Identification of specific phylogenetic properties of HIV-1 M and O integrases

Elenia Toccafondi

Université de Strasbourg

UNIVERSITÉ DE STRASBOURG

École Doctorale des Sciences de la Vie et de la Santé STRASBOURG

ÉCOLE DOCTORALE des Sciences de la Vie et de la Santé

CNRS UPR9002

THÈSE présentée par :

Elenia TOCCAFONDI

soutenue le : **30 Septembre 2022**

pour obtenir le grade de : **Docteur de l'université de Strasbourg**

Discipline/ Spécialité : Aspects moléculaires et cellulaires de la biologie

Identification des propriétés phylogénétiques spécifiques des intégrases M et O du VIH-1

THÈSE dirigée par :

NEGRONI Matteo	Directeur de recherche (CNRS), Université de Strasbourg

RAPPORTEURS EXTERNES :

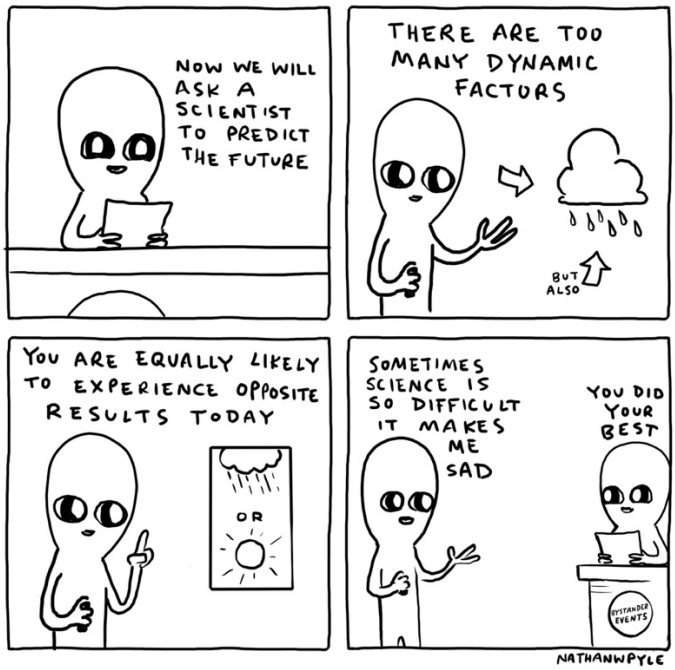
ETIENNE Lucie	Chargée de recherche (CNRS), Université de Lyon
PARISSI Vincent	Directeur de recherche (CNRS), Université de Bordeaux

EXAMINATEUR INTERNE :

DIMITROVA Maria	Professeur, Université de Strasbourg

INVITE – CO-ENCADRANT de la thèse:

LENER Daniela	Maître de conférence, Université de Strasbourg


NOW WE WILL ASK A SCIENTIST TO PREDICT THE FUTURE
THERE ARE TOO MANY DYNAMIC FACTORS
BUT ALSO
YOU ARE EQUALLY LIKELY TO EXPERIENCE OPPOSITE RESULTS TODAY
OR
SOMETIMES SCIENCE IS SO DIFFICULT IT MAKES ME SAD
YOU DID YOUR BEST
DYSTANDLE EVENTS
NATHANWPYLE

ACKNOWLEDGMENTS

First of all, I would like to thank my jury members Dr Lucie Etienne, Dr Maria Dimitrova, and Dr Vincent Parissi for having accepted to be part of my thesis committee and for taking the time of reading and evaluating this manuscript.

Importantly, I would like to thank Dr. Matteo Negroni, my thesis director, for being the great supervisor he had been throughout my PhD. Thank you for giving me the opportunity to do a PhD in your lab and for being present literally every day of my thesis to help me and guide me. I really enjoyed our scientific discussions that were always inspiring as well as full of funny moments. Each time I was lost or demotivated by a bad result, you always gave me the strength and motivation to keep going. Without your help this thesis would not have been possible. Thank you also for being more than just a thesis director, by being someone I could always count on and talk to. I really appreciate everything you did for me during these years.

My sincere gratitude goes to Dr. Daniela Lener, the "godmother" of my project. Your contribution to my thesis on a theoretical and practical level was essential. Despite your busy schedule, you always found the time for being there for me and you never hesitated to help me when I needed it. Having the opportunity to work with you and to learn from you vast knowledge was a fulfilling experience on a professional and personal level. Thank you for your constant support and kindness.

A special thanks goes to the person who thought me for the first time all the cell culture techniques I would have ended up doing everyday during my thesis: Flore. Thank you so much for this and also for your help with my project. Thanks for always bringing a good mood in the lab and for all the nice and funny talks we shared.

Finally, the last member of the 337 room, Alexis. I cannot really imagine my time here without you. Thank you for all the laughs and for all the times you were there when I needed it. A look was enough for understanding if one of us was having a rough day and just a word (usually "burger?") was already enough to feel better. You always told me that I took a risk by picking you as a flat mate, but I knew it would be a perfect match, and it was indeed. You and Laura are the best flat mates I could ever asked for. Thank you for both all the funny and nice memories we created, for all the food and drinks we shared (lots of food and drinks), and also for all the misadventures we went through in these four years. Even spending Christmas day in Strasbourg with Covid was not bad after all with you.

I was lucky enough to meet many special people during these years in Strasbourg and I would like to spend a few words for some of them. Matteoandrea and Edoardo, the first two people that were able to make me feel at home here in Strasbourg. Alessia, it was "friendship at first sight" and we never stop laughing since. Lola, my girl, you know that now we share a "connection" and that this is forever. Giulia, you did not stay long and yet we created so many memories, also, you inspired me in (finally!) learning French and to fall in love with Strasbourg as you did. Charlotte, at the beginning you hated me and three years later you took me on a surprise trip in your hometown (best gift ever), what a plot twist. Javier, we shared so many things during these years, starting from the same building (for living and for working!), then we shared the lockdown time during which we really bonded, then the same birthday that led us to start this beautiful tradition that I hope it will continue. Micol, although we met with the same frequency at which popes are elected, I am grateful for the moments we shared together and I feel like we know each other for a very long time. Mattia, thank you for the short but nice time we shared together and for all the nerd talks. Martina, you made my time at IBMC so much lighter and funnier, and I am grateful for that.

Menzione speciale per i miei amici di una vita (anche se il 2009 era l'altro ieri) Giulia e Tommaso. Siete stati fondamentali durante questi quattro anni e non so quanti minuti di audio vi ho mandato con ogni tipo di mood, eppure li avete ascoltati tutti e siete sempre stati lì pronti ad offrirmi il vostro supporto. Avete persino affrontato il viaggio della speranza che serve per venire a Strasburgo due volte per venirmi a trovare. Grazie di tutto.

Grazie ai miei genitori, Francesca e Claudio, per avermi sempre dato la libertà di fare quello che volevo e per aver creduto in me in ogni step della mia vita, anche quando nemmeno io ci credevo. Grazie alla mia famiglia allargata, Alena, Katiuscia e Roberto, per avermi sempre spronato a lottare e a diventare una persona migliore e per essere i miei più grandi fan.

Infine, grazie Mattia. Grazie per avermi sempre capita (col senno di poi in maniera molto lungimirante) e avermi dato tutto il supporto di cui avevo bisogno e anche di più. Grazie per aver reso questi quattro anni migliori.

RÉSUMÉ DE LA THÈSE

INTRODUCTION

La transmission de virus de l'animal à l'homme est une menace majeure pour la santé humaine, la pandémie de VIH-1 en étant un exemple clair. Chacun des quatre groupes du VIH-1 est issu d'une transmission zoonotique indépendante du virus simiens à l'homme. Les groupes M et N dérivent de SIVcpzPtt (Gao et al., 1999; Keele et al., 2006), tandis que les groupes O et P dérivent de SIVgor (Plantier et al., 2009; D'Arc et al., 2015). Bien que le groupe M et le groupe O partagent des origines géographiques et temporelles similaires (Korber et al., 2000; Lemey et al., 2004; Leoz et al., 2015), ils ont rencontré des succès épidémiologique largement différents. Alors que le groupe M est responsable de la pandémie de SIDA, infectant environ 39 millions de personnes dans le monde, le groupe O a un succès épidémiologique très inférieur, infectant environ 100 000 personnes, principalement dans la région centre-ouest de l'Afrique (Peeters et al., 1997; Mourez et al., 2013). Les raisons de cet écart ne sont que partiellement connues à ce jour, bien que constituant une problématique centrale pour identifier les propriétés critiques permettant la transmission et la diffusion inter-espèces.

Leurs origines zoonotiques différentes et la diversification subséquente des séquences chez l'hôte humain sont responsables de la grande diversité génétique intergroupe entre les groupes M et O qui peut atteindre près de 50 % dans le gène *env* (Santoro and Perno, 2013). Malgré cela, ils ont des phénotypes globalement convergents et, à ce jour, seules quelques différences fonctionnelles ont été mises en évidence entre leurs protéines et leurs enzymes. Parmi ceux-ci, la plus marquée concerne la neutralisation des propriétés antivirales de la protéine cellulaire tetherin, qui est exercée par Vpu dans le VIH-1 M alors qu'elle est partiellement réalisée par Nef dans le cas du VIH-1 O (Kluge et al., 2014; Bego et al., 2016).

La réplication du VIH nécessite l'intégration de l'ARN génomique, rétrotranscrit pour former un ADN double brin, dans le génome de la cellule infectée. Cette étape clé est réalisée par l'intégrase (IN), l'une des trois enzymes virales. IN est une polynucléotidyl-transférase catalisant deux réactions séquentielles de transestérification SN2 dépendantes du magnésium, le traitement en 3' et le transfert de brin (Engelman et al., 1991), conduisant à l'intégration. Cette étape est irréversible et établis l'infection permanente de la cellule cible. IN est constitué de trois domaines reliés par des lieurs flexibles : le domaine N-terminal (NTD), le domaine central catalytique (CCD) et le domaine C-terminal (CTD) (Engelman and Craigie,

1992; Engelman et al., 1993; van Gent et al., 1993). Chacun de ces domaines est spécialisé dans une ou plusieurs fonctions. Le NTD est important pour la multimérisation et la stabilisation de la forme active de l'intégrase (Zheng et al., 1996; Eijkelenboom et al., 1997), qui est un multimère hautement organisé formé de plusieurs dimères de dimères (Passos et al., 2017, 2020). Le CCD est impliqué dans la liaison à l'ADN et contient la triade d'acides aminés responsable de l'activité catalytique de l'enzyme (Kulkosky et al., 1992), mais c'est aussi le domaine impliqué dans la dimérisation des protéines et il est responsable de l'interaction avec le LEDGF/p75, un facteur de l'hôte requis pour le succès de l'infection par le VIH-1 (Busschots et al., 2005). Enfin, le CTD est impliqué dans la liaison de l'ARN/ADN viral à différentes étapes du cycle infectieux (Engelman et al., 1994; Kessl et al., 2016; Elliott et al., 2020; Engelman and Kvaratskhelia, 2022), et dans l'interaction avec la reverse transcriptase virale (Zhu et al., 2004; Wilkinson et al., 2009).

Les intégrases M et O partagent 84% d'identité de séquence ainsi que la même organisation des domaines et les mêmes fonctions. En exploitant la rupture du réseau de coévolution, en construisant des intégrases chimériques entre le groupe M et le groupe O, le laboratoire a pu identifier un motif fonctionnel spécifique au groupe dans le CTD de IN M ($N_{222}K_{240}N_{254}K_{273}$) (Kanja et al., 2020). Le motif est très conservé dans le groupe M et est composé d'une alternance de deux acides aminés chargés positivement (K) et de deux acides aminés amidiques polaires (N). Les deux se sont révélés essentiels à l'intégration, en effet, lorsque le N ou le K sont mutés, l'intégration est abolie (Kanja et al., 2020). D'autres expériences ont montré comment les caractéristiques importantes des deux acides aminés composant le motif étaient, pour le K, leur charge positive, et pour le N, leurs chaînes latérales amidiques. En effet, en les remplaçant par des acides aminés aux caractéristiques similaires (N remplacé par Q ; K remplacé par R), l'intégration n'a pas été affectée (Kanja et al., 2020). Pour cette raison, le motif a été renommé motif "C-terminal lysins amidic" (CLA), par moi-même et les co-auteurs de mon manuscrit de doctorat. Malgré sa conservation *in vivo*, le motif a montré des niveaux importants de flexibilité génétique en culture de cellules. En effet, il était possible de conserver la fonctionnalité lorsque des acides aminés similaires remplaçaient ceux d'origine (QRQR). Il était même suffisant de ne conserver que deux lysines dans le motif, avec les autres positions occupées par des acides aminés amidiques (N, Q) (Kanja et al., 2020) pour maintenir un niveau de fonctionnalité comparable à celui de la protéine sauvage. Cette caractéristique est probablement due au fait que les quatre acides aminés composant le motif CLA, forment une surface chargée positivement, conservée dans les lentivirus (Kanja et al., 2020). Cette découverte, ainsi que la flexibilité génétique montrée par le motif, ont conduit Kanja et ses collègues à émettre l'hypothèse que la fonction possible du motif

pourrait être d'interagir avec un partenaire chargé négativement mais qui ne présente pas une séquence spécifique, comme le squelette de phosphates d'une molécule d'ARN ou d'ADN,

Sur toutes les combinaisons testées possedant 2 K, seules deux ont montré un défaut dans l'étape d'intégration, le phénotype le plus sévère étant à 25% d'intégration par rapport au wt en présence de la séquence d'acides aminés NQKK. De façon surprenante, la séquence NQKK constitue la séquence consensus trouvée aux positions CLA dans l'intégrase du groupe O, soulevant la question de savoir comment le groupe O a pu sélectionner une séquence avec une si faible efficacité apparente. Répondre à cette question était l'objectif principal de ce travail.

Un deuxième objectif lié au motif CLA, bien que moins développé, a également été poursuivi au cours de ma thèse de doctorat. Comme indiqué ci-dessus, une caractéristique essentielle de ce motif sont les charges positives portées par les lysines. En effet, un motif portant quatre K (KKKK) a les mêmes niveaux d'intégration que le wt. Néanmoins, les acides aminés amidiques présents dans le motif se sont également avérés jouer un rôle essentiel. En effet, abolir la nature polaire de l'acide aminé, en remplaçant à la fois N par L, même longueur de chaîne latérale mais pas de polarité, abolit l'intégration, indiquant que la polarité était une caractéristique cruciale dans ces positions. Cependant, la polarité seule n'était pas suffisante, car lorsque les N ont été remplacés par deux T (polaires et avec une chaîne latérale de taille similaire aux asparagines, mais portant un groupe OH au lieu du groupe amidique), l'intégration a chuté à des niveaux à peine détectables (2 à 5 % par rapport à wt IN), ce qui indique que la nature du groupe porté par la chaîne latérale est également importante. Ces résultats avaient été obtenus par Marine pendant son travail de doctorat, mais seulement une partie d'entre eux a été incluse dans l'article publié dans Journal of Virology, pour préserver la clarté du message de cet article qui aurait été diminué par l'inclusion d'un message focalisé sur un deuxième sujet.

Dans les travaux de Kanja et al, il a été effectué une estimation théorique de la contribution de chaque type de défaut observé dans les différentes étapes du cycle infectieux avec les différents mutants (comme une diminution de l'import nucléaire du produit de transcription inverse, par exemple) à l'efficacité globale de l'intégration. Cela a été fait en supposant, par exemple, qu'une diminution de 30 % observée dans l'import nucléaire (pour s'en tenir à l'exemple donné ci-dessus) pour un mutant par rapport à wt IN, si c'est l'unique défaut, devrait entrainer une diminution de 30 % du nombre de provirus générés par rapport à l'enzyme sauvage. En suivant cette approche, le laboratoire est arrivé à la conclusion que les

défauts observés dans l'import nucléaire et dans le traitement en 3' pourraient expliquer la totalité de la diminution d'intégration observée avec chacun des mutants des K du motif. Lorsque la même analyse a été réalisée pour les mutants LKLK et TKTK, alors les défauts observés en import nucléaire et en traitement 3' n'étaient pas suffisants pour rendre compte de l'amplitude de la diminution d'intégration observée avec ces mutants. Pour les trois mutants des résidus K, les valeurs sont (efficacité observée par rapport à wt IN vs fréquence attendue par rapport à wt IN) : 0,03 vs 0,09, 0,24 vs 0,26 et 0,00 vs 0,05, témoignant d'une similitude remarquable entre résultats observés et attendus. À l'opposé, les mutants où le N a été remplacé par L ou T ont montré des valeurs d'intégration observées qui ne correspondaient pas aux valeurs attendues : "0,01 contre 0,27 et 0,05 contre 0,47, pour les mutants L et T, respectivement. Ces résultats suggèrent que des défauts supplémentaires étaient présents avec ces mutants. Pour ces mutants nous avons évalué la plupart des étapes de la génération depuis l'ARN génomique jusqu'à celle de l'ADN proviral, nous nous sommes concentrés sur l'un des rares paramètres que nous n'avions pas pris en compte jusque-là : le choix des sites d'intégration. En effet, dans notre système expérimental, nous évaluons l'efficacité de l'intégration sur la base de l'expression d'un transgène inséré dans l'ADN proviral. Si le choix des sites d'intégration était affecté chez les mutants, cela pourrait influencer notre lecture, introduisant éventuellement un écart entre les résultats attendus et observés. L'interaction de IN avec LEDGF/p75 est particulièrement importante pour le choix des sites d'intégration. Le modèle actuel de ciblage de l'intégration prend en effet en charge un mécanisme en deux phases. Premièrement, CPSF6 permet la libération du complexe RTC/PIC du complexe de pores nucléaires, en liant CA, puis le conduit au-delà de la périphérie nucléaire, vers les régions internes du noyau et, en particulier dans les "speckles-associated domains" (SPAD) (Sowd et al., 2016; Bejarano et al., 2019; Francis et al., 2020; Li et al., 2020), des régions génomiques associées aux speckles nucléaires (Chen et al., 2018; Chen and Belmont, 2019). Ensuite, via la liaison LEDGF/p75 à l'IN, l'intégration se fait préférentiellement dans les corps géniques, sous l'influence potentielle des machineries cellulaires d'épissage de l'ARNm et/ou d'élongation transcriptionnelle (Ciuffi et al., 2005; Gijsbers et al., 2010; Singh et al., 2015). LEDGF/p75 interagit avec l'IN via les domaines CCD et NTD et non le CTD. Néanmoins, un rôle de l'IN CTD, indépendant de LEDGF/p75, dans la liaison à la chromatine et le ciblage d'intégration a été montré (Demeulemeester et al., 2014; Benleulmi et al., 2017; Mauro et al., 2019; Winans et al., 2022). Par conséquent, nous avons commencé à cartographier les sites d'intégration obtenus avec le mutant TKTK (l'intégration avec le mutant LKLK était si faible qu'il n'aurait pas été possible d'obtenir du matériel pour

ces analyses). L'analyse des sites d'intégration a été réalisée en collaboration avec le Dr Marina Lusic, à l'Université de Heidelberg, où j'ai passé un mois pour démarrer le projet.

RESULTATS ET DISCUSSION

Identification des propriétés et des origines phylogénétiques spécifiques des intégrases du VIH-1 M et O

Pour comprendre le rôle du motif CLA dans l'intégrase du groupe O, nous avons remplacé dans IN O la séquence dans les positions CLA par NQNQ (IN O/NQNQ), une séquence qui s'est avérée abolir l'intégration dans les isolats M (Kanja et al., 2020). Dans le groupe O, contrairement à ce qui a été observé pour le groupe M, le remplacement de la séquence d'origine dans les positions CLA par NQNQ n'a pas affecté l'intégration dans les cellules HEK293T et Jurkat. Ce résultat indique que les isolats O ne nécessitent pas la fonction exercée par le motif CLA ou que cette fonction est assurée soit par une autre région de l'intégrase soit par une autre protéine. En construisant des chimères entre les intégrases O et M nous avons pu identifier que c'était le NTD de IN O qui permettait au groupe O de contourner le besoin du motif CLA pour l'intégration. Ensuite, pour identifier les acides aminés responsables de ce phénotype, nous avons aligné les séquences d'acides aminés de IN NTD O et M et avons remarqué que 10 positions différaient entre les deux. Selon le score de la matrice BLOSUM62 (Henikoff and Henikoff, 1992), le remplacement de quatre de ces résidus (Q7, G27, P41, H44) induit des changements plus drastiques dans les propriétés de la protéine par rapport à la substitution des autres résidus. Pour tester si les quatre résidus $Q_7G_{27}P_{41}H_{44}$ du NTD O étaient ceux permettant la complémentation de la fonctionnalité assurée par le motif CLA, nous les avons insérés dans le NTD du IN M qui abrite, aux positions CLA, la séquence consensus des isolats O (IN M/QGPH/NQKK). Ces deux mutations ont été suffisantes pour passer d'une efficacité d'intégration de 25 % de IN M/NQKK à 100 % de wt IN M, à la fois dans les cellules HEK293T et Jurkat. Les mêmes résultats ont été obtenus en remplaçant le NTD M entier par le NTD O (IN M/NTD-O/NQKK). Par conséquent, nous avons conclu que les $Q_7G_{27}P_{41}H_{44}$ sont les acides aminés suffisants pour complémenter fonctionnellement l'absence du motif CLA dans le groupe O et nous avons décidé de désigner ces positions par le motif "N-terminal O group" (NOG). La restauration des niveaux d'intégration lorsque le motif NOG est inséré dans IN M/NQKK est obtenue en augmentant la quantité de produits de transcription inverse (RTP) et en favorisant leur intégration en améliorant le traitement en 3'. Dans les cellules Jurkat, ces effets étaient

concomitants à une augmentation de la stabilité de la capside, qui pourrait potentiellement favoriser les deux processus en augmentant le temps de séjour des acides nucléiques dans la capside (Forshey et al., 2002; Stremlau et al., 2006; Eschbach et al., 2020).

Comment des domaines aussi différents ont-ils pu converger pour assurer des fonctions aussi similaires qu'interchangeables ? L'explication la plus simple est qu'ils sont impliqués dans la même étape mécanistique du cycle infectieux, probablement par une interaction essentielle avec la même molécule. Le laboratoire avait montré, au cours du travail de doctorat de Marine, que les trois premiers résidus du motif ($N_{222}K_{240}N_{254}$) formaient une surface chargée positivement, absente dans le cas de la séquence $N_{222}Q_{240}K_{254}$ (Kanja et al., 2020). Il a été proposé que cette surface interagisse, avec la contribution éventuelle du K273 supplémentaire, avec un partenaire redondant chargé négativement (Kanja et al., 2020) tel que le squelette des molécules d'ADN ou d'ARN. Dans IN O, la présence du motif NOG devrait induire la formation d'une surface chargée positivement, ce qui pourrait conduire l'interaction à impliquer préférentiellement le NTD.

Pour comprendre comment ces deux motifs divergents ont pu émerger, nous avons retracé l'origine et l'évolution des motifs NOG et CLA en examinant les mêmes positions dans les IN des virus qui sont à l'origine des groupes O et M, SIVgor et SIVcpz*Ptt* respectivement. La séquence QGPH, est hautement conservée dans le groupe O et dans SIVgor et on la retrouve également dans l'isolat supposé être le plus proche du fondateur du VIH-1 O, SIVgor BQID2 (D'Arc et al., 2015). Ces observations suggèrent fortement que ce motif a été sélectionné dans le virus du singe et est resté inchangé après transmission à l'homme. Dans le cas de SIVcpz*Ptt*, bien que les résidus trouvés dans les positions CLA soient majoritairement K et N, les mêmes que le motif CLA essentiel dans HIV-1 M, aucun conservation ne se dégage, hormis la conservation de K273. Pour évaluer la possibilité que la séquence NKNK était néanmoins présente dans l'isolat transmis à l'homme et conservé depuis, nous avons comparé les séquences trouvées dans les deux isolats de SIVcpz*Ptt* identifiés comme les plus proches du VIH-1, isolats SIVcpz*Ptt* MB897 et SIVcpz*Ptt* LB715 (Heuverswyn et al., 2007). Dans aucun des deux cas, la séquence était NKNK. Le fait qu'aucune des combinaisons de ces acides aminés n'ait été sélectionnée dans le virus simien pourrait indiquer que la fonction exercée par ce motif n'était pas requise dans les cellules simiennes. Alternativement, on pourrait imaginer que, même si nécessaire, la pression sélective pour le motif CLA n'était pas aussi forte qu'elle semble l'être chez l'homme, permettant la coexistence de plusieurs séquences fonctionnelles chez le virus simien. Dans les deux cas, il est tentant de supposer que l'émergence du motif NKNK faisait partie du processus d'adaptation au nouvel hôte. Ce sujet a soulevé la question de savoir pourquoi la sélection

pour le même motif n'a pas émergé également après le transfert de SIVgor aux humains. En principe, la présence dans ce virus d'un motif déjà fonctionnel (celui du NOG) n'excluait pas l'ajout d'un second motif (comme celui du CLA) qui aurait pu conférer un avantage supplémentaire au virus porteur des deux. Toutefois, lorsque nous avons testé cette possibilité en remplaçant la séquence NQKK par NKNK dans l'isolât O, nous n'avons observé aucune augmentation de l'intégration, apportant une réponse potentielle à la question. Ces résultats, ainsi que l'observation que le groupe O n'a pas besoin du motif CLA pour l'intégration, sont plutôt évocateurs de la présence d'une épistasie dominante des positions qui constituent le motif NOG par rapport à celles constituant le motif CLA.

Néanmoins, nous avons voulu comprendre si un virus avec un IN de SIVcpz*Ptt* mais avec la séquence NKNK dans les positions CLA pouvait avoir été infectieux dans les cellules humaines. Nous avons donc procédé au remplacement de la séquence KKKK dans les positions CLA de l'intégrase MB897 de l'isolat SIVcpz*Ptt* par NKNK et l'avons testée dans la lignée lymphocytaire humaine Jurkat. Ce changement était suffisant pour réduire l'intégration à environ 10 % par rapport à wt IN SIVcpz*Ptt*. Le remplacement des acides aminés dans les positions CLA par NQNQ, condition qui a aboli l'intégration dans IN M, a provoqué une chute de l'intégration à des niveaux indétectables, ainsi qu'une diminution significative des niveaux de transcription inverse. Au bilan, ces résultats indiquent que, dans SIVcpz*Ptt*, la séquence dans les positions CLA est cruciale pour déterminer les niveaux d'intégration, comme pour l'IN M. Contrairement à l'IN M, cependant, dans SIVcpzPtt, la séquence NKNK n'a pas assuré des niveaux élevés d'intégration.

Dans l'ensemble, avec ce travail, nous documentons que les intégrases des groupes M et O du VIH-1 ont développé deux motifs fonctionnels spécifiques au groupe phylogénétique qui peuvent se complémenter mutuellement. Un motif (CLA) est localisé dans le CTD de la protéine du groupe M, l'autre (NOG) dans le NTD des isolats du groupe O.

Le rôle des acides aminés amidiques dans le motif CLA

Pour comprendre si le mutant TKTK conduit à un changement dans le choix des sites d'intégration, une transduction de cellules Jurkat avec des particules virales portant soit le IN M wt soit le M/TKTK a été réalisée. Une bibliothèque enrichie en sites d'intégration de l'ADN génomique a ensuite été préparée, séquencée et analysée. A partir des cellules infectées avec un virus portant une IN wt, un total de 4 674 sites ont été cartographiés, tandis que pour le mutant TKTK, les sites cartographiés ont été 1 375. De manière frappante, le pourcentage de sites d'intégration intra- et intergéniques était différent entre les deux

intégrases, la wt ciblant préférentiellement les sites intragéniques (72%) comme décrit dans la littérature, tandis que le mutant TKTK ne s'intègre dans ces régions que 56% du temps. De plus, en examinant les niveaux d'expression des gènes ciblés, il était clair que le mutant s'intègre avec une fréquence plus élevée dans les gènes à faible expression et/ou réduits au silence par rapport à l'intégrase sauvage. Ces résultats pourraient expliquer l'écart entre les niveaux d'intégration attendus et observés pour le mutant TKTK. L'efficacité d'intégration observée est en effet mesurée grâce à la présence de gènes rapporteurs dans le génome viral modifié par les VLP ($PURO^R$). Par conséquent, notre détection d'événements d'intégration est limitée à ceux qui sont situés dans des régions transcriptionnellement actives. Les niveaux d'intégration plus élevés attendus sur une base théorique, pour le mutant TKTK peuvent donc s'expliquer car une partie importante des événements d'intégration se produisent dans des régions où les gènes rapporteurs ne peuvent pas être transcrits. En effet, on peut estimer que l'efficacité d'intégration observée du mutant TKTK à partir de la cartographie est d'environ 30% par rapport au wt, ce qui correspond mieux à la valeur attendue trouvée lorsque l'intégration a été estimée sur les défauts de traitement 3' et d'importation nucléaire.

