nucleus is generally considered to be a protection for host cell DNA, and viruses in turn have evolved a number of capsid-based strategies to bring their nucleic acids into the nucleus (Ravindran and Tsai, 2016). Viral capsids are accordingly typical molecular patterns found mainly in viruses, and capsid assembly and functions are under strong structural constrains, likely underlying their genetic fragility. Recognition of viral capsid proteins by host proteins, particularly at sites of genetic fragility in capsids, could be an evolutionary favorable mechanism to ensure host protection against viruses that are able to evolve rapidly. A two-step mechanism of viral recognition involving both viral nucleic acids and viral capsid proteins may also represent a general mechanism to maximize specific and sensitive discrimination of viruses over related self-elements. Our work enables exploring viral capsid-based recognition and NONO-cGAS crosstalk in the design of vaccine and therapeutic strategies.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Human subjects
  - Cells lines
- METHOD DETAILS
  - Constructs
  - Cells
  - Virus production
  - Infections
  - Knock-down by shRNA interference
  - Knock-out by gRNA interference
  - IP-10 protein quantification
  - IFN-β and IFN-λ1 proteins quantifications
  - Quantitative bioassay for IFNs
  - HIV Real-time PCR
  - Gene expression quantification
  - Western Blotting
  - Cytoplasmic and nuclear fractionation
  - Recombinant protein expression/purification
  - Microscale Thermophoresis analysis
  - Yeast Two Hybrids Screen
  - Yeast Two Hybrid Assays
  - Immunoprecipitations
  - Immunofluorescence
  - cGAS-associated HIV-DNA Immunoprecipitation
  - Image processing and analysis
  - Electron microscopy
  - Immuno-electron microscopy
  - Microarray Data and Bio-informatics methods

(F) Expression of genes induced in DCs from two healthy donors (WT) and one NONO-deficient patient (KO) after 24 hr of infection with HIV-2 GFP-reporter viruses (33 μL) or lipofection with cGAMP (1.33 μg/mL) shown as in (D).

(G) Log_2 fold change over WT control of genes induced in DCs as in (F) (n = 2 for healthy donors and n = 1 for NONO-deficient patients, Friedman test with Dunn's multiple comparisons).

(H) Working model.

*p < 0.05, **p < 0.01, ***p < 0.0001, ****p < 0.0001; ns, not statistically significant; bars indicate mean.

See also Figure S7.

SUPPLEMENTAL INFORMATION

Supplemental Information includes seven figures and six tables and can be found with this article online at https://doi.org/10.1016/j.cell.2018.08.062.

ACKNOWLEDGMENTS

We thank A. Bhargava for macrophage transduction setup; P. Benaroch, S. Amigorena, D. Littman, and N. De Silva for critical reading of the manuscript; Marius Döring, Karl-Peter Hopfner and Luidmila Andreeva for discussions; Laurent Guéguen for input on evolutionary analyses; Andrea Cimarelli and Shauna Katz; and contributors of publicly available genome sequences. We acknowledge the flow cytometry facility, the imaging facility (PICT-IBiSA, LABEX ANR-10-LBX-0038, ANR-10-IDEX-0001-02 PSL, France-BioImaging, ANR-10-INSB-04), and Audrey Rapinat and David Gentien from the Genomics Platform at Institut Curie. This work was supported by Institut Curie, INSERM, and CNRS and by grants from ACTERIA Foundation, Fondation Schlumberger pour l'Education et la Recherche, Sidaction (VIH2016126002), DIM Biothérapies, European Research Council (309848 HIVINNATE), LABEX VRI (ANR-10-LABX-77), LABEX DCBIOL (ANR-10-IDEX-0001-02 PSL* and ANR-11-LABX-0043 to N.M.), National Health and Medical Research Council (NHMRC) (1048659 and 1050585 to C.S.B. and A.H.F.), FINOVI "recently settled scientist," LABEX ECOFECT (ANR-11-IDEX-0007 and ANR-11-LABX-0048 to L.E.), Fondation pour la Recherche Médicale (Projet Innovant ING20160435028 to L.E. and DEQ20160334938 to L.C.), ANRS (France REcherche Nord & Sud Sida-hiv Hépatites; ECTZ25472 and ECTZ36691 to N.M. and ECTZ19143 to L.E.), and the National Research Agency (ANR-14-CE14-0004-02 to N.M. and ANR-10-IAHU-01 to L.C.).

AUTHOR CONTRIBUTIONS

X.L. and N.M. designed the experiments and wrote the paper. X.L. conducted the experiments related to NONO (innate immune activation in response to HIV, interaction with cGAS, cGAS presence in the nucleus), to nuclear capsid and capsid mutants. M.G. identified and characterized the presence of cGAS in the nucleus of primary MDDCs and conducted cGAS enzymatic assays. C.C. conducted the NONO-capsid interaction experiments using Y2H and MST. A.S. contributed to cDC2 experiments. M.J. conducted electron microscopy experiments. L.P. and L.E. performed the phylogenetic analyses. E.Z. conducted the ChIP experiment. M.M. conducted image analyses. F.N. conducted bioinformatics analyses. P.L., F.D., and B.Z. provided cGAS recombinant protein. G.J.K., C.S.B., and A.H.F. provided reagents and expertise on NONO. M.R., J.A., and L.C. provided samples from NONO-deficient patients.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 21, 2017
Revised: July 5, 2018
Accepted: August 28, 2018
Published: September 27, 2018

REFERENCES

Abdul, F., Filleton, F., Gerossier, L., Paturel, A., Hall, J., Strubin, M., and Etienne, L. (2018). Smc5/6 antagonism by HBx is an evolutionarily conserved function of hepatitis B virus infection in mammals. J. Virol. *92*, e00769-18.

Andreeva, L., Hiller, B., Kostrewa, D., Lässig, C., de Oliveira Mann, C.C., Jan Drexler, D., Maiser, A., Gaidt, M., Leonhardt, H., Hornung, V., and Hopfner, K.P. (2017). cGAS senses long and HMGB/TFAM-bound U-turn DNA by forming protein-DNA ladders. Nature *549*, 394–398.

Berre, S., Gaudin, R., Cunha de Alencar, B., Desdouits, M., Chabaud, M., Naffakh, N., Rabaza-Gairi, M., Gobert, F.X., Jouve, M., and Benaroch, P. (2013). CD36-specific antibodies block release of HIV-1 from infected primary macrophages and its transmission to T cells. J. Exp. Med. *210*, 2523–2538.

Briggs, J.A., Wilk, T., Welker, R., Kräusslich, H.G., and Fuller, S.D. (2003). Structural organization of authentic, mature HIV-1 virions and cores. EMBO J. *22*, 1707–1715.

Casrouge, A., Decalf, J., Ahloulay, M., Lababidi, C., Mansour, H., Vallet-Pichard, A., Mallet, V., Mottez, E., Mapes, J., Fontanet, A., et al. (2011). Evidence for an antagonist form of the chemokine CXCL10 in patients chronically infected with HCV. J. Clin. Invest. *121*, 308–317.

Chen, Q., Sun, L., and Chen, Z.J. (2016). Regulation and function of the cGAS-STING pathway of cytosolic DNA sensing. Nat. Immunol. *17*, 1142–1149.

Chin, C.R., Perreira, J.M., Savidis, G., Portmann, J.M., Aker, A.M., Feeley, E.M., Smith, M.C., and Brass, A.L. (2015). Direct visualization of HIV-1 replication intermediates shows that capsid and CPSF6 modulate HIV-1 intra-nuclear invasion and integration. Cell Rep. *13*, 1717–1731.

Chougui, G., Munir-Matloob, S., Matkovic, R., Martin, M.M., Morel, M., Lahouassa, H., Leduc, M., Ramirez, B.C., Etienne, L., and Margottin-Goguet, F. (2018). HIV-2/SIV viral protein X counteracts HUSH repressor complex. Nat. Microbiol. *3*, 891–897.

Denais, C.M., Gilbert, R.M., Isermann, P., McGregor, A.L., te Lindert, M., Weigelin, B., Davidson, P.M., Friedl, P., Wolf, K., and Lammerding, J. (2016). Nuclear envelope rupture and repair during cancer cell migration. Science *352*, 353–358.

Esbjörnsson, J., Månsson, F., Kvist, A., Isberg, P.E., Nowroozalizadeh, S., Biague, A.J., da Silva, Z.J., Jansson, M., Fenyö, E.M., Norrgren, H., and Medstrand, P. (2012). Inhibition of HIV-1 disease progression by contemporaneous HIV-2 infection. N. Engl. J. Med. *367*, 224–232.

Formstecher, E., Aresta, S., Collura, V., Hamburger, A., Meil, A., Trehin, A., Reverdy, C., Betin, V., Maire, S., Brun, C., et al. (2005). Protein interaction mapping: a Drosophila case study. Genome Res. *15*, 376–384.

Fox, A.H., Nakagawa, S., Hirose, T., and Bond, C.S. (2018). Paraspeckles: where long noncoding RNA meets phase separation. Trends Biochem. Sci. *43*, 124–135.

Gao, D., Wu, J., Wu, Y.T., Du, F., Aroh, C., Yan, N., Sun, L., and Chen, Z.J. (2013). Cyclic GMP-AMP synthase is an innate immune sensor of HIV and other retroviruses. Science *341*, 903–906.

Gentili, M., Kowal, J., Tkach, M., Satoh, T., Lahaye, X., Conrad, C., Boyron, M., Lombard, B., Durand, S., Kroemer, G., et al. (2015). Transmission of innate immune signaling by packaging of cGAMP in viral particles. Science *349*, 1232–1236.

Grewe, B., Hoffmann, B., Ohs, I., Blissenbach, M., Brandt, S., Tippler, B., Grunwald, T., and Uberla, K. (2012). Cytoplasmic utilization of human immunodeficiency virus type 1 genomic RNA is not dependent on a nuclear interaction with gag. J. Virol. *86*, 2990–3002.

Herzner, A.M., Hagmann, C.A., Goldeck, M., Wolter, S., Kübler, K., Wittmann, S., Gramberg, T., Andreeva, L., Hopfner, K.P., Mertens, C., et al. (2015). Sequence-specific activation of the DNA sensor cGAS by Y-form DNA structures as found in primary HIV-1 cDNA. Nat. Immunol. *16*, 1025–1033.

Honeyborne, I., Prendergast, A., Pereyra, F., Leslie, A., Crawford, H., Payne, R., Reddy, S., Bishop, K., Moodley, E., Nair, K., et al. (2007). Control of human immunodeficiency virus type 1 is associated with HLA-B*13 and targeting of multiple gag-specific CD8+ T-cell epitopes. J. Virol. *81*, 3667–3672.

Jacques, D.A., McEwan, W.A., Hilditch, L., Price, A.J., Towers, G.J., and James, L.C. (2016). HIV-1 uses dynamic capsid pores to import nucleotides and fuel encapsidated DNA synthesis. Nature *536*, 349–353.

Janoueix-Lerosey, I., Jollivet, F., Camonis, J., Marche, P.N., and Goud, B. (1995). Two-hybrid system screen with the small GTP-binding protein Rab6. Identification of a novel mouse GDP dissociation inhibitor isoform and two other potential partners of Rab6. J. Biol. Chem. *270*, 14801–14808.

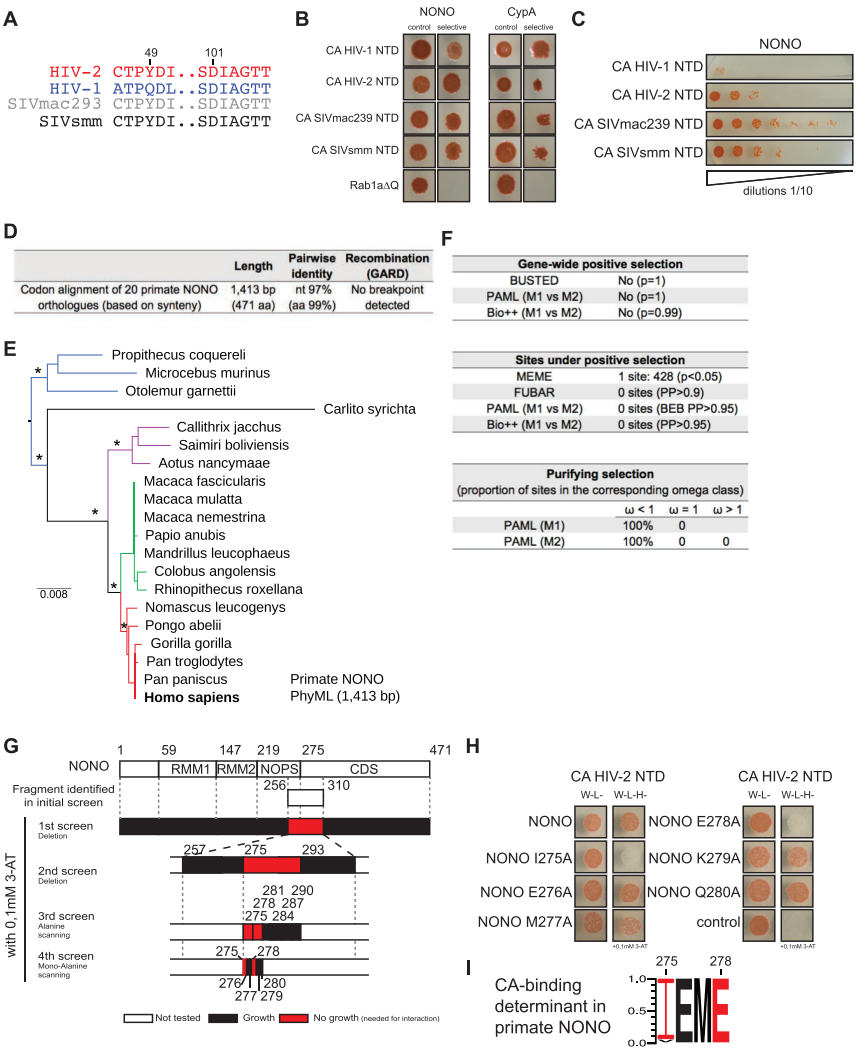


Figure S6. Conservation of NONO-Binding to Capsids of Primate Lentiviruses, Related to Figure 6

(A) Alignment of HIV-1, HIV-2, SIVmac239 and SIVsmm capsid residues implicated in the NONO-HIV-2 CA interaction.

(B) Interactions between the N-terminal domains (NTD) of the capsid (CA) proteins of HIV-1, HIV-2, SIVmac239 or SIVsmm or a negative control protein Rab1aΔQ with NONO and CypA, measured by yeast two-hybrid (n = 4, one representative experiment is shown).

(C) Estimation of the strength of the interactions as in (B) using dilutions (n = 4, one representative experiment is shown).

(D) Overview of the phylogenetic analyses of primate NONO (n = 20 orthologous sequences). Pairwise identity was computed in Geneious, Biomatters. Recombination analysis was performed with GARD

(E) Phylogenic tree of primate NONO performed with PhyML (HKY+G+I model; 1,000 bootstrap replicates). The asterisks show bootstrap values above 900/1,000. The scale bar indicates the number of nucleotide substitutions per site.

(F) Positive selection analyses of NONO. Upper panel, results of three "gene-wide" positive selection analyses (HYPHY BUSTED, PAML Codeml, Bio++); p values from LRT (maximum-likelihood ratio tests) indicate whether the model that allows positive selection better fits the data. Middle panel, results from "site-specific" positive selection analyses (HYPHY MEME and FUBAR, PAML Codeml, Bio++). Statistical thresholds for significance are indicated (PP, posterior probabilities; p, p value). BEB, Bayes Empirical Bayes. See STAR Methods for details. Lower panel, proportion of sites falling into each omega class (omega = dN/dS) computed in PAML Codeml. Both M1 and M2 models show that all sites of primate NONO have a dN/dS < 1, reflecting purifying selection.

(G) Mapping of the binding to the N-terminal domain of the HIV-2 CA in NONO by iterative mutagenesis and two-hybrid screen. Individual screens and strategies as described.

(H) Interaction of NONO (or mutated NONO) with the NTD CA NTD protein of HIV-2, or the negative control (empty vector). Data represent the 4th screen described in (G) (n = 3, one representative experiment is shown).

(I) Sequence logo of primate NONO at the site I275-E278.

Chapter 3

Identification of evolutionarily-relevant modulators of HIV in a dataset derived from a shRNA screen

This work is a collaborative effort with several members of the "Host-pathogen interactions during lentiviral infection" team at the Centre International de Recherche en Infectiologie (CIRI): in particular, Anuj Kumar who has led this work as first author and Andrea Cimarelli as senior author, as well as the other current contributing members: Yuxin Song, Xuan-Nhi Nguyen, Claire da Silva Santos, Li Zhong, Laurent Guéguen, and Lucie Etienne.

Provided that this work is unpublished yet and requires confidentiality, all genes have been anonymized and are referred to as Gene1 through Gene56.

3.1 Material and methods

Initial dataset

To identify novel modulators of HIV-1 infection, we silenced 419 ISGs described in the Interferome database (Rusinova et al. 2013) by shRNAs (3 each) in monocyte-derived THP-1 cells differentiated into a macrophage-like status with PMA and further incubated or not for 24 hours with IFNα (1000U/mL). We then challenged the cells with a Luciferase-

coding HIV-1 vector in single-round infection. Luminescence was measured as a readout for viral infectivity and normalized to the control condition in order to establish effect of each gene upon viral replication. This setup allowed us to retrieve cellular genes whose silencing modulated infection in both non-stimulated and IFNα conditions. We submitted candidate genes from this primary screen to a secondary screen aiming to exclude modulators of IFN signaling rather than of viral infection per se, leading to a list of 56 genes of interest. The steps leading to this dataset are described in Figure 17.

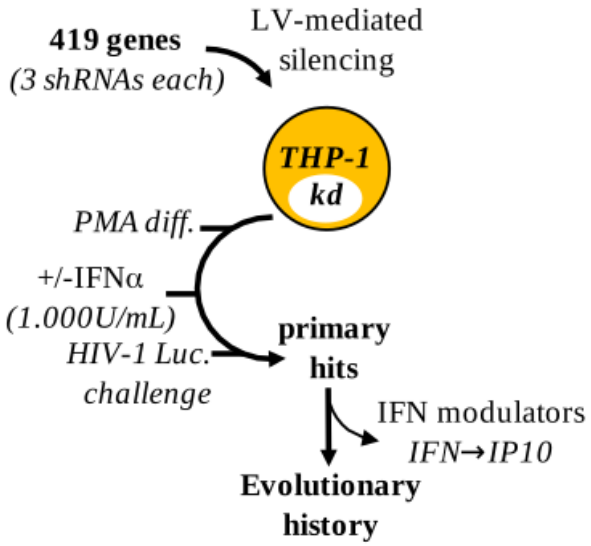


Figure 17: Overview of the different screens leading to the initial dataset

Figure courtesy of A. Cimarelli.

Evolutionary screen to detect genetic innovations in the protein-coding genes of the dataset

We screened the 56 genes of interest using the previously described DGINN pipeline (see Results - Chapter 1 and Picard et al. 2020).

We downloaded the Consensus CoDing Sequences (CCDS) from the CCDS database using the CCDSQuery script provided with DGINN, with one exception for which we manually retrieved the coding sequence from the NCBI databases as no CCDS existed. In cases where multiple CCDS were referenced for one gene, we kept the longest one. Those sequences were then used as the entry query for DGINN. We performed the phylogenetic analyses (step 1-7 of the DGINN pipeline, including duplication and recombination) against the NCBI nr database limited to primate species, with otherwise default parame-

ters (blastn e-value 10^{-4}, identity 70% and coverage 50%, at least 8 species for separation of orthologous groups). The species tree used to identify duplication events is the same as used previously in Results - Chapter 1 and Picard et al. 2020, and is based on the phylogeny of primates established by Perelman et al. 2011 and updated by Pecon-Slattery 2014.

All alignments and phylogenetic trees produced during this first part were then analyzed for marks of positive selection using the five different methods included in DGINN. It uses BUSTED and MEME from the Hyphy package (Pond et al. 2005) to look for gene-wide and site-specific episodic positive selection. We considered genes as under positive selection for a BUSTED p-value < 0.05 and sites for a MEME p-value < 0.10. Codeml from PAML (Yang 2007) and Bio++ (Guéguen et al. 2013) were used to run codon substitution models M1, M2, M7 and M8. M1 and M7 are neutral models not allowing for positive selection and M2 and M8 are their pendant allowing some codons to evolve under positive selection. We derived p-values from the likelihood ratio tests between the two models (M1 vs M2, M7 vs M8) to determine which model is a better fit for the data. We considered genes as under positive selection for $p < 0.05$ and sites for posterior probabilities > 0.90.