Le mutant TKTK présente également des préférences de chromatine différentes, avec une tendance inverse par rapport au wt. En effet, ses sites d'intégration sont moins associés aux caractéristiques signature de chromatine ouverte H3K4me1, H3K27Ac et H3K36me3 (Wang et al., 2007; Roth et al., 2011; Kvaratskhelia et al., 2014; Sowd et al., 2016). H3K4me1 et H3K27Ac sont des signatures de super-enhancer, qui sont des sites cibles préférentiels du VIH-1. Cependant il semble maintenant que cette préférence soit une conséquence de l'abondance des super-enhancer dans les SPAD, l'une des cibles privilégiées pour l'intégration (Bedwell et al., 2021; Singh et al., 2022). H3K36me3, au contraire, est associé aux corps de gènes. De plus, c'est la modification préférentiellement reconnue par le domaine PWWP de LEDGF/p75. Le phénotype du mutant TKTK ne dépend probablement pas de son incapacité à se lier à LEDGF/p75, car il interagit avec le NTD et le domaine CCD de l'IN, tandis que le mutant TKTK a deux mutations ponctuelles dans le CTD. Une étude récente a mis en évidence la façon dont les sites d'intégration du VIH-1 semblent être plus fortement corrélés avec les gènes associés à H3K36me3, plutôt qu'avec les gènes associés à LEDGF/p75 liés à la chromatine (Singh et al., 2022). Cette observation suggère que le H3K36me3 pourrait être la cible de l'intégration du VIH-1 même de manière non corrélée LEDGF/p75. L'existence de ce type de mécanisme pourrait expliquer comment les sites d'intégration du mutant TKTK sont moins associés au marqueur H3K36me3, alors que sa liaison avec LEDGF/p75 n'est pas perturbée. À l'inverse, le mutant TKTK a montré plus

d'affinité pour les marques répressives de la chromatine telles que H3K9me2/3 et H3K27me3 (Wang et al., 2007; Roth et al., 2011; Sowd et al., 2016).

Si la phase initiale du ciblage du site d'insertion est connue, les phases ultérieures, dont le contact avec la chromatine de l'hôte, sont mal caractérisées. Nos résultats, dans la lignée de ceux trouvés dans la littérature, vont dans une direction où l'IN CTD apparaît responsable du choix des sites d'intégration, et en particulier du tethering de la chromatine. Si l'observation peut s'expliquer par un effet direct des mutations CTD, la possibilité qu'un ou plusieurs facteurs cellulaires non encore identifiés soient impliqués n'est pas à exclure. Ils pourraient être impliqués, de manière directe ou indirecte, dans l'orientation de l'événement d'intégration vers des corps de gènes transcriptionnellement actifs. L'identification de ce/ces facteur(s) cellulaire(s) représenterait une étape cruciale pour la poursuite de ce projet.

CONCLUSIONS

Grâce au projet principal de ma thèse de doctorat, qui portait sur l'étude du rôle du motif CLA dans le groupe O, nous avons pu mettre en évidence l'existence de deux motifs fonctionnels groupes spécifiques dans les intégrases des groupes M et O se complétant fonctionnellement : le motif du groupe M CLA ($N_{22}2K_{240}N_{254}K_{273}$), localisé dans le CTD, et le motif du groupe O NOG ($Q_7G_{27}P_{41}H_{44}$), localisé dans le NTD. Ce résultat est important à la lumière de la compréhension du succès épidémiologique du groupe M sur le groupe O. Après transmission inter-espèces à l'homme, les caractéristiques génétiques de chaque groupe sont ce qui, plus que d'autres facteurs, a probablement déterminé le succès de réplication des isolats. Le fait que le groupe M ait optimisé et sélectionné son motif CLA en tant qu'adaptation au nouvel hôte pourrait représenter l'une des nombreuses caractéristiques conférant à ce groupe un avantage de réplication pour la réplication dans l'hôte humain. Comprendre ces mécanismes est important pour mieux comprendre l'histoire derrière les menaces importantes pour la santé humaine telles que la pandémie de VIH-1 groupe M et pour pouvoir la combattre avec des outils optimisés.

Un projet secondaire a été réalisé en collaboration avec le Dr Marina Lusic et son équipe à l'Université de Heidelberg, axé sur la compréhension du rôle des acides aminés amidiques présents dans le motif CLA. L'investigation sur le N du motif a conduit à découvrir un nouveau rôle potentiellement LEDGF/p75-indépendant du IN dans le choix des sites d'intégration. Au total, bien que limité à quelques résultats préliminaires, cette partie de mon travail de thèse a démontré avec succès le rôle de l'IN CTD dans le choix des sites d'intégration. Comprendre le mécanisme derrière ce phénotype pourrait ouvrir de nouvelles perspectives intéressantes

dans la thérapie à base de vecteurs lentiviraux. Il est clair que le mutant TKTK a un grave défaut d'intégration et son intégration dans des régions non transcriptionnellement actives, ce qui rend difficile d'imaginer son emploi pour la thérapie. Cependant, élucider le mécanisme pour lequel le VIH-1 IN choisit ses sites d'intégration, pourrait permettre de le "hacker" pour cibler des régions génomiques spécifiques du génome humain, ce qui pourrait réduire drastiquement la possibilité d'induire une mutagenèse insertionnelle (et potentiellement une oncogenèse) dans les cellules cibles.

Enfin, au cours de ma thèse, j'ai travaillé sur un troisième projet, dont l'objectif était de comprendre le rôle du motif CLA dans le mécanisme de « uncoating » de la capside virale. Bien que ce projet n'ait pas atteint le niveau requis pour prévoir la publication des résultats obtenus, il m'a permis d'acquérir des connaissances techniques et théoriques pertinentes sur le sujet. Pendant que je travaillais dessus, une nouvelle vague d'articles mettant en évidence l'entrée nucléaire de noyaux de capsides intacts ou presque intacts est sortie. Ainsi, lorsque le premier confinement causé par la pandémie de Covid est arrivé, j'ai profité de l'arrêt imposé à mon activité expérimentale, pour exploiter mes connaissances sur l'étape de de « uncoating » pour rédiger une revue sur le sujet, mettant l'accent sur les avancées récentes décrites dans la littérature. La revue a été publiée dans Frontiers in Microbiology et est cosignée avec le Dr Daniela Lener et le Dr Matteo Negroni et peut être trouvée en annexe 1. Un deuxième aspect pour lequel ce projet s'est avéré déterminant pour mon travail de doctorat était qu'il a permis m'a permis de transférer les compétences que j'avais acquises (par exemple, le test EURT) au dernier développement du projet qui a conduit à l'article de recherche présenté dans les résultats de cette thèse.

RÉFÉRENCES

Bedwell, G. J., Jang, S., Li, W., Singh, P. K., and Engelman, A. N. (2021). rigrag: High-resolution mapping of genic targeting preferences during HIV-1 integration in vitro and in vivo. *Nucleic Acids Research* 49, 7330–7346. doi: 10.1093/nar/gkab514.

Bego, M. G., Cong, L., Mack, K., Kirchhoff, F., and Cohen, É. A. (2016). Differential Control of BST2 Restriction and Plasmacytoid Dendritic Cell Antiviral Response by Antagonists Encoded by HIV-1 Group M and O Strains. *Journal of Virology* 90, 10236–10246. doi: 10.1128/jvi.01131-16.

Bejarano, D. A., Peng, K., Laketa, V., Börner, K., Jost, K. L., Lucic, B., et al. (2019). HIV-1 nuclear import in macrophages is regulated by CPSF6-capsid interactions at the nuclear pore complex. *Elife* 8, 1–31. doi: 10.7554/eLife.41800.

Benleulmi, M. S., Matysiak, J., Robert, X., Miskey, C., Mauro, E., Lapaillerie, D., et al. (2017). Modulation of the functional association between the HIV-1 intasome and the nucleosome by histone amino-terminal tails. *Retrovirology* 14, 1–16. doi: 10.1186/s12977-017-0378-x.

Busschots, K., Vercammen, J., Emiliani, S., Benarous, R., Engelborghs, Y., Christ, F., et al. (2005). The Interaction of LEDGF/p75 with Integrase Is Lentivirus-specific and Promotes DNA Binding. *Journal of Biological Chemistry* 280, 17841–17847. doi: 10.1074/jbc.M411681200.

Chen, Y., and Belmont, A. S. (2019). Genome organization around nuclear speckles. *Current Opinion in Genetics and Development* 55, 91–99. doi: 10.1016/j.gde.2019.06.008.

Chen, Y., Zhang, Y., Wang, Y., Zhang, L., Brinkman, E. K., Adam, S. A., et al. (2018). Mapping 3D genome organization relative to nuclear compartments using TSA-Seq as a cytological ruler. *Journal of Cell Biology* 217, 4025–4048. doi: 10.1083/jcb.201807108.

Ciuffi, A., Llano, M., Poeschla, E., Hoffmann, C., Leipzig, J., Shinn, P., et al. (2005). A role for LEDGF/p75 in targeting HIV DNA integration. *Nature Medicine* 11, 1287–1289. doi: 10.1038/nm1329.

D'Arc, M., Ayouba, A., Esteban, A., Learn, G. H., Boué, V., Liegeois, F., et al. (2015). Origin of the HIV-1 group O epidemic in western lowland gorillas. *Proc Natl Acad Sci U S A* 112, E1343–E1352. doi: 10.1073/pnas.1502022112.

Demeulemeester, J., Vets, S., Schrijvers, R., Madlala, P., de Maeyer, M., de Rijck, J., et al. (2014). HIV-1 integrase variants retarget viral integration and are associated with disease progression in a chronic infection cohort. *Cell Host and Microbe* 16, 651–662. doi: 10.1016/j.chom.2014.09.016.

Eijkelenboom, A. P. A. M., van den Ent, F. M. I., Vos, A., Doreleijers, J. F., Hård, K., Tullius, T. D., et al. (1997). The solution structure of the amino-terminal HHCC domain of HIV-2 integrase: a three-helix bundle stabilized by zinc. *Current Biology* 7, 739–746. doi: 10.1016/S0960-9822(06)00332-0.

Elliott, J., Eschbach, J. E., Koneru, P. C., Li, W., Puray-Chavez, M., Townsend, D., et al. (2020). Integrase-RNA interactions underscore the critical role of integrase in HIV-1 virion morphogenesis. *Elife* 9, 1–56. doi: 10.7554/ELIFE.54311.

Engelman, A., Bushman, F. D., and Craigie, R. (1993). Identification of discrete functional domains of HIV-1 integrase and their organization within an active multimeric complex. *EMBO Journal* 12, 3269–3275. doi: 10.1002/j.1460-2075.1993.tb05996.x.

Engelman, A., and Craigie, R. (1992). Identification of conserved amino acid residues critical for human immunodeficiency virus type 1 integrase function in vitro. *Journal of Virology* 66, 6361–6369. doi: 10.1128/jvi.66.11.6361-6369.1992.

Engelman, A., Hickman, A. B., and Craigie, R. (1994). The core and carboxyl-terminal domains of the integrase protein of human immunodeficiency virus type 1 each contribute to nonspecific DNA binding. *Journal of Virology* 68, 5911–5917. doi: 10.1128/jvi.68.9.5911-5917.1994.

Engelman, A., Mizuuchi, K., and Craigie, R. (1991). HIV-1 DNA integration: Mechanism of viral DNA cleavage and DNA strand transfer. *Cell* 67, 1211–1221. doi: 10.1016/0092-8674(91)90297-C.

Engelman, A. N., and Kvaratskhelia, M. (2022). Multimodal Functionalities of HIV-1 Integrase. *Viruses* 14, 926. doi: 10.3390/v14050926.

Eschbach, J. E., Elliott, J. L., Li, W., Zadrozny, K. K., Davis, K., Mohammed, S. J., et al. (2020). Capsid Lattice Destabilization Leads to Premature Loss of the Viral Genome and Integrase Enzyme during HIV-1 Infection. *Journal of Virology* 95. doi: 10.1128/JVI.00984-20.

Forshey, B. M., von Schwedler, U., Sundquist, W. I., and Aiken, C. (2002). Formation of a Human Immunodeficiency Virus Type 1 Core of Optimal Stability Is Crucial for Viral Replication. *Journal of Virology* 76, 5667–5677. doi: 10.1128/jvi.76.11.5667-5677.2002.

Francis, A. C., Marin, M., Singh, P. K., Achuthan, V., Prellberg, M. J., Palermino-Rowland, K., et al. (2020). HIV-1 replication complexes accumulate in nuclear speckles and integrate into speckle-associated genomic domains. *Nature Communications* 11. doi: 10.1038/s41467-020-17256-8.

Gao, F., Bailes, E., Robertson, D. L., Chen, Y., Rodenburg, C. M., Michael, S. F., et al. (1999). Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. *Nature* 397, 436–441. doi: 10.1038/17130.

Gijsbers, R., Ronen, K., Vets, S., Malani, N., de Rijck, J., McNeely, M., et al. (2010). LEDGF hybrids efficiently retarget lentiviral integration into heterochromatin. *Molecular Therapy* 18, 552–560. doi: 10.1038/mt.2010.36.

Henikoff, S., and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* 89, 10915–10919. doi: 10.1073/pnas.89.22.10915.

Heuverswyn, F. van, Li, Y., Bailes, E., Neel, C., Lafay, B., Keele, B. F., et al. (2007). Genetic diversity and phylogeographic clustering of SIVcpzPtt in wild chimpanzees in Cameroon. *Virology* 368, 155–171. doi: 10.1016/j.virol.2007.06.018.

Kanja, M., Cappy, P., Levy, N., Oladosu, O., Schmidt, S., Rossolillo, P., et al. (2020). NKNK: a New Essential Motif in the C-Terminal Domain of HIV-1 Group M Integrases. *Journal of Virology* 94, 1–23. doi: 10.1128/JVI.01035-20.

Keele, B. F., van Heuverswyn, F., Li, Y., Bailes, E., Takehisa, J., Santiago, M. L., et al. (2006). Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. *Science (1979)* 313, 523–526. doi: 10.1126/science.1126531.

Kessl, J. J., Kutluay, S. B., Townsend, D., Rebensburg, S., Slaughter, A., Larue, R. C., et al. (2016). HIV-1 Integrase Binds the Viral RNA Genome and Is Essential during Virion Morphogenesis. *Cell* 166, 1257-1268.e12. doi: 10.1016/j.cell.2016.07.044.

Kluge, S. F., Mack, K., Iyer, S. S., Pujol, F. M., Heigele, A., Learn, G. H., et al. (2014). Nef Proteins of Epidemic HIV-1 Group O Strains Antagonize Human Tetherin. *Cell Host & Microbe* 16, 639–650. doi: 10.1016/j.chom.2014.10.002.

Korber, B., Muldoon, M., Theiler, J., Gao, F., Gupta, R., Lapedes, A., et al. (2000). Timing the Ancestor of the HIV-1 Pandemic Strains. *Science (1979)* 288, 1789–1796. doi: 10.1126/science.288.5472.1789.

Kulkosky, J., Jones, K. S., Katz, R. A., Mack, J. P., and Skalka, A. M. (1992). Residues critical for retroviral integrative recombination in a region that is highly conserved among retroviral/retrotransposon integrases and bacterial insertion sequence transposases. *Molecular and Cellular Biology* 12, 2331–2338. doi: 10.1128/mcb.12.5.2331-2338.1992.

Kvaratskhelia, M., Sharma, A., Larue, R. C., Serrao, E., and Engelman, A. (2014). Molecular mechanisms of retroviral integration site selection. *Nucleic Acids Research* 42, 10209–10225. doi: 10.1093/nar/gku769.

Lemey, P., Pybus, O. G., Rambaut, A., Drummond, A. J., Robertson, D. L., Roques, P., et al. (2004). The molecular population genetics of HIV-1 group O. *Genetics* 167, 1059–1068. doi: 10.1534/genetics.104.026666.

Leoz, M., Feyertag, F., Kfutwah, A., Mauclère, P., Lachenal, G., Damond, F., et al. (2015). The Two-Phase Emergence of Non Pandemic HIV-1 Group O in Cameroon. *PLoS Pathogens* 11, 1–13. doi: 10.1371/journal.ppat.1005029.

Li, W., Singh, P. K., Sowd, G. A., Bedwell, G. J., Jang, S., Achuthan, V., et al. (2020). CPSF6-Dependent Targeting of Speckle-Associated Domains Distinguishes Primate from Nonprimate Lentiviral Integration. *mBio* 11, 1–20. doi: 10.1128/mBio.02254-20.

Mauro, E., Lesbats, P., Lapaillerie, D., Chaignepain, S., Maillot, B., Oladosu, O., et al. (2019). Human H4 tail stimulates HIV-1 integration through binding to the carboxy-terminal domain of integrase. *Nucleic Acids Research* 47, 3607–3618. doi: 10.1093/nar/gkz091.

Mourez, T., Simon, F., and Plantiera, J. C. (2013). Non-M variants of human immunodeficiency virus type. *Clinical Microbiology Reviews* 26, 448–461. doi: 10.1128/CMR.00012-13.

Passos, D. O., Li, M., Jóźwik, I. K., Zhao, X. Z., Santos-Martins, D., Yang, R., et al. (2020). Structural basis for strand-transfer inhibitor binding to HIV intasomes. *Science (1979)* 367, 810–814. doi: 10.1126/science.aay8015.

Passos, D. O., Li, M., Yang, R., Rebensburg, S. v., Ghirlando, R., Jeon, Y., et al. (2017). Cryo-EM structures and atomic model of the HIV-1 strand transfer complex intasome. *Science (1979)* 355, 89–92. doi: 10.1126/science.aah5163.

Peeters, M., Gueye, A., Mboup, S., Bibollet-Ruche, F., Ezaka, E., Mulanga, C., et al. (1997). Geographical distribution of HIV-1 group O viruses in Africa. *AIDS* 11, 493–498.

Plantier, J. C., Leoz, M., Dickerson, J. E., de Oliveira, F., Cordonnier, F., Lemée, V., et al. (2009). A new human immunodeficiency virus derived from gorillas. *Nature Medicine* 15, 871–872. doi: 10.1038/nm.2016.

Roth, S. L., Malani, N., and Bushman, F. D. (2011). Gammaretroviral Integration into Nucleosomal Target DNA In Vivo . *Journal of Virology* 85, 7393–7401. doi: 10.1128/jvi.00635-11.

Santoro, M. M., and Perno, C. F. (2013). HIV-1 Genetic Variability and Clinical Implications. *ISRN Microbiology* 2013, 1–20. doi: 10.1155/2013/481314.

Singh, P. K., Bedwell, G. J., and Engelman, A. N. (2022). Spatial and Genomic Correlates of HIV-1 Integration Site Targeting. *Cells* 11, 655. doi: 10.3390/cells11040655.

Singh, P. K., Plumb, M. R., Ferris, A. L., Iben, J. R., Wu, X., Fadel, H. J., et al. (2015). LEDGF/p75 interacts with mRNA splicing factors and targets HIV-1 integration to highly spliced genes. *Genes and Development* 29, 2287–2297. doi: 10.1101/gad.267609.115.

Sowd, G. A., Serrao, E., Wang, H., Wang, W., Fadel, H. J., Poeschla, E. M., et al. (2016). A critical role for alternative polyadenylation factor CPSF6 in targeting HIV-1 integration to transcriptionally active chromatin. *Proc Natl Acad Sci U S A* 113, E1054–E1063. doi: 10.1073/pnas.1524213113.

Stremlau, M., Perron, M., Lee, M., Li, Y., Song, B., Javanbakht, H., et al. (2006). Specific recognition and accelerated uncoating of retroviral capsids by the TRIM5 restriction factor. *Proceedings of the National Academy of Sciences* 103, 5514–5519. doi: 10.1073/pnas.0509996103.

van Gent, D. C., Vink, C., Groeneger, A. A., and Plasterk, R. H. (1993). Complementation between HIV integrase proteins mutated in different domains. *The EMBO Journal* 12, 3261–3267. doi: 10.1002/j.1460-2075.1993.tb05995.x.

Wang, G. P., Ciuffi, A., Leipzig, J., Berry, C. C., and Bushman, F. D. (2007). HIV integration site selection: Analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. *Genome Research* 17, 1186–1194. doi: 10.1101/gr.6286907.

Wilkinson, T. A., Januszyk, K., Phillips, M. L., Tekeste, S. S., Zhang, M., Miller, J. T., et al. (2009). Identifying and Characterizing a Functional HIV-1 Reverse Transcriptase-binding Site on Integrase. *Journal of Biological Chemistry* 284, 7931–7939. doi: 10.1074/jbc.M806241200.

Winans, S., Yu, H. J., de los Santos, K., Wang, G. Z., KewalRamani, V. N., and Goff, S. P. (2022). A point mutation in HIV-1 integrase redirects proviral integration into centromeric repeats. *Nature Communications* 13, 1474. doi: 10.1038/s41467-022-29097-8.

Zheng, R., Jenkins, T. M., and Craigie, R. (1996). Zinc folds the N-terminal domain of HIV-1 integrase, promotes multimerization, and enhances catalytic activity. *Proceedings of the National Academy of Sciences* 93, 13659–13664. doi: 10.1073/pnas.93.24.13659.

Zhu, K., Dobard, C., and Chow, S. A. (2004). Requirement for Integrase during Reverse Transcription of Human Immunodeficiency Virus Type 1 and the Effect of Cysteine Mutations of Integrase on Its Interactions with Reverse Transcriptase. *Journal of Virology* 78, 5045–5055. doi: 10.1128/jvi.78.10.5045-5055.2004.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

LIST OF ANNEXES

LIST OF ABBREVIATIONS

-sssDNA	minus strand strong stop DNA
+sssDNA	plus strand strong stop DNA
(-)sDNA	minus-strand DNA
(+)sDNA	plus-strand DNA
1LTRc	1 long terminal repeat circle
2LTRc	2 long terminal repeats circle
AIDS	acquired immuno definciency syndrome
ALLINI	allosteric IN inhibitors
APOBEC3	apoliprotein B mRNA-editing enzyme catalytic polypeptide-like 3
BAF	barrier-to-autointegration factor
BER	base excision repair
CA	capsid
cART	combined antiretroviral therapy
CCD	catalytic core domain
CFIm	cleavage factor I mammalian
CIC	conserved intasome core
CLA	C-terminal lysin amidic
CPSF6	cleavage and polyadenylation specificity factor 6
CRF	circulating recombinant form
CSC	cleaved synaptic complex
CTD	c-terminal domain
CypA	cyclophilin A
DRC	Democratic Republic of Congo
ESCRT	endosomial sorting complexes required for transport
FAD	food and drug administration
GAPDH	glyceraldehyde-3-phosphate dehydrogenase
gp	glycoprotein
HAART	highly active antiretroviral treatments
HDAC1	histone deacetylase 1
HDGFL2	hepatoma-derived growth factor like 2
HIV	**human immunodeficiency virus**
HMGA1	high mobility group chromosomal protein A1
HRP	horseradish peroxidase
IBD	integrase binding domain
IFN	interferon
IN	integrase
INI1	integrase interactor 1
IN-LAI	IN-LEDGF allosteric inhibitors
INSTI	IN strand transfer inhibitors
LAD	lamina associated domain
LANL	**Los Alamos**
LEDGIN	LEDGF inhibitors
LEDGF/p75	Lens epithelium-derived growth factor
LTR	long terminal repeat
MA	matrix
MINI	multimeric IN inhibitors
MOI	multiplicity of infection
mRNA	messenger RNA
NC	nucleocapsid
NCINI	noncatalytic site IN inhibitors
Nef	negative regulatory factor
NES	nuclear export signal
nGFP	nuclear green fluorescent protein
NHEJ	non-homologous end joining
NLS	nuclear localization signal
NNRTI	nonnucleoside RT inhibitors
NOG	N-terminal O group
nRFP	nuclear red fluorescent motif
NRTI	nucleoside RT inhibitors
NTD	N-terminal domain
Nup	nucleoporine
PBS	prime binding site
PFV	prototype foamy virus
PHAT	pseudo-HEAT analogous topology
PI	protease inhibitor

PIC	pre-integration complex
PJ	perfect junction
PPT	poly purine tract
PR	protease
PRD	Pro-rich domain
PWWP	Pro-Trp-Trp-Pro
Rev	RNA splicing regulator
RLU	relative light unit
RRE	Rev responsive element
RRM	RNA recognition domain
RT	reverse transcriptase
RTC	reverse transcription complex
RTP	reverse transcription product
SAMHD1	SAM and HD domain-containing protein 1
SH3	Src homology 3
SIV	simian immunodeficiency virus
SL	stem loop
SP	spacer peptide
SPAD	speckles-associated domain
SSC	synaptic core complex
STC	strand transfer complex
SWI/SNF	switch/sucrose non-fermentable
TAR	trans-activating response region
Tat	transactivator protein
TCC	target capture complex
tDNA	target DNA
tMRCA	time of the most common ancestor
TNPO3	transportin 3
Ub	ubiquitin
URF	unique recombinant form
UTR	untranslated region
vDNA	viral DNA
Vif	viral infectivity factor
VLP	virus-like particle
Vpr	viral protein r
Vpu	viral protein unique
vRNA	viral RNA

INTRODUCTION

HIV-1: GENERAL INTRODUCTION

HIV-1 structure and genome

The human immunodeficiency virus (HIV) is a lentivirus responsible for the acquired immunodeficiency syndrome (AIDS) pandemic in humans. Two types of HIV have been defined: type 1 (HIV-1) and type 2 (HIV-2). HIV-1 is further subdivided into four groups: M for major, O for outlier, N for non-M-non-O, and P suspected to stand for “Plantier” the discoverer. HIV-2 instead has been divided in eight groups (A-H). Among all these phylogenetic categories, only HIV-1 group M has established a stable and worldwide infection in the human population, being responsible for 98% of HIV infections in the world.

HIV-1 viral particle

HIV-1 genome is an RNA of positive polarity that must be reverse transcribed, and the reverse transcription product integrated in the genome of the host cell, for viral replication to occur. The diameter of the viral particle is approximately 100-150 nm with an external lipidic bilayer envelope, carried along during budding from the infected cell. For this reason, both viral and cellular proteins can be found at the surface of the viral particles. The viral envelope proteins are necessary for recognition of the cellular receptor CD4 and the following viral entry. Underneath the envelope there is a protein lattice that surrounds an “inner envelope”, a fullerene cone-shaped core that protects the viral genome (Figure 1).

HIV-1 genome structure

The HIV-1 genome is around 9 kb and is found in a viral particle as two single stranded RNA molecules that dimerize near their 5’ extremities. The viral RNA (vRNA) is flanked by non-coding regions called the 5’ and 3’ untranslated regions (UTRs) that contain regulatory sequences mediating different steps of the viral cycle. At the 5' R and U5 are found, while at the 3' there are U3 and R. R contains the trans-activating response region (TAR), and the polyA signal that at the 5’ of the genome adopts a conformation that results in the loss of functionality while, at the 3’ end of the genome, it is functional (Berkhout and Jeang, 1989; Das et al., 1999). Downstream the 5'UTR region, two other important regulatory cis-elements are present, the primer-binding site (PBS) and the core encapsidation signal psi (ψ). PBS is the site from which reverse transcription is started (Rhim et al., 1991; Kleiman, 2002). ψ is

composed by four sequential RNA stem-loops (SL1-4) which are essential for the packaging of the full vRNA into the progeny virus particles, as well as dimerization of the two RNA copies of the viral genome (Kuzembayeva et al., 2014).

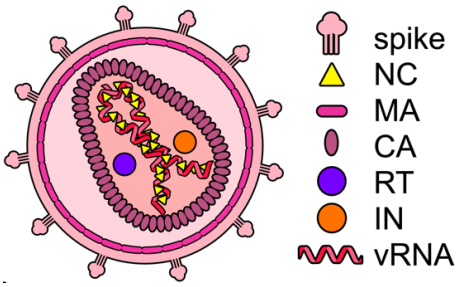


Figure 1. Schematic representation of a HIV-1 viral particle. The HIV-1 viral particle has an external lipidic bilayer covered in envelope "spikes", which are necessary for the viral entry. Inside this, a lattice of matrix (MA) is found, and a mature core, formed by the viral protein capsid (CA), which contains the two copies of the viral RNA genome (vRNA), complexed with the nucleocapsid (NC), and the viral enzymes reverse transcriptase (RT) and integrase (IN). Adapted from Toccafondi et al., 2021.

After reverse transcription and the formation of the double-stranded vDNA the 5' and 3' UTR regions are replaced by two identical regions at the 5' and 3' extremities of the proviral DNA (vDNA) called the long terminal repeats (LTRs), both composed by U3, R, and U5 (Resnick et al., 1984; Das et al., 1994; Zhang and Crumpacker, 2022) (Figure 2). The 5' LTR U3 region contains the promoter sequence, a core enhancer and amodulatory region, which are essential for the provirus transcription (Gaynor, 1992). Between the two LTRs are located the regions encoding the 9 viral proteins (Figure 2), organized in three large open reading frames (*gag, pol, and env*) and a set of individual genes coding for accessory proteins (*vif, vpr, tat, rev, vpu, nef*), which play important roles in the counteraction of the host restriction response and facilitate viral replication.

Viral proteins: Gag, Pol, Env

The viral genome carries three main open reading frames coding for the major structural and functional elements of the virus as well as the viral enzymes: Gag, Pol, and Env (Figure 2).

The *gag* gene codes a polyprotein, Pr55Gag, that, once processed by the viral protease, generates the individual structural proteins: matrix (MA), capsid (CA), spacer peptide 1 (SP1), nucleocapsid (NC), spacer peptide 2 (SP2) and p6 (Bukrinskaya, 2007). The MA protein forms a hexameric protein lattice, just underneath the viral lipid membrane. Indeed, MA anchors the cell membrane through its myristoylated N-ter domain (Dorfman et al., 1994; Spearman et al., 1994; Hill et al., 1996; Qu et al., 2021). This binding is particularly important during the formation of a new viral particle, generating the Pr55Gag precursors lattice. CA is a 24kDa

protein, composed of two globular domains, N-ter and C-ter, connected by a flexible linker. Mature CA is the smallest unit forming the capsid core as it assembles into around 250 hexamers and 12 pentamers, forming the fullerene cone structure inside the viral particle, which is about 120 nm long and 60 nm wide. The stability of the capsid core and its disassembly are two key elements of the infectious cycle, regulated by many factors. Indeed, being the most exposed viral protein during its journey across the cytoplasm, from the cell membrane to the nucleus, CA is the target of several restriction factors, and it mediates the interaction with cellular cofactors of the infection. NC is a small protein, less than 100 aa, containing two highly conserved zinc-finger domains formed by the $CX_2CX_4HX_4C$ sequence (CCHC motif). NC, as domain of Pr55Gag and mature protein, binds the genomic vRNA molecules, encapsidating them inot the budding viruses and stabilizing and compacting them to favor their packaging into the capsid core during the maturation of a new budded viral particle (Aldovini and Young, 1990; Clavel and Orenstein, 1990; Gien et al., 2022). The small peptide p6 has been shown to have multiple functions during the formation and maturation of a new viral particle, as inducing Vpr incorporation into the new particle (Paxton et al., 1993) and helping the budding off of the virion from the host cell membrane (Yu et al., 1995). In HIV-1 M, SP1 and SP2 are 12 and 16 aa long respectively. SP1 is involved in CA maturation and assembly, indeed, its depletion leads to the formation of abnormal Gag and CA lattices (Accola et al., 1998; Gross et al., 2000). SP2 is crucial for incorporation in the viral particles of Pr55Gag and Pr160Gag-Pol precursors (Hill et al., 2006).

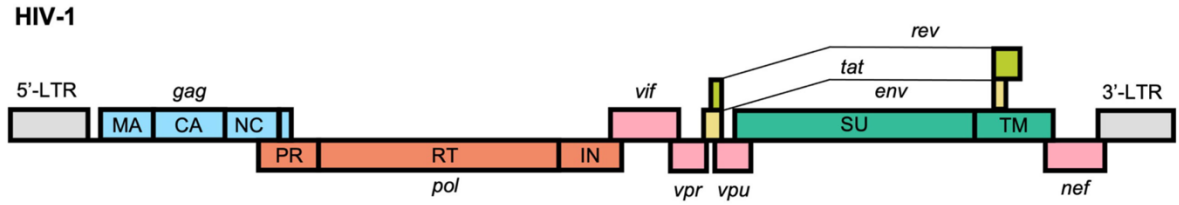


Figure 2. Schematic representation of the HIV-1 proviral genome. The three main retroviral core genes, *gag*, *pol*, and *env*, are shown in blue, orange, and green respectively. Each of them is segmented to show the proteins for which they are coding. The genes encoding the regulatory proteins, *tat* and *rev*, are shown in yellow and lime green respectively, while the genes encoding the other accessory proteins, *vif*, *vpr*, *vpu*, and *nef*, are shown in pink. At the two ends of the genome are drawn in grey, the two identical long-terminal repeats (LTRs). Image from Meissner et al., 2022.

The gene *pol* codes for the viral enzymes: protease (PR), reverse transcriptase (RT) and integrase (IN). These enzymes are first found as part of a large precursor: Pr160Gag-Pol. Gag-Pol is produced by ribosomal frameshifting in the course of translation of the viral RNA, a process that occurs at a frequency of around 5%, determining a stoichiometry of 1:20 for the components of *pol* with respect to *gag* (Jacks et al., 1988). The dimerization of the Gag-Pol precursor, leads to the dimerization of PR, activating the enzyme that first cleaves itself

from the polyprotein and then to proceeds to the release of all the individual components of the precursors (Erickson-Viitanen et al., 1989; Louis et al., 1994; Pettit et al., 2005; Meng et al., 2012). The mature form of the RT is a heterodimer, composed of a p66 and a p51 subunit. p51 is produced by cleavage of one p66 subunit of a p66/p66 homodimer, between the DNA polymerase and the RNase H domains, which is lost. This cleavage triggers a conformational change of p51 that also inactivates its polymerase activity, leaving to this subunit only a structural role of support for p66 (Jacobo-Molina et al., 1993; Wang et al., 1994; Jaeger et al., 1998). The p66 therefore is in charge of the two catalytic activities of the enzyme, the RNA/DNA dependent DNA polymerase, responsible for the conversion of the vRNA into double stranded DNA, and the RNase H activity that degrades the RNA component of RNA/DNA hybrids. This activity is required for removing the genomic RNA, once copied into its complementary sequence, from the nascent DNA. This is essential to allow the transfer of the nascent DNA, whose synthesis is blocked at the 5' end of the RNA genome (minus DNA strong stop) to the second copy of R present at the 3' end of the gRNA. The structure of p66 is similar to the one of a right hand, where the sub-domains "fingers", "thumb" and "palm" are taking contact with the nucleic acid complex (RNA/DNA hybrid). The "palm" contains the active site ($D_{110}D_{185}D_{186}$), that catalyzes the nucleophilic attack of the incoming nucleotide. IN is composed of three structural domains connected by flexible linkers and catalyzes the integration of the vDNA into the chromosomal DNA. It is a multifunctional protein that is also in charge of interacting with several viral and cellular partners. An exhaustive description of the protein structure and its functions is presented below (see *Integrase – Catalytic activity*).

The *env* gene encodes the viral envelope glycoproteins that are first translated as a precursor called gp160. Gp160 is glycosylated in the endoplasmic reticulum and then cleaved by a cellular furin protease in the Golgi apparatus, generating the surface protein (SU), glycoprotein gp120, and the transmembrane (TM) glycoprotein gp41 (Bernstein et al., 1994; Zhu et al., 2006). Gp120 is composed of five conserved domains (C1-5) and five variable regions (V1-5). Gp41 consists of three domains, the N-ter extra-cellular domain, the transmembrane domain and the C-ter cytoplasmatic tail (Douglas et al., 1997). The gp120/gp41 heterodimers assemble in trimers to constitute a "spike" (Figure 1), the functional form of the viral envelope.

Viral proteins: accessory proteins

The viral genome codes also for six additional proteins with regulatory and auxiliary functions. The RNA splicing regulator (Rev) and the transactivator protein (Tat) function as regulatory

proteins of viral replication and cell metabolism. The viral protein unique (Vpu), the viral infectivity factor (Vif), the viral protein r (Vpr), and the negative regulatory factor (Nef), have critical roles in enhancing viral pathogenesis, modulating various parameters as cellular response to infection, counteracting restriction factors and ultimately favoring viral replication.

Rev is a regulatory protein with a role in the processing of the newly transcribed vRNA during infection. It increases the amount of partial or completely unspliced vRNA (Felber et al., 1989) in the cytoplasm. Indeed, Rev, by binding to the Rev response element (RRE) (Kalland et al., 1994; Meyer and Malim, 1994) present on the nascent vRNA, exports the partially spliced or unspliced vRNA from the nucleus. The unspliced vRNA will then either be translated into the Gag and Gag-Pol precursors or incorporated into nascent viral particles as full genomic RNA. Rev is able to perform this function thanks to the concomitant presence of nuclear localization and nuclear export signals.

The trans-activator of transcription (Tat) is a regulatory protein that enhances transcription of HIV-1 provirus by binding to TAR (Rice, 2017). In addition, Tat is also modulating host cell gene expression, by favoring immune suppression, apoptosis and oxidative stress (Badou et al., 2000; Chen et al., 2002; El-Amine et al., 2018), which are processes involved in the progression of HIV-1 infection and its symptoms.

Nef (negative factor) is a myristoylated protein associated with the cytoplasmatic membrane, abundantly expressed during the early phases of the viral infection. It is essential to maintain prominent levels of viral load and to accelerate CD4+ T cells depletion (Kestier et al., 1991; Arien and Verhasselt, 2008; Arhel and Kirchhoff, 2009). To reach this goal, Nef is involved in various aspects of the infectious process, as inducing changes in cellular trafficking complexes and in receptors surface expression (Arien and Verhasselt, 2008; Kirchhoff et al., 2008). Nef is also responsible for affecting survival functions of bystander cells and hamper communication between antigen-presenting cells and T cells (Thoulouze et al., 2006; Arhel and Kirchhoff, 2009; Lenassi et al., 2010).

Vif (virion infectivity factor) is essential for viral replication *in vivo* where it counteracts the effect of the viral restriction factor apoliprotein B mRNA-editing enzyme catalytic polypeptide-like (APOBEC) (Malim and Emerman, 2008; Chiu and Greene, 2009; Malim, 2009; Planelles and Benichou, 2009), which restriction mechanism is discussed more in details below (see *HIV-1 origin and evolution – Adaptation to overcome the cross-species barrier*).

Vpr (viral protein R) is incorporated in viral particles and enhances viral replication by altering both viral and cellular mechanisms. Vpr was found to enhance infection in macrophages (Cohen et al., 1990; Hattori et al., 1990; Balliet et al., 1994; Connor et al., 1995) and CD4+ T cells (Iijima et al., 2004). The protein was found to be part of the pre-integration complex (PIC) and it was therefore believed to facilitate its nuclear import (Popov et al., 1998; Vodicka et al., 1998). However, our new understanding of the uncoating and nuclear import steps (discussed below) questions whether Vpr role as a part of the PIC complex could be related to some other aspects. Vpr induces G2 cell-cycle arrest (He et al., 1995). The G2 phase block enhances viral transcription, since transcription is more active during this cell phase (le Rouzic and Benichou, 2005).

Vpu (viral protein U) is a membrane protein with two main functions. First, it downregulates the expression of CD4 at the surface of the infected cells, by binding to it and recruiting an ubiquitin ligase complex to promote its proteasomal degradation (Willey et al., 1992; van Damme and Guatelli, 2008). Reducing CD4 expression in the infected cell prevents the interaction of viral envelope glycoproteins with it in the Golgi apparatus, allowing a higher number of envelope glycoproteins to be exposed on the nascent viral particles. Second, it facilitates viral particles release by counteracting the cellular restriction factor tetherin (Neil et al., 2007, 2008; van Damme et al., 2008). This role is discussed more in detail below (see *HIV-1 origin and evolution – Adaptation to overcome the cross-species barrier*).

HIV-1 life cycle

The HIV-1 life cycle can be roughly separated into early and late phases (Figure 3). The early phase starts with the viral entry and ends with the integration of the product of reverse transcription to generate a provirus. Transcription and translation from the provirus begin the late phase that ends with the maturation of the budding viral particle.

Early phase

Recognition of the viral receptors by the gp120 marks the beginning of the infection. When the gp120 recognizes the cellular receptor CD4 (Klatzmann et al., 1984), a conformational change occurs, resulting in the generation of a surface on the gp120 that allows its binding to the C-C chemokine receptor type 5 (CCR5) or C-X-C chemokine receptor type 4 (CXCR4) that act as co-receptors for viral entry (Raja et al., 2003). This interaction triggers a further conformational change, this time of both gp120 and gp41 subunits, during which gp41 inserts

its fusion peptide into the cell membrane. The last major conformational change brings the cell and viral membranes into proximity, ultimately leading to membrane fusion (Kielian, 2006). CD4+ cells, therefore, are the target of HIV-1 infections. These cells are helper and regulatory T cells, CD4+ T cells, macrophages, monocytes, and dendritic cells (Weiss, 2002). The viral tropism is determined by which co-receptor is used during viral entry. It can be M-tropic, when the co-receptor used is CCR5, T-tropic, when it is CXCR4, or dual-tropic, when both can be used (Alkhatib et al., 1996; Feng et al., 1996; Zhang et al., 1996). Viral entry ends with the release of the viral capsid core into the cytoplasm of the cell.

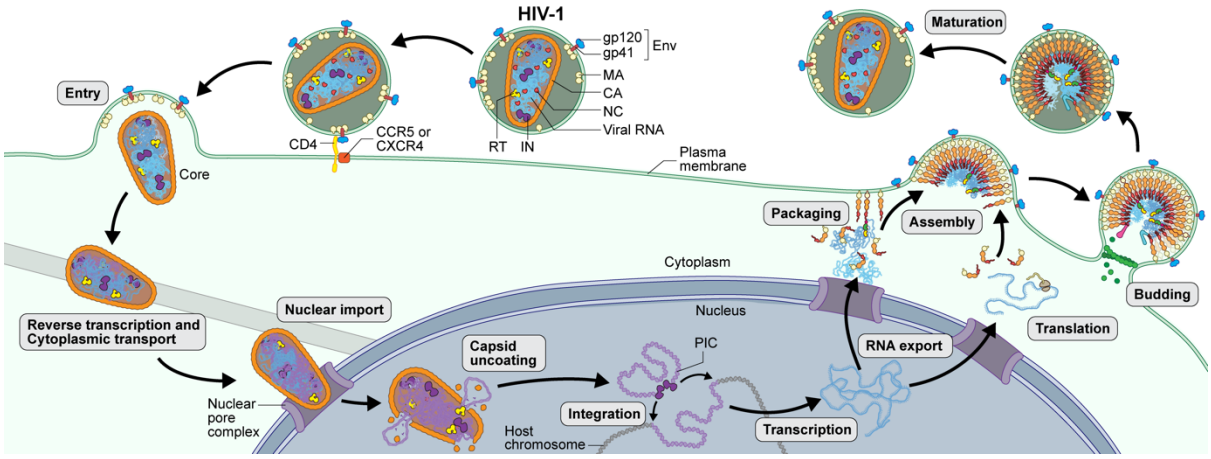


Figure 3. The HIV-1 life cycle. The early phase of the HIV-1 life cycle starts with viral entry, thanks to the binding of the viral envelope proteins to the cellular receptor CD4 and co-receptor CCR5 or CXCR4. After membrane fusion, the capsid core is released into the cytoplasm and starts its journey towards the nucleus. Once imported in the nucleus and reverse transcription is complete, the formation of the pre-integration complex (PIC) allows integration into the host genome to happen. The late phase starts with transcription of the provirus, export of the mRNAs, and translation of the viral proteins. The assembly of the new viral particle at the cellular membrane leads to its budding from the infected cell. The released viral particle will then go through a maturation process, at the end of which a mature viral particle, ready to start a new replication cycle, is formed. Source of the image: https://scienceofhiv.org/wp/life-cycle/.

Contrary to what previously believed, recent studies support a model where the intact capsid core enters the nucleus, and, only after nuclear entry, the uncoating step takes place, near the sites of integration. In the same way, recent works highlighted how reverse transcription is most likely terminated only once into the nucleus (Burdick et al., 2020; Dharan et al., 2020; Francis et al., 2020a; Selyutina et al., 2020). The current most supported model describes that the capsid core, once released into the cytoplasm, starts to move toward the nucleus exploiting the microtubules system. Once reached the nuclear pore complex, an active nuclear import process starts. Indeed, the CA lattice interacts with several proteins that are either part of the nuclear pore complex or are located close to it. Among them there are the nucleoporins 153 (Nup153) and 358 (Nup358), cyclophilin A (CypA), transportin 3 (TNPO3), and the cleavage and splicing factor 6 (CPSF6). All these host proteins are believed to support viral infection and to compose the main pathway used by the core to get to the nucleus and

to allow the integration of the pre-proviral DNA into actively transcribed regions. It is still not clear whether partial uncoating or remodeling of the capsid core is necessary in order for this to happen, but it is more and more believed that an intact or almost intact capsid is transported into the nucleoplasm, as said before.

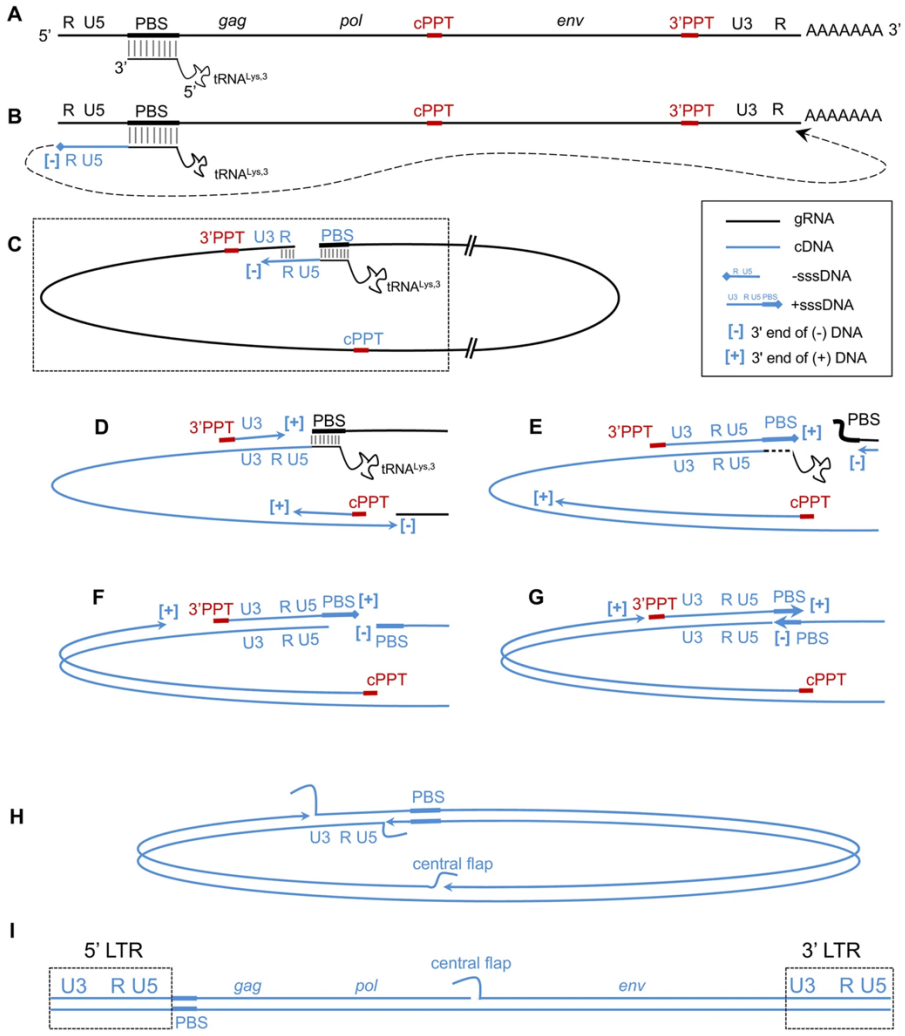


Figure 4. Reverse transcription process of HIV-1. A Schematic representation of the viral genome (vRNA, in black). The R, U5 and U3 regions are shown, as well as the main genes. The central PPT (cPPT) and the 3'PPT are shown in red. The PBS and the tRNALys3 binding to it are depicted. **B** Synthesis of the viral DNA (vDNA, in blue) starts with the (-)sDNA synthesis which shortly generates the -sssDNA, followed by the first strand transfer (dotted line). **C** After strand transfer thanks to the annealing of the identical R sequences, the (-)sDNA synthesis is continued (the square shows the part that will be represented in panels D-G). **D** The (-)sDNA synthesis pursues towards interior regions of the genome. The cPPT and 3'PPT are resistant to RNase H cleavage and prime the synthesis of the (+)sDNA. **E** The (+)sDNA synthesis, that started at the 3'PPT, reaches the tRNALys3 and copies 18 nt, removing the rest. This generates the +sssDNA. **F** The PBS sequence is synthetizes also in the (-)sDNA, allowing the second strand transfer, as shown in panel **G**, to happen. **H** Synthesis of the (+)sDNA is completed, leading to the formation of the double stranded LTRs and the central flap. **I** Schematic representation of completed double-stranded vDNA, flanked at each end by the LTRs. Image from Cappy et al., 2017.

During the journey of the core toward the nucleus, reverse transcription is started, triggered by the higher concentration of nucleotides present in the cell environment. Reverse transcription begins using as primer the $tRNA^{Lys3}$, packaged in the core, annealed to the primer binding site (PBS) sequence adjacent to the 3' end of the 5' UTR of the viral genome

(Figure 4). This binding is facilitated and stabilized by the NC (Auxilien et al., 1999; Tisne, 2005). RT recognizes this RNA/RNA primer template complex and starts to synthetize the minus-strand DNA [(-)sDNA] from the 3' end of the $tRNA^{Lys3}$. Meanwhile, the RNase H domain degrades the template RNA in the RNA/DNA hybrid created by the polymerization of the DNA. This process pauses (minus strand strong stop DNA, -sssDNA) when the RT reaches the 5' of the gRNA, after copying the U5 and R regions (Figure 4). The presence of these complementary regions on the (-)sDNA and the dissociation of the RNA fragments cleaved by the RNase H, allow the transfer of the nascent DNA to the R region present at the 3' extremity of the gRNA (Figure 4). Here, RT can re-start to reverse transcribe the rest of the genome, until it reaches the PBS sequence (the same where reverse transcription started). During the completion of the synthesis of the (-)sDNA, the RNA template is completely degraded, excepts for the highly conserved polypurine tracts 3'PPT and cPPT, which are RNase H resistant. These two sites are then used as starting points to synthetize the plus-strand DNA [(+)sDNA] (Figure 4). The synthesis that started at the 3'PPT, as it happened for the minus strand before, pauses (plus strand strong stop DNA, +sssDNA) at the 5' extremity of the (-)sDNA, after copying 18 nt at the 3' end of the $tRNA^{Lys3}$, which is then cleaved. This pause is followed by a second strand transfer event, the plus-strand transfer, which is started because of the complementarity between the (-)sDNA PBS sequence and the (+)sDNA PBS (the copied 18nt at the 3' end of the $tRNA^{Lys3}$) and propagated thanks to the strand displacement due to the (+)sDNA synthesis started at the cPPT (Figure 4). The plus-strand synthesis can then be finished, and reverse transcription end with the formation of a double-stranded vDNA flanked by two LTRs. When both reverse transcription and uncoating steps are completed, the vDNA forms the pre-integration complex (PIC) together with, at least, a tetramer of IN, and several other proteins. Then, IN catalyzes two subsequent reactions: the 3' processing and the strand transfer. Both of them are the same enzymatic reaction, which is a magnesium dependent S_N2 transesterification reaction, although they differ for their substrates. In the 3' processing one substrate is a water molecule through which IN hydrolyzes vDNA ends adjacent to conserved CA-3' dinucleotides, creating two reactive CA-3'OH ends (Brown et al., 1989; Roth et al., 1989; David Pauza, 1990; Engelman et al., 1991). These ends are used to attack, in a staggered way on both DNA strands, the 5'-phosphate groups in a major groove of the host chromosomal DNA, leading to transesterification and covalent joining of vDNA to the chromosome (Fujiwara and Mizuuchi, 1988; Brown et al., 1989; Engelman et al., 1991). This results in an integration intermediate that needs to be repaired by the host cell enzymes, yielding the integrated vDNA flanked by the duplication of the chromosome cut sequence (Maertens et al., 2022).

Late phase

Once integrated, the vDNA is called provirus and with its transcription starts the late phase of the viral cycle. The majority of integration events lead to active transcription and the generation of new viral particles (Eisele and Siliciano, 2012). However, another fate of the provirus is to be transcriptionally silent, starting a latent infection. It was suggested that the latent reservoir can be originated from infections of active CD4+ T cells that than transition to their resting state (Finzi et al., 1999; Pace et al., 2012; Shan et al., 2017), but the mechanisms behind the establishment of latency are not clear. When the provirus is transcribed it behaves as a regular gene. The U3 region at the 5' of the provirus contains the promoter, as well as other regions which can affect transcription. General cellular transcription factors bind to one of these regions and recruit the RNA polymerase II at the promoter. Other regions are recognized by NF-κB and AP-1, which activate transcription. Another important transcriptional regulatory region is the Tat responsive element (TAR). Tat is one of the first viral proteins to be translated. Once it accumulates, it binds to the TAR stem-loop present at the 5’ extremity of the 5'UTR of newly nascent mRNA to be translated (Dingwall et al., 1990). This is a crucial step as it allows to end the promoter proximal pause of the RNA polymerase II more efficiently. Indeed, the binding of Tat to TAR leads to the recruitment of the positive transcriptional elongation factor (P-TEFb) (Zhu et al., 1997), which phosphorylates the C-ter of the RNA polymerase II inducing its transition into the elongation phase (Cujec et al., 1997; Taube et al., 1999). The transcription product has different fates: it can either be unspliced, or partially or completely spliced. The unspliced and partially spliced products are exported from the nucleus by Rev, which binds to the RRE sequence present on these mRNA. The unspliced mRNA is either incorporated in the new viral particle, constituting the viral genome (vRNA), or is translated into the two different polyproteins precursors: Gag and Gag-Pol. Once the polyprotein precursors acummulate, the CA C-ter domains of different precursor interact with each other and allow the oligomerization of the precursors in the cytoplasm of the cell (Lanman et al., 2003; Eichorst et al., 2018). NC of Gag recognize and bind to the ψ sequence of two copies of the complete and unspliced vRNA (Aldovini and Young, 1990; Jowett et al., 1993; Sundquist et al., 2012). Accumulation to the cell membrane of the Gag and Gag-Pol multimers, leads to the formation of the immature Gag lattice, with MA pointing toward the exterior and, proceeding toward the interior, CA, SP1, NC, SP2, and p6 domain, respectively. The lattice forces a curvature at the plasma membrane and the formation of a spherical particle that ultimately buds from the cell (Rose et al., 2020). To bud HIV-1 exploits the endosomal sorting complexes required for transport (ESCRT), by recruiting Tsg101 and ALIX factors, through the p6 domain of Gag (Garrus et al.,

2001; Martin-Serrano et al., 2001; Strack et al., 2003). During and/or after the newly formed viral particle release, multimerization of Gag-Pol precursor activates the viral protease, which proceeds to an ordered sequence of cuts that cleave the Gag and Gag-Pol precursors into their individual components (Pettit et al., 1994, 2005). The first cut happens between SP1 and NC, allowing the formation of the ribonucleoprotein complex of the vRNA bound to NC. Once released, MA and CA dissociate and spontaneously re-assemble to form respectively the matrix lattice and the capsid core, which contains the two copies of vRNA complexed with the NC, along with the RT and IN proteins. Several host and viral accessory proteins are also found in the mature viral particle, which is now able to start a new viral cycle (Pornillos and Ganser-Pornillos, 2019).

Disease progression and therapy

AIDS is a main threat for human health with more than 39 million people living with this disease in 2020. The main symptom of this disease is a severe immune deficiency with a drop in the amount of lymphocytes T CD4+. While antiretroviral treatment in the recent years has significantly reduced the number of AIDS related deaths, the countries where the higher mortality and morbidity is found are also the one where is more difficult to have access to therapy. HIV-1 is transmitted by sexual, percutaneous, and perinatal routes (Hladik and McElrath, 2008; Cohen et al., 2011). Surprisingly, the efficiency of sexual virus transmission, which represent the 80% of all infections, is poor and it is mostly caused by a single "founder" virus (Keele et al., 2008). However, a single infection can be enough to induce AIDS progression and lead to the death of the infected person, through a rapid and progressive elimination of memory CD4+ helper T cells in lymphoid tissues (Brenchley et al., 2004).

Once the infection is established there are four subsequential phases of disease progression. (1) The eclipse phase, which occurs in the first couple of weeks after transmission. This phase is asymptomatic, but the virus starts to circulate and replicate in the infected individual, although at undetectable levels and without an associated immune response (Coffin and Swanstrom, 2013). (2) The acute phase starts 2 to 4 weeks post-infection and is characterized by a significant increase in the viral load as well as the onset of the immune response. Also, a depletion of CD4+ T cells concomitant with the appearance of the first flu-like symptoms, is observed (Schacker et al., 1996; Veazey et al., 1998; Brenchley and Douek, 2008; Lackner et al., 2012; Coffin and Swanstrom, 2013). (3) Once the viral load is stabilized, starts the chronic phase, which can last 1 up to more than 20 years. This phase can be asymptomatic or with very few symptoms. (4) When the viral load increases again and the CD4+ T cell count

starts to be exceptionally low, the immune system is severely compromised, and the infected individual is exposed to pathogen-related diseases and/or cancer development (Lackner et al., 2012; Coffin and Swanstrom, 2013). This is the final AIDS phase which, if not treated, can lead to the death of the infected individual.