In-depth phylogenetic and positive selection analyses on the Gene53 candidate

For the gene of interest, Gene53, we manually retrieved primate sequences from available primate genomes using Blast on the nr database of the NCBI, with the human reference sequence for the longest isoform as query. Using DGINN, we aligned sequences with PRANK (Loytynoja and Goldman 2008) with a codon model and without forcing insertions to be skipped (prank -F -codon; version 150803). We reconstructed the phylogenetic tree using PhyML (v3.2, Guindon et al. 2010) with HKY+G+I model and 1000 bootstraps for statistical support of the branches. We performed positive selection analyses using the five methods included in DGINN as described in the previous section, with the difference that sites were considered under positive selection for posterior probabilities > 0.95 for the Codeml and Bio++ methods (as opposed to 0.90 during the evolutionary screen).

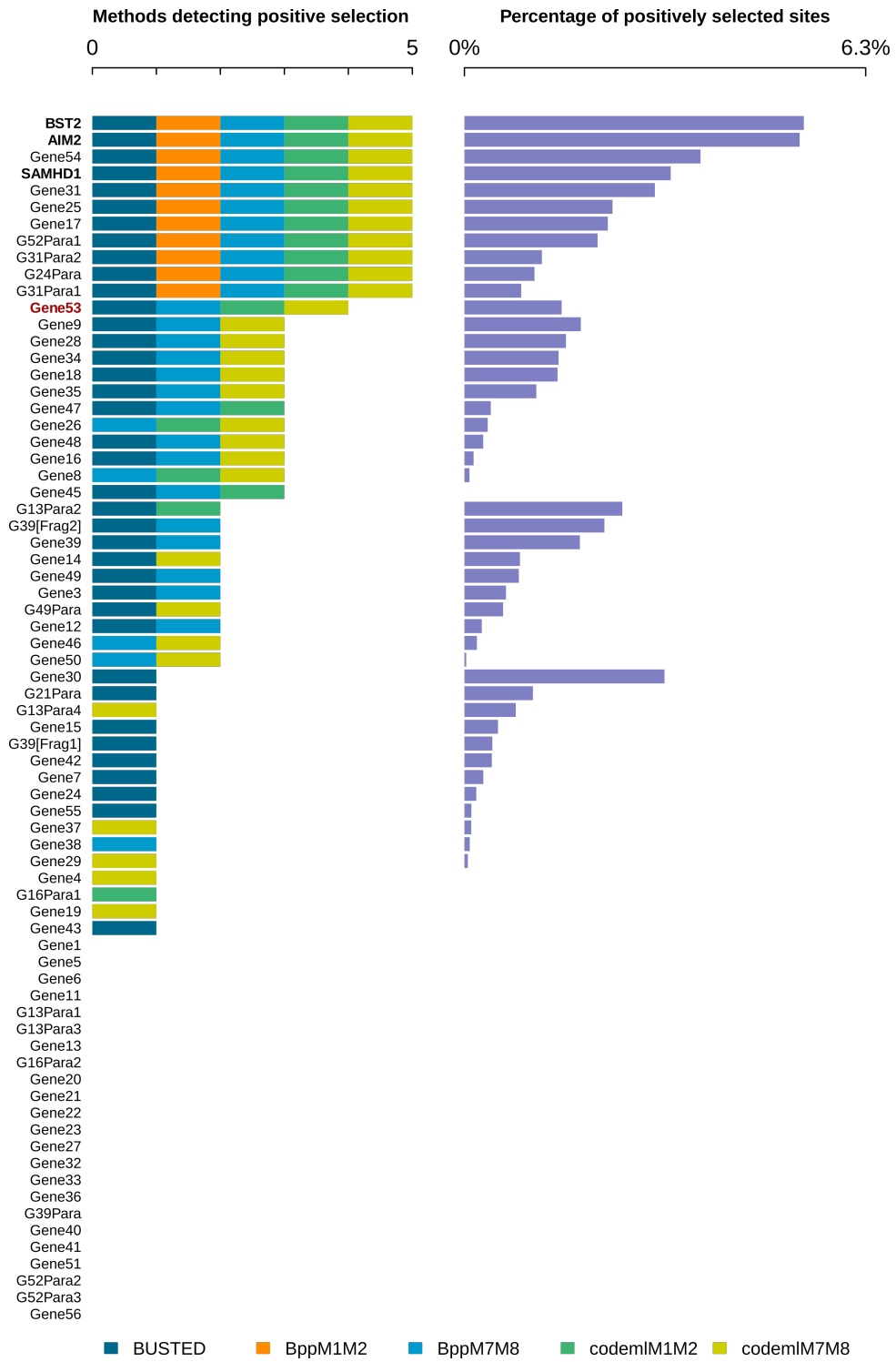


Figure 18: DGINN results on 55 primate genes and their paralogs

Left panel, number of methods detecting significant positive selection for each alignment; each method is color-coded (embedded legend). Right panel, percentage of positively selected sites (by at least one method) over the length of the alignment. Genes are ordered by descending number of methods detecting positive selection then descending percentages of positively selected sites. Gene53 is highlighted in red.

3.2 Results

Numerous modulators of HIV infection in macrophage-like cells are under positive selection

We analyzed our initial dataset of 56 genes for hallmarks of genetic conflicts using DGINN (see Material and Methods and Results - Chapter 1). DGINN retrieved the primate sequences for those 56 genes through Blast, aligned them with Mafft and PRANK, and separated the retrieved homologs into orthology groups through a combination of long-branch parsing and tree reconciliation using Treerecs. At this step, we retrieved several paralogs for genes belonging to multigenic families. Those analyses yielded twenty-five groups of DGINN-retrieved paralogs for eight of the query genes. We eliminated ten groups DGINN had improperly segregated, and that in consequence presented a mix of different paralogs within the same alignment. These instances were all observed with the same query, Gene31, which belongs to a large multigenic family of thirteen members. The fifteen remaining, properly segregated paralogs were. annotated G(X)Para(Y), with G(X) referring to the query gene used at the beginning of the DGINN analysis, and Y the number of the paralog in case more than one had been retrieved. After attribution of ortholog groups, DGINN checked for the presence of recombination events on the alignments using GARD. The fragments of genes found to present recombination breakpoints were annotated G(X)[Frag(X)] (Figure 18). TRIM5 was initially part of our dataset, but we could not get the complete results for positive selection analyses. This is likely due to the fact that we retrieved 162 sequences for TRIM5, including several TRIMcyp sequences. The combination of this massive alignment, with the evolutionary complexity induced by the presence of the TRIMcyp sequences, likely explained our difficulty in obtaining results for codeml M1 vs M2 and M7 vs M8 for this gene. Given that the role of TRIM5 in HIV infection has been extensively characterized, and its evolutionary history is well known (Johnson and Sawyer 2009), we took out this gene from further analyses. We, however, decided to keep the three paralogs retrieved by DGINN with the TRIM5 query, for which we had complete results (annotated TRIM5Para1 through 3, see Figure 18).

Out of the fifty-five remaining genes and their fifteen paralogs, twenty-three were not detected by any methods of positive selection, fifteen by one method, nine by two methods and ten by three methods (Figure 18, left side). Twelve genes were found under strong positive selection by four or five out of five methods, which indicates robust detection of

the positive selection signal. Of those twelve, four were DGINN-retrieved paralogs: two were paralogs of Gene31, which is itself detected by five methods (Figure 18), one was a paralog of TRIM5, a gene well known to evolve under strong positive selection, and the last one was a paralog of Gene24 which, in opposition to the previous examples, is only detected by one out of five methods.This left eight genes out of fifty five which were detected under strong positive selection. Amongst those, some are restriction factors of HIV AIM2, BST2 and SAMHD1, known to be under strong positive selection (highlighted in Figure 18).

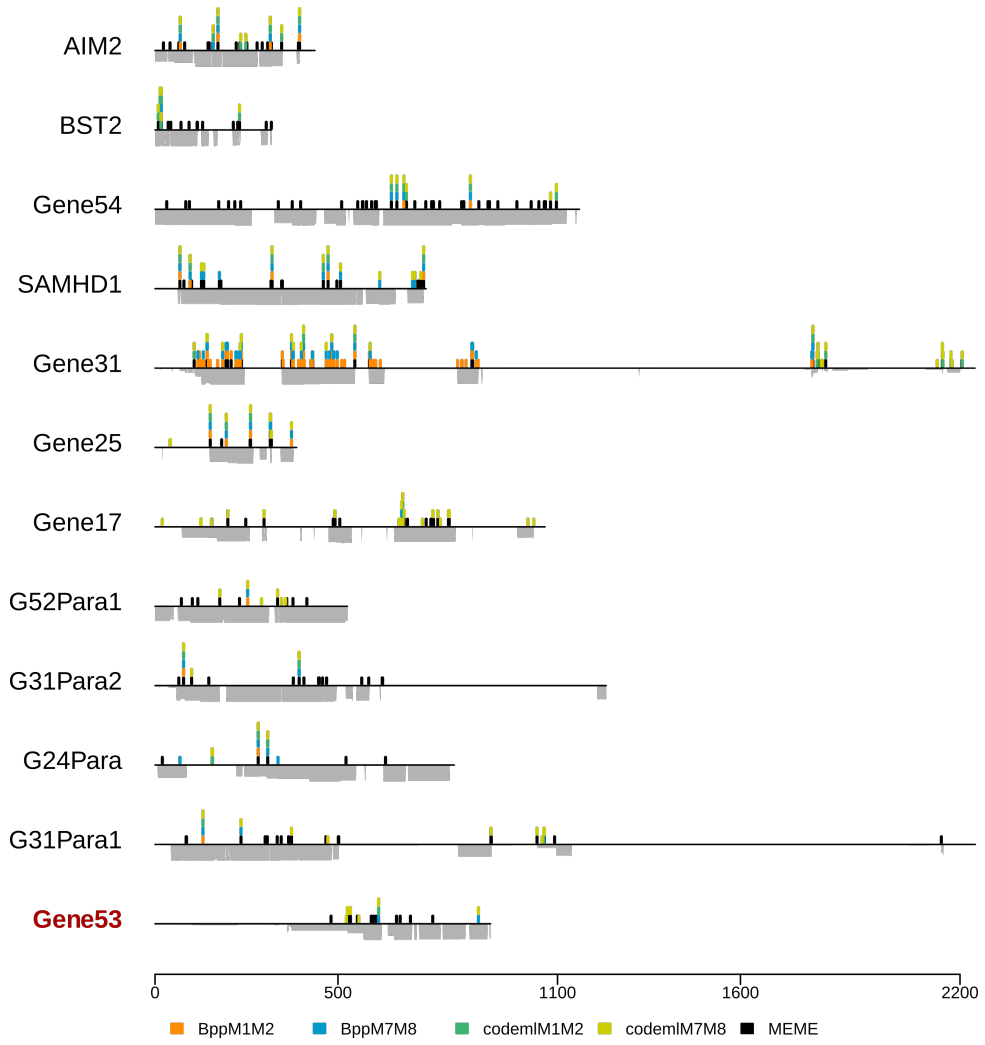


Figure 19: Positive selection patterns on best hits

Genes detected by at least four methods were included in this representation. Positively selected sites are represented as a spike at their position on the alignment. Height of the peak is proportional to the number of methods that have identified the site as being under positive selection (posterior probabilities > 0.90 for Bio++ and PAML codeml, and p-value < 0.10 for MEME), with each method being represented by a different color (embedded legend). HYPHY MEME sites were only mapped if the gene was detected as under positive selection by BUSTED ($p < 0.05$). For each gene, alignment coverage is represented under the line representing the length of the alignment in light grey. Gene53 is highlighted in red.

We also calculated the percentage of positively selected sites (PSS) over the total number of positions in the alignment (Figure 18, right side). Overall, genes did not present a clear pattern in terms of rates of PSS, as we had already observed previously (Picard et al. 2020). Genes detected by at least four methods presented overall higher rates of PSS over the total length of the alignment (from % in G31Para1 up to 5.32% in the BST2 alignment). However, genes detected by only one or two methods could in some instances present similarly high rates of PSS, as is evidenced by Gene30 (3.14% PSS). This hike in PSS rates was observed in genes detected as under positive selection by the combination of the BUSTED and MEME methods (see Material and Methods). Indeed, MEME allows for ω to vary over both branches and sites (branch-site method), which makes it more sensitive to amino-acid changes present in only very few branches. We observed that this bias often led to the detection of sites that appeared dubious upon closer inspection. This highlights the utility of combining different methods to confirm the signatures of positive selection observed on each site.

Gene53 is a potential restriction factor of HIV-1

We further studied the selection profiles established by DGINN on the twelve best hits (genes detected by four or more methods of positive selection, Figure 19). Four were paralogs retrieved by DGINN and did not belong to the initial dataset. They were, as such, not of direct interest as potential novel viral modulators of HIV infection, except to indicate the potential presence of another hallmark of genetic conflict, duplication (see Introduction). Amongst the eight genes from the initial dataset, AIM2, BST2 and SAMHD1's roles in HIV infection have been extensively described previously. Out of the remaining five genes, Gene53 has started to garner attention about its potential antiviral role (Figure 19). Gene53 also belongs to a large multigenic family of which numerous members have been shown to have diverse antiviral roles, which further underlies its potential involvement in the restriction of HIV-1. Experiments conducted in parallel by Anuj Kumar in the lab showed that knocking down Gene53 in primary macrophages increased HIV-1 infectivity by 10 folds but did not affect HIV-2 or SIVmac infectivity (Figure 20A). Furthermore, this knock-down also led to an increase of virion production for a replication competent strain of HIV (NL-ADA), as measured by the increase in Reverse Transcriptase (Figure 20B).

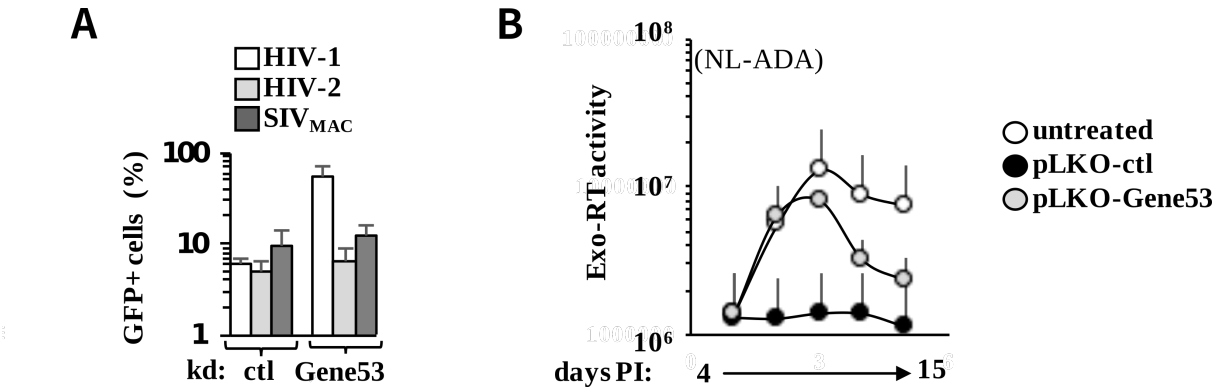


Figure 20: Gene53 is a specific modulator of HIV-1 infection

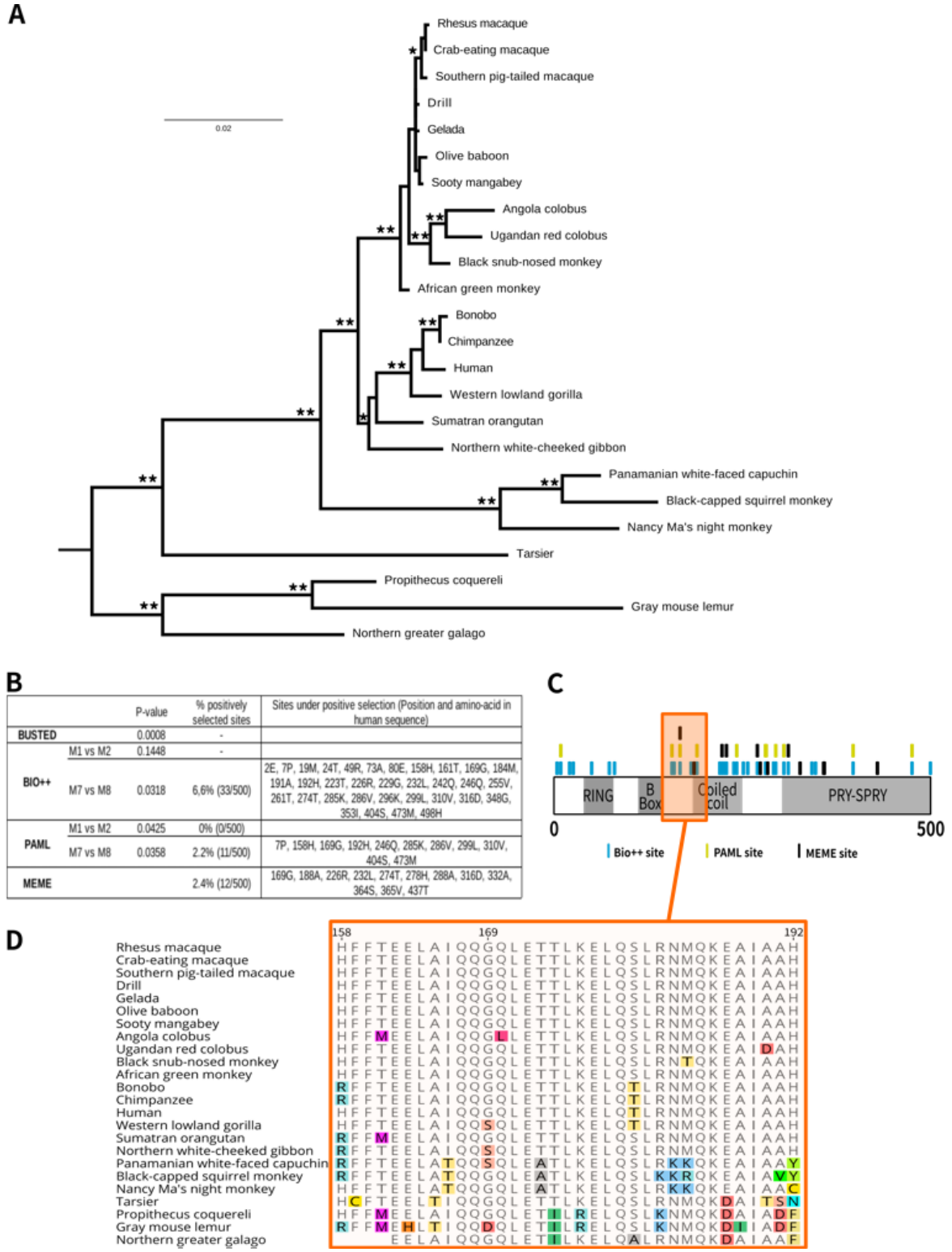
(A) Monocyte-derived macrophages were challenged with shRNA-expressing HIV-1-based lentiviral vectors to silence gene expression. Three to four days post silencing, cells were challenged with the indicated single round of infection competent GFP-coding viruses to examine the effects of gene silencing on the early phases of the viral life cycle. The extent of infection was measured three days later by flow cytometry. **(B)** as in A but silenced cells were challenged with a multiplicity of infection (MOI) of 0,1 of a replication-competent HIV-1 strain (NL-ADA). Viral spread through culture was determined through the measure of the accumulation of exogenous-RT activity in the supernatant of infected cells at the indicated days post-infection. Panels present the mean and standard error of 3 to 4 independent experiments. Figure courtesy of A. Cimarelli.

Combined together, the experimental results and the DGINN phylogenetic results suggest that Gene53 is a strong candidate for a novel HIV-1 restriction factor.

Gene53 has been evolutionarily constrained in primates

To retrace the evolutionary history of Gene53 in more depth and eliminate all the biases possibly introduced by the automatic retrieval of homologs by DGINN (such as the over-representation of sequences from one species), we retrieved the nucleotide sequences from 24 primate species from the NCBI databases (Figure 21A). The alignment and phylogeny generated from these curated sequences were assessed for the presence of genetic innovations using DGINN following the same steps as for the screening. In this curated analysis, we confirmed that Gene53 presents no sign of recombination during primate evolution. We also confirmed the positive selection screen, as four methods out of five detected positive selection at the gene level ($p < 0.05$, Figure 21B), and only Bio++ M1 vs M2 did not ($p = 0.1448$, Figure 21B). We found that up to 6.6% of sites have evolved under significant positive selection (with Bio++ M7 vs M8, Figure 21B). These sites are distributed over the whole length of the coding sequence and do not appear to cluster exclusively in a particular domain, but did present a slighlty higher concentration around the central part of the protein (Figure 21C). Interestingly, one site (169G) was in common across all three methods, highlighting its possible functional importance (Figure 21B, C,

D).



		P-value	% positively selected sites	Sites under positive selection (Position and amino-acid in human sequence)
BUSTED		0.0008	-	
BIO++	M1 vs M2	0.1448	-	
	M7 vs M8	0.0318	6,6% (33/500)	2E, 7P, 19M, 24T, 49R, 73A, 80E, 158H, 161T, 169G, 184M, 191A, 192H, 223T, 226R, 229G, 232L, 242Q, 246Q, 255V, 261T, 274T, 285K, 286V, 296K, 299L, 310V, 316D, 348G, 353I, 404S, 473M, 498H
PAML	M1 vs M2	0.0425	0% (0/500)	
	M7 vs M8	0.0358	2.2% (11/500)	7P, 158H, 169G, 192H, 246Q, 285K, 286V, 299L, 310V, 404S, 473M
MEME			2.4% (12/500)	169G, 188A, 226R, 232L, 274T, 278H, 288A, 316D, 332A, 364S, 365V, 437T

Figure 21: Evidence of site-specific positive selection in Gene53 during primate evolution

Phylogenetic analyses of primate Gene53 were performed on twenty-four orthologous nucleotide sequences that were aligned with PRANK. **A**, Phylogeny was performed with PhyML with an HKY+G+I model and 1,000 bootstrap replicates as statistical support. Bootstrap values above 800/1,000 (*) and 900/1,000 (**) are shown above the branches. The scale bar indicates the number of nucleotide substitutions per site. **B**, Positive selection analyses of Gene53 with four different methods (BUSTED, Bio++, PAML codeml and MEME, see Methods) and their associated p-values from the Likelihood Ratio Tests. **C**, Repartition of the sites detected as under positive selection by the different methods over a schematic representation of GENE and its domains. **D**, Snapshot of the alignment from positions 158 to 192, including position 169G, with sequences ordered to follow the phylogenetic tree in panel A, produced using Geneious Prime.

3.3 Conclusion

In conclusion, we performed an evolutionary screen which allowed us to identify five genes, including Gene53, as under strong positive selection, thus potentially encoding for VIPs. Since Gene53 has recently been shown to have antiviral activities against different viruses, including, as shown in this work, against HIV, this suggests Gene53 has potentially been engaged in an host-virus genetic conflict. We further confirmed that Gene53 is evolving under positive selection through in-depth curated analyses, pinpointing the potential functional importance of the 169G site.

Of note here, Gene31 appears to present multiple hallmarks of genetic conflict: it is part of a large multigenic family (consisting of thirteen paralogs) of receptors, and our results show that itself and at least two of its paralogs appear to have evolved under a strong pressure of positive selection. Given those indications, it appears that retracing the complete evolutionary history of this multigenic family could yield potentially interesting results to identify a long-term genetic conflict between its members and their ligands. We could not automatically resolve the broad strokes of this evolutionary history using DGINN: while it managed to retrieve a large number of homologs (741 sequences), it failed to properly segregate them to different ortholog groups. This is consistent with our previous observation that genes belonging to large multigenic families with complex evolutionary histories and high degree of similarities are harder to properly segregate into ortholog groups (Picard et al. 2020).

Chapter 4

Identification of evolutionarily-relevant modulators of HIV in a dataset derived from a transcriptomic screen

This work was the main project of my PhD for the molecular biology aspects. I performed the wet lab experiments related to the functional characterization of our protein of interest with the technical help of Clara Dahoui. Other collaborators include Laurent Guéguen, Lucie Etienne and Andrea Cimarelli. All results are preliminary.

4.1 Material and methods

Initial dataset

A transcriptomics analysis of genes differentially expressed in macrophages infected *in vitro* by HIV-1 had previously been performed in the lab. Macrophages were obtained from the blood of three donors, and were either infected with HIV-1 or with a mock vector. Relative expression of mRNAs was determined by microarray and by averaging the mean levels of expression for each of the three infected conditions against the mean for each of the mock conditions. We focused on genes that were upregulated in infected macrophages by a minimum Fold Change (FC) of 2, and we further filtered them by selecting those which are reported as responding to Type I Interferon (see Introduction - chapter 3) by a

minimum FC of 10. This yielded us a list of 60 Interferon-stimulated genes likely to play a role in lentiviral transmission and early infection.

Evolutionary screen to detect genetic innovations in the protein-coding genes of the dataset

We screened the 60 genes of interest using the previously described DGINN pipeline (see Results - Chapter 1 and Picard et al. 2020).

We downloaded the Consensus CoDing Sequences (CCDS) from the CCDS database using the CCDSQuery script provided with DGINN. In cases where multiple CCDS were referenced for one gene, we kept the longest one. Those sequences were then used as the entry query for DGINN. We performed the phylogenetic analyses (step 1-7 of the DGINN pipeline, including duplication and recombination) against the NCBI nr database limited to primate species, with otherwise default parameters (blastn e-value 10^{-4}, identity 70% and coverage 50%, at least 8 species for separation of orthologous groups). The species tree used to identify duplication events is the same as used previously in Results - Chapter 1 and Picard et al. 2020, and is based on the phylogeny of primates established by Perelman et al. 2011 and updated by Pecon-Slattery 2014.

All alignments and phylogenetic trees produced during this first part were then analyzed for marks of positive selection using the five different methods included in DGINN. It uses BUSTED and MEME from the Hyphy package (Pond et al. 2005) to look for gene-wide and site-specific episodic positive selection. We considered genes as under positive selection for a BUSTED p-value < 0.05 and sites for a MEME p-value < 0.10. Codeml from PAML (Yang 2007) and Bio++ (Guéguen et al. 2013) were used to run codon substitution models M1, M2, M7 and M8. M1 and M7 are neutral models not allowing for positive selection and M2 and M8 are their pendant allowing some codons to evolve under positive selection. We derived p-values from the likelihood ratio tests between the two models (M1 vs M2, M7 vs M8) to determine which model is a better fit for the data. We considered genes as under positive selection for $p < 0.05$ and sites for posterior probabilities > 0.95.

In-depth phylogenetic and positive selection analyses on TMEM140

We manually retrieved primate sequences for TMEM140 from available primate genomes using tBlastn on the nr database of the NCBI, with the human protein sequence as

Table 3: Species name for the six chosen orthologs.

Species name	Common name	DGINN nomenclature	Clade
Macaca mulatta	Rhesus macaque	macMul	Old World Monkeys
Chlorocebus sabaeus	African green monkey	chlSab	Old World Monkeys
Homo sapiens	Human	homSap	Hominoids
Nomascus leucogenys	Northern white-cheeked gibbon	nomLeu	Hominoids
Cebus capucinus	Panamanian white-faced capuchin	cebCap	New World Monkeys
Otolemur garnettii	Northern greater galago	otoGar	Prosimians

query. Using DGINN, we then aligned sequences with PRANK (Loytynoja and Goldman 2008) with a codon model and without forcing insertions to be skipped (prank -F -codon; version 150803). We reconstructed the phylogenetic tree using PhyML (v3.2, Guindon et al. 2010) with HKY+G+I model and 1000 bootstraps for statistical support of the branches. We performed positive selection analyses using the five methods included in DGINN as described in the previous section.

Plasmids and reagents

Six primate orthologues of TMEM140 with an N-terminal HA tag were synthetized and cloned in a pcDNA3.1+ expression vector (Genewiz). The species are detailed in table 3. The following plasmids to produce HIV-1 pseudoparticles have been described before: viral mini-genome coding GFP (pNaldini GFP), pHIV-1 Gag-Pol (8.2), pVSVg Env.

For Western blot analyses, the following antibodies were used: anti-Tubulin (Mouse, T5168, Sigma), anti-HA (Mouse, H9658-2mL, Sigma), anti-mouse HRP conjugate (P0260, Dako). All were diluted to $1/5000^{\text{th}}$.

Plasmid transfection

HEK 293T cells were grown in complete DMEM medium until 40-60% confluence was reached in 6 or 12-well plates, then transfected either by calcium phosphate/HBS (HEPES-buffered saline) or by TransIT®-LT1 (Mirus). Cells were collected 48 to 72h post-transfection.

Viral production

HIV-1-like viral particles encoding the Green Fluorescent Protein (GFP) were produced through transfection of HEK 293T cells by calcium phosphate/HBS of pNaldini GFP, pHIV-1 Gag-Pol and pVSVg Env. Supernatant was collected, filtered after 48h and

conserved at −80°C pending utilization. The amount of viral supernatant necessary to reach a rate of 30% infected cells was determined by testing a range of volumes (2, 5, 10 and 50 μL) on HEK 293T cells and titration of GFP-positive cells through FACS.

Measure of viral infectivity through FACS

Cells infected with the pseudoviral particles were resuspended in 2% PFA (ParaFormAldehyde) in PBS, three days post-infection, to measure accumulation of GFP in target cells. Measures were carried on BD FACSCanto™ II.

Inhibition of the proteasome degradation pathway

293T cells were transfected with 1200ng of plasmid encoding for either homSap or macMul TMEM140, as previously described. After 24h, medium was changed for medium + DMSO or medium + Mg132 (10 μM) and left to incubate for 2, 15 or 20h prior to cell collection.

4.2 Results

Update of the original dataset

We analyzed our initial dataset of sixty genes for hallmarks of genetic conflicts using DGINN, in the same manner as described in the previous chapter (see Material and Methods, Results - Chapter 1 and Results - Chapter 3). We excluded three genes from the initial dataset: IFITM1, HLA-F and SP110, bringing the number of genes screened down to fifty-seven.

IFITM1 was excluded due to DGINN failing to retrieve homologs spanning the complete phylogeny of primates, as we observed an overrepresentation of Old World Monkeys (OWM) species in the retrieved sequences. This caused major difficulties in the segregation of ortholog groups by DGINN, an issue compounded by the presence of numerous IFITM3 pseudogenes in OWM.

The involvement of HLA-F (HLA class I histocompatibility antigen, alpha chain F) in the Major Histocompatibility Complex (MHC) makes it particularly unsuitable for automated evolutionary analyses. Numerous alleles are maintained in the population through balancing selection to enhance the MHC's ability to recognize numerous pathogens (as reviewed in Quintana-Murci 2019). This caused DGINN to retrieve all those alleles, with

major overrepresentation of some species.

Concerning SP110, we were unable to obtain the complete results for positive selection analyses. Only the combination of the BUSTED and MEME methods could run. We considered that having only one method out of five did not allow for enough robustness in our analyses and consequently excluded this gene from subsequent analyses. In future analyses, it would be interesting to perform positive selection analyses on a hand curated alignment to ensure complete results on the entire dataset.

Screening results

DGINN analyses initially yielded 28 groups of DGINN-retrieved paralogs for twelve of the query genes. We excluded two of those groups due to insufficient species representation for robust positive selection analyses. A further eight were excluded because they overlapped. Indeed, several of the query genes are paralogs from the same multigenic families and were therefore retrieved several times in DGINN.

Out of the 57 remaining genes and their eighteen paralogs, eleven were not detected by any methods of positive selection, thirteen by one method, four by two methods and thirteen by three methods (Figure 22, left side). Thirty-four genes were found under positive selection by four or five methods, indicating robust detection. This represents almost half of all the analyzed genes. Of those 34, eight were DGINN-retrieved paralogs: the combination of belonging to a multigenic family and having evolved under positive selection make them likely to be engaged in a genetic conflict. This is the case of the APOBEC3 family, which we have discussed previously: the initial query of APOBEC3A allowed the retrieval of two paralogs, APOBEC3B and APOBEC3G, both of which are detected by at least four methods of positive selection (Figure 22). A similar pattern can be observed for GBP5 and its paralogs GBP7, 2 and 3, which we discussed in Results - Chapter 1.

Out of the remaining 26, several have already been functionally characterized, notably for their antiviral functions. From the subset of those which had not been, we were particularly interested in the TransMEMbrane protein 140 (TMEM140). This gene was detected under positive selection by all of the five methods, and presented the highest rate of PSS over its alignment length (11.8%, Figure 22, right side). This high rate does not appear to be linked to an overdetection of PSS by MEME, and the pattern of positive selection for TMEM140 show that numerous sites are detected by three to four

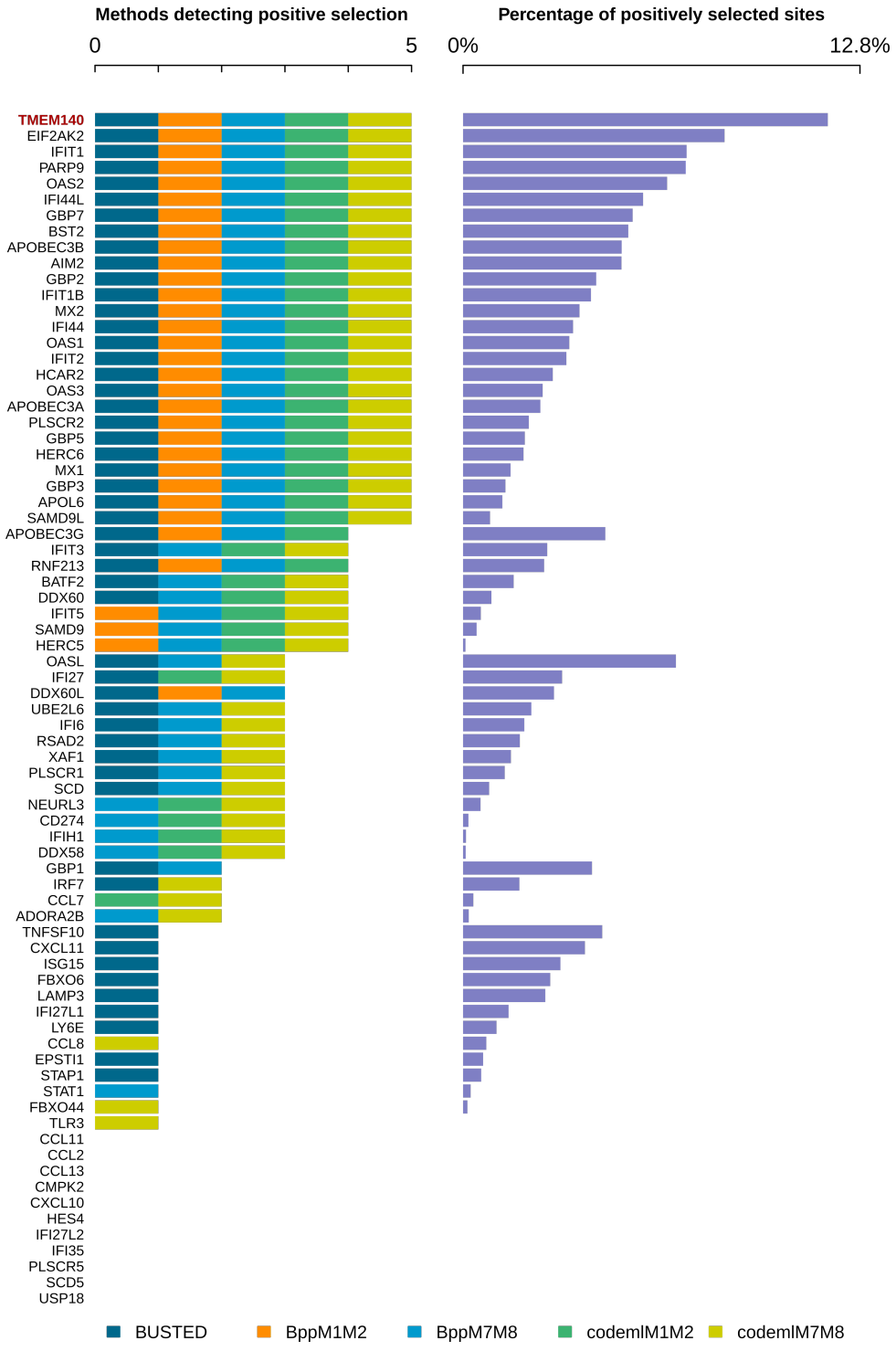


Figure 22: DGINN results on 57 primate genes and their paralogs

Left panel, number of methods detecting significant positive selection for each alignment; each method is color-coded (embedded legend). Right panel, percentage of positively selected sites (by at least one method) over the length of the alignment. Genes are ordered by descending number of methods detecting positive selection then descending percentages of positively selected sites. TMEM140 is highlighted in red.

methods (Figure 23). Such a strong and widespread signal is reminiscent of canonical arms-race genes such as APOBEC3s, EIF2AK2 (which encodes PKR) and IFIT1 found

in this screen. Combined with the context of our screen, which is based on ISGs that affect HIV-1 replication in macrophages, this evidence that TMEM140 has evolved under positive selection makes it a strong candidate as a potential HIV-1 VIP.

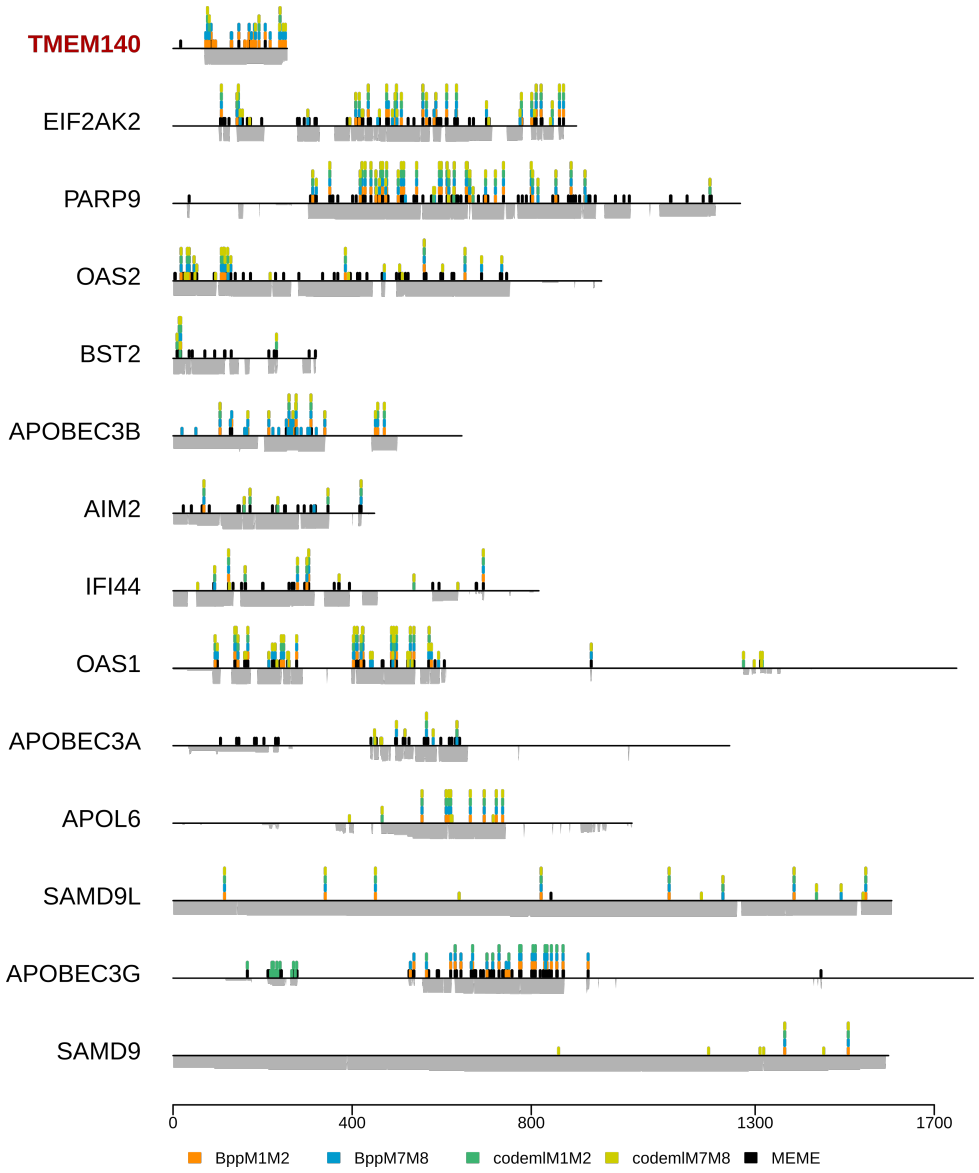


Figure 23: Positive selection patterns on a subset of the best hits

A subset of fourteen genes detected by at least four methods were included in this representation. Positively selected sites are represented as a spike at their position on the alignment. Height of the peak is proportional to the number of methods that have identified the site as being under positive selection (posterior probabilities > 0.95 for Bio++ and PAML codeml, and p-value < 0.10 for MEME), with each method being represented by a different color (embedded legend). HYPHY MEME sites were only mapped if the gene was detected as under positive selection by BUSTED ($p < 0.05$). For each gene, alignment coverage is represented under the line representing the length of the alignment in light grey. TMEM140 is highlighted in red.