Several drugs anti-HIV-1 have been approved since the beginning of the pandemic. These compounds target some of the main steps of the viral cycle to block the replication, as viral entry, reverse transcription, integration, and the maturation step. The viral entry inhibitors are classified in two categories, those that block the binding to the cellular receptor and co-receptor, and those that interfere with membrane fusion (Kuritzkes, 2009). The reverse transcription inhibitors are grouped in the nucleoside RT inhibitors (NRTI) and the nonnucleoside RT inhibitors (NNRTI). NRTI are chain terminator nucleoside analogues that compete with the natural deoxynucleotides for incorporation by the RT, in the nascent DNA chain. Once incorporated, they abort DNA synthesis (Menéndez-Arias, 2008). The NNRTI, instead, exploit steric hindrance to inhibit polymerization, by binding to the RT, and blocking conformational changes that are required to transition from one conformational state of the RT to another during polymerization (Kohlstaedt et al., 1992). The inhibitors of protease (PI) mimick the natural substrate of the PR and therefore they compete for binding to the active site of the enzyme, preventing viral particle maturation (Sundquist et al., 2012). The inhibitors of integration are discussed more in detail below (see *Integrase – Integrase as a target of antiviral therapy)*, but can also be grouped in two main categories: the strand transfer inhibitors (INSTI) that bind to the catalytic site, and the allosteric IN inhibitors (ALLINI) that block IN dimerization by binding to the surface of dimerization between two monomers of IN. However, all currently available drugs can select resistant variants that have indeed been the main cause of treatment failure and justify the constant search for new drugs.

The most common therapeutic treatment nowadays is the combined antiretroviral therapy (cART), formerly called the highly active antiretroviral treatments (HAART), where the NRTIs, NNRTIs, PIs, and INSTIs drugs are used simultaneously, in various qualitative and quantitative combinations (Cihlar and Fordyce, 2016). INSTIs, in particular, showed to be successful and the most recent guidelines recommend using a second-generation INSTI (dolutegravir or bictegravir) in combination with two NRTIs (Saag et al., 2020). To date, more than 23 HIV-1 drugs combinations have been approved by the Food and Drug Administration (FAD). They are formulated in a single daily pill, and, over the years, they have become more potent and with less adverse effects.

INTEGRASE

Structure

IN is constituted of three structural domains, the N-terminal domain (NTD), the catalytic core domain (CCD), and the C-terminal domain (CTD) (Drelich et al., 1992; Bushman et al., 1993; Vink et al., 1993; Andrake and Skalka, 1996) (Figure 5). Each domain is essential for the catalytic activity of the enzyme, but each also has a specific role in the protein.

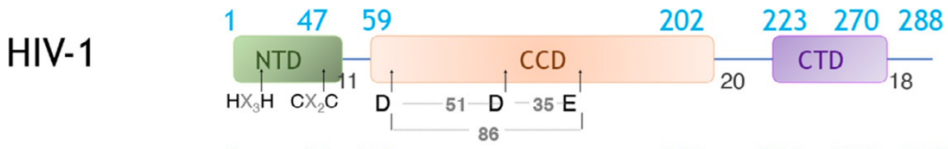


Figure 5. Schematic representation of HIV-1 integrase. The HIV-1 integrase (IN) three domains, NTD (in gren), CCD (in orange), and CTD (in purple) are shown. The first and the last amino acid positions of each domain is shown in blue. Amino acids of the conserved functional domains among retroviruses (HHCC in the NTD, and DDE in the CCD) are shown. Image from Passos et al., 2021.

The NTD is 47 aa long and is folded in a three-helix bundle (Cai 1997). It contains a highly conserved motif, among retroviruses, harboring two histidines (H) and cysteines (C), the $H_{12}H_{16}C_{40}C_{43}$ motif (Figure 5). This motif coordinates a Zn^{2+} and it is important for proper folding of the domain into its three-helix bundle (Cai 1997), IN multimerization, as well as for the catalytic activity (Burke et al., 1992; Engelman and Craigie, 1992; Bushman et al., 1993; Zheng et al., 1996; Lee et al., 1997). Indeed, when mutated, integration is abolished (Zheng et al., 1996).

The CCD spans 143 aa and is folded like an RNase H domain (Dyda et al., 1994). It harbors the catalytic triad, three non-contiguous amino acids ($D_{64}D_{116}E_{152}$) (Figure 5), highly conserved among all retroviral IN and also bacterial transposase (Engelman and Craigie, 1992; Kulkosky et al., 1992; van Gent et al., 1992). The triad coordinates two divalent metal ions (Mg^{2+} in physiological conditions or Mn^{2+} *in vitro*), essential for both enzymatic reactions exerted by IN. Different regions and residues of the CCD are binding the vDNA (Heuer and Brown, 1997; Esposito and Craigie, 1998; Chen et al., 2006), while others are taking contact with the host DNA (Engelman et al., 1994; Harper et al., 2001; Passos et al., 2017). The CCD contains the dimerization and multimerization surfaces, essential for protein functionality (Berthoux et al., 2007; Serrao et al., 2012), as well as the interaction surface with one of the most important host factors for HIV-1 replication, the lens epithelium growth factor (LEDGF/p75) (Cherepanov et al., 2005a, 2005b; Rahman et al., 2007), which role will be further discussed below (see *Integrase – Choice of the integration sites*).

The CTD consist of 76 aa and it is composed of 5 beta-sheet, which are structured in a Src homology 3 (SH3)-like β-barrel fold (Eijkelenboom et al., 1995; Lodi et al., 1995). The CTD is the least conserved domain among the three. It is rich in basic aa, which bind the vDNA and were more recently shown to bind also the vRNA (Lutzke and Plasterk, 1998; Kessl et al., 2016; Passos et al., 2017; Elliott et al., 2020; Rocchi et al., 2022). The CTD is involved in IN multimerization and binding the RT (Engelman et al., 1993; van Gent et al., 1993; Jenkins et al., 1996; Wilkinson et al., 2009; Tekeste et al., 2015; Rocci et al., 2022).

The domains are connected by flexible linkers, with the one connecting the NTD to the CCD being ~10 aa long and the one between the CCD and the CTD being ~20 aa long. It is important to mention that, while the NTD-CCD linker is flexible, the CCD-CTD one is structured in an alpha-helix, reducing the flexibility of this linker (Ballandras-Colas et al., 2017; Passos et al., 2017). Also, a flexible tail of ~18 aa is present at the end of the CTD (Dar et al., 2009; Mohammed et al., 2011).

Integration is mediated by the pre-integration complex (PIC), a large nucleoprotein complex (Bowerman et al., 1989; Farnet and Haseltine, 1990) containing multimers of IN bound to the linear ends of the ~10 Kbp vDNA, the CA (Bedwell and Engelman, 2020), as well as host proteins, including barrier-to-autointegration factor (BAF) (Lee and Craigie, 1998; Lin and Engelman, 2003), high mobility group protein A1 (HMGA1) (Farnet and Bushman, 1997), lens epithelium-derived growth factor (LEDGF/p75) (Llano et al., 2004), and histones (Machida et al., 2020; Winans and Goff, 2020). Within the PIC, IN is found in its active form, a high ordered oligomer, which when complexed with the vDNA forms the intasome (Miller et al., 1997; Chen et al., 1999; Passos et al., 2017, 2020). The number of protomers present in an intasome varies in function of the virus considered and it is dictated by the composition of its CCD-CTD linker (Ballandras-Colas et al., 2016). For example, this linker is about 50 residues long in PFV, allowing the CTDs of the active IN protomers (which are the ones in charge of the catalytic activity) to bind the vDNA. In lentiviruses this linker is about 15-20 residues long and is structured in an alpha-helical conformation (Chen et al., 2000; Ballandras-Colas et al., 2017), creating a structural constraint that precludes the CTDs of the active molecules to bind the vDNA, consequently increasing the numbers of minimum molecules needed for this interaction to occur. The stoichiometry of HIV-1 intasome is still debated, with structures obtained by cryo-EM and negative-stain electron microscopy, containing between 4 and 16 IN molecules (Passos et al., 2017; Cook et al., 2020; Li et al., 2020a). Despite intasomes from different retroviruses vary in their IN composition, they all share the same conserved intasome core (CIC), which coincides with the PFV intasome structure. In this structure, two CCDs are flanked on one side by the vDNA and on the other side by another CCD, which does not bind

the vDNA, with which they dimerize (Figure 6). The NTDs of the two CCDs involved in binding the vDNA are crossed over the acid nucleic molecules, and their CTDs are flanking the vDNA (Figure 6). Therefore, the minimal number of IN monomers in lentiviruses is four (to generate a tetramer) with two molecules involved in the catalytic activities and the other two supporting the binding to the vDNA as well as the inter-molecular interactions. The vDNA complexed with the IN molecules is extensively interacting with all three IN domains.

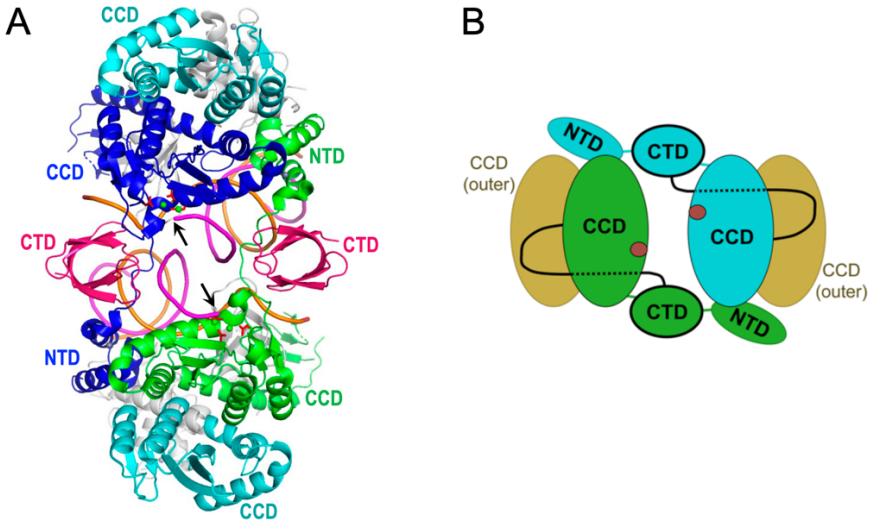


Figure 6. HIV-1 intasome core structure. A The intasome core of HIV-1 is composed by 4 IN monomers. The catalytic CCDs are shown in blue and green, taking contact with the vDNA (in magenta and orange). Their NTDs are crossing over the vDNA (in blue and green). Both their CTDs (in pink) are stabilizing the binding to the vDNA. Each catalytic CCD is dimerizing with another CCD (in light blue) from another IN monomer. PDB accession code: 6PUT. Adapted from Engelman and Kvaratskhelia, 2022. **B** Schematic representation of the intasome core. The two catalytic protomers are shown in green and cyan. The catalytic site localization in the CCDs is shown with a red dot. Each CCD is dimerizing with an outer CCD on the opposite side of the vDNA. Adapted from Lesbats et al., 2016.

Catalytic activity

IN is a polynucleotidyl transferase that catalyzes two sequential magnesium-dependent S_N2 transesterification reactions, the 3' processing and the strand transfer (Engelman et al., 1991), leading to integration, which is an irreversible step establishing a permanent infection of the target cell (Figure 7). In the course of 3' processing a water molecule is used as a nucleophile to hydrolyze vDNA ends adjacent to conserved CA-3' dinucleotides, revealing two reactive 3' CA_{OH} ends (Brown et al., 1989; Roth et al., 1989; David Pauza, 1990; Engelman et al., 1991) (Figure 7). The two divalent metal ions, coordinated by the DDE triad, play the key role in neutralizing the negative charge of the substrate phosphodiester bond and assisting the deprotonation of the attacking nucleophile (Hare et al., 2012). After binding to a major groove of the target DNA, the same mechanism allows the strand transfer step to happen, but, this time, IN uses the reactive 3'OH vDNA ends as nucleophile to cut the chromosomal DNA in a

staggered fashion and simultaneously joining them to the 5’-phosphate groups of the cut double-stranded chromosomal DNA (Fujiwara and Mizuuchi, 1988; Brown et al., 1989; Engelman et al., 1991) (Figure 7). Strand transfer can only happen once 3' processing is completed, since the hydrolysis of the dinucleotide would be sterically incompatible with transesterification with the chromosomal DNA.

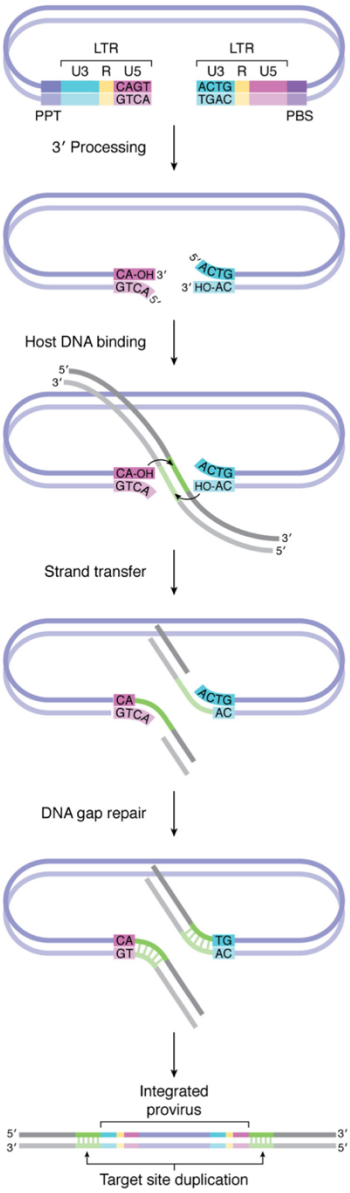


Figure 7. Catalytic activity of the viral integrase. The reverse transcription product, the double-stranded vDNA, is shown in lavender. At each ends it contains a copy of the LTR, composed by U3 (in blue), R (in yellow), and U5 (in pink). The 5'LTR is followed by the PBS sequence (purple box), while the 3'LTR is preceded by the PPT sequence (lavender box). During 3' processing IN hydrolyzes, at each strand 3' end, the GT dinucleotide, adjacent to a conserved CA dinucleotide. The cleaved ends are used to promote vDNA transfer to the host target DNA (in grey with targeted green sequence). Both reactions proceed via S_N2 transesterification at phosphorus atoms and require a pair of divalent metal cations (Mg^{2+} or Mn^{2+}) as cofactors. The gap created by the strand transfer is then repaired by the host machinery, yielding a target site duplication flanking the provirus. Image from Engelman, 2019.

The conformations of the intasome vary depending on the step of the process of integration process considered (Figure 8). The initial stable synaptic complex (SSC) is formed by a multimer of IN and a pair of vDNA ends that, after cleaving, switches to the cleaved synaptic complex (CSC). When this complex binds the host genomic DNA the target capture complex (TCC) is formed. Finally, the covalent joining of viral and chromosomal DNA results in the strand transfer complex (STC) (Hare et al., 2012; Engelman and Cherepanov, 2017) (Figure 8).

The product of the strand transfer is a hemi-integrated form with unjoined 5' vDNA ends that need to be repaired by the host cell enzymes, ending with the integrated vDNA being flanked by a 5-bp long sequence duplication of the cleaved chromosomal sequence (Vincent et al., 1990; Vink et al., 1990; Maertens et al., 2022). In order for the host cell machinery to be able to repair the gap left from the strand transfer step, the STC complex must disassemble. Indeed, it was shown that, when expressed in an ectopic way, IN is ubiquitinated and eliminated in a proteasome-dependent way (Mulder and Muesing, 2000; Devroe et al., 2003; Llano et al., 2004). Similarly, von Hippel-Lindau binding protein 1, a cellular subunit of the prefoldin chaperone, was shown to be essential for HIV-1 replication and to be implicated in proteasome-mediated IN degradation (Mousnier et al., 2007). After STC disassembles from the host DNA, three independent enzymes are necessary to repair the gaps and complete integration by covalently joining the 5' vDNA ends to chromosomal DNA: a DNA polymerase, a 5' flap endonuclease, and a ligase. Cell-based studies highlighted the involvement in this DNA repair step of enzymes of the BER (base excision repair) pathway of oxidative DNA damage (Espeseth et al., 2011; Yoder et al., 2011) and of the nonhomologous end-joining (NHEJ) pathway (Li et al., 2001; Knyazhanskaya et al., 2019). The STC disassembly and the following DNA repair performed by the cell machinery, however, are the steps of the integration process less characterized.

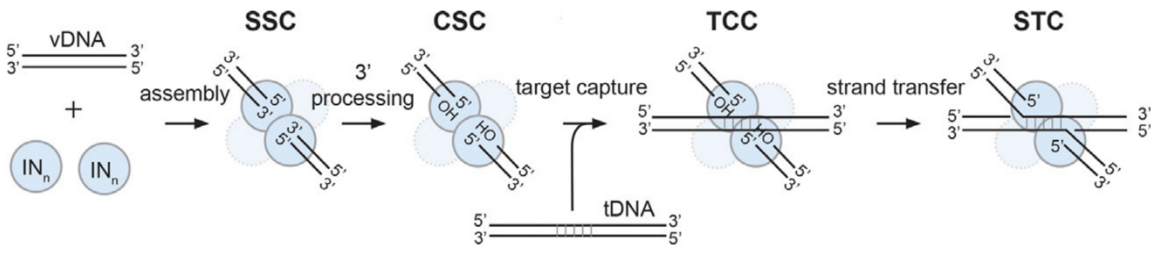


Figure 8. Intasome complex conformations. Schematic representation of the sequential conformations assumed by the intasome throughout the viral cycle. The viral DNA (vDNA) and the required amount of IN monomers assembly to form the stable synaptic complex (SSC). Then IN cleaves two nucleotides from the 3' ends of vDNA, forming the cleaved synaptic complex (CSC). With the capture of the target DNA (tDNA) the complex forms the target capture complex (TCC). Finally, IN catalyzes the strand transfer to form the post-catalytic strand transfer complex (STC) where the vDNA and the tDNA are still bound to the IN. Image from Passos et al., 2021.

Unintegrated forms of vDNA

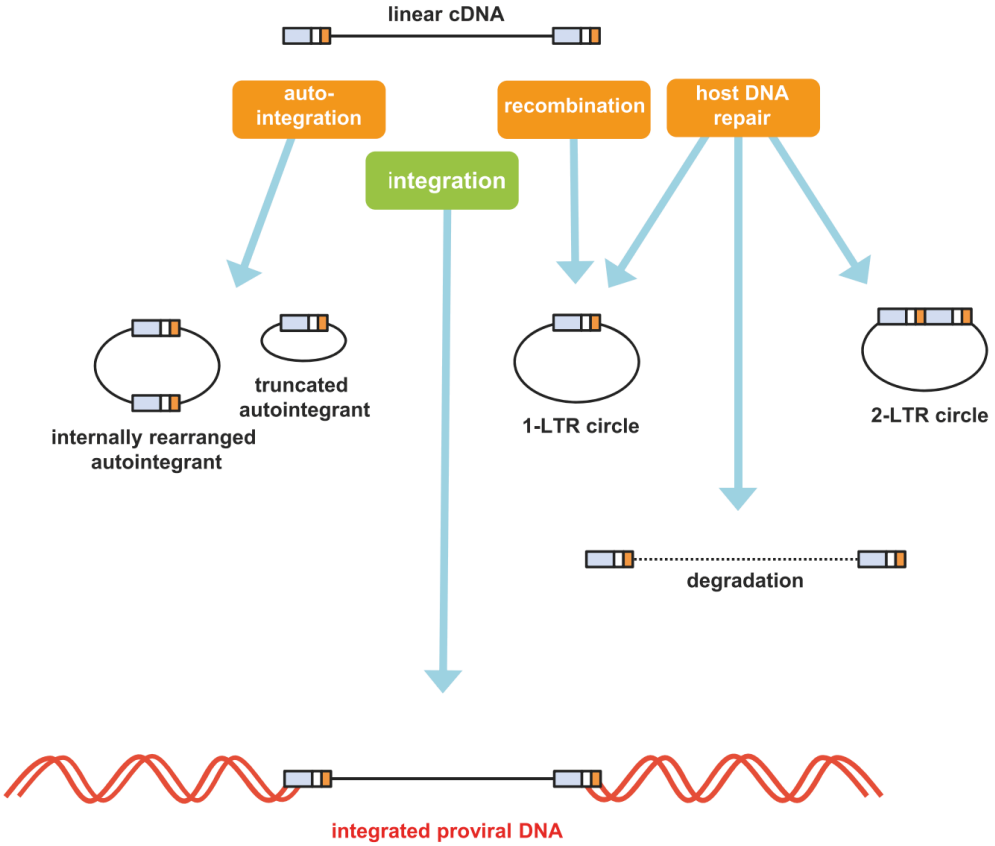


Figure 9. Unintegrated forms of vDNA. The linear vDNA, the product of reverse transcription, is susceptible to face different fates other than integration (green box) into the host genome. The final product of each path (specify in the orange boxes) is shown. Image from Sloan and Wainberg, 2011.

Various forms of non-integrated vDNA are also found in the infected cells. They can result either from events of autointegration or by the host cell DNA repair machinery. The non-integrated forms can be linear (the unintegrated product of reverse transcription) or circular (Figure 9). Circularization by the non-homologous end joining (NHEJ) forms the 2LTR circles (2LTRc) (Miller et al., 1995; Li et al., 2001). Another circular form, with only one LTR, 1LTR circles (1LTRc) can result from defective reverse transcription (Miller et al., 1995), auto-integration or homologous recombination of 2LTRc (Cara and Klotman, 2006). Both circular forms can alternatively be formed from vDNA that was or not processed, leading in the case of 2LTRc to the obtention of a "perfect junction", when the vDNA is not processed, or "imperfect junction", when it is, instead, processed. 2LTRc, due to their unique LTR-LTR junction and the fact that they are present exclusively in the nucleus where the NHEJ machinery is located, are of particular interest for research purposes as good marker for the nuclear import of the RTC/PIC complex. 2LTRc are also used as a marker of integration default (Engelman et al., 1995, 1997; Wiskerchen and Muesing, 1995; Leavitt et al., 1996; Lu et al., 2005; Johnson et al., 2013). In fact, although representing a small percentage of the

total vDNA forms, they are generally considered to be inversely proportional to the amount of proviruses since a hypothetical default in the integration step would leave more substrate material for the formation of 2LTRc.

The physiological role of the unintegrated DNA is still controversial. For long time these molecules were considered to be dead-end products, then it was shown that unintegrated forms of vDNA can be transcriptionally active (Brussel and Sonigo, 2004; Wu, 2004; Sloan et al., 2010; Chan et al., 2016; Meltzer et al., 2018), therefore having a potential role in viral replication. However, 2LTRc were found to be quickly silenced in the nucleus by host proteins (Zhu et al., 2018; Geis and Goff, 2019; Machida et al., 2020; Dupont et al., 2021; Geis et al., 2022). Interestingly, it was also highlighted how 2LTRc can constitute a viral genome reservoir for integration (Thierry et al., 2015; Richetta et al., 2019).

Choice of the integration sites

The genomic region where the provirus is integrated impacts viral replication, by inducing active viral expression or transcriptional silencing and, therefore, (Jordan et al., 2001; Maldarelli et al., 2014; Anderson and Maldarelli, 2018). While IN is highly specific for the interaction with its vDNA, it does not have the same selectivity for the host DNA. However, integration sites are not randomly distributed among the host genome, but they are preferentially located in genomic regions with high gene density and histone modifications associated with active chromatin (Lusic and Siliciano, 2017). To make its way across the nuclear landscape and specifically target active regions of the host DNA, HIV-1 exploits numerous host factors as the lens epithelium-derived growth factor (LEDGF/p75) and the cleavage and polyadenylation specificity factor 6 (CPSF6) with which it interacts via IN and CA, respectively (Engelman and Singh, 2018).

LEDGF/p75 is a ubiquitous chromosome associated transcriptional co-factor that belongs to the hepatoma-derived growth factor-related protein (HRP) family (Nishizawa et al., 2001). It is composed of two globular domains, a N-ter PWWP chromatin reader domain (Izumoto et al., 1997; Qin and Min, 2014), and a downstream PHAT domain composed of two helix-hairpin-helix motifs that is also the integrase binding domain (IBD) (Cherepanov et al., 2004) (Figure 10). The PWWP belongs to the Tudor family, and it has a preference for histone H3 tri-methylated on K36 (H3K36me3), an epigenetic mark associated with transcription elongation (Pradeepa et al., 2012; Eidahl et al., 2013; van Nuland et al., 2013; Wang et al., 2020). The IBD binds the dimerization surface of two IN. More in details, its residues I365 and D366 bind to a pocket in the CCD dimerization interface (Cherepanov et al., 2005a, 2005b).

Other IBD residues, K401, K402 and N405 are taking contact with IN NTD (Hare et al., 2009b), meaning that LEDGF/p75 is taking contact with the CCD and the NTD, the two domains involved in multimerization and intasome assembly. Hence, it is not surprising that adding LEDGF/p75 to IN *in vitro* increases both dimer stabilizaty and catalytic activity (Cherepanov et al., 2003; Turlure et al., 2006; Hayouka et al., 2007; McKee et al., 2008; Hare et al., 2009a, 2009b; Tsiang et al., 2011). While the IBD is in charge of the interaction with the IN, it is the PWWP domain that direct the IN towards the integration sites, indeed, by swapping it with the chromatin-binding domain of another factor, it was possible to redirect integration sites towards the regions bound by that factor (Ferris et al., 2010; Gijsbers et al., 2010; Silvers et al., 2010).

Cellular depletion of LEDGF/p75 causes a reduction in integration levels (Emiliani et al., 2005; Llano et al., 2006; Vandekerckhove et al., 2006; Marshall et al., 2007; Shun et al., 2007; Schrijvers et al., 2012; Wang et al., 2012), as well as a decrease of proviruses in the mid-regions of gene bodies (Shun et al., 2007; Singh et al., 2015; Sowd et al., 2016) with a concomitant increase in transcription start sites (Ciuffi et al., 2005; Marshall et al., 2007; Shun et al., 2007), and GC-rich regions (Ciuffi et al., 2005). LEDGF/p75 interacts with numerous mRNA splicing factors (Pradeepa et al., 2012; Singh et al., 2015) and is able to overcome the transcription block created by nucleosomes *in vitro* (LeRoy et al., 2019), suggesting that integration targeting might involve the cellular mRNA splicing and/or transcriptional machineries. Accordingly, a correlation between integration sites and genes with numerous introns was found (Singh et al., 2015, 2022; Sowd et al., 2016).

LEDGF/p75 also modulates the IN-intrinsic chromatin-binding property to H4 histone tail (Lapaillerie et al., 2021). Indeed, for integration to occur IN has to interact with the nucleosome and this interaction is happening between the IN-CTD and the H4 histone tail (Benleulmi et al., 2017; Mauro et al., 2019). A recent work showed how this interaction is enhanced and redirected towards LEDGF/p75-enriched sites in presence of LEDGF/p75 (Lapaillerie et al., 2021).

Interestingly, when LEDGF/p75 was depleted, integration still occurred preferentially in gene-dense regions with a higher frequency than expected on a random basis (Marshall et al., 2007; Shun et al., 2007; Schrijvers et al., 2012; Singh et al., 2015), suggesting that other proteins, either binding IN or not, might play a role in it. The hepatoma-derived growth factor like 2 (HDGFL2) is the only HRP family member, other than LEDGF/p75, that possesses a functional IBD (Figure 10). It binds to IN (Cherepanov et al., 2004) and play a minor role in directing integration into highly transcribed genes (Schrijvers et al., 2012; Wang et al., 2012).

A combined depletion LEDGF/p75 and HDGFL2 further decreased integration in transcriptional units, compared to LEDGF/p75 depletion alone, but the preference was not completely lost (Wang et al., 2012).

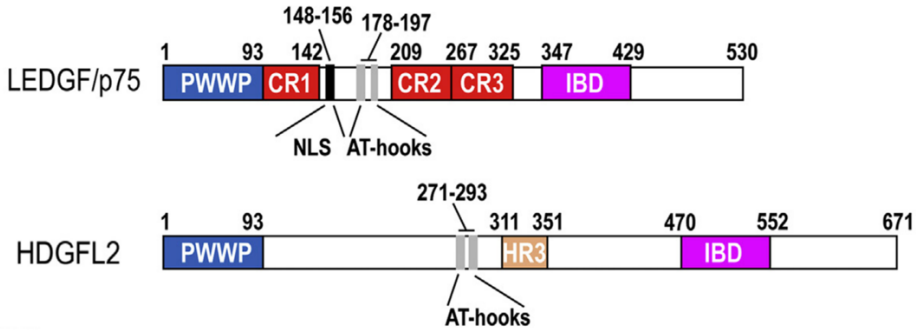


Figure 10. LEDGF/p75 and HDGFL2 schematic representation. LEDGF/p75 and HDGFL2 are shown in their domain organization with numbers indicating domain boundaries and interdomain lengths. The IBD domain present in both proteins is the one taking contact with HIV-1 IN. Image from Engelman and Maertens, 2018.