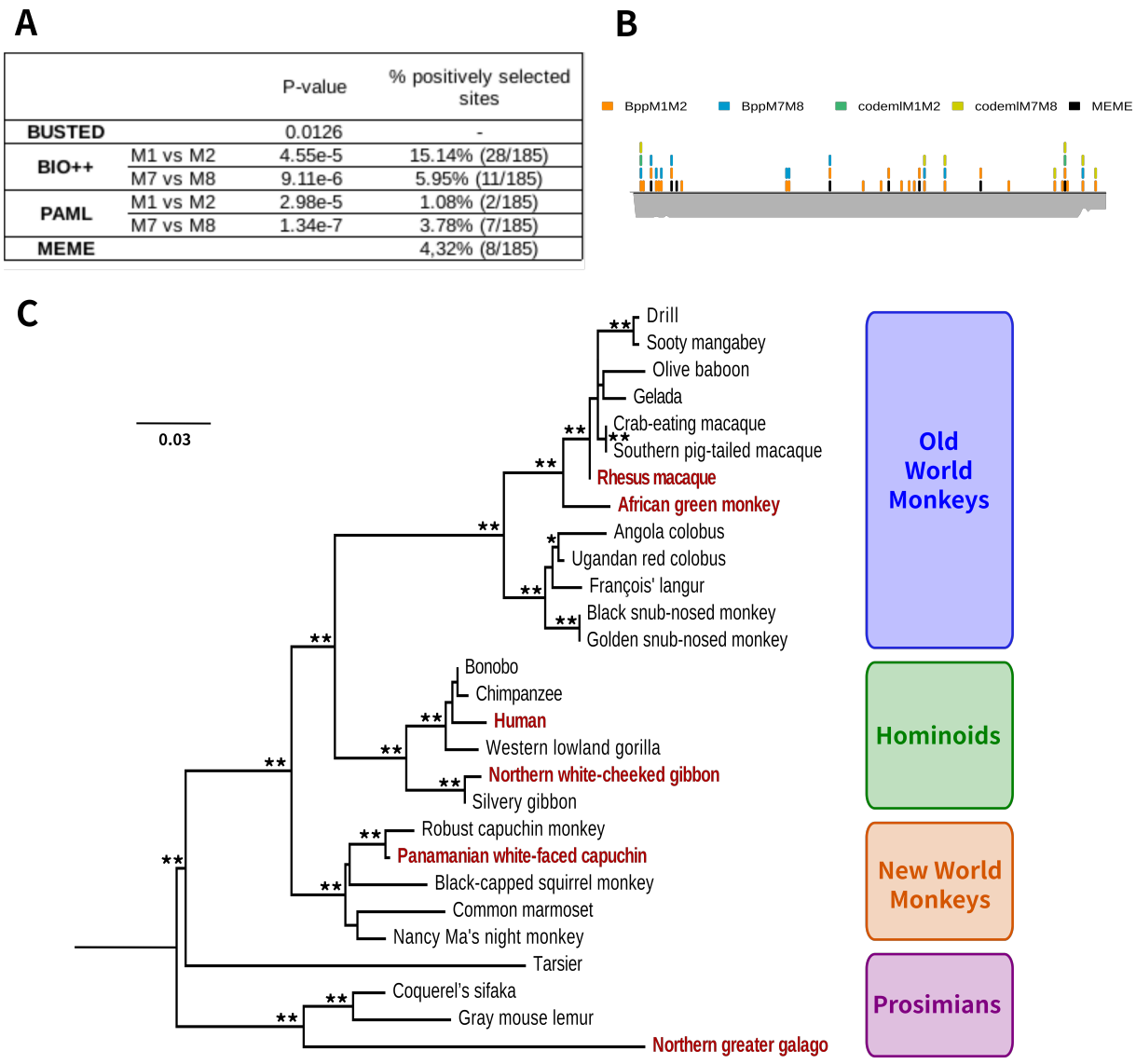
TMEM140 has been under strong adaptive evolution in primates

To retrace the evolutionary history of TMEM140 in more depth and eliminate all the biases possibly introduced by the automatic retrieval of homologs by DGINN, we retrieved the nucleotide sequences from 28 primate species from the NCBI databases. The alignment and phylogeny generated from these curated sequences were assessed for the presence of genetic innovations using DGINN following the same steps as for the screening. In this curated analysis, we confirmed that TMEM140 presents no sign of recombination during primate evolution. We also confirmed the positive selection screen, all five methods detected positive selection at the gene level ($p < 0.05$, Figure 24A). We found that up to 15.14% of sites have evolved under significant positive selection, with a maximum number of 28 out of 185 codons in the TMEM140 sequence identified as PSS by Bio++ M1 vs M2 (Figure 24A). These sites are distributed over the whole length of the coding sequence, with a cluster at each extremity of the protein, and a more spread-out concentration in the central region of the protein (Figure 21C). Interestingly, almost all the PSS were detected by more than one method, which suggest strong selective pressures at these positions (Figure 24B). If TMEM140 has evolved under genetic conflict due to a virus protein, these PSS could therefore be the functional determinants of the selective pressure.

To test this hypothesis, we cloned six primate orthologs of TMEM140 so as to have the highest variability possible at the PSS. We also focused on spanning as much of the primate phylogeny as possible (Figure 24C). To this end, we selected TMEM140 from two species of the Old World Monkey clade, the rhesus macaque and the African green monkey, two species from the Hominoid clade, human and the Northern white-cheeked gibbon, and one New World Monkeys and one Prosimians (Panamarian white-faced capuchin and Northern greater galago, respectively) (Figure 24C). Our aim was to perform heterologous viral infectivity assays (as reviewed in Sawyer and Elde 2012; Sironi et al. 2015) to check if any of those orthologs had an effect on HIV, and if that effect varied in a species-specific manner.

Macaque and capuchin TMEM140 do not affect HIV-1 replication in 293T cells, but primate TMEM140 bear species-specific differences in their protein stability

To determine the effect of TMEM140 on the early phases of HIV-1 infection, we first



		P-value	% positively selected sites
BUSTED		0.0126	-
BIO++	M1 vs M2	4.55e-5	15.14% (28/185)
	M7 vs M8	9.11e-6	5.95% (11/185)
PAML	M1 vs M2	2.98e-5	1.08% (2/185)
	M7 vs M8	1.34e-7	3.78% (7/185)
MEME			4,32% (8/185)

Figure 24: Evidence of site-specific positive selection in TMEM140 during primate evolution

Phylogenetic analyses of primate TMEM140 were performed on twenty-eight orthologous nucleotide sequences that were aligned with PRANK. **A**, Positive selection analyses of TMEM140 with four different methods (BUSTED, Bio++, PAML codeml and MEME, see Methods) and their associated p-values from the Likelihood Ratio Tests. **B**, Selection pattern of TMEM140 based on in-depth positive selection analyses. **C**, Phylogeny was performed with PhyML with an HKY+G+I model and 1,000 bootstrap replicates as statistical support. Bootstrap values above 800/1,000 (*) and 900/1,000 (**) are shown above the branches. The scale bar indicates the number of nucleotide substitutions per site. Six orthologs of TMEM140, spanning as much as possible of the primate phylogeny, were selected for further functional characterization, based on their variability at positively selected sites. They are highlighted in red.

transfected 293T cells with the plasmids containing the HA-tagged TMEM140 ortholog of each of the six species (see Table 3 for the nomenclature). We found that expression of macMulTMEM140 and cebCapTMEM140 was much stronger than for the other four orthologs (Figure 25A). This suggests that those two orthologous proteins are more stable. We also observed two bands of higher molecular weight for those two orthologs, suggesting that high concentrations of TMEM140 lead to the acquisition of post-translational

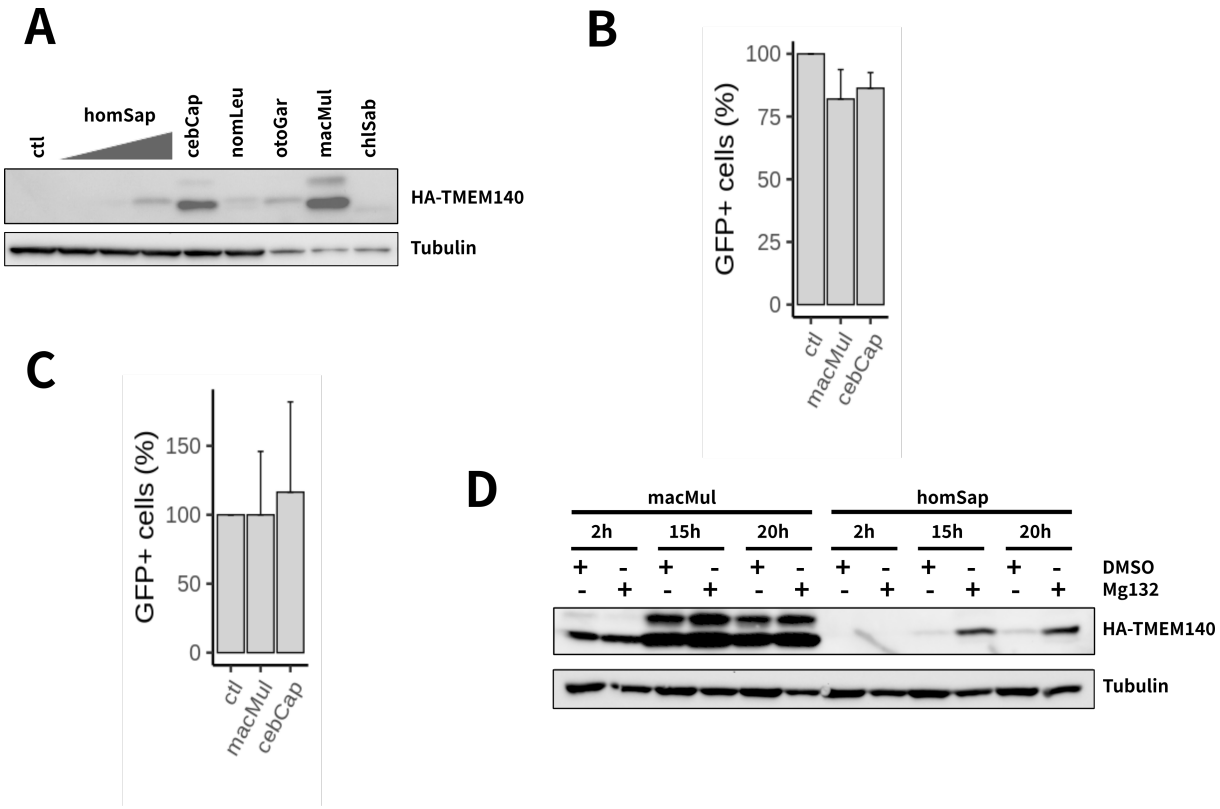


Figure 25: Preliminary results show that primate TMEM140s have no effect on HIV-1 replication during early or late phases, but reveal species-specific variation in its protein stability.

A, 293T cells were transfected with six different orthologs of TMEM140 and lysed 24h post-transfection for analysis by Western Blot to measure the overexpression of TMEM140. 1.2μg of each plasmid was transfected, except for homSapTMEM140 which was transfected with 0.3, 0.6 and 1.2μg. **B**, 293T cells were transfected with either an empty vector, or macMulTMEM140 or cebCapTMEM140 and challenged 24h later with GFP-coding HIV-1 pseudoparticles to examine the effects of TMEM140 overexpression on the early phases of the viral life cycle. The extent of infection was measured three days later by flow cytometry. Results are the mean and standard error of two independent experiments with two replicates each, normalized to the control (empty vector condition normalized to 100%). **C**, 293T cells were transfected with the indicated TMEM140 orthologs along with the plasmids necessary for the production of single-round GFP-coding HIV-1 pseudoparticles. 10μL of supernatant were then collected and used to infect fresh 293T cells to examine the effects of TMEM140 overexpression on the late phases of the viral life cycle. The extent of infection was measured three days later by flow cytometry. Results are the mean and standard error of four independent experiments, normalized to the control (empty vector condition normalized to 100%). **D** 293T cells were transfected with either macMulTMEM140 or homSapTMEM140 and treated with MG132 then lysed after the indicated lengths of time to measure the effect of proteasomal degradation on TMEM140 expression levels. DMSO was used as control in the experiment. Tubulin serves as loading control.

modifications such as ubiquitination or phosphorylation, or (and more likely) that those posttranslational modifications are only observable at high concentrations (Figure 25A). We also observed that chlSabTMEM140 was of lower molecular weight, which is due to a premature STOP codon in the sequence (Figure 25A). The two highly expressed macMulTMEM140 and cebCapTMEM140 were then tested in the context of HIV infection. We observed no significant effect of either of those orthologs on the early phases (Figure

25B) or late phases of HIV replication in 293T cells (Figure 25B-C). To address the issue of stability, we treated cells transfected with macMulTMEM140 or homSapTMEM140 with an inhibitor of the proteasome, MG132. We observed higher expression of each of the orthologs at 15h and 20h post-transfection, suggesting that TMEM140 is indeed degraded through the proteasome (Figure 25D).

4.3 Conclusion

We found that TMEM140 has evolved under strong positive selection and has likely been engaged in a genetic conflict during its evolutionary history in primates. However, our preliminary analyses suggest that lentiviruses do not appear to have driven this adaptation, as we did not observe any effect of TMEM140 on its replication or species-specificity. It is worth noting, though, that all our experiments were carried out in 293T cells, which are not be the most relevant cellular model for observing a potential effect of TMEM140 on HIV replication. Indeed, 293T cells may not express protein partners necessary to mediate the activity of TMEM140, for example. As such, further experiments could be carried out in different cell types, such as the macrophage-like cells THP-1, or even in primary macrophages themselves. Another experimental perspective would be to test fully replication-competent HIV virus, instead of a pseudovirus using VSVg as entry factor. Moreover, testing other lentiviral strains would help to dissect species-specific differences. Alternatively, other viruses may be responsible for the positive selection observed on TMEM140. Indeed, although very little is known on TMEM140, it was shown to have an antiviral effect against HSV-1 through its inhibition of the UL31/UL34 complex mediating nuclear egress of HSV-1 nucleocapsid (Guan et al. 2014). Our six orthologs of TMEM140 could therefore be tested to ascertain whether the PSS we identified play a role in a species-specific response against HSV-1.

Another aspect worthy of further investigation is the differential stability between the six orthologs. Stability has been shown to be an important factor in APOBEC3H antiviral effect against HIV-1, both at the intrapopulation level (OhAinle et al. 2008; Refsland et al. 2014) and the inter-species level (Zhang et al. 2017), and has been lost twice in human evolution through the acquisition of independent amino-acid mutations (OhAinle et al. 2008). The PSS we detected may play a similar role in the stability of TMEM140 across species.

Discussion

DGINN provides an important addition to the available resources for the detection of genetic innovations

The pipeline presented in this study was born from careful consideration of the available resources for automated evolutionary analyses, specifically those geared toward the identification of genetic conflicts between hosts and viruses. This field has garnered considerable attention in the past, as highlighted by the growing literature combining evolutionary and functional analyses of VIPs (e.g. Sawyer et al. 2005; Elde et al. 2009; Mitchell et al. 2012; Fregoso et al. 2013; McCarthy et al. 2015; McLaren et al. 2015; Xu et al. 2016). The recent emergence of the SARS-CoV-2 strain, and its resultant pandemic, has provided even stronger impetus for the ability to quickly identify and characterize VIPs, as potential targets for therapeutic treatments. A recent study established an interaction map for SARS-CoV-2 through mass-spectometry (Gordon et al. 2020), and they included the results of positive selection in primates for the 332 human genes they identified. It is, for example, our hope that DGINN can be used in the future to facilitate such high-throughput analyses for the scientific community.

Indeed, DGINN presents several advantages that we feel may be of help to the field. We have focused on making it as easy as possible to use for biologists, for example through the use of a parameter file to handle most parameters, but also by providing a docker version, freeing the user from having to install the numerous softwares included in the pipeline. Its flexibility will allow users to perform any subset of analyses of their choice, from fine-tuned analyses on manually curated alignments to the rapid screening of large datasets.

Moreover, DGINN fully automates every step of the process, which few available tools do (Sahm et al. 2017; Fuchs et al. 2017). The only files needed to start an analysis are the coding sequence of the gene of interest, and, ideally, a cladogram of the species of interest for the proper detection and attribution of duplication groups. This means large scale screens can be performed with minimal input from the user. In the lab, we are intending to use it to screen all genes responding to Type I Interferon.

We are also aiming at characterizing the evolutionary histories of the 332 VIPs of SARS-CoV-2 identified by Gordon and colleagues in primates, to retrace their human history, bats, the virus' natural reservoir hosts, and mammals, which includes species that

are either susceptible or resistant to the virus, as well as symptomatic and asymptomatic. The identification of which of those VIPs might have been involved in host-pathogen arms-race will allow researchers to select evolutionarily-relevant genes for further functional characterization. Moreover, identifying species specifies in the context of infection by this virus, either regarding resistance/susceptibility status or the expression of symptoms will further allow to pinpoint key players. Such host factors are at the forefront of the virus-host interface and may represent species-barriers to viral spillovers and primary drug targets (Sawyer and Elde 2012).

Another important feature of DGINN is its ability to detect and assign duplication groups. Duplication events are generally not accounted for in previously available pipelines, though they are an important potential hallmark of genetic conflict. Indeed, numerous restriction factors are part of multigene families (APOBEC3s, IFITMs, IFITs, MX1/2...). DGINN's ability to retrieve different paralogs of a multigene family allows for the reconstitution of a more complete evolutionary history of each gene, and lead to the identification of multiple VIPs in one analysis as exemplified in our Validation dataset from Chapter 1 and our screens in Chapters 3-4.

However, the detection of duplication events is also a part that has the potential for most improvements. Indeed, we observed that, in the case of large multigenic families and complex evolutionary histories, DGINN sometimes failed in its attribution of duplicated groups. In some occasions, different paralogs were mixed in the same duplication group. In others, one paralog was split in two different groups erroneously. Longer branches (due, for example, to unbalanced sampling) can also cause erroneous group attributions. We observed this on three multigenic families in our various analyses in primates, with various degrees of severity: the APOBEC3 family (Results – Chapter 1 and 4), the Gene31 family (Results – Chapter 3) and the IFITM family (Results – Chapter 4). In the first instance, the errors in group attribution did not significantly impact our ability to perform analyses on various members of the multigenic family. In the second one, DGINN retrieved a large number of the thirteen members of the multigenic families, but failed to properly reconstitute groups for all but three of them. Finally, the impact on the last family was so important that no robust positive selection analyses could be conducted. All of those points will be a focus of improvement in future versions of DGINN.

Finally, DGINN incorporates several methods for the detection of positive selection,

which few other pipelines do. A major goal in subsequent versions will be to integrate new methods as they become available. For example, the future version of Bio++ should handle polymorphisms in positive selection analyses, which would mitigate the bias introduced when DGINN retrieves several alleles of one gene in some species.

Evolutionary analyses can inform and guide the characterization of novel VIPs

DGINN was used in the lab to identify evolutionary-relevant genes of interest for further functional characterization. Its integration of five different methods for the detection of positive selection allows us to "rank" screened genes and to focus on the ones detected by the largest number of methods first, and then progress in a descending manner along the list as we eliminate genes with already described anti-retroviral activities and lentiviral VIPs. With this approach, we identified Gene53, a VIP with a strong restrictive ability against HIV-1, and TMEM140, which evolves under very strong positive selection, in a manner reminiscent of canonical arms race genes such as APOBEC3s and TRIM5.

DGINN also allowed us to establish the selection profile of Gene53, both during the screening and on a curated alignment. This profile can further guide functional characterization at different levels. First, the sites detected under positive selection might be part of the virus-host interface (Sawyer et al. 2005; Mitchell et al. 2012), where the selective pressure exerted by the virus on the host protein would have been the highest. If not directly involved in the interaction with the virus, they might also be responsible for conformational changes which would favor or hinder to access to the interaction site, as is the case for the Mx1 L4 loop in primates (Colón-Thillet et al. 2019). Here, the fact that DGINN integrates multiple methods of positive selection is again beneficial: the more methods detect a site, the more likely it is to be under strong positive selection, and thus, of functional importance. Considering that Colon-Thillet and colleagues demonstrate the strong effect of epistasis between PSS on the ability of proteins to acquire stronger antiviral abilities, being able to identify the sites most likely to produce such effects would be very informative to retrace the evolutionary histories of antiviral proteins.

Those positively-selected sites can also be used to guide functional characterization of species-specificity (Sironi et al. 2015). The presence of different amino-acids between different species might indicate their importance for the maintenance of the species-barrier to other species' viruses. The profile of positive selection can then be used to select the orthologs to test by sampling for the highest amino-acid diversity at the positively selected sites.

However, it is worth noting that not all VIPs are evolving under positive selection

or bear hallmarks of genetic conflict. Our results on the CA-interacting protein NONO (Results – Chapter 2) showed no sign of positive selection in primates, and it was highly conserved in this clade. This suggests that the ability of NONO to recognize the lentiviral capsid is a recent evolutionary acquisition, and has not been present long enough for a genetic conflict to arise between HIV and NONO. Another non-exclusive possibility is that NONO targets a region of the viral CA which is critical for viral viability, making counter-adaptation harder for the virus as it would be costly in terms of fitness. This example shows that evolutionary analyses, though extremely useful, would not be able to identify such host proteins which have acquired their ability to interact with virus in recent evolutionary history and/or in such cases where one of the virus or the host protein is constrained in its evolution by other essential functions (Enard et al. 2016; Abdul et al. 2018).

Overall, using evolutionary analyses allows to better decorticate the functions of VIPs, and DGINN streamlines performing such analyses. It also allows researchers to completely reverse the usual workflow in characterizing genes under genetic conflict, and makes combining evolutionary and functional analyses easier. Instead of identifying potential VIPs by functional analyses then progress to mixed evolutionary and functional characterization, researchers will be able to first identify potential genes of interest by screening their evolutionary histories, in both a quicker and cheaper way. They will then be able to follow up on the best candidates with functional and mixed evolutionary and functional analyses.

Bibliography

Bibliography

Abascal, F., R. Zardoya, and M. J. Telford. 2010. "TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations". *Nucleic Acids Research* 38 (Web Server issue): W7–13. doi:10.1093/nar/gkq291.

Abdul, F., F. Filleton, L. Gerossier, A. Paturel, J. Hall, M. Strubin, and L. Etienne. 2018. "Smc5/6 Antagonism by HBx Is an Evolutionarily Conserved Function of Hepatitis B Virus Infection in Mammals". *Journal of Virology* (): JVI.00769–18. doi:10.1128/JVI.00769-18.

Ablasser, A., I. Hemmerling, J. L. Schmid-Burgk, R. Behrendt, A. Roers, and V. Hornung. 2014. "TREX1 Deficiency Triggers Cell-Autonomous Immunity in a cGAS-Dependent Manner". *The Journal of Immunology* 192, no. 12 (): 5993–5997. doi:10.4049/jimmunol.1400737.

Ajawatanawong, P., and S. L. Baldauf. 2013. "Evolution of protein indels in plants, animals and fungi". *BMC evolutionary biology* 13 (): 140. doi:10.1186/1471-2148-13-140.

Alexopoulou, L., A. C. Holt, R. Medzhitov, and R. A. Flavell. 2001. "Recognition of double-stranded RNA and activation of NF-kappaB by Toll-like receptor 3". *Nature* 413, no. 6857 (): 732–738. doi:10.1038/35099560.

Anisimova, M., R. Nielsen, and Z. Yang. 2003. "Effect of Recombination on the Accuracy of the Likelihood Method for Detecting Positive Selection at Amino Acid Sites". *Genetics* 164:1229–1236.

Ansari, M. A., S. Dutta, M. V. Veettil, D. Dutta, J. Iqbal, B. Kumar, A. Roy, L. Chikoti, V. V. Singh, and B. Chandran. 2015. "Herpesvirus Genome Recognition Induced Acetylation of Nuclear IFI16 Is Essential for Its Cytoplasmic Translocation, Inflam-

masome and IFN- Responses". *PLoS pathogens* 11, no. 7 (): e1005019. doi:10.1371/journal.ppat.1005019.

Arenas, M. 2015. "Trends in substitution models of molecular evolution". *Frontiers in Genetics* 6 (). doi:10.3389/fgene.2015.00319.

Ayouba, A., L. Duval, F. Liégeois, S. Ngin, S. Ahuka-Mundeke, W. M. Switzer, E. Delaporte, F. Ariey, M. Peeters, and E. Nerrienet. 2013. "Nonhuman primate retroviruses from Cambodia: High simian foamy virus prevalence, identification of divergent STLV-1 strains and no evidence of SIV infection". *Infection, Genetics and Evolution* 18 (): 325–334. doi:10.1016/j.meegid.2013.04.015.

Bailes, E., F. Gao, F. Bibollet-Ruche, V. Courgnaud, M. Peeters, P. A. Marx, B. H. Hahn, and P. M. Sharp. 2003. "Hybrid origin of SIV in chimpanzees". *Science (New York, N.Y.)* 300, no. 5626 (): 1713. doi:10.1126/science.1080657.

Bansal, M. S., M. Kellis, M. Kordi, and S. Kundu. 2018. "RANGER-DTL 2.0: rigorous reconstruction of gene-family evolution by duplication, transfer and loss". Ed. by Kelso, J. *Bioinformatics* 34, no. 18 (): 3214–3216. doi:10.1093/bioinformatics/bty314.

Barin, F., S. M'Boup, F. Denis, P. Kanki, J. S. Allan, T. H. Lee, and M. Essex. 1985. "Serological evidence for virus related to simian T-lymphotropic retrovirus III in residents of west Africa". *Lancet (London, England)* 2, no. 8469 (): 1387–1389. doi:10.1016/s0140-6736(85)92556-5.

Barre-Sinoussi, F., J. Chermann, F. Rey, M. Nugeyre, S. Chamaret, J. Gruest, C. Dauguet, C. Axler-Blin, F. Vezinet-Brun, C. Rouzioux, W. Rozenbaum, and L. Montagnier. 1983. "Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS)". *Science* 220, no. 4599 (): 868–871. doi:10.1126/science.6189183.

Bell, S. M., and T. Bedford. 2017. "Modern-day SIV viral diversity generated by extensive recombination and cross-species transmission". Ed. by Silvestri, G. *PLOS Pathogens* 13, no. 7 (): e1006466. doi:10.1371/journal.ppat.1006466.

Benson, D. A., M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, J. Ostell, K. D. Pruitt, and E. W. Sayers. 2018. "GenBank". *Nucleic Acids Research* 46 (D1): D41–D47. doi:10.1093/nar/gkx1094.

Biberfeld, P., A. Porwit-Ksiazek, B. Böttiger, L. Morfeldt-Månsson, and G. Biberfeld. 1985. "Immunohistopathology of lymph nodes in HTLV-III infected homosexuals with persistent adenopathy or AIDS". *Cancer Research* 45, no. 9 (): 4665s–4670s.

Bieniasz, P. D. 2006. "Late budding domains and host proteins in enveloped virus release". *Virology* 344, no. 1 (): 55–63. doi:10.1016/j.virol.2005.09.044.

Bina, M. 2008. "The Genome Browser at UCSC for Locating Genes, and Much More!" *Molecular Biotechnology* 38, no. 3 (): 269–275. doi:10.1007/s12033-007-9019-2.

Bininda-Emonds, O. R. P. 2005. "transAlign: using amino acids to facilitate the multiple alignment of protein-coding DNA sequences". *BMC bioinformatics* 6 (): 156. doi:10.1186/1471-2105-6-156.

Bouckaert, R., T. G. Vaughan, J. Barido-Sottani, S. Duchêne, M. Fourment, A. Gavryushkina, J. Heled, G. Jones, D. Kühnert, N. De Maio, M. Matschiner, F. K. Mendes, N. F. Müller, H. A. Ogilvie, L. du Plessis, A. Popinga, A. Rambaut, D. Rasmussen, I. Siveroni, M. A. Suchard, C.-H. Wu, D. Xie, C. Zhang, T. Stadler, and A. J. Drummond. 2019. "BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis". *PLoS computational biology* 15 (4): e1006650. doi:10.1371/journal.pcbi.1006650.

Boutet, E., D. Lieberherr, M. Tognolli, M. Schneider, P. Bansal, A. J. Bridge, S. Poux, L. Bougueleret, and I. Xenarios. 2016. "UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View". *Methods in Molecular Biology (Clifton, N.J.)* 1374:23–54. doi:10.1007/978-1-4939-3167-5_2.

Bowzard, J. B., W. G. Davis, V. Jeisy-Scott, P. Ranjan, S. Gangappa, T. Fujita, and S. Sambhara. 2011. "PAMPer and tRIGer: ligand-induced activation of RIG-I". *Trends in Biochemical Sciences* 36, no. 6 (): 314–319. doi:10.1016/j.tibs.2011.03.003.

Braun, E., D. Hotter, L. Koepke, F. Zech, R. Groß, K. M. Sparrer, J. A. Müller, C. K. Pfaller, E. Heusinger, R. Wombacher, K. Sutter, U. Dittmer, M. Winkler, G. Simmons, M. R. Jakobsen, K.-K. Conzelmann, S. Pöhlmann, J. Münch, O. T. Fackler, F. Kirchhoff, and D. Sauter. 2019. "Guanylate-Binding Proteins 2 and 5 Exert Broad Antiviral Activity by Inhibiting Furin-Mediated Processing of Viral Envelope Proteins". *Cell Reports* 27, no. 7 (): 2092–2104.e10. doi:10.1016/j.celrep.2019.04.063.

Breyne, S. de, R. Soto-Rifo, M. López-Lastra, and T. Ohlmann. 2013. "Translation initiation is driven by different mechanisms on the HIV-1 and HIV-2 genomic RNAs". *Virus Research* 171, no. 2 (): 366–381. doi:10.1016/j.virusres.2012.10.006.

Briggs, J. A. G., M. N. Simon, I. Gross, H.-G. Kräusslich, S. D. Fuller, V. M. Vogt, and M. C. Johnson. 2004. "The stoichiometry of Gag protein in HIV-1". *Nature Structural & Molecular Biology* 11, no. 7 (): 672–675. doi:10.1038/nsmb785.

Bruno, M., M. Mahgoub, and T. S. Macfarlan. 2019. "The Arms Race Between KRAB-Zinc Finger Proteins and Endogenous Retroelements and Its Impact on Mammals". *Annual Review of Genetics* 53:393–416. doi:10.1146/annurev-genet-112618-043717.

Bruns, A. M., and C. M. Horvath. 2015. "LGP2 synergy with MDA5 in RLR-mediated RNA recognition and antiviral signaling". *Cytokine* 74, no. 2 (): 198–206. doi:10.1016/j.cyto.2015.02.010.

Busset, J., C. Cabau, C. Meslin, and G. Pascal. 2011. "PhyleasProg: a user-oriented web server for wide evolutionary analyses". *Nucleic Acids Research* 39 (suppl): W479–W485. doi:10.1093/nar/gkr243.

Cagan, A., C. Theunert, H. Laayouni, G. Santpere, M. Pybus, F. Casals, K. Prüfer, A. Navarro, T. Marques-Bonet, J. Bertranpetit, and A. M. Andrés. 2016. "Natural Selection in the Great Apes". *Molecular Biology and Evolution* 33, no. 12 (): 3268–3283. doi:10.1093/molbev/msw215.

Cagliani, R., S. Riva, M. Biasin, M. Fumagalli, U. Pozzoli, S. Lo Caputo, F. Mazzotta, L. Piacentini, N. Bresolin, M. Clerici, and M. Sironi. 2010. "Genetic diversity at endoplasmic reticulum aminopeptidases is maintained by balancing selection and is associated with natural resistance to HIV-1 infection". *Human Molecular Genetics* 19, no. 23 (): 4705–4714. doi:10.1093/hmg/ddq401.

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden. 2009. "BLAST+: architecture and applications". *BMC Bioinformatics* 10 (1): 421. doi:10.1186/1471-2105-10-421.

Carmona, L. M., and D. G. Schatz. 2017. "New insights into the evolutionary origins of the recombination-activating gene proteins and V(D)J recombination". *The FEBS Journal* 284, no. 11 (): 1590–1605. doi:10.1111/febs.13990.

Carroll, L., and J. Tenniel. 1871. *Through the Looking-Glass, and What Alice Found There*. New York: Macmillan.

Chen, B. 2019. "Molecular Mechanism of HIV-1 Entry". *Trends in Microbiology* 27 (10): 878–891. doi:10.1016/j.tim.2019.06.002.

Colón-Thillet, R., E. Hsieh, L. Graf, R. N. McLaughlin, J. M. Young, G. Kochs, M. Emerman, and H. S. Malik. 2019. "Combinatorial mutagenesis of rapidly evolving residues yields super-restrictor antiviral proteins". Ed. by Regoes, R. R. *PLOS Biology* 17, no. 10 (): e3000181. doi:10.1371/journal.pbio.3000181.

Compton, A. A., T. Bruel, F. Porrot, A. Mallet, M. Sachse, M. Euvrard, C. Liang, N. Casartelli, and O. Schwartz. 2014. "IFITM proteins incorporated into HIV-1 virions impair viral fusion and spread". *Cell Host & Microbe* 16, no. 6 (): 736–747. doi:10.1016/j.chom.2014.11.001.

Compton, A. A., V. M. Hirsch, and M. Emerman. 2012. "The Host Restriction Factor APOBEC3G and Retroviral Vif Protein Coevolve due to Ongoing Genetic Conflict". *Cell Host & Microbe* 11, no. 1 (): 91–98. doi:10.1016/j.chom.2011.11.010.

Comte, N., B. Morel, D. Hasic, L. Guéguen, B. Boussau, V. Daubin, S. Penel, C. Scornavacca, M. Gouy, A. Stamatakis, E. Tannier, and D. P. Parsons. 2019. *Treerecs: an integrated phylogenetic tool, from sequences to reconciliations*. Preprint. Bioinformatics. doi:10.1101/782946.

Conte, M. G., S. Gaillard, N. Lanau, M. Rouard, and C. Périn. 2008. "GreenPhylDB: a database for plant comparative genomics". *Nucleic Acids Research* 36 (Database issue): D991–998. doi:10.1093/nar/gkm934.

Cooper, J. C., C. J. Leonard, B. S. Pedersen, C. M. Carey, A. R. Quinlan, N. C. Elde, and N. Phadnis. 2019. "Endless Conflicts: Detecting Molecular Arms Races in Mammalian Genomes". *bioRxiv*. doi:10.1101/685321.

Crow, K. D. 2006. "What Is the Role of Genome Duplication in the Evolution of Complexity and Diversity?" *Molecular Biology and Evolution* 23, no. 5 (): 887–892. doi:10.1093/molbev/msj083.

D'arc, M., A. Ayouba, A. Esteban, G. H. Learn, V. Boué, F. Liegeois, L. Etienne, N. Tagg, F. H. Leendertz, C. Boesch, N. F. Madinda, M. M. Robbins, M. Gray, A. Cournil, M. Ooms, M. Letko, V. A. Simon, P. M. Sharp, B. H. Hahn, E. Delaporte, E. Mpoudi Ngole, and M. Peeters. 2015. "Origin of the HIV-1 group O epidemic in western lowland gorillas". *Proceedings of the National Academy of Sciences* 112, no. 11 (): E1343–E1352. doi:10.1073/pnas.1502022112.

Darwin, C. 1859. *On the origin of species*.

Daugherty, M. D., and H. S. Malik. 2012. "Rules of Engagement: Molecular Insights from Host-Virus Arms Races". *Annual Review of Genetics* 46, no. 1 (): 677–700. doi:10.1146/annurev-genet-110711-155522.

Daugherty, M. D., and S. E. Zanders. 2019. "Gene conversion generates evolutionary novelty that fuels genetic conflicts". *Current Opinion in Genetics & Development* 58-59 (): 49–54. doi:10.1016/j.gde.2019.07.011.

Davis, Z. H., E. Verschueren, G. M. Jang, K. Kleffman, J. R. Johnson, J. Park, J. Von Dollen, M. C. Maher, T. Johnson, W. Newton, S. Jäger, M. Shales, J. Horner, R. D. Hernandez, N. J. Krogan, and B. A. Glaunsinger. 2015. "Global Mapping of Herpesvirus-Host Protein Complexes Reveals a Transcription Strategy for Late Genes". *Molecular Cell* 57, no. 2 (): 349–360. doi:10.1016/j.molcel.2014.11.026.

Dennis, M. Y., and E. E. Eichler. 2016. "Human adaptation and evolution by segmental duplication". *Current Opinion in Genetics & Development* 41 (): 44–52. doi:10.1016/j.gde.2016.08.001.

Didelot, X., and M. C. J. Maiden. 2010. "Impact of recombination on bacterial evolution". *Trends in Microbiology* 18, no. 7 (): 315–322. doi:10.1016/j.tim.2010.04.002.

Diebold, S. S., M. Montoya, H. Unger, L. Alexopoulou, P. Roy, L. E. Haswell, A. Al-Shamkhani, R. Flavell, P. Borrow, and C. Reis e Sousa. 2003. "Viral infection switches non-plasmacytoid dendritic cells into high interferon producers". *Nature* 424, no. 6946 (): 324–328. doi:10.1038/nature01783.

Diner, B. A., K. K. Lum, J. E. Toettcher, and I. M. Cristea. 2016. "Viral DNA Sensors IFI16 and Cyclic GMP-AMP Synthase Possess Distinct Functions in Regulating Viral Gene Expression, Immune Defenses, and Apoptotic Responses during Herpesvirus Infection". *mBio* 7 (6). doi:10.1128/mBio.01553-16.

Doyle, T., C. Goujon, and M. H. Malim. 2015. "HIV-1 and interferons: who's interfering with whom?" *Nature Reviews Microbiology* 13, no. 7 (): 403–413. doi:10.1038/nrmicro3449.

Dufayard, J.-F., L. Duret, S. Penel, M. Gouy, F. Rechenmann, and G. Perrière. 2005. "Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases". *Bioinformatics (Oxford, England)* 21, no. 11 (): 2596–2603. doi:10.1093/bioinformatics/bti325.

Edgar, R. C. 2004. "MUSCLE: multiple sequence alignment with high accuracy and high throughput". *Nucleic Acids Research* 32, no. 5 (): 1792–1797. doi:10.1093/nar/gkh340.

Egan, A., A. Mahurkar, J. Crabtree, J. H. Badger, J. M. Carlton, and J. C. Silva. 2008. "IDEA: Interactive Display for Evolutionary Analyses". *BMC Bioinformatics* 9 (1): 524. doi:10.1186/1471-2105-9-524.

Egli, A., D. M. Santer, D. O'Shea, D. L. Tyrrell, and M. Houghton. 2014. "The impact of the interferon-lambda family on the innate and adaptive immune response to viral infections". *Emerging Microbes & Infections* 3, no. 7 (): e51. doi:10.1038/emi.2014.51.

Eickbush, M. T., J. M. Young, and S. E. Zanders. 2019. "Killer Meiotic Drive and Dynamic Evolution of the wtf Gene Family". Ed. by Larracuente, A. *Molecular Biology and Evolution* 36, no. 6 (): 1201–1214. doi:10.1093/molbev/msz052.

Elde, N. C., S. J. Child, A. P. Geballe, and H. S. Malik. 2009. "Protein kinase R reveals an evolutionary model for defeating viral mimicry". *Nature* 457, no. 7228 (): 485–489. doi:10.1038/nature07529.