The choice of the integration sites does not concern only the characteristic of the target regions (e.g., intragenic or intergenic), but also their localization inside the chromatin architecture of the nucleus. Nuclear import of the RTC/PIC is a key step in the infectious cycle of lentiviruses and it is the entry point for the viral genomic material into the nucleus. The ability of lentiviruses to infect non-dividing cells is mediated by CA and its role in RTC/PIC nuclear import (Lewinski et al., 2006). Several host factors participate to this process by interacting with CA, as Nup358 (Schaller et al., 2011), Nup153 (Buffone et al., 2018; Bejarano et al., 2019), CPSF6 (Lee et al., 2012; Price et al., 2012, 2014; Bhattacharya et al., 2014) and cyclophilin A (CypA) (Franke et al., 1994; Gamble et al., 1996). HIV-1 integration occurs preferentially in peripheral chromatin regions of the nucleus, close to nuclear pores. Indeed, depletion of Nup153, a component of the nuclear pore complex, shifted the integration sites to more central nuclear regions (Koh et al., 2013; Marini et al., 2015), suggesting that nuclear entry and integration are two closely related steps of the viral cycle. Even before entering the nuclear pore complex, the interaction with CypA ensures that a nuclear import pathway involving Nup358 and Nup153 is used (Schaller et al., 2011). Once on the nuclear side, however, it is mostly CPSF6 that takes the relay.

CPSF6 is an SR-like protein with a N-ter RNA recognition motif (RRM), a central Pro-rich domain (PRD), and a C-ter RS-like domain (RSLD) (Rüegsegger et al., 1998; Dettwiler et al., 2004). It binds the CA through its PRD domain (Lee et al., 2012; Price et al., 2012), which has more affinity for CA hexamers rather than CA monomers (Bhattacharya et al., 2014; Price et al., 2014). CPSF6 is part of the cleavage factor I mammalian (CFIm) complex, which is in charge of regulating polyadenylation in the 3'UTR of mRNAs (Rüegsegger et al., 1996; Gruber

et al., 2012; Martin et al., 2012). Nevertheless, it was shown that its role as an HIV-1 co-factor is independent from the other subunits of the complex (Rasheedi et al., 2016). CPSF6 can be found in both cytoplasm and nucleus, and, in the latter, is found preferentially in nuclear paraspeckles (Dettwiler et al., 2004). CPSF6 binds to the same CA region of Nup153 (Price et al., 2014), and it is thought to compete for CA binding and its release from the nuclear pore complex (Bejarano et al., 2019). When CA mutants (N74D, A77V) that cannot bind CPSF6 were evaluated, an accumulation of proviruses in lamina-associated domains (LADs) with a concomitant reduction in speckles-associated regions (SPADs) was observed (Chin et al., 2015; Achuthan et al., 2018; Burdick et al., 2020; Francis et al., 2020b; Li et al., 2020b). LADs are gene-sparse heterochromatin regions located at the periphery of the nucleus, while SPADs are genomic regions associated to nuclear speckles (Chen et al., 2018; Chen and Belmont, 2019), which are nuclear domains located in interchromatin regions enriched in pre-mRNA splicing factors, as CPSF6, and transcriptional factors (Spector and Lamond, 2011; Galganski et al., 2017). These domains are surrounded by gene-enriched and transcriptionally active chromatin regions, as SPADs, which explain them being the preferential target of HIV-1 for integration.

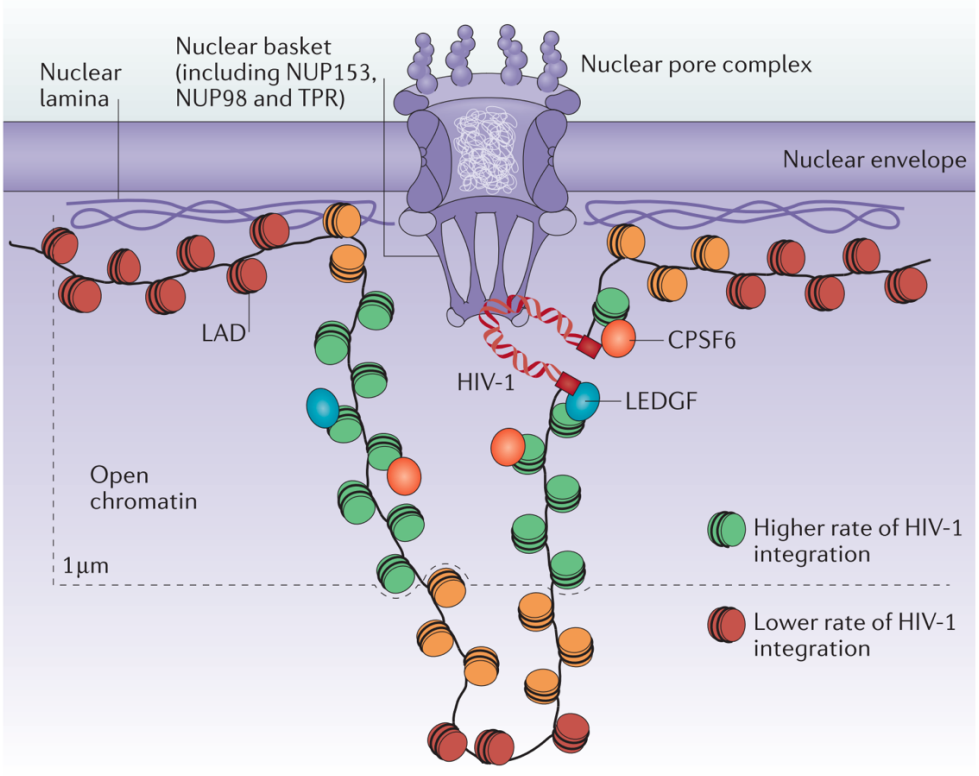


Figure 11. Nuclear landscape of HIV-1 integration. Integration of HIV-1 is happening preferential in open and actively transcribed chromatin. For this HIV-1 is guided through the nuclear landscape mostly by two cellular cofactors: CPSF6 and LEDGF/p75 (LEDGF). CPSF6 binds to the capsid (CA) and guides the pre-integration complex (PIC, in red) towards more internal region, away from the lamina-associated domain (LAD). Then, LEDGF bind to the HIV-1 and tether the PIC to active chromatin. Image from Lusic and Siliciano, 2017.

The current model for the choice of the integration sites is supporting a two-phase mechanism (Bedwell and Engelman, 2020; Singh et al., 2022) (Figure 11). First, CPSF6 allows the release of the RTC/PIC complex from the nuclear pore complex, by binding CA, and then leads it beyond nuclear periphery, towards inner regions of the nucleus and, in particular into SPADs. Then, via LEDGF/p75 binding the viral IN, integration is happening preferentially into gene bodies, under a potential influence of the cellular mRNA splicing and/or transcriptional elongation machineries.

Non-catalytic activities

IN is a pleiotropic protein and, when mutated, can affect several steps of the viral cycle. IN mutations are separated in two classes, class I and class II, based on the different phenotypes they can give. Class I mutations are strictly connected to the inability of these mutants to integrate, leading to an accumulation of non-integrated forms of vDNA, and they usually correlate with amino acidic substitutions in the catalytic triad, or with substitutions proximal to the catalytic site. Class II mutations, instead, affect one or more steps of the viral cycle while retaining, at least partially, the catalytic activity of IN (Engelman, 1999). The most common phenotype observed in these mutants is a reduction in the reverse transcription products (RTPs). However, the variety of phenotypes class II mutations can cause and the different mechanisms behind them, recently led to a new subclassifications of them into classes IIa, IIb, and IIc (Engelman and Kvaratskhelia, 2022), that will be discussed later.

One of the most relevant non-catalytic activities of IN is during viral particle morphogenesis (Engelman et al., 1995; Quillent et al., 1996; Kessl et al., 2016; Elliott et al., 2020). HIV-1 IN interacts with the vRNA in a specific way, by binding preferentially to structural elements, like the TAR stem-loop (Kessl et al., 2016). *In vitro*, IN effectively shows vRNA bridging activity exclusively in its tetrameric form and not in the monomeric or dimeric ones, indicating that IN multimerization is necessary for this to happen. Different works highlighted how the amino acids involved in this binding are all located in the CTD and are specifically K246, K266, R269, and K273 (Johnson et al., 2013; Kessl et al., 2016; Elliott et al., 2020). Disruption of this interaction leads to the formation of eccentric particles, where the vRNA is mislocated outside of the capsid core (Kessl 2016, Elliot 2020, Jurado 2013). These particles are non-infectious since the vRNA, as well as viral proteins, not being protected from the capsid core, are quickly eliminated once in the target cell, causing defects in reverse transcription and all the subsequent steps (Fontana et al., 2015; Madison et al., 2017).

It is this non-catalytic activity of IN that led to further diversification of class II mutations (Engelman and Kvaratskhelia, 2022). The disruption of the IN-vRNA interaction, indeed, can occur via three mechanisms, which also define the three subclasses. Class IIa mutations are characterized by a specific and direct disruption of IN-vRNA binding, without affecting other functional aspects of the protein or protein multimerization. Substitutions of residues K264, K266, R269 or K273 belong to this subclass. Class IIb mutations are those that disrupt the multimerization of IN and, consequently, prevent binding to the vRNA, while affecting also other IN functions. Class IIc mutants are characterized by a poor efficiency in viral release and processing, resulting in an amount of IN that is insufficient to correctly incorporate vRNA into virion cores. Overall, these observations suggest that the defaults in early steps of the viral cycle can be collectively caused by malfunctions in viral morphogenesis when IN-vRNA interaction is impaired.

However, not all class II mutants have impaired levels of mature viral proteins (Masuda et al., 1995; Leavitt et al., 1996; Wu et al., 1999) and the IN appears to have a more direct role in other steps of the viral cycle, like, for example, reverse transcription. Magnetic resonance spectroscopy identified the aa in the CTD of IN that interact with the RT (Tekeste et al., 2015). This interaction favors reverse transcription by increasing its processivity (Dobard et al., 2007; Tekeste et al., 2015; Rocchi et al., 2022). A study showed how the integrase can influence the uncoating, by regulating the CypA-CA interaction, helping to maintain the correct stability of the core (Briones et al., 2010). IN was suggested to have a role in the nuclear import, by finding different NLS signals on the protein, as well as showing its interaction with different host proteins involved in the nuclear import of the RTC/PIC (Gallay et al., 1997; Bouyac-Bertoia et al., 2001; Zaitseva et al., 2009; Jayappa et al., 2011; de Houwer et al., 2014)

A recent study highlighting the role of IN in regulating the switch from the post-catalytic STC to transcriptionally active proviruses, led to the creation of a third class of IN mutations, class III mutants (Knyazhanskaya et al., 2019). Knyazhanskaya and colleagues showed that the IN double mutant E212A/L213A, which is not able to interact with Ku70 (Anisenko et al., 2017), causes an important delay in the repair of the strand transfer hemi-integrant (Knyazhanskaya et al., 2019). Contrary to class II mutants, these mutations retain partial infectivity. A similar phenotype was observed also with the quadruple mutant K258R/K264R/K266R/K274R, which at early times post-integration, showed to have a transcription defect caused by altered levels histone modifications at the provirus LTR (Winans and Goff, 2020).

Cellular co-factors binding IN

Apart from the interaction with LEDGF/p75, described above, the viral IN was able to turn other host factors into viral co-factor and, by interacting directly with them, ensure a functional integration.

Among them there is SMARCB1, initially known as IN interactor 1 (INI1) (Kalpana et al., 1994), a member of the chromatin remodeling complex SWI/SNF. This host factor is incorporated into the nascent viral particle, and it has been shown to positively affect several steps of the viral cycle, as reverse transcription, integration, transcription, and particle assembly/release (Yung et al., 2001; Ariumi et al., 2006; Sorin et al., 2006; Lesbats et al., 2011; Mathew et al., 2013; la Porte et al., 2016). The mechanism through which this factor acts on reverse transcription, in particular, was elucidated. INI1 mediates the incorporation of Sin3a-associated protein 18 kDa (SAP18) and histone deacetylase 1 (HDAC1) into viral particles, which directly favors reverse transcription. Viral particles carrying inactive HDAC1 (H141A), in fact, were not able to synthetize vDNA (Sorin et al., 2009). INI1 was shown to bind the IN tetramer complexed with LEDGF/p75, stabilizing it and preventing non-specific interactions and auto integration, until capture of the target DNA, which displaces INI1 (Maillot et al., 2013) (Figure 12). Furthermore, the IN-INI1 interaction influences integration activity that is enhanced thanks to the presence of the SWI/SNF complex (Lesbats et al., 2011).

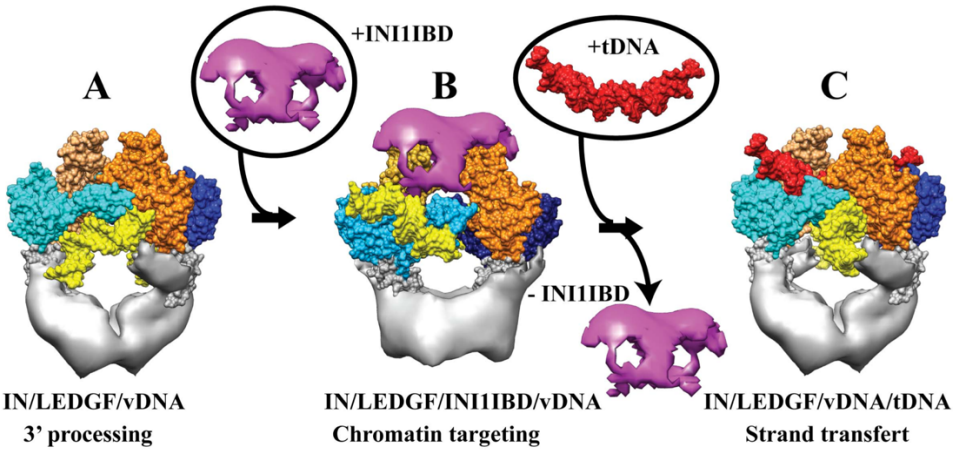


Figure 12. Structure and function of the IN/LEDGF/INI1IBD/vDNA complex. A Two dimers of IN (light and dark, with each monomer shown in blue and gold) are complexed with the vDNA (yellow) and LEDGF/p75 (LEDGF, grey) in a conformation that is compatible with 3' processing. **B** When INI1IBD (positions 174-289, in purple) is added to the complex the conformation is not compatible with 3' processing anymore, but it is compatible with chromatin targeting. **C** The removal of INI1IBD and the capture of the tDNA lead to the formation of the strand transfer complex. Image from Maillot et al., 2013.

As mentioned above, BAF and HMGA1 were identified for being part of the PIC complex. BAF (barrier to autointegration factor) is a host protein stimulating the integration *in vitro* (Chen and Engelman, 1998; Lin and Engelman, 2003). The suggested mechanism is that by

binding the vDNA it alters its structure, facilitating integration (Skoko et al., 2009). HMGA1 (high mobility group chromosomal protein A1) is a DNA-binding protein regulating chromatin structure and transcription. Similarly to BAF, it stimulates integration *in vitro*, by binding to the vDNA and making it more compact and prone to strand transfer (Farnet and Bushman, 1997).

Integrase as a target of antiviral therapy

The essential role of IN in HIV-1 life cycle and the lack of homologous protein in human cells, make this protein the ideal target for anti-viral therapy. However, the scientific community was initially skeptical about the efficacy that drugs against the IN could have. This came from the observation that although the stoichiometry in the viral particle of IN, PR and RT is the same (around 120), while PR and RT are able to catalyze more than 12,000 chemical reactions each, IN is catalyzing only four chemical reactions, posing the problem of how a molecule can efficiently block these reactions with such an excess of enzyme present (Engelman, 2019). What was not yet known at that time, however, was the fact that while 3' processing is happening right after, if not at the same time, of reverse transcription (Miller et al., 1997; Munir et al., 2013), the strand transfer happens only hours or even days later (Cardozo et al., 2017), creating a large window of time for a molecule to block this step.

A class of inhibitors that target specifically this "weak point" of the viral cycle, blocking strand transfer, are the IN strand transfer inhibitors (INSTIs). INSTIs are inhibitors acting specifically on the catalytical site of IN and have the same effects as class I IN mutations, with the only differences that they inhibit specifically the strand transfer step (Hazuda et al., 2000), while class I mutations block both catalytic steps. There are currently five molecules approved by United States Food and Drug Administration (FDA): Raltegravir and Elvitegravir, from first generation INSTIs, and second generations INSTIs Dolutegravir, Bictegravir, and Cabotegravir. These molecules are able to target IN only when it is complexed with the vDNA, competing with it to bind the active site (Hare et al., 2010), and they result to be particularly efficient thanks to their unusually long binding half-life. All INSTIs share the same structure and mechanism. In their core, three electronegative atoms (normally oxygen atoms) chelate the metal cations from the IN catalytic triad (Grobler et al., 2002; Hare et al., 2010), while a halogenated benzene ring, connected to the core via a flexible linker, interacts with the reactive end of the vDNA displacing it from the IN active site (Hare et al., 2010).

The appearance of several resistance mutations (Y143C, G140S/Q148H, N155H) (Cooper et al., 2008) after the use of first-generation INSTIs led to the development of the second-

generation ones. These molecules are also composed of two ligands with the same properties as the first generations drugs but improved on some aspects. First, the flexible linker is longer allowing the halogenated benzene ring to have a stronger interaction with the vDNA. Second, the core was enlarged generating more surface contacts with IN active site. In general, they are more efficient in binding to the intasome and occupying the catalytic site (Hare et al., 2011; DeAnda et al., 2013; Zhao et al., 2016, 2017), as well as causing fewer and less efficient cases of resistance compared to the first-generation compounds (Quashie et al., 2012), although it still occurs (Anstett et al., 2017).

Another class of IN inhibitors has been developed more recently, the allosteric IN inhibitors (ALLINIs). Subclasses of ALLINIs, based on their mechanism of action, are LEDGF-interaction site inhibitors (LEDGINs) (Christ et al., 2010), noncatalytic site IN inhibitors (NCINIs) (Balakrishnan et al., 2013), IN-LEDGF allosteric inhibitors (IN-LAI) (le Rouzic et al., 2013), and multimeric IN inhibitors (MINI) (Sharma et al., 2014). ALLINI molecules have a backbone made by pyridine-like structure harboring a carboxylic and *t*-butoxy group connected by a carbon arm. ALLINIs bind to the dimerization surface, far from the catalytic active site, leading to formation of IN aggregates (Gupta et al., 2014; Sharma et al., 2014; Feng et al., 2016), which are unable to assemble with vDNA (Kessl et al., 2012), consequently blocking the integration step and HIV-1 infection (Christ et al., 2010; Tsiang et al., 2012; Balakrishnan et al., 2013; Desimmie et al., 2013; Jurado et al., 2013; le Rouzic et al., 2013).

However, while ALLINI compounds do have an effect on the choice of the integration site, suggesting that they indeed disrupt IN-LEDGF/p75 binding (Sharma et al., 2014; Feng et al., 2016; Vranckx et al., 2016), it does not seem as the main mechanism of action to block HIV-1 infection, since the main effect of this class of compounds can be observed in the viral particles formation and maturation (Kessl et al., 2012; Desimmie et al., 2013; Jurado et al., 2013; le Rouzic et al., 2013; Sharma et al., 2014; Fontana et al., 2015), similarly to IN class II mutations (Engelman et al., 1995; Desimmie et al., 2013; Jurado et al., 2013; Gupta et al., 2016). In fact, in the presence of ALLINIs, eccentric viral particles are formed with the viral genetic material localized outside the capsid core (Engelman et al., 1995; Balakrishnan et al., 2013; Johnson et al., 2013; Jurado et al., 2013; Fontana et al., 2015), causing its rapid degradation once the particle enters into a cell (Madison et al., 2017). While being powerful compounds, ALLINIs seem to have a lower genetic barrier to resistance compared to INSTIs (Christ et al., 2010, 2012; Tsiang et al., 2012; Fenwick et al., 2014; Sharma et al., 2014).

Nevertheless, the therapeutic approach of combining INSTIs and ALLINIs remain a very powerful tool to counteract HIV-1 infection and both their clinical development is still relevant.

Early steps of the AIDS pandemic

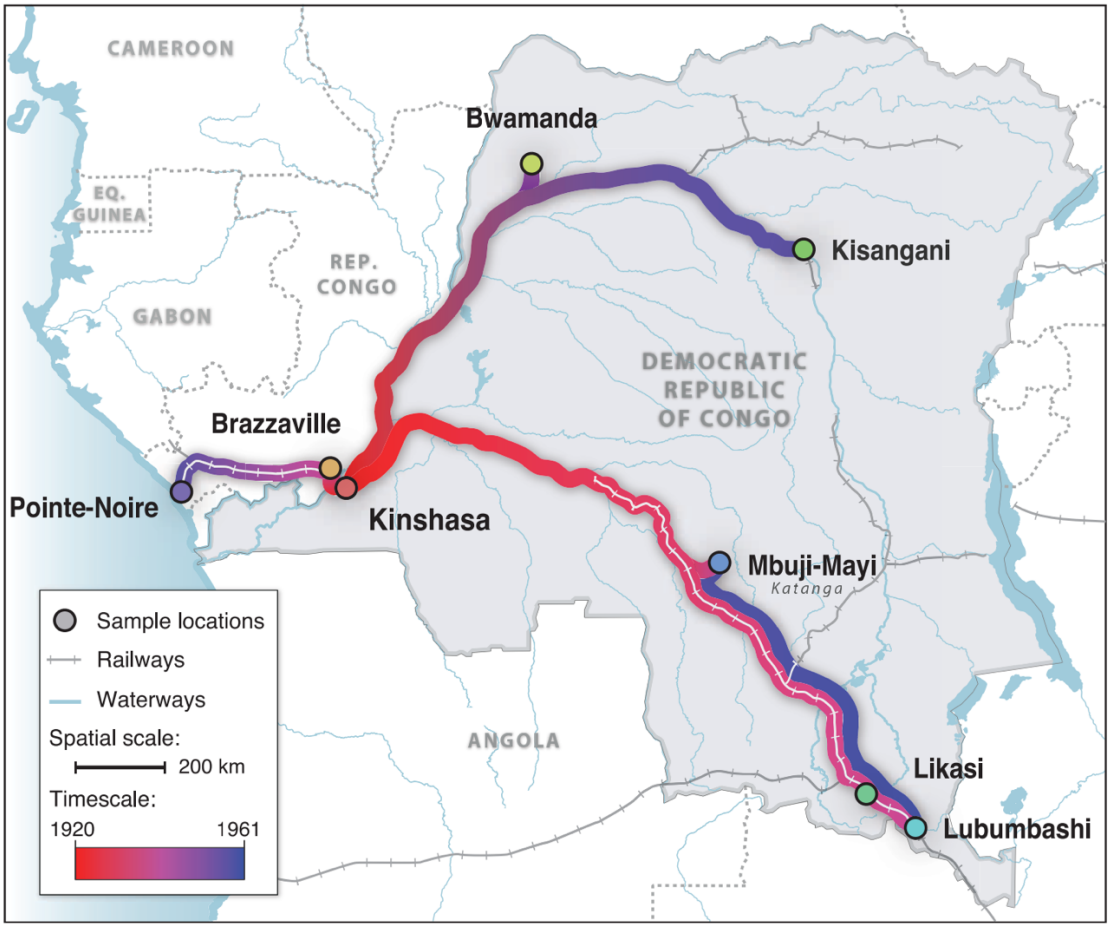


Figure 13. Early spread dynamic of HIV-1 group M. The reconstruction of the first decades of HIV-1 group M spread in the Democratic Republic of Congo (in grey) is shown in the figure. The virus appeared in Kinshasa, to then spread via waterways and railways in the surrounding towns. Image from Faria et al., 2014.

The first reports of AIDS and the identification of HIV-1 and 2 date back to the 80's in the United States (Gottlieb et al., 1981; Barré-Sinoussi et al., 1983; Gallo et al., 1983). Early after, it was found that the disease was well established in the heterosexual population of Cameroon (van de Perre et al., 1984), suggesting that the disease had a more complex and important pandemic history. Indeed, the beginning of the AIDS pandemic dates back way before its discovery and started in the West-Africa region, as it is supported by the fact that the genetic diversity in the Democratic Republic of Congo (DRC) (Vidal et al., 2000; Kalish et al., 2004), known back then as Zaire, in the Republic of Congo (Bikandou et al., 2004; Niama et al., 2006), in Cameroon and in Gabon (Pandrea et al., 2002; Carr et al., 2010), is more complex than in the rest of the world (Vidal et al., 2000; Kalish et al., 2004). If we started to

gain more information about the epidemiology of the pandemic in the late 1980s, very little is known about the initial phase. The oldest HIV-1 sequences, in fact, were found in Kinshasa (the capital of DRC) and are dated 1959-1960 (Zhu et al., 1998; Worobey et al., 2008), while the tMRCA (time to the most common ancestor) of group M is estimated to be in the first decades of the 1900s (Korber et al., 2000; Worobey et al., 2008), meaning that the virus spread for 50-70 years before it was discovered. Further studies confirmed that Kinshasa could be the place where the M pandemic originated (Vangroenweghe, 2001; Faria et al., 2014), a thesis also supported by the fact that Kinshasa is the place where we find the most HIV-1 genetic diversity to date (Vidal et al., 2000; Kalish et al., 2004) (Figure 13).

From Kinshasa the virus had spread initially to the closest towns (Bikandou et al., 2004; Niama et al., 2006; Faria et al., 2014), and then to the other towns located in central Africa thanks to the transportation system present at that time (Quinn, 1994; Gray et al., 2009; Faria et al., 2014) (Figure 13). These movements are thought to have taken place between 1920-1950 and represent the first wave of exponential growth of group M (Faria et al., 2014). After 1960, group M entered a second faster phase of exponential growth (Faria et al., 2014) and it is during this period that the divergence between groups M and O started to emerge. In fact, groups M and O share the same tMRCA (around 1920) (Korber et al., 2000; Lemey et al., 2004; Worobey et al., 2008) and during the first exponential wave of group M, group O showed to have a similar rate of growth. Later, group O did not go through the same increase in growth rate as group M, remaining confined to the West-Africa region (Peeters et al., 1997). This is interesting because if a difference in the early phases after human transmission could be explained by the genetic differences of the two viral populations, the gap that occurred only some decades later could not be explained on a genetic level, since it would be highly improbable that a positive mutation could have occurred at the same time in all the M strains, which were already spread in several countries. It is most likely, instead, that the second wave was caused by ambient and social factors, as, for example, the passage from a more restricted patient group composed by high-risk subjects to a more wide-spread and heterogeneous one. New analysis of group O epidemiology showed how it also went through a second exponential phase, but only later (1970-1990) compared to group M and on a smaller scale in terms of absolute numbers (Leoz et al., 2015). Finally, after the two exponential phases, group M encountered a stabilization of the pandemic in the 1980-1990 that is still present today (Nzilambi et al., 1988; Mulanga-Kabeya et al., 1998; Faria et al., 2014).

The cross-species transmissions that led to HIV

Several non-human primate species are infected with SIV, each of them being species-specific (Aghokeng et al., 2010). All of the known transmissions from apes to humans, which have originated HIV types and groups, happened in the same geographical area, the Congo River basin (Peeters et al., 1997; Ayouba et al., 2000; Vallari et al., 2011), and are believed to have occurred because of hunting of primates and/or capture of apes as pets (Hahn et al., 2000).

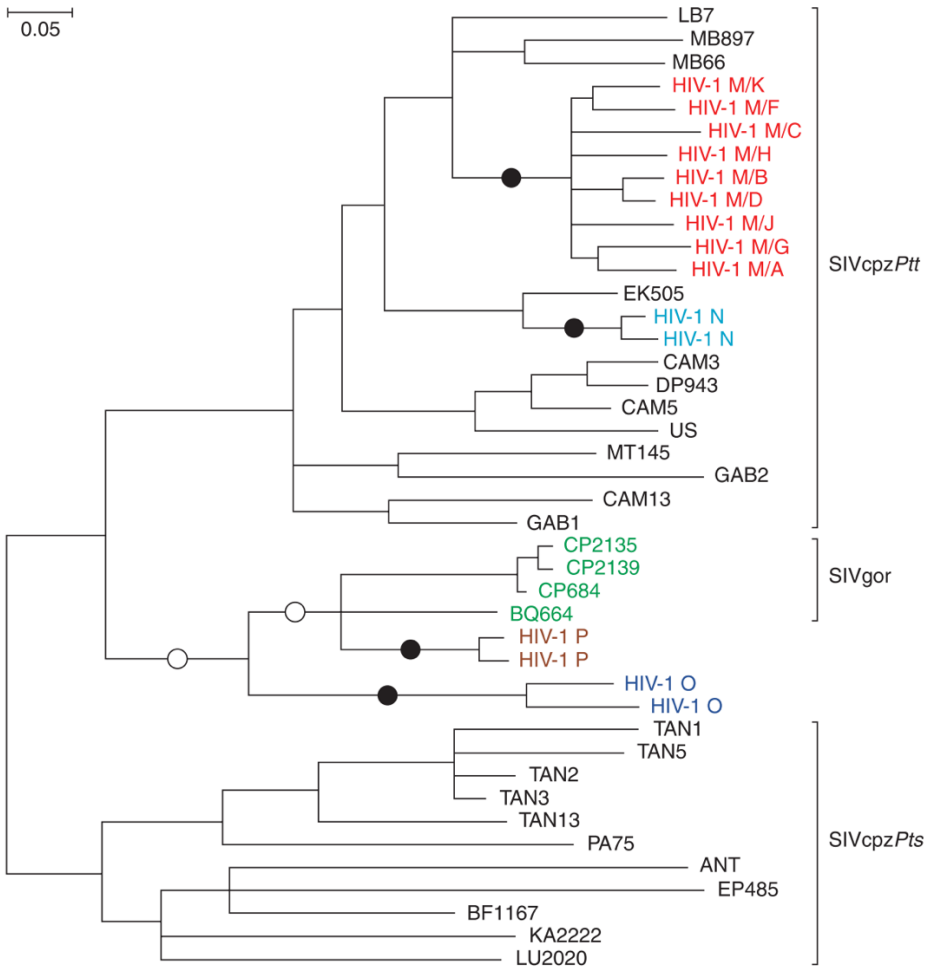


Figure 14. HIV-1 origins. The phylogenetic relationship for a region of the *pol* gene among SIVcpz*Ptt*, SIVcpz*Pts*, SIVgor and the four HIV-1 groups (M, N, O, P) are shown. SIVcpz sequences are shown in black, SIVgor in green, HIV-1 group M in red, group N in light blue, group P in brown, group O in blue. White circles represent the possible chimpanzee-to-gorilla transmissions. Black circles represent the possible chimpanzee/gorilla-to-human transmissions. Image from Sharp and Hahn, 2011.

It is in the south-east region of Cameroon, that chimpanzee populations, in particular of the sub-species *Pan troglodytes troglodytes*, were found to be infected with the simian immunodeficiency virus (SIV) most closely related to group M (Keele et al., 2006; Heuverswyn et al., 2007) (Figure 14). This local transmission probably arrived in Kinshasa via ferry following the Sangha River (Sharp and Hahn, 2011), in fact, fluvial connections during that time were

quite frequent because of the exploitation of rubber and ivory during the German colonization of Cameroon (de Sousa et al., 2012). SIVcpz*Ptt* was transmitted to humans in another independent cross-species event originating group N (Gao et al., 1999; Keele et al., 2006) (Figure 14). Group N is more recent than group M, with its tMRCA estimated to be around 1963 (Wertheim and Worobey, 2009). While group M was already pandemic, as previously discussed, group N was identified in 1998 (Simon et al., 1998) and it has been limited to less than 20 cases, all of them found in Cameroon (Ayouba et al., 2000; Roques et al., 2004). SIVcpz*Ptt* itself is the result of an inter-species transmission and a recombination event between the ancestors of at least two SIV lineages: the SIV infecting the red capped mangabey (SIVrcm) and the SIV infecting Cercopithecus species, including greater spot-nosed (*C. nictitans*), mustached (*C. cephus*), and mona (*C. mona*) monkeys (SIVgsn/mus/mon) (Bailes, 2003). In particular, the 5' half of the genome (5' LTR*, gag, pol*) is more closely related to SIVrcm, while the 3' part of the genome (*vpu, tat, rev, env*, *nef,* 3'LTR) is phylogenetically closer to SIVgsn/mus/mon. A more recent work, however, highlighted how the SIVcpz*Ptt* genome origin might be more complicated than this (Bell and Bedford, 2017). Bell and colleagues found phylogenetic evidence to support that the SIVcpz*Ptt* genome portion including the IN gene, *vif*, and *vpr,* is equally related to SIVrcm and SIVmnd-2, which infects mandrils. Also, they found that the 5' portion of SIVcpz*Ptt* genome, including the 5'LTR, *gag,* the PR and RT genes, and the 3'LTR, do not correlate with any SIV genome known (Bell and Bedford, 2017) (Figure 15). It is possible that further sampling of lentiviruses lineages could resolve the identity of this genome portions ancestor, while it is not to exclude that SIVcpz*Ptt* has sufficiently diverged from this putative ancestor and therefore it is not identifiable.

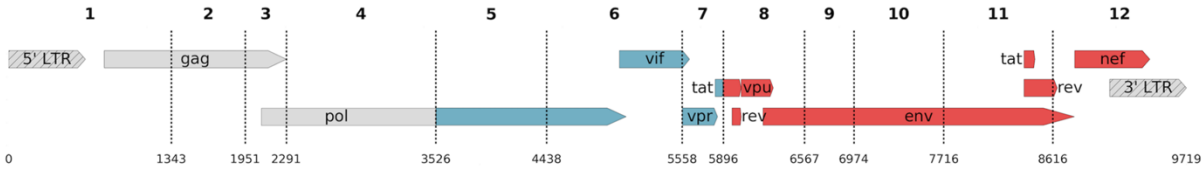


Figure 15. SIVcpz*Ptt* mosaic genome origins. A schematic representation of the SIVcpz*Ptt* genome is shown, with breakpoints used for the phylogenetic analysis indicated by dashed lines and the genome position. Segments from 1 to 4, plus the 3'LTR, do not correlate with any SIV genome known to date. Segments 5 to 7 equally correlate with SIVrcm and SIVmnd-2. Segments 8 to 12 correlate with SIVgsn/mus/mon. Image from Bell and Bredford, 2017.

SIVcpz*Ptt* cross-species events are not limited to the two that originated HIV-1 groups M and N in human, but they also include other species. The SIV infecting gorillas (*Gorilla gorilla gorilla*), SIVgor, in fact, originated from a single zoonotic transmission from SIVcpz*Ptt*, estimated to have happened 100-200 years ago (Takehisa et al., 2009; D'Arc et al., 2015) (Figure 14). SIVgor is much less prevalent in gorillas than SIVcpz in chimpanzees (Neel et al.,

2010; D'Arc et al., 2015), although it could be that clusters of SIVgor are present in parts of Africa that have not been evaluated yet. Chimpanzees are known to hunt other small monkeys, and this can easily explain how they came across SIV in the first place, on the other hand, gorillas are herbivores, and they tend to avoid physical encounters with other primates (Nishihara, 1995; Stanford and Nkurununungi, 2003), making it more difficult to understand how the cross-species transmission could have occurred. It is true, though, that the chimpanzee and gorilla populations overlap in some areas, and it is in these locations that at least one transmission event must have occurred (D'arc et al., 2015). Two independent zoonotic transmission from SIVgor to human originated groups O and P (Keele et al., 2006; Heuverswyn et al., 2007; Plantier et al., 2009) (Figure 14). Group O is the second most widespread group with around 100,000 thousand cases, mostly concentrated in Cameroon where it is endemic (Mourez et al., 2013), with some sporadic cases in other African countries (Peeters et al., 1997), the United States (Rayfield et al., 1996), and Europe (Charneau et al., 1994; Loussert-Ajaka et al., 1995). Group P was discovered in 2009 in a woman from Cameroon living in France (Plantier et al., 2009) and there was only another case reported so far of another person from Cameroon (Vallari et al., 2011). It is difficult to infer its tMRCA since only these two sequences are available.

SIVmmb infecting sooty mangabey originated HIV-2 (Gao et al., 1992; Chen et al., 1997) that remained confined to West-Africa, and it is recently getting "replaced" by HIV-1 (van der Loeff et al., 2006; Hamel et al., 2007). Eight lineages of HIV-2 exist, each of them originated from an independent zoonotic transmission from SIVsmm to human, called groups A-H (Sharp and Hahn, 2011). Among the eight cross-transmission events, however, 6 of them resulted in a single observed infection (Gao et al., 1992; Chen et al., 1997), with just groups A and D able to established sustained transmission chains and presently circulating (Pieniazek et al., 1999; Damond et al., 2001; Ishikawa et al., 2001; Visseaux et al., 2016). The natural history of HIV-2 differs from that of HIV-1 in both disease progression and transmission. In fact, most of the people infected with HIV-2 do not develop AIDS (Rowland-Jones and Whittle, 2007; van der Loeff et al., 2010). Also, HIV-2 is less infectious than HIV-1, both in horizontal and vertical transmissions, and that can be explained by its lower viral load (Popper et al., 2000; Berry et al., 2002).

It is interesting to note that among the four subspecies of chimpanzee existing, only two of them appear to be infected with SIV. The other subspecies infected is the eastern chimpanzee (*Pan troglodytes schweinfurthii*). This could be the outcome of two possible scenarios: either the virus infected the common ancestor of the central-eastern clade, then therefore inherited by the two subspecies, either their infection occurred after the subspecies

division and it is comparatively recent (Santiago et al., 2002; Switzer et al., 2005; Keele et al., 2006; Heuverswyn et al., 2007; Li et al., 2012; Schmidt et al., 2019; Pawar et al., 2022). SIVcpz*Pts* is located in central Africa with a prevalence similar to the one of SIVcpz*Ptt* (Keele et al., 2006, 2009; Rudicell et al., 2010). No transmission to human of this virus is known to have happened and several hypotheses could explain this. SIVcpz*Ptt* and SIVcpz*Pts* are quite divergent with a genetic diversity of 30% and 50% respectively in their Gag-Pol and Env genes (vanden Haesevelde et al., 1996). Also, there could be a different frequency in ape-human encounters of this sub-species compared to *Pan troglodytes troglodytes*. Finally, of course, some zoonotic transmission to human might have happened, but led to a dead-end infection and/or could have not been detected yet. It is also not to exclude that new zoonotic transmission of SIVcpz*Pts*, as well as the other primate lentiviruses, could happen again in the future.

Adaptation to overcome the cross-species barrier

A hallmark of adaption from apes to humans is found at position 30 of the MA protein. While in HIV-1 this position is occupied by a conserved basic amino acid (R or K) within all HIV-1 groups, excepts group P, for which only one of the two sequences available showed to have this mutation (Wain et al., 2007; Plantier et al., 2009; Vallari et al., 2011), in SIVcpz*Ptt* and SIVgor, it was originally occupied by a M (Wain et al., 2007; Takehisa et al., 2009). This suggest that this mutation might have played a key role in the adaption to the new host. This observation was further confirmed with two experiments. In one, a chimpanzee was infected with HIV-1 and, after propagation *in vivo*, a reversion of this position to a methionine was found (Wain et al., 2007). In another experiment they observed how a virus with M30 is replicating more effectively in CD4+ T chimpanzee lymphocytes than of a virus carrying the M30R mutation (Wain et al., 2007). Similarly, a SIVcpz*Ptt* isolate carrying the M30R mutation was shown to significantly better replicate in human CD4+ T cells than the wt (M30) (Sato et al., 2018). However, when tested in a different experimental setting as a humanized mice, the same SIVcpz*Ptt* mutant showed the same replication kinetics as the wt (Sato et al., 2018). Although the exact role of this amino acid is yet to be understood, it is clear that this position was under strong selective pressure once transmitted to the new host.

The adaptation steps to which a protein is submitted are reflected in the genome and correlates with a change in the rate of evolution. The rate of evolution of a given protein is not necessarily constant over the time, resulting in what are called rate shifts. In viral proteins, rate shifts are frequently associated to infection of new hosts, by cross-species transmission.

In these occasions, the ability of a protein to adapt to a new environment is particularly important. In a recent work, the analyses of rate shift along the entire genome of all the HIV and SIV sequences available, highlighted that the genes coding for the accessory proteins have a higher rate shift than the other genomic regions (Gelbart and Stern, 2020), suggesting a significant role of these proteins in species jump. They also highlighted how group M and group O went through independent models of human adaptation. In particular, they observed how group O had about twice the rate shift of group M (Gelbart and Stern, 2020). This could be partially explained by the fact that group O originated from gorillas, which are phylogenetically more distant from humans than are chimpanzees, likely requiring more extensive adaptation to cross the new species barrier (Hasegawa et al., 1985; Ruvolo, 1997). To date, the adaptation path of several viral proteins was traced back and, indeed, most of the proteins which went through major changes are the accessory proteins. This is mostly because they are the proteins in charge of facing the host restriction factors, which are often species-specific.

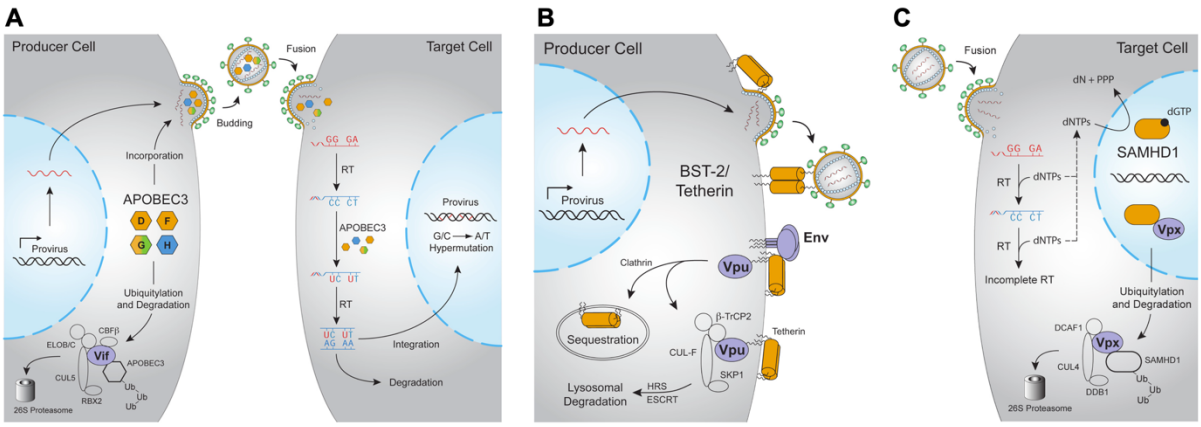


Figure 16. Lentiviral restriction factors. A APOBEC3 proteins are encapsidated into nascent HIV-1 viral particle. Once reverse transcription takes place in the target cell, APOBEC proteins catalyze the deamination of cytosines to uracils in vDNA, causing hypermutations in the vDNA. The viral protein Vif counteracts APOBEC3 restriction in the producer cell by recruiting an E3 ubiquitin (Ub) ligase complex to polyubiquitinate APOBEC3 proteins leading to their degradation by the 26S proteasome. **B** Tetherin block viral production by tethering budding virions to the cell surface. HIV-1 Vpu can either sequestering tetherin in internal compartments (and also shown in the figure HIV-2 Env) or recruit an E3 ubiquitin ligase complex to ubiquitinate and target tetherin for degradation in lysosomes (Vpu only). **C** SAMHD1 blocks reverse transcription by depletion of dNTPs. HIV-2/SIV Vpx overcomes the SAMHD1 restriction by recruiting an E3 ubiquitin (Ub) ligase complex to ubiquitinate and target SAMHD1 for degradation by the 26S proteasome. Image from Harris et al., 2012.

Noteworthy, the viral response to the restriction factor tetherin is a distinctive mark of cross-species adaptation. Tetherin is anchored to the cell membrane and is formed by a cytoplasmic N-terminal region, a trans-membrane domain, a coiled-coiled extra cellular domain, and a C-terminal glycosylphosphatidylinositol (GPI) anchor. It is normally not highly expressed in primary CD4+ T cells, but its expression is induced by type I interferons (Neil et al., 2007), which are triggered by HIV-1 infection (Soper et al., 2018). This host protein is able

to block the release of new viral particles from the infected cell membrane by "tethering" the viral membrane to the cellular one (Figure 16). This action is countered in SIVcpz*Ptt* and SIVgor by the viral protein Nef that targets the cytoplasmic domain of tetherin, inducing its degradation (Jia et al., 2009; Zhang et al., 2009). After zoonotic transmission from SIVcpz*Ptt* to SIVgor, Nef adaptation was most likely smooth, since chimpanzee and gorilla's tetherin differ for only two amino acids (Sauter et al., 2009). When SIVcpz*Ptt* passed to humans, though, it found a more complicated situation. In fact, as viruses evolve in order to counteract restriction factors, the latter are rapidly changing to fight back the virus. Genes encoding for restriction factors, especially, have a fast-paced evolution history (Sawyer et al., 2004, 2005; McNatt et al., 2009; Lim et al., 2010), constituting the barrier that need to be overcome by zoonotic viral transmissions. An example of this adaptation in response to the pression exerted by viral protein is the human tetherin, which has a five-codon deletion in its cytoplasmic domain (Sauter et al., 2009). This means that when SIVcpz*Ptt* and SIVgor passed to humans their Nef protein, it was not able to counteract the human tetherin. As mentioned above, SIVcpz*Ptt* originated from SIVrcm and SIVgsn/mus/mon and it most likely inherited its anti-tetherin activity from both of them, in particular through SIVrcm Nef, and SIVgsn/mus/mon Vpu (Schindler et al., 2006). However, only Nef activity was selected and evolved to be more efficient against chimpanzee tetherin over Vpu, probably for sequence similarities between the cytoplasmatic domains of the monkeys tetherin (Sauter et al., 2010). Therefore, once in humans, one possibility of adaptation was switching back to using Vpu as SIVcpz*Ptt* monkey ancestors (Sauter et al., 2009; Schmökel et al., 2011). This is, indeed, what happened in HIV-1 groups M and N, where Vpu is in charge of this counteraction (Neil et al., 2008; van Damme et al., 2008). HIV-1 M and N Vpu targets the tetherin membrane-spanning domain (Iwabu et al., 2009; Rong et al., 2009) and induces its proteasomal and/or lysosomal degradation (Douglas et al., 2009; Goffinet et al., 2009; Mangeat et al., 2009) (Figure 16). However, while group M Vpu was able to gain a potent anti-tetherin activity (Sauter et al., 2009), group N Vpu not only was not as highly effective (Sauter et al., 2009; Kirchhoff, 2010), but it also lost its ability to down-modulate CD4. In fact, the other activity of Vpu, is to degrade CD4 receptors in the infected cell in order to avoid the internalization of excreted viruses, and it is conducted equally in SIVcpz, SIVgor and all the HIV-1 groups (Sauter et al., 2009). Group O Nef protein is able to down-modulate tetherin at the cell-surface, but with just a mild effect on virus release (Sauter et al., 2009; Kluge et al., 2014; Bego et al., 2016), nevertheless this group is still able to successfully infect the human host. Yet, another pathway to counteract tetherin was taken by HIV-2 that, by inheriting the anti-tetherin activity from SIVmmb that was carried by Nef, found itself helpless against human tetherin. HIV-2 group A then evolved

envelope glycoprotein gp41 to exert this activity (le Tortorec and Neil, 2009) (Figure 16), while in the other HIV-2 groups no anti-tetherin functionality is observed. Overall, the loss of the cytoplasmic domain of human tetherin seem to have represented an important barrier to overstep for non-human primate lentiviruses, that only group M was able to successfully surmount, raising the speculations that this could be one of the reasons behind its global success compared to the other groups (Gupta and Towers, 2009; Sauter et al., 2009; Kirchhoff, 2010; Bego et al., 2016; Sato et al., 2018).

Adaptation is not always a straightforward process and some evolutionary advantages, as gaining a new functionality, might come at a cost. This is what happened with the ability to counteract the restriction factors APOBEC3G (A3G) in SIVcpz*Ptt* and HIV-1. APOBEC3 proteins are a family of cytosine deaminases that catalyze the deamination of cytosine to uracil in the ssDNA substrate (Harris and Dudley, 2015; McDaniel et al., 2020). There are 7 APOBEC3 proteins encoded by the human genome (A3A-D, A3F-H), and 4 of them (A3D, A3F, A3G, A3H) have been shown to restrict HIV-1 infection. A3G, in particular, is the one with the strongest restriction phenotype against HIV-1 (Meissner et al., 2022). A3G blocks viral infection by being packaged into the assembling virion (Mariani et al., 2003; Stopak et al., 2003) and, once a new viral cycle starts, it inhibits vDNA synthesis during reverse transcription (Holmes et al., 2007; Miyagi et al., 2007; Bishop et al., 2008), and catalyzes deamination of cytidine to uridine during negative-strand transfer DNA synthesis (Conticello et al., 2005) (Figure 16). The hypermutations (G-to-A) accumulation lead to either degradation of the reverse transcription product, by cellular uracil DNA glycosylase (Okada and Iwatani, 2016), or, after integration, to an inactive provirus (Kirchhoff, 2010). SIVcpz*Ptt* and HIV-1 Vif are able to efficiently counteract A3G, by interacting with it and recruiting a ubiquitin ligase complex to start the proteasomal degradation of A3G, therefore preventing its encapsidation into the assembling virions (Mangeat et al., 2003; Bishop et al., 2008) (Figure 16). SIVcpz*Ptt* Vif resulted from a recombination event, which made it more efficient against its species-specific A3G, but also to human A3G, conferring to the virus an advantage for cross-species transmission (Etienne et al., 2013; Sato et al., 2018). However, the same recombination event led to the concomitant loss of SIVrcm *vpx* gene, making SIVcpz, and consequently HIV-1, not able to antagonize the SAM and HD domain-containing protein 1 (SAMHD1) (Etienne et al., 2013) (Figure 17).

SAMHD1 is an antiviral factor that limits viral reverse transcription by decreasing the intracellular concentration of dNTPs (Lahouassa et al., 2012) (Figure 16). HIV-2 and SIVsmm Vpx induces SAMHD1 proteasomal degradation by tying it to the ubiquitin-proteasome system (Hrecka et al., 2011; Laguette et al., 2011). The SIVcpz*Ptt* ancestor, SIVrcm, as

mentioned above, also encodes Vpx and it is able to counteract SAMHD1. SIVgsn/mus/mon, instead, do not possess a *vpx* gene, but they are still able to escape to SAMHD1 restriction through their Vpr protein (Lim et al., 2012). Vpx originated from a gene duplication of Vpr in a lentiviral precursor virus, thus, the two proteins have similar sequences and share similar, but not identical, functions (Lim et al., 2012). Noteworthy, HIV-1 Vpr is able to both facilitate infection in macrophages (Balliet et al., 1994; Connor et al., 1995) and induces cell-cycle arrest. In HIV-2 and SIVsmm these two functions are split, with Vpr inducing the cell-cycle arrest in G2 phase, and Vpx enhancing the infection in macrophages (Yu et al., 1991; Goujon et al., 2006, 2008).

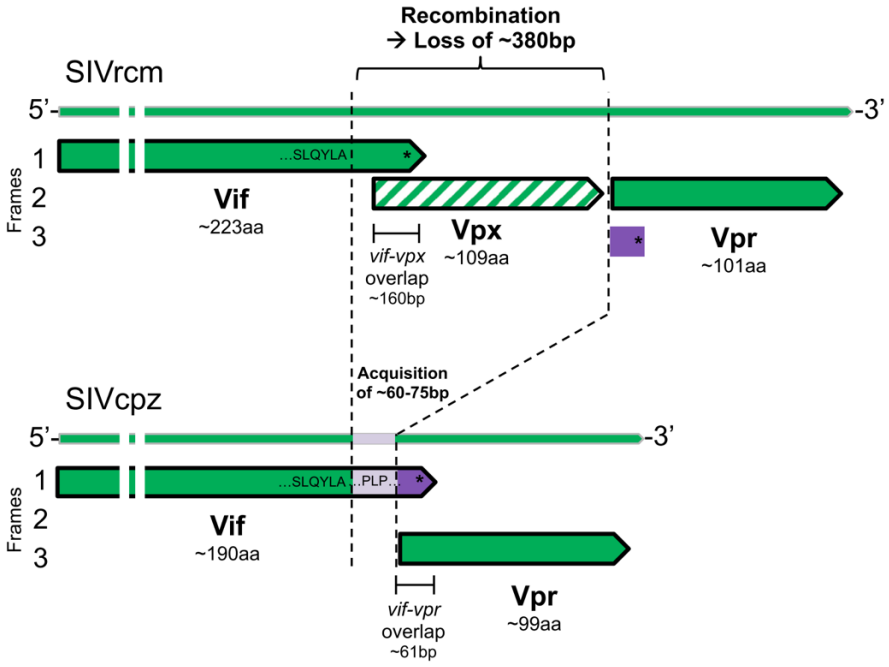


Figure 17. SIVcpz Vif resulted from recombination. The same genome region is shown for SIVrcm (on top) and SIVcpz (below). The 3 reading frames are indicated on the left and proteins ORFs are represented as large arrow (protein length is indicated below each protein name). In green are represented protein portions which were present in SIVrcm and inherited by SIVcpz; in white and green stripes are represented sequences unique to SIVrcm; in light purple are represented new amino acids acquired by SIVcpz; in dark purple are represented sequences that were not expressed in SIVrcm but that are expressed in SIVcpz. The dashed lines indicate the breakpoint of the recombination event. Asterisks indicate stop codons. Image from Etienne et al., 2013.

The fact that SIVcpz*Ptt* is not able to counteract SAMHD1, but can more efficiently counteract A3G restriction, suggests that the selective pressure to counteract the latter was higher than the one to counteract SAMHD1 and conferred to the virus an advantage in the adaptation to the new host (chimpanzee). Furthermore, SIVcpz*Ptt* Vif efficiently antagonizes human A3G, opening to tempting speculations such as that the adaptation SIV went through in chimpanzee, facilitated somehow the cross-species jump that started the pandemic (Etienne et al., 2013).

HIV-1 genetic diversity

HIV has a high genetic variability and is evolving around one million times faster than mammalian DNA (Lemey et al., 2006). This high genetic diversity is due in part to its phylogenetic origins and in part to its high mutation rate caused by the lack of a proof-reading mechanism of the viral RT and recombination happening during reverse transcription (Roberts et al., 1988). These aspects, when combined to the high rate of replication of the virus (Ho DD et al., 1995), are contributing to increase the genetic diversity of the HIV-1 population both intra- and inter- patient (Korber et al., 2001), constituting one of the main obstacles for virus eradication.

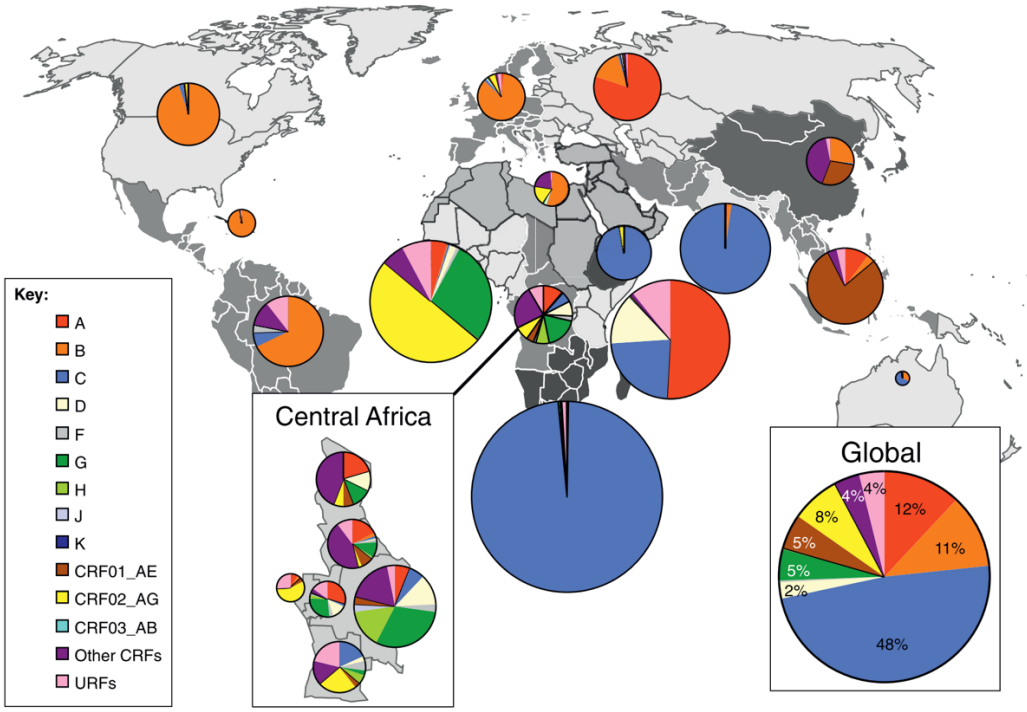


Figure 18. Global distribution of HIV-1 group M subtypes and recombinant forms. The size of the pie chart on every region represents the relevance of the percentage of people living with HIV over the total population. A detail of the Central Africa region distribution is shown. Global distribution, with percentage values, is shown. Image from Hemelaar, 2012.

For all these reasons, from the beginning of the pandemic until now, the intra-group genetic diversity has severely increased, requiring further classifications. Within group M, 10 subtypes are recognized: A, B, C, D, F, H, J, K, L (Worobey et al., 2008; Bbosa et al., 2019; Yamaguchi et al., 2019). Subtype C is the most widespread, being responsible for around half of the world infections and it is prevalent in Africa (Hemelaar et al., 2011) (Figure 18). This subtype originated in DRC to then spread, probably via migrant labor, to South Africa (Jochelson et al., 1991; Hemelaar et al., 2011; Faria et al., 2014), where it is most prevalent today (Figure 18). The different subtypes were shown to have different evolutionary rates (Abecasis et al., 2009), and the intra- and inter-subtypes genetic diversity is increasing with

time (Rambaut et al., 2001). Overall, the genetic distance within a subtype can be between 15 and 20%, whereas across subtypes can reach 35% (Hemelaar et al., 2006).

Group O classifications changed with time. Isolates O were first divided in 3 clades (Quiñones-Mateu et al., 1998; Roques et al., 2002), then 5 clusters (Yamaguchi et al., 2002), and moreover in two lineages (Tebit et al., 2010). At the moment the accepted classification is in two sub-groups: the head, or H strain, and the tail, or T strain (Leoz et al., 2015). The H strains is formed by three subclusters, H1, H2 and H3, while the T strain is composed of two subclusters, T1 and T2 (Leoz et al., 2015). The H strain is the most dominant clade and exhibits the greatest variability (Roques et al., 2002; Leoz et al., 2015). Groups P and N are constituted by only few cases and their isolates are all closely related (Roques et al., 2004; Vallari et al., 2011).