Enard, D., L. Cai, C. Gwennap, and D. A. Petrov. 2016. "Viruses are a dominant driver of protein adaptation in mammals". *eLife* 5:e12469.

Engelman, A. N., and P. K. Singh. 2018. "Cellular and molecular mechanisms of HIV-1 integration targeting". *Cellular and molecular life sciences: CMLS* 75, no. 14 (): 2491–2507. doi:10.1007/s00018-018-2772-5.

Espert, L., G. Degols, C. Gongora, D. Blondel, B. R. Williams, R. H. Silverman, and N. Mechti. 2003. "ISG20, a new interferon-induced RNase specific for single-stranded RNA, defines an alternative antiviral pathway against RNA genomic viruses". *The Journal of Biological Chemistry* 278, no. 18 (): 16151–16158. doi:10.1074/jbc.M209628200.

Etienne, L. 2015. "Virus-Host Evolution and Positive Selection". In *Encyclopedia of AIDS*, ed. by Hope, T. J., Stevenson, M., and Richman, D., 1–13. New York, NY: Springer New York.

Etienne, L., F. Bibollet-Ruche, P. H. Sudmant, L. I. Wu, B. H. Hahn, and M. Emerman. 2015. "The Role of the Antiviral APOBEC3 Gene Family in Protecting Chimpanzees against Lentiviruses from Monkeys". Ed. by Aiken, C. *PLOS Pathogens* 11, no. 9 (): e1005149. doi:10.1371/journal.ppat.1005149.

Etienne, L., B. H. Hahn, P. M. Sharp, F. A. Matsen, and M. Emerman. 2013. "Gene Loss and Adaptation to Hominids Underlie the Ancient Origin of HIV-1". *Cell Host & Microbe* 14, no. 1 (): 85–92. doi:10.1016/j.chom.2013.06.002.

Etienne, L., E. Nerrienet, M. LeBreton, G. T. Bibila, Y. Foupouapouognigni, D. Rousset, A. Nana, C. F. Djoko, U. Tamoufe, A. F. Aghokeng, E. Mpoudi-Ngole, E. Delaporte, M. Peeters, N. D. Wolfe, and A. Ayouba. 2011. "Characterization of a new simian immunodeficiency virus strain in a naturally infected Pan troglodytes troglodytes chimpanzee with AIDS related symptoms". *Retrovirology* 8 (): 4. doi:10.1186/1742-4690-8-4.

Faust, T. B., J. M. Binning, J. D. Gross, and A. D. Frankel. 2017. "Making Sense of Multifunctional Proteins: Human Immunodeficiency Virus Type 1 Accessory and Regulatory Proteins and Connections to Transcription". *Annual Review of Virology* 4, no. 1 (): 241–260. doi:10.1146/annurev-virology-101416-041654.

Felsenstein, J. 1981. "Evolutionary trees from DNA sequences: A maximum likelihood approach". *Journal of Molecular Evolution* 17, no. 6 (): 368–376. doi:10.1007/BF01734359.

Feng, Q., M. A. Langereis, and F. J. M. van Kuppeveld. 2014. "Induction and suppression of innate antiviral responses by picornaviruses". *Cytokine & Growth Factor Reviews* 25, no. 5 (): 577–585. doi:10.1016/j.cytogfr.2014.07.003.

Fletcher, W., and Z. Yang. 2010. "The Effect of Insertions, Deletions, and Alignment Errors on the Branch-Site Test of Positive Selection". *Molecular Biology and Evolution* 27, no. 10 (): 2257–2267. doi:10.1093/molbev/msq115.

Freed, E. O. 2015. "HIV-1 assembly, release and maturation". *Nature Reviews Microbiology* 13, no. 8 (): 484–496. doi:10.1038/nrmicro3490.

Fregoso, O. I., J. Ahn, C. Wang, J. Mehrens, J. Skowronski, and M. Emerman. 2013. "Evolutionary Toggling of Vpx/Vpr Specificity Results in Divergent Recognition of the Restriction Factor SAMHD1". Ed. by Luban, J. *PLoS Pathogens* 9, no. 7 (): e1003496. doi:10.1371/journal.ppat.1003496.

Fuchs, J., M. Hölzer, M. Schilling, C. Patzina, A. Schoen, T. Hoenen, G. Zimmer, M. Marz, F. Weber, M. A. Müller, and G. Kochs. 2017. "Evolution and Antiviral Specificities of Interferon-Induced Mx Proteins of Bats against Ebola, Influenza, and Other RNA Viruses". Ed. by Williams, B. R. G. *Journal of Virology* 91, no. 15 (). doi:10.1128/JVI.00361-17.

Gal-Ben-Ari, S., I. Barrera, M. Ehrlich, and K. Rosenblum. 2018. "PKR: A Kinase to Remember". *Frontiers in Molecular Neuroscience* 11:480. doi:10.3389/fnmol.2018.00480.

Ganser, B. K., S. Li, V. Y. Klishko, J. T. Finch, and W. I. Sundquist. 1999. "Assembly and analysis of conical models for the HIV-1 core". *Science (New York, N.Y.)* 283, no. 5398 (): 80–83. doi:10.1126/science.283.5398.80.

Gartner, S., P. Markovits, D. M. Markovitz, R. F. Betts, and M. Popovic. 1986. "Virus isolation from and identification of HTLV-III/LAV-producing cells in brain tissue from a patient with AIDS". *JAMA* 256, no. 17 (): 2365–2371.

Gifford, R. J. 2012. "Viral evolution in deep time: lentiviruses and mammals". *Trends in Genetics* 28, no. 2 (): 89–100. doi:10.1016/j.tig.2011.11.003.

Gizzi, A. S., T. L. Grove, J. J. Arnold, J. Jose, R. K. Jangra, S. J. Garforth, Q. Du, S. M. Cahill, N. G. Dulyaninova, J. D. Love, K. Chandran, A. R. Bresnick, C. E. Cameron, and S. C. Almo. 2018. "A naturally occurring antiviral ribonucleotide encoded by the human genome". *Nature* 558 (7711): 610–614. doi:10.1038/s41586-018-0238-4.

Gordon, D. E., et al. 2020. "A SARS-CoV-2 protein interaction map reveals targets for drug repurposing". *Nature* (). doi:10.1038/s41586-020-2286-9.

Goubau, D., M. Schlee, S. Deddouche, A. J. Pruijssers, T. Zillinger, M. Goldeck, C. Schuberth, A. G. Van der Veen, T. Fujimura, J. Rehwinkel, J. A. Iskarpatyoti, W. Barchet, J. Ludwig, T. S. Dermody, G. Hartmann, and C. Reis e Sousa. 2014. "Antiviral immunity via RIG-I-mediated recognition of RNA bearing 5'-diphosphates". *Nature* 514 (7522): 372–375. doi:10.1038/nature13590.

Guan, Y., L. Guo, E. Yang, Y. Liao, L. Liu, Y. Che, Y. Zhang, L. Wang, J. Wang, and Q. Li. 2014. "HSV-1 nucleocapsid egress mediated by UL31 in association with UL34 is impeded by cellular transmembrane protein 140". Notes 7, *Virology* 464-465 (): 1–10. doi:10.1016/j.virol.2014.06.034.

Guéguen, L., S. Gaillard, B. Boussau, M. Gouy, M. Groussin, N. C. Rochette, T. Bigot, D. Fournier, F. Pouyet, V. Cahais, A. Bernard, C. Scornavacca, B. Nabholz, A. Haudry, L. Dachary, N. Galtier, K. Belkhir, and J. Y. Dutheil. 2013. "Bio++: Efficient Extensible Libraries and Tools for Computational Molecular Evolution". *Molecular Biology and Evolution* 30, no. 8 (): 1745–1750. doi:10.1093/molbev/mst097.

Guindon, S., J.-F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel. 2010. "New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0". *Systematic Biology* 59, no. 3 (): 307–321. doi:10.1093/sysbio/syq010.

Guindon, S., A. G. Rodrigo, K. A. Dyer, and J. P. Huelsenbeck. 2004. "Modeling the site-specific variation of selection patterns along lineages". *Proceedings of the National Academy of Sciences of the United States of America* 101, no. 35 (): 12957–12962. doi:10.1073/pnas.0402177101.

Hasegawa, M., H. Kishino, and T.-a. Yano. 1985. "Dating of the human-ape splitting by a molecular clock of mitochondrial DNA". *Journal of Molecular Evolution* 22, no. 2 (): 160–174. doi:10.1007/BF02101694.

Hawkins, J. A., M. E. Kaczmarek, M. A. Müller, C. Drosten, W. H. Press, and S. L. Sawyer. 2019. "A metaanalysis of bat phylogenetics and positive selection based on genomes and transcriptomes from 18 species". *Proceedings of the National Academy of Sciences* 116, no. 23 (): 11351–11360. doi:10.1073/pnas.1814995116.

Heger, A., and C. P. Ponting. 2008. "OPTIC: orthologous and paralogous transcripts in clades". *Nucleic Acids Research* 36 (Database issue): D267–270. doi:10.1093/nar/gkm852.

Heil, F., H. Hemmi, H. Hochrein, F. Ampenberger, C. Kirschning, S. Akira, G. Lipford, H. Wagner, and S. Bauer. 2004. "Species-specific recognition of single-stranded RNA via toll-like receptor 7 and 8". *Science (New York, N.Y.)* 303, no. 5663 (): 1526–1529. doi:10.1126/science.1093620.

Hemmi, H., O. Takeuchi, T. Kawai, T. Kaisho, S. Sato, H. Sanjo, M. Matsumoto, K. Hoshino, H. Wagner, K. Takeda, and S. Akira. 2000. "A Toll-like receptor recognizes bacterial DNA". *Nature* 408, no. 6813 (): 740–745. doi:10.1038/35047123.

Hess, N. J., S. Jiang, X. Li, Y. Guan, and R. I. Tapping. 2017. "TLR10 Is a B Cell Intrinsic Suppressor of Adaptive Immune Responses". *Journal of Immunology (Baltimore, Md.: 1950)* 198 (2): 699–707. doi:10.4049/jimmunol.1601335.

Hotter, D., M. Bosso, K. L. Jønsson, C. Krapp, C. M. Stürzel, A. Das, E. Littwitz-Salomon, B. Berkhout, A. Russ, S. Wittmann, T. Gramberg, Y. Zheng, L. J. Martins, V. Planelles, M. R. Jakobsen, B. H. Hahn, U. Dittmer, D. Sauter, and F. Kirchhoff. 2019. "IFI16 Targets the Transcription Factor Sp1 to Suppress HIV-1 Transcription and Latency Reactivation". *Cell Host & Microbe* 25 (6): 858–872.e13. doi:10.1016/j.chom.2019.05.002.

Hsin, J.-P., Y. Lu, G. B. Loeb, C. S. Leslie, and A. Y. Rudensky. 2018. "The effect of cellular context on miR-155-mediated gene regulation in four major immune cell types". *Nature Immunology* 19 (10): 1137–1145. doi:10.1038/s41590-018-0208-x.

Hu, G., and L. Kurgan. 2019. "Sequence Similarity Searching". *Current Protocols in Protein Science* 95, no. 1 (): e71. doi:10.1002/cpps.71.

Hurst, L., and N. Smith. 1998. "The evolution of concerted evolution". *Proceedings of the Royal Society of London. Series B: Biological Sciences* 265, no. 1391 (): 121–127. doi:10.1098/rspb.1998.0272.

Hurst, T. P., A. Aswad, T. Karamitros, A. Katzourakis, A. L. Smith, and G. Magiorkinis. 2019. "Interferon-Inducible Protein 16 (IFI16) Has a Broad-Spectrum Binding Ability Against ssDNA Targets: An Evolutionary Hypothesis for Antiretroviral Checkpoint". *Frontiers in Microbiology* 10 (). doi:10.3389/fmicb.2019.01426.

Isaacs, A., and J. Lindenmann. 1987. "Virus interference. I. The interferon". *Journal of Interferon Research* 7, no. 5 (): 429–438. doi:10.1089/jir.1987.7.429.

Ishikawa, H., Z. Ma, and G. N. Barber. 2009. "STING regulates intracellular DNA-mediated, type I interferon-dependent innate immunity". *Nature* 461, no. 7265 (): 788–792. doi:10.1038/nature08476.

Ivashkiv, L. B., and L. T. Donlin. 2014. "Regulation of type I interferon responses". *Nature Reviews Immunology* 14, no. 1 (): 36–49. doi:10.1038/nri3581.

Jacox, E., C. Chauve, G. J. Szöllősi, Y. Ponty, and C. Scornavacca. 2016. "ecceTERA: comprehensive gene tree-species tree reconciliation using parsimony: Table 1." *Bioinformatics* 32, no. 13 (): 2056–2058. doi:10.1093/bioinformatics/btw105.

Jäger, S., P. Cimermancic, N. Gulbahce, J. R. Johnson, K. E. McGovern, S. C. Clarke, M. Shales, G. Mercenne, L. Pache, K. Li, H. Hernandez, G. M. Jang, S. L. Roth, E. Akiva, J. Marlett, M. Stephens, I. D'Orso, J. Fernandes, M. Fahey, C. Mahon, A. J. O'Donoghue, A. Todorovic, J. H. Morris, D. A. Maltby, T. Alber, G. Cagney, F. D. Bushman, J. A. Young, S. K. Chanda, W. I. Sundquist, T. Kortemme, R. D. Hernandez, C. S. Craik, A. Burlingame, A. Sali, A. D. Frankel, and N. J. Krogan. 2012. "Global landscape of HIV–human protein complexes". *Nature* 481, no. 7381 (): 365–370. doi:10.1038/nature10719.

Jakobsen, M. R., R. O. Bak, A. Andersen, R. K. Berg, S. B. Jensen, J. Tengchuan, T. Jin, A. Laustsen, K. Hansen, L. Ostergaard, K. A. Fitzgerald, T. S. Xiao, J. G. Mikkelsen, T. H. Mogensen, and S. R. Paludan. 2013. "IFI16 senses DNA forms of the lentiviral replication cycle and controls HIV-1 replication". *Proceedings of the National Academy of Sciences of the United States of America* 110, no. 48 (): E4571–4580. doi:10.1073/pnas.1311669110.

Jin, L., P. M. Waterman, K. R. Jonscher, C. M. Short, N. A. Reisdorph, and J. C. Cambier. 2008. "MPYS, a novel membrane tetraspanner, is associated with major histocompatibility complex class II and mediates transduction of apoptotic signals". *Molecular and Cellular Biology* 28, no. 16 (): 5014–5026. doi:10.1128/MCB.00640-08.

Johnson, L. S., S. R. Eddy, and E. Portugaly. 2010. "Hidden Markov model speed heuristic and iterative HMM search procedure". *BMC bioinformatics* 11 (): 431. doi:10.1186/1471-2105-11-431.

Johnson, W. E., and S. L. Sawyer. 2009. "Molecular evolution of the antiretroviral TRIM5 gene". *Immunogenetics* 61, no. 3 (): 163–176. doi:10.1007/s00251-009-0358-y.

Jordan, G., and N. Goldman. 2012. "The Effects of Alignment Error and Alignment Filtering on the Sitewise Detection of Positive Selection". *Molecular Biology and Evolution* 29, no. 4 (): 1125–1139. doi:10.1093/molbev/msr272.

Jukes, T. H., and C. R. Cantor. 1969. "Evolution of protein molecules". *Mammalian protein metabolism* 3 (21): 132.

Katoh, K., K. Misawa, K.-i. Kuma, and T. Miyata. 2002. "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform". *Nucleic Acids Research* 30, no. 14 (): 3059–3066.

Keele, B. F. 2006. "Chimpanzee Reservoirs of Pandemic and Nonpandemic HIV-1". *Science* 313, no. 5786 (): 523–526. doi:10.1126/science.1126531.

Keele, B. F., J. H. Jones, K. A. Terio, J. D. Estes, R. S. Rudicell, M. L. Wilson, Y. Li, G. H. Learn, T. M. Beasley, J. Schumacher-Stankey, E. Wroblewski, A. Mosser, J. Raphael, S. Kamenya, E. V. Lonsdorf, D. A. Travis, T. Mlengeya, M. J. Kinsel, J. G. Else, G. Silvestri, J. Goodall, P. M. Sharp, G. M. Shaw, A. E. Pusey, and B. H. Hahn. 2009. "Increased mortality and AIDS-like immunopathology in wild chimpanzees infected with SIVcpz". *Nature* 460, no. 7254 (): 515–519. doi:10.1038/nature08200.

Kerur, N., M. V. Veettil, N. Sharma-Walia, V. Bottero, S. Sadagopan, P. Otageri, and B. Chandran. 2011. "IFI16 acts as a nuclear pathogen sensor to induce the inflammasome in response to Kaposi Sarcoma-associated herpesvirus infection". *Cell Host & Microbe* 9, no. 5 (): 363–375. doi:10.1016/j.chom.2011.04.008.

Khan, A. S., J. Bodem, F. Buseyne, A. Gessain, W. Johnson, J. H. Kuhn, J. Kuzmak, D. Lindemann, M. L. Linial, M. Löchelt, M. Materniak-Kornas, M. A. Soares, and W. M. Switzer. 2018. "Spumaretroviruses: Updated taxonomy and nomenclature". *Virology* 516 (): 158–164. doi:10.1016/j.virol.2017.12.035.

Klatt, N. R., G. Silvestri, and V. Hirsch. 2012. "Nonpathogenic Simian Immunodeficiency Virus Infections". *Cold Spring Harbor Perspectives in Medicine* 2, no. 1 (): a007153–a007153. doi:10.1101/cshperspect.a007153.

Kosakovsky Pond, S. L., A. F. Y. Poon, S. Zárate, D. M. Smith, S. J. Little, S. K. Pillai, R. J. Ellis, J. K. Wong, A. J. Leigh Brown, D. D. Richman, and S. D. W. Frost. 2008. "Estimating selection pressures on HIV-1 using phylogenetic likelihood models". *Statistics in Medicine* 27, no. 23 (): 4779–4789. doi:10.1002/sim.3192.

Kosakovsky Pond, S. L., and S. D. W. Frost. 2005. "Not So Different After All: A Comparison of Methods for Detecting Amino Acid Sites Under Selection". *Molecular Biology and Evolution* 22, no. 5 (): 1208–1222. doi:10.1093/molbev/msi105.

Kosiol, C., T. Vinař, R. R. da Fonseca, M. J. Hubisz, C. D. Bustamante, R. Nielsen, and A. Siepel. 2008. "Patterns of Positive Selection in Six Mammalian Genomes". Ed. by Schierup, M. H. *PLoS Genetics* 4, no. 8 (): e1000144. doi:10.1371/journal.pgen.1000144.

Krupp, A., K. R. McCarthy, M. Ooms, M. Letko, J. S. Morgan, V. Simon, and W. E. Johnson. 2013. "APOBEC3G polymorphism as a selective barrier to cross-species transmission and emergence of pathogenic SIV and AIDS in a primate host". *PLoS pathogens* 9 (10): e1003641. doi:10.1371/journal.ppat.1003641.

Kulikova, T., R. Akhtar, P. Aldebert, N. Althorpe, M. Andersson, A. Baldwin, K. Bates, S. Bhattacharyya, L. Bower, P. Browne, M. Castro, G. Cochrane, K. Duggan, R. Eberhardt, N. Faruque, G. Hoad, C. Kanz, C. Lee, R. Leinonen, Q. Lin, V. Lombard, R. Lopez, D. Lorenc, H. McWilliam, G. Mukherjee, F. Nardone, M. P. G. Pastor, S. Plaister, S. Sobhany, P. Stoehr, R. Vaughan, D. Wu, W. Zhu, and R. Apweiler. 2007. "EMBL Nucleotide Sequence Database in 2006". *Nucleic Acids Research* 35 (Database issue): D16–20. doi:10.1093/nar/gkl913.

Kumar, S., J. H. Morrison, D. Dingli, and E. Poeschla. 2018. "HIV-1 Activation of Innate Immunity Depends Strongly on the Intracellular Level of TREX1 and Sensing of Incomplete Reverse Transcription Products". *Journal of Virology* 92 (16). doi:10.1128/JVI.00001-18.

Lazear, H. M., J. W. Schoggins, and M. S. Diamond. 2019. "Shared and Distinct Functions of Type I and Type III Interferons". *Immunity* 50 (4): 907–923. doi:10.1016/j.immuni.2019.03.025.

Lee, R. van der, L. Wiel, T. J. van Dam, and M. A. Huynen. 2017. "Genome-scale detection of positive selection in nine primates predicts human-virus evolutionary conflicts". *Nucleic Acids Research* 45, no. 18 (): 10634–10648. doi:10.1093/nar/gkx704.

Lefkowitz, E. J., D. M. Dempsey, R. C. Hendrickson, R. J. Orton, S. G. Siddell, and D. B. Smith. 2018. "Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV)". *Nucleic Acids Research* 46 (D1): D708–D717. doi:10.1093/nar/gkx932.

Lesbats, P., A. N. Engelman, and P. Cherepanov. 2016. "Retroviral DNA Integration". *Chemical Reviews* 116 (20): 12730–12757. doi:10.1021/acs.chemrev.6b00125.

Li, T., B. A. Diner, J. Chen, and I. M. Cristea. 2012. "Acetylation modulates cellular distribution and DNA sensing ability of interferon-inducible protein IFI16". *Proceedings of the National Academy of Sciences of the United States of America* 109, no. 26 (): 10558–10563. doi:10.1073/pnas.1203447109.

Locatelli, S., and M. Peeters. 2012. "Cross-species transmission of simian retroviruses: how and why they could lead to the emergence of new diseases in the human population". *AIDS* 26, no. 6 (): 659–673. doi:10.1097/QAD.0b013e328350fb68.

Loytynoja, A., and N. Goldman. 2008. "Phylogeny-Aware Gap Placement Prevents Errors in Sequence Alignment and Evolutionary Analysis". *Science* 320, no. 5883 (): 1632–1635. doi:10.1126/science.1158395.

Luo, X., X. Wang, Y. Gao, J. Zhu, S. Liu, G. Gao, and P. Gao. 2020. "Molecular Mechanism of RNA Recognition by Zinc-Finger Antiviral Protein". *Cell Reports* 30 (1): 46–52.e4. doi:10.1016/j.celrep.2019.11.116.

Magadum, S., U. Banerjee, P. Murugan, D. Gangapur, and R. Ravikesavan. 2013. "Gene duplication as a major force in evolution". *Journal of Genetics* 92, no. 1 (): 155–161. doi:10.1007/s12041-013-0212-8.

Mahesh, G., and R. Biswas. 2019. "MicroRNA-155: A Master Regulator of Inflammation". *Journal of Interferon & Cytokine Research: The Official Journal of the International Society for Interferon and Cytokine Research* 39 (6): 321–330. doi:10.1089/jir.2018.0155.

Malathi, K., B. Dong, M. Gale, and R. H. Silverman. 2007. "Small self-RNA generated by RNase L amplifies antiviral innate immunity". *Nature* 448, no. 7155 (): 816–819. doi:10.1038/nature06042.

Malathi, K., T. Saito, N. Crochet, D. J. Barton, M. Gale, and R. H. Silverman. 2010. "RNase L releases a small RNA from HCV RNA that refolds into a potent PAMP". *RNA (New York, N.Y.)* 16, no. 11 (): 2108–2119. doi:10.1261/rna.2244210.

Malathi, K., M. A. Siddiqui, S. Dayal, M. Naji, H. J. Ezelle, C. Zeng, A. Zhou, and B. A. Hassel. 2014. "RNase L interacts with Filamin A to regulate actin dynamics and barrier function for viral entry". *mBio* 5, no. 6 (): e02012. doi:10.1128/mBio.02012-14.

Mandell, D. T., J. Kristoff, T. Gaufin, R. Gautam, D. Ma, N. Sandler, G. Haret-Richter, C. Xu, H. Aamer, J. Dufour, A. Trichel, D. C. Douek, B. F. Keele, C. Apetrei, and I. Pandrea. 2014. "Pathogenic Features Associated with Increased Virulence upon Simian Immunodeficiency Virus Cross-Species Transmission from Natural Hosts". *Journal of Virology* 88, no. 12 (): 6778–6792. doi:10.1128/JVI.03785-13.

Margolis, S. R., S. C. Wilson, and R. E. Vance. 2017. "Evolutionary Origins of cGAS-STING Signaling". *Trends in Immunology* 38, no. 10 (): 733–743. doi:10.1016/j.it.2017.03.004.

Markova-Raina, P., and D. Petrov. 2011. "High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 Drosophila genomes". *Genome Research* 21, no. 6 (): 863–874. doi:10.1101/gr.115949.110.

Martin, D. P., P. Lemey, and D. Posada. 2011. "Analysing recombination in nucleotide sequences". *Molecular Ecology Resources* 11, no. 6 (): 943–955. doi:10.1111/j.1755-0998.2011.03026.x.

McCarthy, K. R., A. Kirmaier, P. Autissier, and W. E. Johnson. 2015. "Evolutionary and Functional Analysis of Old World Primate TRIM5 Reveals the Ancient Emergence of Primate Lentiviruses and Convergent Evolution Targeting a Conserved Capsid Interface". *PLoS pathogens* 11, no. 8 (): e1005085. doi:10.1371/journal.ppat.1005085.

McDonald, M. J., D. P. Rice, and M. M. Desai. 2016. "Sex speeds adaptation by altering the dynamics of molecular evolution". *Nature* 531, no. 7593 (): 233–236. doi:10.1038/nature17143.

McLaren, P. J., A. Gawanbacht, N. Pyndiah, C. Krapp, D. Hotter, S. F. Kluge, N. Götz, J. Heilmann, K. Mack, D. Sauter, D. Thompson, J. Perreaud, A. Rausell, M. Munoz, A. Ciuffi, F. Kirchhoff, and A. Telenti. 2015. "Identification of potential HIV restriction factors by combining evolutionary genomic signatures with functional analyses". *Retrovirology* 12, no. 1 (). doi:10.1186/s12977-015-0165-5.

McLaughlin, R. N., and H. S. Malik. 2017. "Genetic conflicts: the usual suspects and beyond". *The Journal of Experimental Biology* 220, no. 1 (): 6–17. doi:10.1242/jeb.148148.

Merkl, P. E., M. H. Orzalli, and D. M. Knipe. 2018. "Mechanisms of Host IFI16, PML, and Daxx Protein Restriction of Herpes Simplex Virus 1 Replication". *Journal of Virology* 92 (10). doi:10.1128/JVI.00057-18.

Mesev, E. V., R. A. LeDesma, and A. Ploss. 2019. "Decoding type I and III interferon signalling during viral infection". *Nature Microbiology* 4 (6): 914–924. doi:10.1038/s41564-019-0421-x.

Meyerson, N. R., P. A. Rowley, C. H. Swan, D. T. Le, G. K. Wilkerson, and S. L. Sawyer. 2014. "Positive selection of primate genes that promote HIV-1 replication". *Virology* 454-455 (): 291–298. doi:10.1016/j.virol.2014.02.029.

Mitchell, P. S., M. Emerman, and H. S. Malik. 2013. "An evolutionary perspective on the broad antiviral specificity of MxA". *Current Opinion in Microbiology* 16, no. 4 (): 493–499. doi:10.1016/j.mib.2013.04.005.

Mitchell, P. S., C. Patzina, M. Emerman, O. Haller, H. S. Malik, and G. Kochs. 2012. "Evolution-Guided Identification of Antiviral Specificity Determinants in the Broadly Acting Interferon-Induced Innate Immunity Factor MxA". *Cell Host & Microbe* 12, no. 4 (): 598–604. doi:10.1016/j.chom.2012.09.005.

Miyakawa, K., S. Matsunaga, M. Yokoyama, M. Nomaguchi, Y. Kimura, M. Nishi, H. Kimura, H. Sato, H. Hirano, T. Tamura, H. Akari, T. Miura, A. Adachi, T. Sawasaki, N. Yamamoto, and A. Ryo. 2019. "PIM kinases facilitate lentiviral evasion from SAMHD1 restriction via Vpx phosphorylation". *Nature Communications* 10 (1): 1844. doi:10.1038/s41467-019-09867-7.

Motwani, M., S. Pesiridis, and K. A. Fitzgerald. 2019. "DNA sensing by the cGAS–STING pathway in health and disease". *Nature Reviews Genetics* 20, no. 11 (): 657–674. doi:10.1038/s41576-019-0151-1.

Mücksch, F., V. Laketa, B. Müller, C. Schultz, and H.-G. Kräusslich. 2017. "Synchronized HIV assembly by tunable PIP2 changes reveals PIP2 requirement for stable Gag anchoring". *eLife* 6. doi:10.7554/eLife.25287.

Mukhtar, M. S., A.-R. Carvunis, M. Dreze, P. Epple, J. Steinbrenner, J. Moore, M. Tasan, M. Galli, T. Hao, M. T. Nishimura, S. J. Pevzner, S. E. Donovan, L. Ghamsari, B. Santhanam, V. Romero, M. M. Poulin, F. Gebreab, B. J. Gutierrez, S. Tam, D. Monachello, M. Boxem, C. J. Harbort, N. McDonald, L. Gai, H. Chen, Y. He, European Union Effectoromics Consortium, J. Vandenhaute, F. P. Roth, D. E. Hill, J. R. Ecker, M. Vidal, J. Beynon, P. Braun, and J. L. Dangl. 2011. "Independently Evolved Virulence Effectors Converge onto Hubs in a Plant Immune System Network". *Science* 333, no. 6042 (): 596–601. doi:10.1126/science.1203659.

Nakano, Y., H. Aso, A. Soper, E. Yamada, M. Moriwaki, G. Juarez-Fernandez, Y. Koyanagi, and K. Sato. 2017. "A conflict of interest: the evolutionary arms race between mammalian APOBEC3 and lentiviral Vif". *Retrovirology* 14, no. 1 (): 31. doi:10.1186/s12977-017-0355-4.

Negishi, H., T. Taniguchi, and H. Yanai. 2018. "The Interferon (IFN) Class of Cytokines and the IFN Regulatory Factor (IRF) Transcription Factor Family". *Cold Spring Harbor Perspectives in Biology* 10 (11). doi:10.1101/cshperspect.a028423.

Nielsen, R., and Z. Yang. 1998. "Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene". *Genetics* 148, no. 3 (): 929–936.

Nielsen, R. 2005. "Molecular Signatures of Natural Selection". *Annual Review of Genetics* 39, no. 1 (): 197–218. doi:10.1146/annurev.genet.39.073003.112420.

Novikova, M., Y. Zhang, E. O. Freed, and K. Peng. 2019. "Multiple Roles of HIV-1 Capsid during the Virus Replication Cycle". *Virologica Sinica* 34, no. 2 (): 119–134. doi:10.1007/s12250-019-00095-3.

Núñez, R., M. Budt, S. Saenger, K. Paki, U. Arnold, A. Sadewasser, and T. Wolff. 2018. "The RNA Helicase DDX6 Associates with RIG-I to Augment Induction of Antiviral Signaling". *International Journal of Molecular Sciences* 19, no. 7 (): 1877. doi:10.3390/ijms19071877.

O'Neill, L. A. J., D. Golenbock, and A. G. Bowie. 2013. "The history of Toll-like receptors - redefining innate immunity". *Nature Reviews. Immunology* 13 (6): 453–460. doi:10.1038/nri3446.

OhAinle, M., J. A. Kerns, M. M. H. Li, H. S. Malik, and M. Emerman. 2008. "Antiretroelement activity of APOBEC3H was lost twice in recent human evolution". *Cell Host & Microbe* 4, no. 3 (): 249–259. doi:10.1016/j.chom.2008.07.005.

Ohlmann, T., C. Mengardi, and M. López-Lastra. 2014. "Translation initiation of the HIV-1 mRNA". *Translation (Austin, Tex.)* 2, no. 2 (): e960242. doi:10.4161/2169074X.2014.960242.

Ohto, U., T. Shibata, H. Tanji, H. Ishida, E. Krayukhina, S. Uchiyama, K. Miyake, and T. Shimizu. 2015. "Structural basis of CpG and inhibitory DNA recognition by Toll-like receptor 9". *Nature* 520, no. 7549 (): 702–705. doi:10.1038/nature14138.

Ortiz-Barrientos, D., J. Engelstädter, and L. H. Rieseberg. 2016. "Recombination Rate Evolution and the Origin of Species". *Trends in Ecology & Evolution* 31, no. 3 (): 226–236. doi:10.1016/j.tree.2015.12.016.

Orzalli, M. H., N. A. DeLuca, and D. M. Knipe. 2012. "Nuclear IFI16 induction of IRF-3 signaling during herpesviral infection and degradation of IFI16 by the viral ICP0 protein". *Proceedings of the National Academy of Sciences of the United States of America* 109, no. 44 (): E3008–3017. doi:10.1073/pnas.1211302109.

Ostlund, G., T. Schmitt, K. Forslund, T. Köstler, D. N. Messina, S. Roopra, O. Frings, and E. L. L. Sonnhammer. 2010. "InParanoid 7: new algorithms and tools for eukaryotic orthology analysis". *Nucleic Acids Research* 38 (Database issue): D196–203. doi:10.1093/nar/gkp931.

Ott, D. E. 2008. "Cellular proteins detected in HIV-1". *Reviews in Medical Virology* 18, no. 3 (): 159–175. doi:10.1002/rmv.570.

Palmer, W. H., J. D. Hadfield, and D. J. Obbard. 2018. "RNA-Interference Pathways Display High Rates of Adaptive Protein Evolution in Multiple Invertebrates". *Genetics* 208, no. 4 (): 1585–1599. doi:10.1534/genetics.117.300567.

Pandrea, I., and C. Apetrei. 2010. "Where the Wild Things Are: Pathogenesis of SIV Infection in African Nonhuman Primate Hosts". *Current HIV/AIDS Reports* 7, no. 1 (): 28–36. doi:10.1007/s11904-009-0034-8.

Pandrea, I., D. L. Sodora, G. Silvestri, and C. Apetrei. 2008. "Into the wild: simian immunodeficiency virus (SIV) infection in natural hosts". *Trends in Immunology* 29, no. 9 (): 419–428. doi:10.1016/j.it.2008.05.004.

Paparisto, E., M. W. Woods, M. D. Coleman, S. A. Moghadasi, D. S. Kochar, S. K. Tom, H. P. Kohio, R. M. Gibson, T. J. Rohringer, N. R. Hunt, E. J. Di Gravio, J. Y. Zhang, M. Tian, Y. Gao, E. J. Arts, and S. D. Barr. 2018. "Evolution-Guided Structural and Functional Analyses of the HERC Family Reveal an Ancient Marine Origin and Determinants of Antiviral Activity". Ed. by Kirchhoff, F. *Journal of Virology* 92, no. 13 (): e00528–18. doi:10.1128/JVI.00528-18.

Paulsen, I., and A. von Haeseler. 2006. "INVHOGEN: a database of homologous invertebrate genes". *Nucleic Acids Research* 34 (Database issue): D349–353. doi:10.1093/nar/gkj100.

Pearson, W. R., and D. J. Lipman. 1988. "Improved tools for biological sequence comparison". *Proceedings of the National Academy of Sciences of the United States of America* 85, no. 8 (): 2444–2448. doi:10.1073/pnas.85.8.2444.

Pearson, W. R. 2013. "An Introduction to Sequence Similarity ("Homology") Searching". *Current Protocols in Bioinformatics* 42, no. 1 (): 3.1.1–3.1.8. doi:10.1002/0471250953.bi0301s42.

Pecon-Slattery, J. 2014. "Recent Advances in Primate Phylogenomics". *Annual Review of Animal Biosciences* 2, no. 1 (): 41–63. doi:10.1146/annurev-animal-022513-114217.

Perelman, P., W. E. Johnson, C. Roos, H. N. Seuánez, J. E. Horvath, M. A. M. Moreira, B. Kessing, J. Pontius, M. Roelke, Y. Rumpler, M. P. C. Schneider, A. Silva, S. J. O'Brien, and J. Pecon-Slattery. 2011. "A Molecular Phylogeny of Living Primates". Ed. by Brosius, J. *PLoS Genetics* 7, no. 3 (): e1001342. doi:10.1371/journal.pgen.1001342.

Pérez-Losada, M., M. Arenas, J. C. Galán, F. Palero, and F. González-Candelas. 2015. "Recombination in viruses: mechanisms, methods of study, and evolutionary consequences". *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases* 30 (): 296–307. doi:10.1016/j.meegid.2014.12.022.

Picard, L., Q. Ganivet, O. Allatif, A. Cimarelli, L. Guéguen, and L. Etienne. 2020. *DGINN, an automated and highly-flexible pipeline for the Detection of Genetic INNovations on protein-coding genes*. Preprint. Evolutionary Biology. doi:10.1101/2020.02.25.964155.

Pichlmair, A., O. Schulz, C. P. Tan, T. I. Näslund, P. Liljeström, F. Weber, and C. Reis e Sousa. 2006. "RIG-I-mediated antiviral responses to single-stranded RNA bearing 5'-phosphates". *Science (New York, N.Y.)* 314, no. 5801 (): 997–1001. doi:10.1126/science.1132998.

Pichlmair, A., O. Schulz, C.-P. Tan, J. Rehwinkel, H. Kato, O. Takeuchi, S. Akira, M. Way, G. Schiavo, and C. Reis e Sousa. 2009. "Activation of MDA5 requires higher-order RNA structures generated during virus infection". *Journal of Virology* 83, no. 20 (): 10761–10769. doi:10.1128/JVI.00770-09.

Pond, S. L. K., S. D. W. Frost, and S. V. Muse. 2005. "HyPhy: hypothesis testing using phylogenies". *Bioinformatics* 21, no. 5 (): 676–679. doi:10.1093/bioinformatics/bti079.

Pornillos, O., and B. K. Ganser-Pornillos. 2019. "Maturation of retroviruses". *Current Opinion in Virology* 36 (): 47–55. doi:10.1016/j.coviro.2019.05.004.

Posada, D., and K. A. Crandall. 2002. "The Effect of Recombination on the Accuracy of Phylogeny Estimation". *Journal of Molecular Evolution* 54, no. 3 (): 396–402. doi:10.1007/s00239-001-0034-9.

Privman, E., O. Penn, and T. Pupko. 2012. "Improving the Performance of Positive Selection Inference by Filtering Unreliable Alignment Regions". *Molecular Biology and Evolution* 29, no. 1 (): 1–5. doi:10.1093/molbev/msr177.

Psarras, A., P. Emery, and E. M. Vital. 2017. "Type I interferon-mediated autoimmune diseases: pathogenesis, diagnosis and targeted therapy". *Rheumatology (Oxford, England)* 56 (10): 1662–1675. doi:10.1093/rheumatology/kew431.

Quach, H., D. Wilson, G. Laval, E. Patin, J. Manry, J. Guibert, L. B. Barreiro, E. Nerrienet, E. Verschoor, A. Gessain, M. Przeworski, and L. Quintana-Murci. 2013. "Different selective pressures shape the evolution of Toll-like receptors in human and African great ape populations". *Human Molecular Genetics* 22, no. 23 (): 4829–4840. doi:10.1093/hmg/ddt335.

Quintana-Murci, L. 2019. "Human Immunology through the Lens of Evolutionary Genetics". *Cell* 177, no. 1 (): 184–199. doi:10.1016/j.cell.2019.02.033.

Quintana-Murci, L., and A. G. Clark. 2013. "Population genetic tools for dissecting innate immunity in humans". *Nature Reviews. Immunology* 13, no. 4 (): 280–293. doi:10.1038/nri3421.

Radwan, J., W. Babik, J. Kaufman, T. L. Lenz, and J. Winternitz. 2020. "Advances in the Evolutionary Understanding of MHC Polymorphism". *Trends in genetics: TIG* 36, no. 4 (): 298–311. doi:10.1016/j.tig.2020.01.008.

Ranwez, V., S. Harispe, F. Delsuc, and E. J. P. Douzery. 2011. "MACSE: Multiple Alignment of Coding SEquences Accounting for Frameshifts and Stop Codons". Ed. by Murphy, W. J. *PLoS ONE* 6, no. 9 (): e22594. doi:10.1371/journal.pone.0022594.

Refsland, E. W., J. F. Hultquist, E. M. Luengas, T. Ikeda, N. M. Shaban, E. K. Law, W. L. Brown, C. Reilly, M. Emerman, and R. S. Harris. 2014. "Natural polymorphisms in human APOBEC3H and HIV-1 Vif combine in primary T lymphocytes to affect viral G-to-A mutation levels and infectivity". *PLoS genetics* 10, no. 11 (): e1004761. doi:10.1371/journal.pgen.1004761.