As mentioned above, among the causes of the high variability of HIV, the RT plays a key role. The enzyme does not have a proof-reading activity, introducing 1 to 3 mutations per genome per cycle (Preston et al., 1988; Smyth et al., 2012). Furthermore, during the reverse transcription process, two events of strand transfer are necessary for the formation of a complete vDNA molecule with a duplicated complete LTR (van Wamel and Berkhout, 1998). Further strand transfer events can occur in internal regions of the genome. These last strand transfer events, in particular, when occurring in heterozygous viruses, can amplify genetic diversity, by combining several different polymorphisms present in the two copies of genomic RNA present in the virion (Hu and Temin, 1990; Chen et al., 2009).

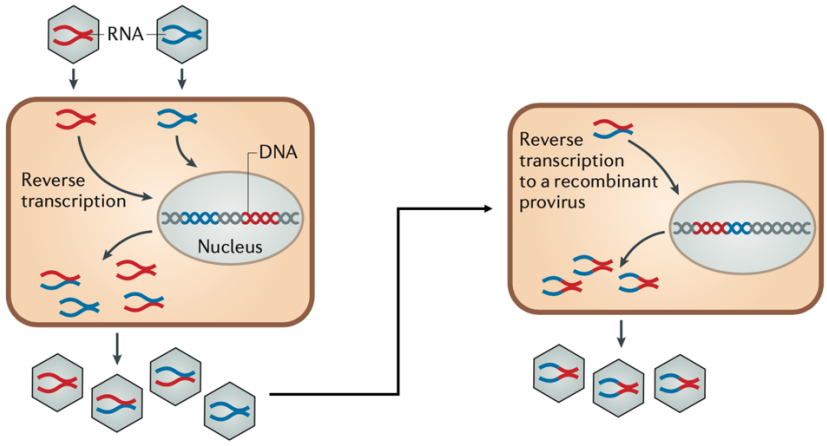


Figure 19. Generation of recombinant viruses. When two genetically distinct strains co-infect the same cell this could lead to the formation of heterozygous viral particles, which, when infecting a new cell, will generate recombinant viruses during reverse transcription. Adapted from Simon-Loriere and Holmes, 2011.

Recombination is one of the main reasons for the high genetic variation found in HIV-1 and it can involve viruses from different subtypes or, albeit less frequently, groups (Peeters et al., 1999; Rousseau et al., 2007) co-infecting the same cell (Figure 19). Recombination is so

extensive in HIV-1 M group that a classification for the recombinant form was also necessary. The recombinants between different subtypes are called either circulating recombinant forms (CRFs), when found in more than three epidemiologically unrelated individuals and fully sequenced, or unique recombinant forms (URFs), when these criteria are not fulfilled (Robertson et al., 2000). There are more than 98 circulating CRFs and they can be formed by two fragments of two different subtypes (e.g. CRF01_AE, CRF02_AG) or form a more complex mosaic of recombination events, as to comprise up to five or six different subtypes (e.g. CRF04_cpx). Some of these recombinant forms appeared shortly after the zoonotic transmission to human, like the CRF01_AE and the CRF02_AG. With more subtypes cocirculating all over the world, recombinant forms are growing at an impressive rate, doubling the number of reported CRFs in the last ten years (Bbosa et al., 2019).

The fast rate at which HIV is evolving, and especially recombination, poses an important challenge for epidemiologically-based investigations as well as accurate diagnosis and antiretroviral treatments. As already mentioned, recombination happens when a viral particle is carrying two vRNA containing divergent sequences, which are created as a consequence of superinfections of two or more distinct subtypes and/or groups. In the heterozygous viral particles recombination takes place during the synthesis of the minus DNA strand. In this step, the nascent DNA switches from one copy to another in a process known as copy choice, leading to the formation of a new recombinant proviral DNA and, therefore, genomic RNA. What triggers the template switch is still debated. Initially, it was believed that the switch occurred when the RT found a damage on the RNA and was therefore forced to change template, as explained by the "forced copy choice" model (Coffin, 1979). Then, it was proposed that pausing of the RT, caused by the difficult incorporation of a nucleotide for different reasons as, for example, the presence of an RNA structure, could promote template switching (Wu et al., 1995; Suo and Johnson, 1997). Further studies highlighted how recombination breakpoints did not necessarily correlate with RT pausing sites, but with highly structured regions of the acceptor RNA template, the one onto which the synthesis is transferred (Negroni and Buc, 2000). These models are not mutually exclusive and all of them might explain recombination events occurring at different positions of the genome. No matter the trigger, template switch is more likely to occur when the two RNA templates share a certain degree of sequence identity near the breakpoint (Baird et al., 2006b, 2006a). Nevertheless, recombinants involving phylogenetically distant isolates were reported, as, for example, recombinants between isolates from group HIV-1 group M and O (de Oliveira et al., 2017, 2018) indicating that recombination presents a certain degree of tolerance to genetic divergence.

Recombination has important consequences on viral evolution, since it can combine fragments of genomes carrying several mutations previously selected in each isolate, creating new potentially advantageous combinations. In the same way, recombinant viruses can also result to be non-infectious, because of an incompatibility of the assembled fragments and/or the rupture of the coevolution network that was present in each isolate. The viral genome encodes for highly conserved residues, which usually are in charge of providing features that are essential throughout the whole viral population, and less- or non-conserved amino acids. Non-conserved amino acids, however, can be as essential as the conserved ones as being part of the coevolution network present in one isolate. The mutations and the consequent compensatory mutations carried in one protein or even among different proteins cross-talk to each other in order to maintain functionality. This is why, non-conserved amino acids become particularly important in the light of recombinant viruses. Their different distribution and combinations can lead to the acquisition of new advantageous functionalities and/or resistance to therapy (Mansky, 2002), as well as favoring the elimination of deleterious mutations and/or damaged DNA (Simon-Loriere and Holmes, 2011).

Although in theory recombination events could happen along the entire genome of HIV-1, the breakpoints frequency distributions in circulating viruses clearly shows that they map preferentially in specific hot spots (Fan et al., 2007). This non-random distribution of recombination events is caused by limitations of the recombination event *per se* (as sequence similarity or the RNA structure), but also of the consequent selective pressure for specific recombinants forms, which are first of all functional and can also be more advantageous for the virus, rather than others. One of the reasons behind this selection can be that recombination is breaking coevolution networks, formed by the genetic interplay between conserved as well as non-conserved residues of the same or different protein, present in the genome (Galli et al., 2010; Woo et al., 2014).

OBJECTIVES OF THE STUDY

By exploiting the rupture of the above-mentioned co-evolution network, by building chimeric integrases between group M and group O, the laboratory was able to identify a group-specific functional motif in the CTD of IN M ($N_{222}K_{240}N_{254}K_{273}$) (Kanja et al., 2020). The motif is highly conserved in group M, as observed when more than seven thousand sequences from this group were aligned (Figure 20A). The motif is composed by an alternance of two positively charged (K) and two polar amidic amino acids (N) and both were shown to be essential for integration, indeed, when either the N or the K are mutated, integration is abolished (Kanja et al., 2020) (Figure 20B, C). Further experiments showed how the important features of the two amino acids composing the motif were, for the K, their positive charge, and for the N, their amidic side chains. Indeed, by replacing them with amino acids with similar characteristics (K replaced by R; N replaced by Q), integration was not affected (Kanja et al., 2020) (Figure 20D). For this reason, the motif was renamed in the article manuscript on the main project of my thesis work, by myself and co-authors, the C-terminal lysins amidic (CLA) motif.

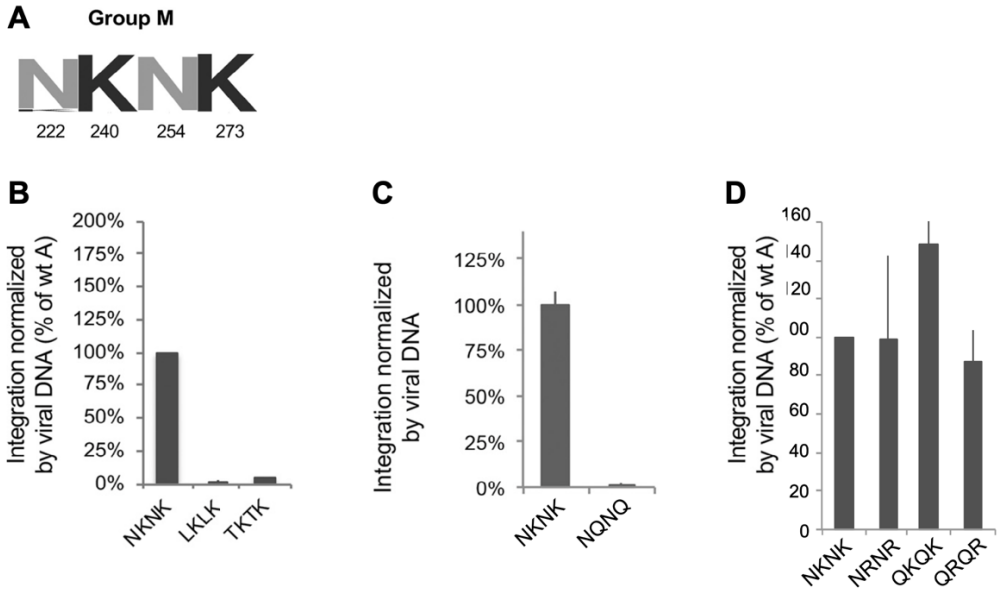


Figure 20. The CLA motif is highly conserved and essential for integration in group M. A Conservation logo obtained with WebLogo of positions 222, 240, 254, 273 in group M. **B** Integration levels for the N mutants (LKLK and TKTK) of the motif relative to the wt (NKNK). **C** Integration levels for the K mutant (NQNQ). **D** Integration levels for the conservative mutant of the K (NRNR), of the N (QKQK) and of the double mutants (QRQR). Panels A-C adaptated from Kanja et al., 2020. Panel D adapted from Kanja, 2017 (doctoral thesis).

Despite its conservation *in vivo*, however, the motif showed to have important levels of genetic flexibility. Indeed, it was not only possible to retain functionality when similar amino acids replaced the original ones (QRQR, Figure 20), but functionality was also retained to wt levels when 2 or more lysines were present in any of the four positions composing the motif, with the remaining ones being occupied by amidic amino acids (N, Q) (Kanja et al., 2020)

(Figure 21A). This feature is probably due to the fact that the four amino acids composing the CLA motif, form a positively charged surface, conserved in lentivirus (Kanja et al., 2020) (Figure 22). This finding, together with the genetic flexibility showed by the motif, led Kanja and colleagues to hypothesize that the possible function of the motif could be to interact with a negatively charged partner, most likely with a repetitive negative charge, since swapping the positive charges inside the motif did not affect the phenotype in most of the cases.

Out of all the combination with 2 K tested, only two showed to have a default in the integration step, with the most severe phenotype being at 25% of integration levels of the wt in the presence of the NQKK amino acidic sequence (Figure 21B). Surprisingly, the NQKK sequence constitutes the consensus sequence found at the CLA positions in integrase from group O (Article Figure 1), raising the question on how could group O have selected such a sequence. Answering to this question was one of the objectives of this work (**OBJECTIVE 1**).

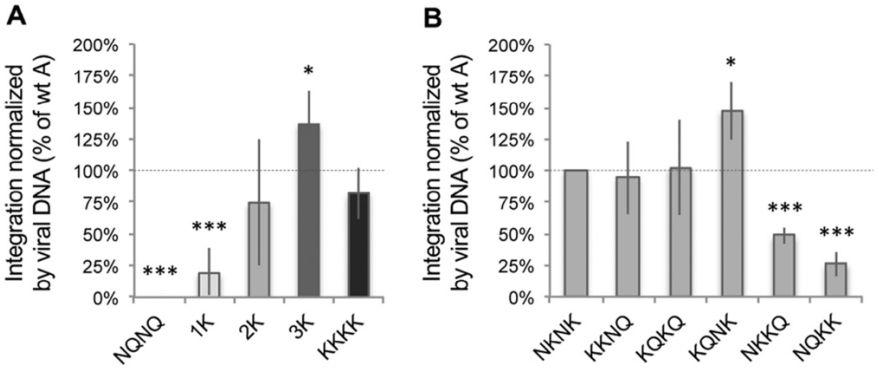


Figure 21. Genetic flexibility of the CLA motif. A Integration levels for the CLA motif mutants carrying 0 (NQNQ), 1, 2, 3 or 4 lysines (K) among the four positions. **B** Integration levels for every mutant with 2 K tested. Adapted from Kanja et al., 2020.

As stated above, an essential feature of this motif are the positive charges carried by the lysins. Indeed, a motif carrying four K (KKKK) has the same integration levels as the wt (Figure 21A). Nevertheless, the amidic amino acids present in the motif also showed to play an essential role (Figure 20B). Despite the fact that the observed phenotype is the same when either the K or the N are mutated, the analyses of different pre-integration steps to understand whether the default observed was specific to the integration step or not, gave contrasting results. Indeed, while when the K were mutated, the default observed in 3' processing and in the 2LTRc levels was coherent with the observed loss of integration efficiency, this was not the case when the N were mutated (Table 1). For both mutants, LKLK and TKTK, the integration default observed were more severe than what expected, suggesting that the N might be involved in further pre-integration steps that were not checked yet. This observation constituted one of the starting points of this work (**OBJECTIVE 2**).

To conclude, the main objectives of this thesis work are two:

1. To understand whether the CLA motif could have an essential role in IN O, where a conserved sequence that is barely functional in IN M is present at its CLA positions.
2. To investigate the role of the N in the CLA motif, which appear to be involved in other pre-integration steps not checked yet.

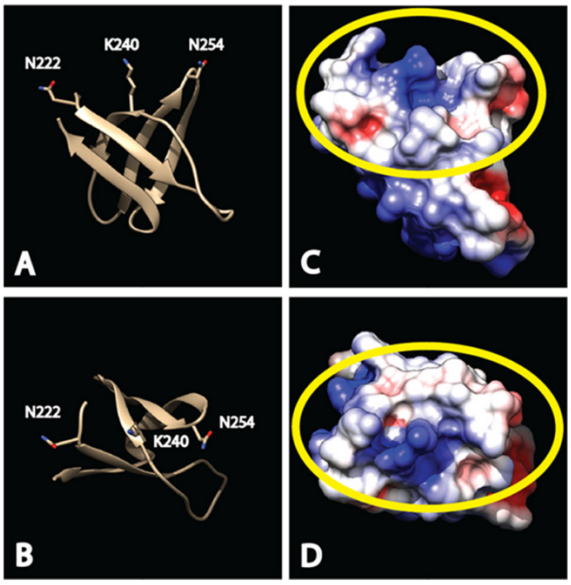


Figure 22. The CLA motif forms a positive charged surface. Structural analysis of IN M CTD showed how the first three positions of the CLA motif (the K273 was not resolved as it is located in a too flexible region) form a positive charged surface. **A**, **B** Side view (A) and top view (B) of the IN CTD. **C**, **D** Surface electrostatic potential representation of IN CTD. Adapted from Kanja et al., 2020.

METHODS

Cell lines

HEK293T and Jurkat cells were obtained from the American Type Culture Collection (ATCC). HEK293T were cultured in DMEM while Jurkat were cultured in RPMI. Both mediums were completed with 10% fetal bovine serum and 1% PenStrep. HEK293T and Jurkat culture conditions were at 37°C in 5% CO_2.

Viral strains and sequence alignments

The primary HIV-1 isolates used in this study were: isolate HXB2 (GenBank accession number: K03455.1), isolate A2 (GenBank accession number: AF286237) from group M, subtype A2, (named "isolate M" in this study) obtained from the NIH AIDS Research and Reference Reagent Program; isolate RBF206 (GenBank accession number: KU168298) and isolate BCF120 (GenBank accession number: KU168297) both from group O, kindly provided by J.C. Plantier (CHU Rouen, France). Isolates AF286237 and RBF206 were chosen because they were used in the work that originated the present study (Kanja et al., 2020). Isolate BCF120 was chosen as the isolate O carrying the consensus sequence in the two motifs considered in this work. The SIVcpz*Ptt* isolate employed in this work is the MB897 (GenBank accession number: EF535994) and it was chosen being one of the two isolates which are the most phylogenetically related to group M.

For the creation of the conservation logos, by using WebLogo (http://weblogo.threeplusone.com) (Schneider and Stephens, 1990; Crooks et al., 2004), we performed sequence alignments for all the sequences covering the NOG and CLA motif positions of isolates from HIV-1 group M, group O, group P, group N, SIVcpz*Ptt*, SIVcpz*Pts*, SIVrcm, SIVmnd-2, and SIVgor. The alignement was performed with the CLC Genomics Workbench 22. All the sequences were obtained from the Los Alamos National Laboratory HIV database (https://www.hiv.lanl.gov/content/index).

Plasmid and molecular cloning

The plasmid p8.91-MB previously described (Kanja et al., 2020), was used as backbone for all cloning procedures. Therefore, all our constructs have the *gag* and the protease-coding sequences from HXB2 (group M). RT and IN coding-sequences, instead, varied. In isolate M the RT and IN was from isolate A2, while in isolates O it was either from isolate O206 or O120. In the chimpanzee isolate the RT and IN was from MB897. All IN mutant coding sequences were inserted between the *BspE*I and *Sal*I restriction sites of p8.91-MB by Gibson assembly.

The plasmid used to produce the genomic RNA of the VLPs, carrying the two reporter genes used to evaluate integration efficiency (nGFP/nRFP and PURO^{R}), is a modified version of the previously-described pSRP (Kanja et al., 2020) where the nuclear RFP was replaced by the nuclear GFP, giving the pSGP.

The pEUrep-RNA (da Silva Santos et al., 2016) was kindly provided by Andrea Cimarelli. The plasmid is coding for an mRNA containing the RNA packaging sequence (Ψ) and the cDNA of the Firefly luciferase followed by a polyA signal.

Two plasmids, both previously described (Kanja et al., 2020), were employed for the creation of standard curves in the quantitative PCR assays. The pJet-1LTR for the detection of late RTPs and the pGenuine2LTR for the evaluation of the 3' processing efficiency.

Transfection and VLPs collection

To produce virus-like particles (VLPs) HEK293T cells were seeded the day before and co-transfected with the plasmid coding for the vesicular stomatitis virus glycoprotein (VSV-G) (Naldini et al., 1996), the plasmid carrying HIV-1 Gag-Pol gene (p8.91MB with different IN) and the plasmid with the modified viral genome with the reporter genes to follow the infection (pSGP). For the EURT assay the pEU-repRNA plasmid, coding for the EU-repRNA, was either added to the mix or used in the place of pSGP. All transfections were done by using 5 µg of total DNA and polyethyleneimine (PEI, Polyscience) following the manufacturer's instructions. The medium was changed after 6 h and VLPs were collected and filtered with a 0.45 µm filter after 48-72 hours. The amount of VLPs was estimated by quantifying the p24 via ELISA (Fujirebio).

Western blot analysis

The same volume of VLPs was concentrated by centrifuging them through a 20% sucrose for 2 h at 20,000 g and at 4°C. Pellets were resuspended in 3x Laemmli buffer and viral proteins were separated on a CriterionTM TGX Strain-Free 4-15% gradient gel (Bio-Rad) and then blotted on a PVDF membrane. To evaluate Pr55Gag proteolytic processing, polyproteins and mature capsid proteins were detected by probing the membrane with a mouse monoclonal anti-CA primary antibody (NIH AIDS Reagent Program) and a secondary anti-mouse HRP-conjugated antibody (Millipore). ECL reagent (Bio-Rad) was added to the membrane and images were taken with Bio-Rad Chemidoc Touch and analyzed with the Image Lab software (Bio-Rad). The Pr55Gag processing efficiency was expressed as the ratio

of mature CA signal on the total CA signal (unprocessed, partially processed and fully processed CA proteins).

Cell fractionations

HEK293T and Jurkat were transduced by spinoculation with polybrene (Sigma-Aldrich) and an amount of VLPs corresponding to a nominal MOI of 0,1. 24-hpt cells were resuspended in fractionation buffer, incubated on ice, and passed multiple times in a 27-gauge needle to separate the nuclei from the cytoplasm. After centrifugation, the supernatant was collected and stored as the cytoplasmic fraction, while nuclei were washed with fractionation buffer, passed multiple times through a 26-gauge needle, and centrifuged again. After centrifugation, the pellet was stored as the nuclear fractions. Fractions were split, one part was used to perform a Western Blot to check their quality (protocol as above with the following primary antibodies: Mouse mAb GAPDH (#GT239, GeneTex), Mouse mAb Nucleolin (#E5M7Km Cell Signaling Technology), and one part was used to extract DNA and quantify late reverse transcription products (protocol in the next paragraph).

Quantitative PCR for viral DNA and its forms

HEK293T or Jurkat cells were transduced by spinoculation with polybrene (Sigma-Aldrich) and an amount of VLPs corresponding to a nominal MOI of 1. Prior to infection, VLPs were incubated with Benzonase nuclease (Sigma-Aldrich) to remove non-internalized DNA. 24-hours post-transduction (hpt) cells were collected, and total DNA was extracted with DNeasy Blood & Tissue Kit (QIAGEN). All qPCR assays were designed with the Taqman® hydrolysis probe technology using the IDT Primers and Probes design software (IDT), with dual quencher probes (one internal ZEN™ quencher and one 3' Iowa Black™ FQ quencher). qPCRs were performed with the iTaq Universal Probes Supermix (Bio-Rad) on a CFX96 (Bio-Rad) thermal cycler according to the manufacturer's protocols. Standard curves and analysis were conducted with the CFX Manager (Bio-Rad).

Late reverse transcription products were quantified with oligos amplifying the U5-Psi junction. This was normalized by the amount of genomic DNA that was quantified by amplifying an exon of the actin gene. Absolute quantification was performed by creating a standard curve with known quantities of pJet-1LTR for RTPs and the genome extracted from a known quantity of cells for actin quantification.

To evaluate the 3' processing efficiency we first quantified the quantity of 2LTR circles (2LTRc) with oligos and probe annealing to the 2LTRc junction and then we evaluated the nature of this junction (perfect or imperfect, which are respectively the unprocessed and processed 3' ends) with oligos and probes annealing specifically only to the perfect junction. The imperfect junction ratio was subsequently calculated as 1-perfect junction, where 1 represents the total amount of 2LTRc. For both 2LTRc and perfect junction quantification, the same standard curve was used done with pGenuine2LTR. All the oligos and probes used for the qPCR assays can be found in Table S.

Evaluation of integration

HEK293T or Jurkat cells were transduced by spinoculation with polybrene and an amount of VLPs corresponding to a nominal MOI of 0.01. 24hpt puromycin was added to HEK293T with a final concentration of 0.6 µg/ml and integration was measured by counting the puromycin-resistant clones 1-week post-transduction. As previously shown, this method is comparable to the classical Alu-gag quantitative PCR method (Kanja et al., 2020). For Jurkat cells integration was measured by FACS 72-hpt by counting the percentage of nGFP positive cells. This time was chosen after having established that no signal would be detected using a catalytically inactive IN (D116A), to exclude the possibility that the signal of our constructions would derive from episomal forms of the viral DNA. Since integration depends on the availability of the RTPs and since reverse transcription is affected by the viral IN, in both HEK293T and Jurkat, results were normalized by the reverse transcription efficiency evaluated by qPCR as described above. Namely, the amount X_1 of RTP was estimated for sample 1, for example. The number of puro resistant clones (P_X) for HEK293T cells or the intensity of the nGFP signal (F_1) by FACS for Jurkat cells, was computed for the same sample. The normalized integration values were then computed as P_1/X_1 or F_1/X_1.

Assessment of the capsid stability

As described above VLPs employed in this assay contain either two RNAs, the EUrep-RNA and the SGP-RNA, or the EUrep-RNA alone. The corresponding quantity of VLPs of a nominal MOI of 0.5 was used to transduce either HEK293T or Jurkat cells. 8-hpt cell protein extract was obtained, and Luciferase assay was performed with the Luciferase Assay System (Promega). Luminescence (RLU) was normalized for protein concentration measured by the Bradford assay and therefore expressed as RLU per mg of protein extract (RLU/mg).

Structure and molecular modelling

The NTD and CTD structures of IN M show in the manuscript belong to PDB 6PUT (Naldini et al., 1996; Passos et al., 2020). The NTD and CTD structures of IN O were obtained from molecular modelling of isolate O120 made with AlphaFold2 (Jumper et al., 2021; Varadi et al., 2022) by Patrice Gouet. Pictures used in the manuscript were obtained with PyMOL2.5.

Statistical analysis

All statistical tests were performed on at least three independent experiments (n is indicated in every figure legend) using Prism 9. ANOVA with Tukey's multiple comparisons correction was used when more than three groups were compared. An unpaired t-test was used when two samples were directly compared

Chromatin immunoprecipitation sequencing[1]

$10x10^6$ million Jurkat cells were grown in a 15 cm cell culture dish. Medium was removed and cells washed once in PBS (+1 mM sodium butyrate for H3K27ac IP to block deacetylases). Cells were fixed with 1% formaldehyde/PBS (1 mM sodium butyrate for K27ac) for 7 min at RT followed by quenching with 0.125 M glycine/PBS for 7 min at RT. After removing all liquid, 2 times 5 ml ice cold PBS was added to scrape cells and cells were pelleted by centrifugation (1200xg, 7 min). After washing with 10 ml of cold PBS, pellet was resuspended in swelling buffer (10 mM HEPES/KOH pH 7.9, 85 mM KCl, 1mM EDTA, 0.5% IGEPAL CA-630, 1x protease inhibitor cocktail (Roche)) and incubated for 10 min rotating at 4°C. Pellet was dounced 10x before centrifugation 10 min, 3500xg for 10 min at 4°C. One extra wash with swelling buffer without IGEPAL CA-630 was performed before resuspending the nuclei in cold sonication buffer (TE pH=8, 0.1% SDS, protease inhibitor tablet). Sonication was done with a Covaris S220 Focused Ultrasonicator for 18 min (Duty cycle 20%, Intensity 5, Cycles/burst 200). DNA size was followed by agarose gel. DNA fragments should be between 200-500 bp. Triton-X was added to the lysate to a final concentration of 1% and incubated for 10 min on ice. Lysate was cleared by centrifugation at 18000xg 4°C for 5 minutes. Magna ChIP Protein A and G magnetic beads (Millipore) were washed twice with TE 0.1% SDS and 1% TritonX and added to the lysate to preclear for 1h at 4°C rotating. 2-8 mg of chromatin were incubated with corresponding amounts of antibody overnight at 4°C. 1% of chromatin was saved as input. The used antibodies were the following: H3K36me3

(ab9050, Abcam), H3K27ac (ab4729, Abcam), H3K27me3 (C36B11, Cell Signaling), H3K4me1 (ab8895, Abcam), H3K9me3 (ab8898, Abcam), H3K9me2 (ab1220, Abcam), Rabbit control IgG (ab46540, Abcam), Mouse control IgG2a (ab18413, Abcam).

The next day, protein A and G magnetic beads were washed twice with sonication buffer plus 1% TritonX and incubated for 2 h at 4°C with the lysates. Beads were washed twice 10 min with cold buffer I (150 mM NaCl, 1% Triton X-100, 0.1% SDS, 2 mM EDTA, protease inhibitor cocktail (Roche)), once 10 min with cold buffer II (10 mM Tris/HCl pH 7.5, 250 mM LiCl, 1% IGEPAL CA-630, 0.7% Deoxycholate, 1 mM EDTA, protease inhibitor cocktail (Roche)), twice 10 min with cold TET buffer (10 mM Tris/HCl pH7.5, 1 mM EDTA, 0.1% Tween-20, protease inhibitor cocktail (Roche)) and eluted with TE buffer, 1% SDS, 100 mM NaCl. 0.5 mg/ml Proteinase K were added, and samples were incubated for 2 h at 55°C and overnight at 65°C. The next day, 0.33 mg/ml RNAse A (Thermo Fisher Scientific) was added and incubated for 1 h at 37°C. Supernatant was removed from the magnetic beads and DNA was purified with AMpure beads XP clean up according to manufacturer's instructions. Concentrations were determined by Qubit Fluorometer and enrichment was determined by qPCR on a CFX96 C1000 Touch Thermal Cycler (BioRad) with SsoFast™ EvaGreen® Supermix (BioRad) and the following primers: Human Negative Control Primer Set 2 (ActiveMotif), Human Positive Control Primer Set GAPDH-1 (ActiveMotif), Human Positive Control Primer Set MYT1 (ActiveMotif), Human Positive Control Primer Set ACTB-2 (ActiveMotif), SimpleChIP® Human Sat2 Repeat Element Primers (CellSignaling). ChIP libraries were prepared using NEBNext® Ultra™ II DNA Library Prep Kit for Illumina® (NEB) and NEBNext® Multiplex Oligos for Illumina® (Index Primers Set 1) (NEB) according to manufacturer's instructions. Libraries were sequenced at c.ATG sequencing core facility at Tübingen University on a NextSeq instrument 2x75 bp.