Ritz, K. R., M. A. Noor, and N. D. Singh. 2017. "Variation in Recombination Rate: Adaptive or Not?" *Trends in Genetics* 33, no. 5 (): 364–374. doi:10.1016/j.tig.2017.03.003.

Rokas, n., and n. Holland. 2000. "Rare genomic changes as a tool for phylogenetics". *Trends in Ecology & Evolution* 15, no. 11 (): 454–459. doi:10.1016/s0169-5347(00)01967-4.

Ronquist, F., M. Teslenko, P. van der Mark, D. L. Ayres, A. Darling, S. Höhna, B. Larget, L. Liu, M. A. Suchard, and J. P. Huelsenbeck. 2012. "MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space". *Systematic Biology* 61, no. 3 (): 539–542. doi:10.1093/sysbio/sys029.

Rothenburg, S., E. J. Seo, J. S. Gibbs, T. E. Dever, and K. Dittmar. 2009. "Rapid evolution of protein kinase PKR alters sensitivity to viral inhibitors". *Nature Structural & Molecular Biology* 16, no. 1 (): 63–70. doi:10.1038/nsmb.1529.

Rusinova, I., S. Forster, S. Yu, A. Kannan, M. Masse, H. Cumming, R. Chapman, and P. J. Hertzog. 2013. "INTERFEROME v2.0: an updated database of annotated interferon-regulated genes". *Nucleic Acids Research* 41 (D1): D1040–D1046. doi:10.1093/nar/gks1215.

Sahm, A., M. Bens, M. Platzer, and K. Szafranski. 2017. "PosiGene: automated and easy-to-use pipeline for genome-wide detection of positively selected genes". *Nucleic Acids Research* 45, no. 11 (): e100–e100. doi:10.1093/nar/gkx179.

Sauter, D., and F. Kirchhoff. 2019. "Key Viral Adaptations Preceding the AIDS Pandemic". *Cell Host & Microbe* 25, no. 1 (): 27–38. doi:10.1016/j.chom.2018.12.002.

— . 2018. "Multilayered and versatile inhibition of cellular antiviral factors by HIV and SIV accessory proteins". *Cytokine & Growth Factor Reviews* 40 (): 3–12. doi:10.1016/j.cytogfr.2018.02.005.

Sawyer, S. L., and N. C. Elde. 2012. "A cross-species view on viruses". *Current Opinion in Virology* 2, no. 5 (): 561–568. doi:10.1016/j.coviro.2012.07.003.

Sawyer, S. L., L. I. Wu, M. Emerman, and H. S. Malik. 2005. "Positive selection of primate TRIM5alpha identifies a critical species-specific retroviral restriction domain". *Proceedings of the National Academy of Sciences of the United States of America* 102, no. 8 (): 2832–2837. doi:10.1073/pnas.0409853102.

Schmidt, J. M., M. de Manuel, T. Marques-Bonet, S. Castellano, and A. M. Andrés. 2019. "The impact of genetic adaptation on chimpanzee subspecies differentiation". Ed. by Gojobori, T. *PLOS Genetics* 15, no. 11 (): e1008485. doi:10.1371/journal.pgen.1008485.

Schneider, A., A. Souvorov, N. Sabath, G. Landan, G. H. Gonnet, and D. Graur. 2009. "Estimates of Positive Darwinian Selection Are Inflated by Errors in Sequencing, Annotation, and Alignment". *Genome Biology and Evolution* 1 (): 114–118. doi:10.1093/gbe/evp012.

Schneider, W. M., M. D. Chevillotte, and C. M. Rice. 2014. "Interferon-Stimulated Genes: A Complex Web of Host Defenses". *Annual Review of Immunology* 32, no. 1 (): 513–545. doi:10.1146/annurev-immunol-032713-120231.

Scornavacca, C., K. Belkhir, J. Lopez, R. Dernat, F. Delsuc, E. J. P. Douzery, and V. Ranwez. 2019. "OrthoMaM v10: scaling-up orthologous coding sequence and exon alignments with more than one hundred mammalian genomes": 7.

Seong, S.-Y., and P. Matzinger. 2004. "Hydrophobicity: an ancient damage-associated molecular pattern that initiates innate immune responses". *Nature Reviews Immunology* 4, no. 6 (): 469–478. doi:10.1038/nri1372.

Sharp, P. M., and B. H. Hahn. 2011. "Origins of HIV and the AIDS Pandemic". *Cold Spring Harbor Perspectives in Medicine* 1, no. 1 (): a006841–a006841. doi:10.1101/cshperspect.a006841.

Shoemaker, J. S., and W. M. Fitch. 1989. "Evidence from nuclear sequences that invariable sites should be considered when sequence divergence is calculated". *Molecular Biology and Evolution* 6, no. 3 (): 270–289. doi:10.1093/oxfordjournals.molbev.a040550.

Shultz, A. J., and T. B. Sackton. 2019. "Immune genes are hotspots of shared positive selection across birds and mammals". *eLife* 8 (): e41815. doi:10.7554/eLife.41815.

Siddell, S. G., P. J. Walker, E. J. Lefkowitz, A. R. Mushegian, M. J. Adams, B. E. Dutilh, A. E. Gorbalenya, B. Harrach, R. L. Harrison, S. Junglen, N. J. Knowles, A. M. Kropinski, M. Krupovic, J. H. Kuhn, M. Nibert, L. Rubino, S. Sabanadzovic, H. Sanfaçon, P. Simmonds, A. Varsani, F. M. Zerbini, and A. J. Davison. 2019. "Additional changes to taxonomy ratified in a special vote by the International Committee on Taxonomy of Viruses (October 2018)". *Archives of Virology* 164, no. 3 (): 943–946. doi:10.1007/s00705-018-04136-2.

Siegal, F. P., N. Kadowaki, M. Shodell, P. A. Fitzgerald-Bocarsly, K. Shah, S. Ho, S. Antonenko, and Y. J. Liu. 1999. "The nature of the principal type 1 interferon-producing cells in human blood". *Science (New York, N.Y.)* 284, no. 5421 (): 1835–1837. doi:10.1126/science.284.5421.1835.

Sironi, M., R. Cagliani, D. Forni, and M. Clerici. 2015. "Evolutionary insights into host–pathogen interactions from mammalian sequence data". *Nature Reviews Genetics* 16, no. 4 (): 224–236. doi:10.1038/nrg3905.

Smith, S. D., M. W. Pennell, C. W. Dunn, and S. V. Edwards. 2020. "Phylogenetics is the New Genetics (for Most of Biodiversity)". *Trends in Ecology & Evolution* 35, no. 5 (): 415–425. doi:10.1016/j.tree.2020.01.005.

Stamatakis, A. 2014. "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies". *Bioinformatics (Oxford, England)* 30, no. 9 (): 1312–1313. doi:10.1093/bioinformatics/btu033.

Steinegger, M., and J. Söding. 2017. "MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets". *Nature Biotechnology* 35 (11): 1026–1028. doi:10.1038/nbt.3988.

Steinway, S. N., R. Dannenfelser, C. D. Laucius, J. E. Hayes, and S. Nayak. 2010. "JCoDA: a tool for detecting evolutionary selection". *BMC Bioinformatics* 11 (1): 284. doi:10.1186/1471-2105-11-284.

Stern, A., A. Doron-Faigenboim, E. Erez, E. Martz, E. Bacharach, and T. Pupko. 2007. "Selecton 2007: advanced models for detecting positive and purifying selection using a Bayesian inference approach". *Nucleic Acids Research* 35 (Web Server): W506–W511. doi:10.1093/nar/gkm382.

Stetson, D. B., J. S. Ko, T. Heidmann, and R. Medzhitov. 2008. "Trex1 Prevents Cell-Intrinsic Initiation of Autoimmunity". *Cell* 134, no. 4 (): 587–598. doi:10.1016/j.cell.2008.06.032.

Stolzer, M., H. Lai, M. Xu, D. Sathaye, B. Vernot, and D. Durand. 2012. "Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees". *Bioinformatics (Oxford, England)* 28, no. 18 (): i409–i415. doi:10.1093/bioinformatics/bts386.

Strebel, K. 2013. "HIV accessory proteins versus host restriction factors". *Current Opinion in Virology* 3, no. 6 (): 692–699. doi:10.1016/j.coviro.2013.08.004.

Su, F., H.-Y. Ou, F. Tao, H. Tang, and P. Xu. 2013. "PSP: rapid identification of orthologous coding genes under positive selection across multiple closely related prokaryotic genomes". *BMC Genomics* 14 (1): 924. doi:10.1186/1471-2164-14-924.

Sun, L., J. Wu, F. Du, X. Chen, and Z. J. Chen. 2013. "Cyclic GMP-AMP synthase is a cytosolic DNA sensor that activates the type I interferon pathway". *Science (New York, N.Y.)* 339, no. 6121 (): 786–791. doi:10.1126/science.1232458.

Sun, W., Y. Li, L. Chen, H. Chen, F. You, X. Zhou, Y. Zhou, Z. Zhai, D. Chen, and Z. Jiang. 2009. "ERIS, an endoplasmic reticulum IFN stimulator, activates innate immune signaling through dimerization". *Proceedings of the National Academy of Sciences of the United States of America* 106, no. 21 (): 8653–8658. doi:10.1073/pnas.0900850106.

Suyama, M., D. Torrents, and P. Bork. 2006. "PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments". *Nucleic Acids Research* 34 (Web Server issue): W609–612. doi:10.1093/nar/gkl315.

Svardal, H., A. J. Jasinska, C. Apetrei, G. Coppola, Y. Huang, C. A. Schmitt, B. Jacquelin, V. Ramensky, M. Müller-Trutwin, M. Antonio, G. Weinstock, J. P. Grobler, K. Dewar, R. K. Wilson, T. R. Turner, W. C. Warren, N. B. Freimer, and M. Nordborg. 2017. "Ancient hybridization and strong adaptation to viruses across African vervet monkey populations". *Nature Genetics* 49, no. 12 (): 1705–1713. doi:10.1038/ng.3980.

Szöllősi, G. J., E. Tannier, V. Daubin, and B. Boussau. 2015. "The inference of gene trees with species trees". *Systematic Biology* 64, no. 1 (): e42–62. doi:10.1093/sysbio/syu048.

Takaoka, A., Z. Wang, M. K. Choi, H. Yanai, H. Negishi, T. Ban, Y. Lu, M. Miyagishi, T. Kodama, K. Honda, Y. Ohba, and T. Taniguchi. 2007. "DAI (DLM-1/ZBP1) is a cytosolic DNA sensor and an activator of innate immune response". *Nature* 448, no. 7152 (): 501–505. doi:10.1038/nature06013.

Takehisa, J., M. H. Kraus, A. Ayouba, E. Bailes, F. Van Heuverswyn, J. M. Decker, Y. Li, R. S. Rudicell, G. H. Learn, C. Neel, E. M. Ngole, G. M. Shaw, M. Peeters, P. M. Sharp, and B. H. Hahn. 2009. "Origin and Biology of Simian Immunodeficiency Virus in Wild-Living Western Gorillas". *Journal of Virology* 83, no. 4 (): 1635–1648. doi:10.1128/JVI.02311-08.

Talavera, G., and J. Castresana. 2007. "Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments". *Systematic Biology* 56, no. 4 (): 564–577. doi:10.1080/10635150701472164.

Tan, G., M. Muffato, C. Ledergerber, J. Herrero, N. Goldman, M. Gil, and C. Dessimoz. 2015. "Current Methods for Automated Filtering of Multiple Sequence Alignments

Frequently Worsen Single-Gene Phylogenetic Inference". *Systematic Biology* 64, no. 5 (): 778–791. doi:10.1093/sysbio/syv033.

Tartour, K., R. Appourchaux, J. Gaillard, X.-N. Nguyen, S. Durand, J. Turpin, E. Beaumont, E. Roch, G. Berger, R. Mahieux, D. Brand, P. Roingeard, and A. Cimarelli. 2014. "IFITM proteins are incorporated onto HIV-1 virion particles and negatively imprint their infectivity". *Retrovirology* 11, no. 1 (): 103. doi:10.1186/s12977-014-0103-y.

Tatusov, R. L., N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, S. Smirnov, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin, and D. A. Natale. 2003. "The COG database: an updated version includes eukaryotes". *BMC bioinformatics* 4 (): 41. doi:10.1186/1471-2105-4-41.

Tavaré, S. 1986. "Some probabilistic and statistical problems in the analysis of DNA sequences". *Lectures on mathematics in the life sciences* 17 (2): 57–86.

Thompson, J. D., B. Linard, O. Lecompte, and O. Poch. 2011. "A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives". *PloS One* 6, no. 3 (): e18093. doi:10.1371/journal.pone.0018093.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice". *Nucleic Acids Research* 22 (22): 4673–4680.

Unterholzner, L., S. E. Keating, M. Baran, K. A. Horan, S. B. Jensen, S. Sharma, C. M. Sirois, T. Jin, E. Latz, T. S. Xiao, K. A. Fitzgerald, S. R. Paludan, and A. G. Bowie. 2010. "IFI16 is an innate immune sensor for intracellular DNA". *Nature Immunology* 11, no. 11 (): 997–1004. doi:10.1038/ni.1932.

Van Valen, L. 1973. "A new evolutionary law". *Evol. Theory*, no. 1: 1–30.

Wan, D., W. Jiang, and J. Hao. 2020. "Research Advances in How the cGAS-STING Pathway Controls the Cellular Inflammatory Response". *Frontiers in Immunology* 11 (): 615. doi:10.3389/fimmu.2020.00615.

Wang, X., Y. Li, L.-F. Li, L. Shen, L. Zhang, J. Yu, Y. Luo, Y. Sun, S. Li, and H.-J. Qiu. 2016. "RNA interference screening of interferon-stimulated genes with antiviral activities against classical swine fever virus using a reporter virus". *Antiviral Research* 128 (): 49–56. doi:10.1016/j.antiviral.2016.02.001.

Wernersson, R., and A. G. Pedersen. 2003. "RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences". *Nucleic Acids Research* 31, no. 13 (): 3537–3539. doi:10.1093/nar/gkg609.

Weßling, R., P. Epple, S. Altmann, Y. He, L. Yang, S. R. Henz, N. McDonald, K. Wiley, K. C. Bader, C. Gläßer, M. S. Mukhtar, S. Haigis, L. Ghamsari, A. E. Stephens, J. R. Ecker, M. Vidal, J. D. Jones, K. F. Mayer, E. Ver Loren van Themaat, D. Weigel, P. Schulze-Lefert, J. L. Dangl, R. Panstruga, and P. Braun. 2014. "Convergent Targeting of a Common Host Protein-Network by Pathogen Effectors from Three Kingdoms of Life". *Cell Host & Microbe* 16, no. 3 (): 364–375. doi:10.1016/j.chom.2014.08.004.

Wong, K. M., M. A. Suchard, and J. P. Huelsenbeck. 2008. "Alignment Uncertainty and Genomic Analysis". *Science* 319, no. 5862 (): 473–476. doi:10.1126/science.1151532.

Worobey, M., M. Gemmel, D. E. Teuwen, T. Haselkorn, K. Kunstman, M. Bunce, J.-J. Muyembe, J.-M. M. Kabongo, R. M. Kalengayi, E. Van Marck, M. T. P. Gilbert, and S. M. Wolinsky. 2008. "Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960". *Nature* 455, no. 7213 (): 661–664. doi:10.1038/nature07390.

Wu, J., and Z. J. Chen. 2014. "Innate Immune Sensing and Signaling of Cytosolic Nucleic Acids". *Annual Review of Immunology* 32, no. 1 (): 461–488. doi:10.1146/annurev-immunol-032713-120156.

Wu, J., L. Sun, X. Chen, F. Du, H. Shi, C. Chen, and Z. J. Chen. 2013. "Cyclic GMP-AMP is an endogenous second messenger in innate immune signaling by cytosolic DNA". *Science (New York, N.Y.)* 339, no. 6121 (): 826–830. doi:10.1126/science.1229963.

Wu, N., X.-N. Nguyen, L. Wang, R. Appourchaux, C. Zhang, B. Panthu, H. Gruffat, C. Journo, S. Alais, J. Qin, N. Zhang, K. Tartour, F. Catez, R. Mahieux, T. Ohlmann, M. Liu, B. Du, and A. Cimarelli. 2019. "The interferon stimulated gene 20 protein (ISG20) is an innate defense antiviral factor that discriminates self versus non-self translation". *PLoS pathogens* 15 (10): e1008093. doi:10.1371/journal.ppat.1008093.

Xu, L., D. Yu, Y. Fan, L. Peng, Y. Wu, and Y.-G. Yao. 2016. "Loss of RIG-I leads to a functional replacement with MDA5 in the Chinese tree shrew". *Proceedings of the National Academy of Sciences* 113, no. 39 (): 10950–10955. doi:10.1073/pnas.1604939113.

Yang, Z. 2007. "PAML 4: Phylogenetic Analysis by Maximum Likelihood". *Molecular Biology and Evolution* 24, no. 8 (): 1586–1591. doi:10.1093/molbev/msm088.

Yang, Z. 1994. "Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods". *Journal of Molecular Evolution* 39, no. 3 (): 306–314. doi:10.1007/BF00160154.

Yin, X., S. Langer, Z. Zhang, K. M. Herbert, S. Yoh, R. König, and S. K. Chanda. 2020. "Sensor Sensibility—HIV-1 and the Innate Immune Response". *Cells* 9, no. 1 (): 254. doi:10.3390/cells9010254.

Yoneyama, M., M. Kikuchi, K. Matsumoto, T. Imaizumi, M. Miyagishi, K. Taira, E. Foy, Y.-M. Loo, M. Gale, S. Akira, S. Yonehara, A. Kato, and T. Fujita. 2005. "Shared and unique functions of the DExD/H-box helicases RIG-I, MDA5, and LGP2 in antiviral innate immunity". *Journal of Immunology (Baltimore, Md.: 1950)* 175, no. 5 (): 2851–2858. doi:10.4049/jimmunol.175.5.2851.

Yoneyama, M., M. Kikuchi, T. Natsukawa, N. Shinobu, T. Imaizumi, M. Miyagishi, K. Taira, S. Akira, and T. Fujita. 2004. "The RNA helicase RIG-I has an essential function in double-stranded RNA-induced innate antiviral responses". *Nature Immunology* 5, no. 7 (): 730–737. doi:10.1038/ni1087.

Zanders, S. E., and R. L. Unckless. 2019. "Fertility Costs of Meiotic Drivers". *Current biology: CB* 29 (11): R512–R520. doi:10.1016/j.cub.2019.03.046.

Zhang, X., T. Zhou, J. Yang, Y. Lin, J. Shi, X. Zhang, D. A. Frabutt, X. Zeng, S. Li, P. J. Venta, and Y.-H. Zheng. 2017. "Identification of SERINC5-001 as the Predominant Spliced Isoform for HIV-1 Restriction". Ed. by Ross, S. R. *Journal of Virology* 91, no. 10 (). doi:10.1128/JVI.00137-17.

Zhang, Z., B. Yuan, M. Bao, N. Lu, T. Kim, and Y.-J. Liu. 2011. "The helicase DDX41 senses intracellular DNA mediated by the adaptor STING in dendritic cells". *Nature Immunology* 12, no. 10 (): 959–965. doi:10.1038/ni.2091.

Zhong, B., Y. Yang, S. Li, Y.-Y. Wang, Y. Li, F. Diao, C. Lei, X. He, L. Zhang, P. Tien, and H.-B. Shu. 2008. "The adaptor protein MITA links virus-sensing receptors to IRF3 transcription factor activation". *Immunity* 29, no. 4 (): 538–550. doi:10.1016/j.immuni.2008.09.003.

Zhou, Y., C. He, L. Wang, and B. Ge. 2017. "Post-translational regulation of antiviral innate signaling". *European Journal of Immunology* 47, no. 9 (): 1414–1426. doi:10.1002/eji.201746959.

Zhu, P., J. Liu, J. Bess, E. Chertova, J. D. Lifson, H. Grisé, G. A. Ofek, K. A. Taylor, and K. H. Roux. 2006. "Distribution and three-dimensional structure of AIDS virus envelope spikes". *Nature* 441, no. 7095 (): 847–852. doi:10.1038/nature04817.

Supplementary materials

Jacquet S, Jegado B, Picard L, Moratorio G, Etienne L. Evolution virale lors de changements d'environnement ou d'hôte : points clés du symposium « Evolution virale ».
Virologie 2018; 22(6) : 273-6 doi:10.1684/vir.2018.0753