ChIP-Seq data processing[1]

Reads were trimmed by TrimGalore (v0.4) with maximum allowed error rate 0.3 and default filtering parameters. Trimmed reads were aligned to the human genome assembly hg38 using Bowtie2 (v2.3) with default settings for paired-end sequencing (Langmead et al., 2009). Peaks were called by MACS2 (v2.1) on the merged replicates in each condition using the —nomodel option, broad cut-off=0.1, and false-discovery rate threshold 0.05 (Zhang et al., 2008). See Supplementary Table S2 for QC values.

Super-enhancers were defined using H3K27ac peaks through the findPeaks function from HOMER (v4.10) using the '-style super -o auto' parameters (Heinz et al., 2010).

All profile plots and metagene plots were generated using soGGi (v1.20) from RPKM-normalized bigwigs generated by bamCompare (Ramírez et al., 2016).

Integration site sequencing[1]

Infected Jurkat cells were harvested at 72 hpt and DNA was isolated with Qiagen blood and tissue kit according to manufacturer's instructions. DNA was sonicated with Covaris S220 Focused Ultrasonicator 450 s (Duty factor 10%, 200 cycles per burst) and size was monitored by agarose gel analysis.

The LM PCR protocol was adapted from (Serrao et al., 2016). Sonicated gDNA ends were repaired using End-It™ DNA End-Repair Kit (Epicentre) according to manufacturer's instructions. Sample was purified with PCR purification Kit (MacheryNagel) according to manufacturer's instructions. A-tailing was performed using NEBNext® dA-Tailing Module (NEB). The reaction was incubated in a thermal cycler for 30 minutes at 37°C and purified with PCR purification Kit (MN) according to manufacturer's instructions. An asymmetric double stranded linker was annealed overnight at 12°C (800 U T4 ligase, 10x ligase buffer, 1.5 µM linker). After purification with PCR purification kit or 0.9x AMpure XP clean up beads according to manufacturer's instructions, first nested PCR with linker and LTR specific primers was performed. 25 µl PCR reaction was set up with 1 µM LTR specific primer, 0.2 µM linker specific primer, 5x buffer, 2.5 mM dNTPs, Phusion polymerase (NEB) and 100 ng ligation reaction. The thermal cycler program was the following: 94°C 2 min, - 94°C 15', 64°C 30', 72°C 30' – 25x, 72°C 10 min. PCR reactions were purified with PCR purification kit (MN) according to manufacturer's instructions (or 0.9x AMpure XP clean up beads). The second nested PCR with the same linker specific primer and an inner LTR specific primer containing Illumina sequencing index was performed under the same conditions as the first PCR, with increased cycle number (30 cycles). PCR reactions were purified with PCR purification kit (MN) according to manufacturer's instructions (or 0.9x AMpure XP clean up beads). Quality of the library was analyzed by Bioanalyzer and NEBNext® Library Quant Kit for Illumina® (NEB). IS libraries were sent for sequencing to c.ATG sequencing core facility at Tübingen University and sequenced on a MiSeq instrument 2x150 bp.

Integration site determination[1]

A BLAT-based pipeline was created to process the LM-PCR raw data. The method was adapted to be used on both single-end (SE) and paired-end (PE) sequencing. Reads with LTR sequence (PE on the first pair/SE) or with Linker sequence (PE, on the second pair) were selected (allowing for 2 mismatches) and trimmed with Cutadapt (v3.2) (Martin and Wang, 2011) to improve alignment. Resulting reads shorter than 15bp were excluded. Trimmed reads were converted to fasta format and aligned to a chimeric genome (hg38 and HIV-1 genome) using BLAT (parameters: -stepSize=6 -minIdentity=97 -maxIntron=0 -minScore=15) (Kent, 2002).

Only BLAT results that align at least 30 bp (SE) or 10bp (PE) with the genome and where the alignment start was from 0 and the first 5bp were kept (PE/SE). Uniquely mapped reads were kept for further processing steps in both PE and SE. Non-standard chromosomes and internal integrations on the HIV-1 genome were excluded. In the case of multi-mapped reads on SE, only BLAT results where the difference between the longest aligned portion and the second longest were higher or equal to 25bp were kept. On PE, pairs shown to be in the 1 kb vicinity were considered properly paired and kept.

Integrations were considered to be duplicates if the distance in between them was less or equal to 10 bp. PE (N=1,771) and SE (N=2,822) integration sets were merged (N = 4,590).

Integrations were annotated to the nearest gene using ChIPpeakAnno (v3.24.2) and the GRCh38 annotation package EnsDb.Hsapiens.v86 (v2.99) (Lihua J Zhu et al., 2010).

Gene ontology analysis (made on genes with genic integrations) was performed using clusterProfiler (v3.18.1) (Yu et al., 2012).

[1] Methods performed in collaboration with Dr. Marina Lusic's team (adapted from Rheinberger et al., in preparation)

RESULTS

OBJECTIVE 1: DIFFERENT ZOONOTIC TRANSMISSION EVENTS LED TO DISTINC MECHANISTIC PATHS TO ENSURE INTEGRATION IN HIV-1 ISOLATES OF GROUPS M AND O

HIV-1 group M originated from a zoonotic transmission of SIVcpz*Ptt* to human (Gao et al., 1999; Heuverswyn et al., 2007), believed to have happened in the first decades of the 1900 in the Cameroon region (Sharp and Hahn, 2011; Hemelaar, 2012). Since then, it spread, first in Africa, and then across the rest of the world, establishing the current AIDS pandemic, which caused millions of victims and still counts millions of infected individuals.

HIV-1 group O originated from SIVgor (D'arc et al., 2015) and was first identified in a Cameroonian patient living in Belgium in the 1990s (Gürtler et al., 1994; vanden Haesevelde et al., 1996), but it is estimated to be circulating in the human population since the 1920s (Korber et al., 2000; Lemey et al., 2004; Leoz et al., 2015). Nowadays, group O is dominant in three countries of West and Central Africa (Cameroon, Gabon, and Equatorial Guinea), but its prevalence is decreasing steadily.

The genetic diversity between group O and M varies with the region of the genome, but is overall high, being 33% in *gag*, 27% in *pol* and reaching its highest value, 44% in the *env* gene. Also, the accessory proteins (Vif, Vpr, Tat, Vpu) present sequence divergence between the two groups, with a mean divergence of 18%. Their integrases M and O differ for 16% of their amino acids. They share the same structural organization in domains, and they possess the same functional properties.

In the CTD of IN M, the laboratory has previously identified the CLA motif ($N_{222}K_{240}N_{254}K_{273}$), which highly conserved and essential for integration in this group (Kanja et al., 2020). At the same positions of IN from group O another highly conserved sequence is present ($N_{222}Q_{240}K_{254}K_{273}$), composed by amino acids that have similar characteristics to those found in group M. Strikingly, this specific sequence (NQKK) is the one conferring the worst phenotype in IN M, when all the possible combinations with 2 K in the CLA positions were tested (Figure 21). This observation raised the question of whether the amino acids occupying the CLA positions could exert the same role in group O as they do in group M. Investigating this constituted the starting point of my PhD main project. The results obtained on this matter allowed us to write and submit for publication an original work, which I inserted below. The work is followed by further results that were not included in the work mentioned, but that complete the results therein described and that constitute an interesting potential starting point for further works.

ARTICLE

Identifying phylogenetic-specific properties and origins of HIV-1 integrases M/O: a snapshot on multifunctional proteins evolution

Elenia Toccafondi[1], Marine Kanja[1], Flore Winter[1], Daniela Lener[1]* & Matteo Negroni[1, 2]*

[1]Architecture et Réactivité de l'ARN-UPR 9002, IBMC, CNRS, Université de Strasbourg, F-67000 Strasbourg, France

[2]Interdisciplinary Thematic Institute (ITI) InnoVec, Université de Strasbourg, 67000 Strasbourg, France.

*Correspondence: d.lener@ibmc-cnrs.unistra.fr, m.negroni@ibmc-cnrs.unistra.fr

ABSTRACT

Transmissions of simian viruses to humans gave rise to the different groups of HIV-1. We recently identified a functional motif (CLA), in the C-terminal domain of integrase, essential for integration in group M. Here, we found that the motif is dispensable for group O isolates, because of the presence, in their N-terminal domain of another specific motif, NOG, which is mutually interchangeable with the CLA motif. While the NOG motif is already highly conserved in the simian ancestors of group O, SIVgor, in SIVcpz*Ptt*, HIV-1 M ancestor, no conservation for the CLA motif is found, suggesting that it was selected after transmission to humans. Functional characterization of NOG-motif-containing integrases traces the mechanistic paths followed by these two viruses to ensure efficient integration, improving our understanding of the viruses evolution and of their multifunctional proteins in human infections.

KEYWORDS: integrase, integration, hiv-1, group M, group O, SIVcpz*Ptt*, SIVcpz, SIVgor, evolution, multifunctional proteins

INTRODUCTION

Transmission of viruses from animals to human is a main threat to human health, with the HIV-1 pandemic being a clear example of this. The four HIV-1 groups, in fact, all originated from an independent zoonotic transmission of simian viruses to humans. Group M and group N both derive from SIVcpz*Ptt* (Gao et al., 1999; Keele et al., 2006), while group O and P derive from SIVgor (D'Arc et al., 2015; Plantier et al., 2009). Although group M and group O share similar geographic and temporal origins (Korber et al., 2000; Lemey et al., 2004; Leoz et al., 2015), they encountered a largely different epidemiological success. While group M is the responsible for the AIDS pandemic, infecting around 39 million people all over the world, group O has a largely lower epidemiological success, infecting around 100 thousand people mostly in the west-central region of Africa (Mourez et al., 2013; Peeters et al., 1997). The bases for this discrepancy are only partially known to date, although they constitute a central question to identify critical properties allowing cross-species transmission and diffusion. Their different zoonotic origin and the subsequent sequence diversification in the human host are responsible for the large intergroup genetic diversity between groups M and O that can reach almost 50% in the *env* gene (Santoro and Perno, 2013). Despite this, they have globally convergent phenotypes and, to date, only few functional differences have been highlighted between their proteins and enzymes. Among those, the most marked one concerns the counteraction of the antiviral properties of the cellular protein tetherin, that is exerted by Vpu in HIV-1 M while it is partially carried out by Nef in the case of HIV-1 O (Bego et al., 2016; Kluge et al., 2014).

HIV replication requires the integration of the reverse transcribed genomic RNA into the genome of the infected cell. This key step is catalyzed by the integrase (IN), one of three viral enzymes. Integrases M and O share 84% of sequence identity as well as the same domain organizations and the same functions. IN is constituted by three domains connected by flexible linkers: the N-terminal domain (NTD), the catalytic core domain (CCD), and the C-terminal domain (CTD) (Engelman and Craigie, 1992; Engelman et al., 1993; van Gent et al., 1993). Each of these domains is specialized in one or more functions. The NTD is important for the multimerization and stabilization of the active form of the integrase (Eijkelenboom et al., 1997; Zheng et al., 1996), which is a highly organized multimer formed by several dimers of dimers (Passos et al., 2017, 2020). The CCD is involved in DNA binding and contains the amino acidic triad responsible for the catalytic activity of the enzyme (Kulkosky et al., 1992), but it is also the domain involved in protein dimerization and it is in charge of the interaction with LEDGF/p75, a fundamental host factor required for the successful infection by HIV-1 (Busschots et al., 2005). Finally, the CTD is involved in binding viral RNA/DNA at different

steps of the infectious cycle (Elliott et al., 2020; Engelman and Kvaratskhelia, 2022; Engelman et al., 1994; Kessl et al., 2016), and in the interaction with the viral reverse transcriptase (Wilkinson et al., 2009; Zhu et al., 2004).

It is in the CTD of the IN from group M that we previously identified a functional motif, constituted by four non-contiguous amino acids (positions 222, 240, 254, and 273) (Kanja et al., 2020). We will refer to the consensus sequence $N_{222}K_{240}N_{254}K_{273}$ of integrases M (that is the one yielding the highest levels of integration in group M while also assuring the highest levels of reverse transcription) as the "CLA (C-terminal lysine-amidic) motif" and to the same positions, irrespectively of the amino acids harbored, as the CLA positions. Despite its high conservation *in vivo*, the positions of the four residues could be permutated within the motif, in most cases, without affecting the efficiency of integration in cell culture (Kanja et al., 2020). In fact, as long as at least two lysine are present within the motif and the remainders are amidic residues (N or Q), functionality is retained to wt levels in most of the possible combinations (Kanja et al., 2020). The combination where the integration efficiency was the most affected, dropping to 25% of the wt, is NQKK. We previously determined the structure of the CTD for this variant showing that the protein folded into a structure similar to that of the wt CTD, but with a different distribution of charges at its surface (Kanja et al., 2020). Astoundingly, this exact aminoacidic sequence constitutes the consensus sequence of group O CLA positions. This observation raised the question of how such a poorly efficient sequence could have been selected in group O. Verifying this constituted the starting point of this work.

MATERIALS AND METHODS

Cell lines

HEK293T and Jurkat cells were obtained from the American Type Culture Collection (ATCC). HEK293T were cultured in DMEM while Jurkat were cultured in RPMI. Both mediums were completed with 10% fetal bovine serum and 1% PenStrep. HEK293T and Jurkat culture conditions were at 37°C in 5% CO_2.

Viral strains and sequence alignments

The primary HIV-1 isolates used in this study were: isolate HXB2 (GenBank accession number: K03455.1), isolate A2 (GenBank accession number: AF286237) from group M, subtype A2, (named "isolate M" in this study) obtained from the NIH AIDS Research and

Reference Reagent Program; isolate RBF206 (GenBank accession number: KU168298) and isolate BCF120 (GenBank accession number: KU168297) both from group O, kindly provided by J.C. Plantier (CHU Rouen, France). Isolates AF286237 and RBF206 were chosen because they were used in the work that originated the present study (Kanja et al., 2020). Isolate BCF120 was chosen as the isolate O carrying the consensus sequence in the two motifs considered in this work. The SIVcpz*Ptt* isolate employed in this work is the MB897 (GenBank accession number: EF535994) and it was chosen being one of the two isolates which are the most phylogenetically related to group M.

For the creation of the conservation logos, by using WebLogo (http://weblogo.threeplusone.com) (Crooks et al., 2004; Schneider and Stephens, 1990), we performed sequence alignments for all the sequences covering the NOG and CLA motif positions of isolates from HIV-1 group M (7,684), group O (50), SIVcpz*Ptt* (14) and SIVgor (8). All the sequences were obtained from the Los Alamos National Laboratory HIV database (https://www.hiv.lanl.gov/content/index).

Plasmid and molecular cloning

The plasmid p8.91-MB previously described (Kanja et al., 2020), was used as backbone for all cloning procedures. Therefore, all our constructs have the *gag* and the protease-coding sequences from HXB2 (group M). RT and IN coding-sequences, instead, varied. In isolate M the RT and IN was from isolate A2, while in isolates O it was either from isolate O206 or O120. In the chimpanzee isolate the RT and IN was from MB897. All IN mutant coding sequences were inserted between the *BspE*I and *Sal*I restriction sites of p8.91-MB by Gibson assembly. The plasmid used to produce the genomic RNA of the VLPs, carrying the two reporter genes used to evaluate integration efficiency (nGFP and $PURO^R$), is a modified version of the previously-described pSRP (Kanja et al., 2020) where the nuclear RFP was replaced by the nuclear GFP, giving the pSGP.

The pEUrep-RNA (da Silva Santos et al., 2016) was kindly provided by Andrea Cimarelli. The plasmid is coding for an mRNA containing the RNA packaging sequence (Ψ) and the cDNA of the Firefly luciferase followed by a polyA signal.

Two plasmids, both previously described (Kanja et al., 2020), were employed for the creation of standard curves in the quantitative PCR assays. The pJet-1LTR for the detection of late RTPs and the pGenuine2LTR for the evaluation of the 3' processing efficiency.

Transfection and VLPs collection

To produce virus-like particles (VLPs) HEK293T cells were co-transfected with the plasmid coding for the vesicular stomatitis virus glycoprotein (VSV-G) (Naldini et al., 1996), the plasmid carrying HIV-1 Gag-Pol gene (p8.91MB with different IN) and the plasmid with the modified viral genome with the reporter genes to follow the infection (pSGP). For the EURT assay the pEU-repRNA plasmid, coding for the EU-repRNA, was either added to the mix or used in the place of pSGP. All transfections were done by using 5 µg of total DNA and polyethyleneimine (PEI, Polyscience) following the manufacturer's instructions. The medium was changed after 6 h and VLPs were collected and filtered with a 0.45 µm filter after 48-72 hours. The amount of VLPs was estimated by quantifying the p24 via ELISA (Fujirebio).

Western blot analysis

The same volume of VLPs was concentrated by centrifuging them through a 20% sucrose for 2 h at 20,000 g and at 4°C. Pellets were resuspended in 3x Laemmli buffer and viral proteins were separated on a CriterionTM TGX Strain-Free 4-15% gradient gel (Bio-Rad) and then blotted on a PVDF membrane. To evaluate Pr55Gag proteolytic processing, polyproteins and mature capsid proteins were detected by probing the membrane with a mouse monoclonal anti-CA primary antibody (NIH AIDS Reagent Program) and a secondary anti-mouse HRP-conjugated antibody (Millipore). ECL reagent (Bio-Rad) was added to the membrane and images were taken with Bio-Rad Chemidoc Touch and analyzed with the Image Lab software (Bio-Rad). The Pr55Gag processing efficiency was expressed as the ratio of mature CA signal on the total CA signal (unprocessed, partially processed and fully processed CA proteins).

Quantitative PCR for viral DNA and its forms

HEK293T or Jurkat cells were transduced by spinoculation with polybrene (Sigma-Aldrich) and an amount of VLPs corresponding to a nominal MOI of 1. Prior to infection, VLPs were incubated with Benzonase nuclease (Sigma-Aldrich) to remove non-internalized DNA. 24-hours post-transduction (hpt) cells were collected, and total DNA was extracted with DNeasy Blood & Tissue Kit (QIAGEN). All qPCR assays were designed with the Taqman® hydrolysis probe technology using the IDT Primers and Probes design software (IDT), with dual quencher probes (one internal ZEN™ quencher and one 3' Iowa Black™ FQ quencher). qPCRs were performed with the iTaq Universal Probes Supermix (Bio-Rad) on a CFX96 (Bio-

Rad) thermal cycler according to the manufacturer's protocols. Standard curves and analysis were conducted with the CFX Manager (Bio-Rad).

Late reverse transcription products were quantified with oligos amplifying the U5-Psi junction. This was normalized by the amount of genomic DNA that was quantified by amplifying an exon of the actin gene. Absolute quantification was performed by creating a standard curve with known quantities of pJet-1LTR for RTPs and the genome extracted from a known quantity of cells for actin quantification.

To evaluate the 3' processing efficiency we first quantified the quantity of 2LTR circles (2LTRc) with oligos and probe annealing to the 2LTRc junction and then we evaluated the nature of this junction (perfect or imperfect, which are respectively the unprocessed and processed 3' ends) with oligos and probes annealing specifically only to the perfect junction. The imperfect junction ratio was subsequently calculated as 1-perfect junction, where 1 represents the total amount of 2LTRc. For both 2LTRc and perfect junction quantification, the same standard curve was done with pGenuine2LTR. All the oligos and probes used for the qPCR assays can be found in Table S.

Evaluation of integration

HEK293T or Jurkat cells were transduced by spinoculation with polybrene and an amount of VLPs corresponding to a nominal MOI of 0.01. 24hpt puromycin was added to HEK293T with a final concentration of 0.6 µg/ml and integration was measured by counting the puromycin-resistant clones 1-week post-transduction. As previously shown, this method is comparable to the classical Alu-gag quantitative PCR method (Kanja et al., 2020). For Jurkat cells integration was measured by FACS 72-hpt by counting the percentage of cells expressing the nGFP. This time was chosen after having established that no signal would be detected using a catalytically inactive IN (D116A), to exclude the possibility that the signal of our constructions would derive from episomal forms of the viral DNA. Since integration depends on the availability of the RTPs and since reverse transcription is affected by the viral IN, in both HEK293T and Jurkat, results were normalized by the reverse transcription efficiency evaluated by qPCR as described above. Namely, the amount X_1 of RTP was estimated for sample 1, for example. The number of puro resistant clones (P_X) for HEK cells or the number of nGFP positive cells (F_1) by FACS for Jurkat cells, was computed for the same sample. The normalized integration values were then computed as P_1/X_1 or F_1/X_1.

Assessment of the capsid stability

As described above VLPs employed in this assay contain either two RNAs, the EUrep-RNA and the SGP-RNA, or the EUrep-RNA alone. The corresponding quantity of VLPs of a nominal MOI of 0.5 was used to transduce either HEK293T or Jurkat cells. 8-hpt cell protein extract was obtained, and Luciferase assay was performed with the Luciferase Assay System (Promega). Luminescence (RLU) was normalized for protein concentration measured by the Bradford assay and therefore expressed as RLU per mg of protein extract (RLU/mg).

Structure and molecular modelling

The NTD and CTD structures of IN M show in the manuscript belong to PDB 6PUT (Passos et al., 2020). The NTD and CTD structures of IN O were obtained from molecular modelling of isolate O120 made with AlphaFold2 (Jumper et al., 2021; Varadi et al., 2022) by Patrice Gouet. Pictures used in the manuscript were obtained with PyMOL2.5.

Quantification and statistical analysis

All statistical tests were performed on at least three independent experiments (n is indicated in every figure legend) using Prism 9. ANOVA with Tukey's multiple comparisons correction was used when more than three groups were compared. An unpaired t-test was used when two samples were directly compared.

RESULTS

The CLA motif is dispensable in isolates of group O

While, as mentioned above, the influence on integration of the amino acids that occupy the CLA positions has been well characterized for group M, their effect is unknown for group O isolates. To shed light on this aspect, we used two isolates from this group, BCF120 and RBF206 (named hereafter O120 and O206, respectively) that present in the CLA positions either the consensus sequence of group O (NQKK, isolate O120, Figure 1A) or a different sequence (KQKQ, isolate O206 that was chosen as outlier). In both isolates we replaced the sequence in the CLA positions by NQNQ (O120/NQNQ and O206/NQNQ, Figure 1B and 1C), a sequence that was shown to abolish integration in isolates M (Kanja et al., 2020). In sharp contrast to what observed for group M, for both isolates the replacement of the original sequence in the CLA positions by NQNQ did not affect integration neither in HEK293T nor in

Jurkat cells (Figures 1B and 1C). The same replacement in isolate M AF286237 (referred hereafter as "isolate M"), used as a control, led to undetectable levels of integration (Figure 1D). These observations indicate that isolates O do not require the function exerted by the CLA motif or that this function is endorsed either by another region of the integrase or by another protein.

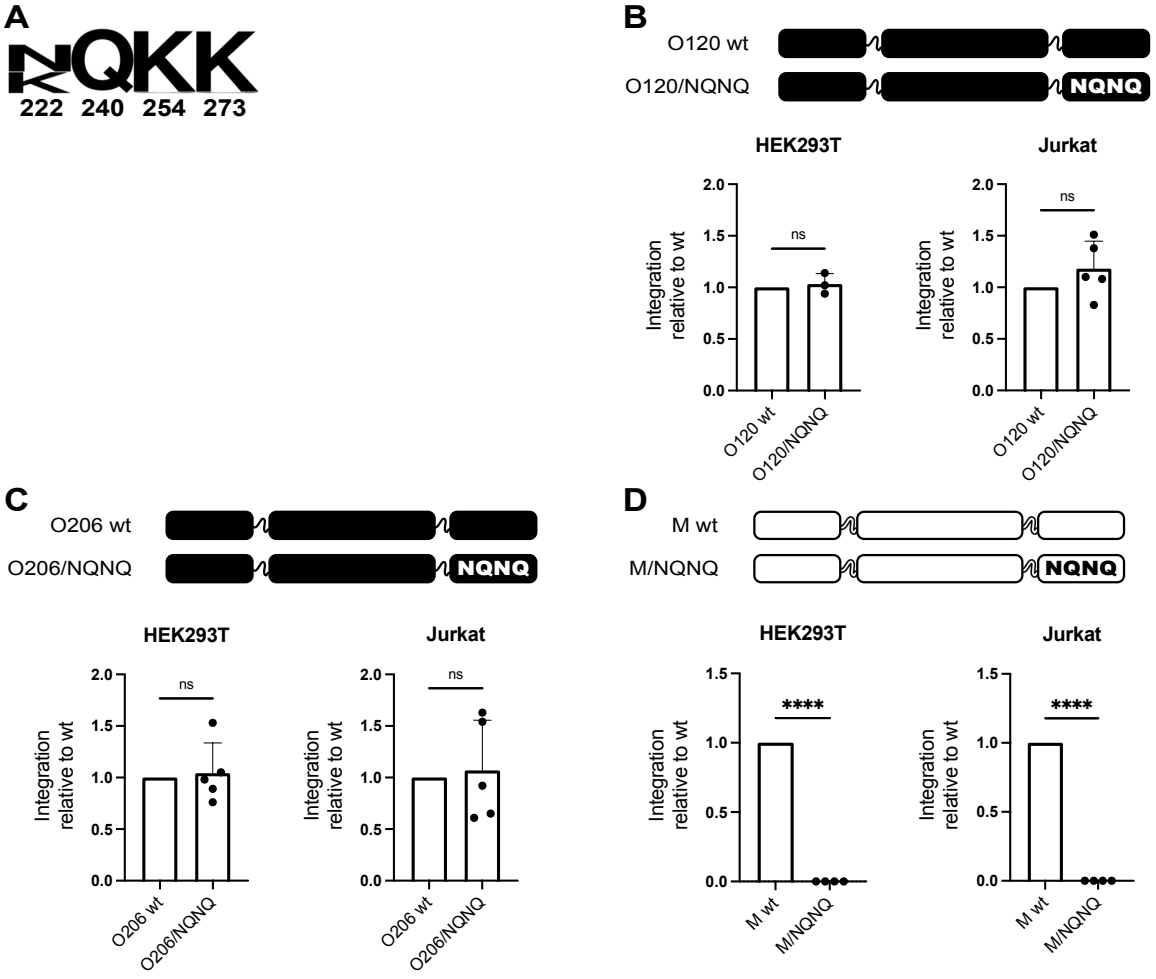


Figure 1. The CLA motif is dispensable in isolates of group O. A Sequence conservation logo of the CLA motif positions in isolates of group O. **B-D** Top of each panel: schematic representation of IN tested for integration. Color code is white for isolates M and black for isolates O. When mutated with respect to the sequence of the wt, the amino acids of the CLA motif are shown in capital letters. Bottom of each panel: normalized levels of integration relative to the level of the wt IN. **B** n=3 for HEK293T and n=5 for Jurkat. **C** n=5. **D** n=4. Data are shown as the average ± SD. ****p ≤ 0.0001. ns, not significant (two-tailed, unpaired Student's t-test).

The NTD of isolates O complements the function of the CLA motif of isolates M

We first investigated whether another region of group O integrases exerts the same function of the integrases M CLA motif. To this end, we replaced, in O206/NQNQ, five large regions

with the homologous ones of isolate M and measured integration in HEK293T (Figure 2A). In isolate M, the replacement of the NKNK sequence, the CLA motif, by NQNQ was sufficient to abolish integration (Figure 1B) indicating that no other region complements the default in the CLA motif in this group. Therefore, if a region that, in IN O206/NQNQ ensures the functions of the CLA motif is present, when it will be replaced by the homologous region of isolate M, integration should no longer occur.

Integrase is a pleiotropic protein. As such, if mutated, it can influence different steps of the infectious cycle, several of which can affect the generation of proviral DNA. Among these are reverse transcription and, when IN is still part of the Gag-Pol precursor, Pr160Gag-Pol proteolytic processing, a step required to obtain a mature infectious particle. Mutating IN could therefore affect the formation of a provirus even at various levels before integration of the pre-proviral DNA into the cell genome. For these reasons, for each mutant generated in this work, we evaluated, besides the formation of proviral DNA, the efficiency of reverse transcription and that of Pr55Gag proteolytic processing. Furthermore, if less reverse transcription products (RTPs) are produced with a mutant, less proviral DNAs will be generated even if the mutant is not affected in the step of integration itself. For this reason, to measure the efficiency of integration *per se* we expressed the levels of integration normalized by the amount of late reverse transcription products (see Material and Methods) throughout the study.

The estimates of the efficiency of integration for the chimeras shown in Figure 2A clearly indicate that integration was abolished for two of them (chimeras O206/NTD-M/NQNQ and O206/CCD1-M/NQNQ, Figure 2B), corresponding to the chimeras where either the NTD or the N-terminal part of the CCD were replaced by the homologous regions of isolate M. The effect on these two mutants was specific for integration since proteolytic processing of Pr55Gag was unaffected with respect to wt IN O206 in all chimeras as well as in O206/NQNQ (Figures S1A and S1B) while reverse transcription was reduced to approximately 60% of wt IN O206, although in a comparable manner across the chimeras (Figure 2B).

The inability of O206/NTD-M/NQNQ and O206/CCD1-M/NQNQ to produce proviral DNAs could be due to the absence of the functionality provided by the equivalent of the CLA motif or to other defects such as, for example, protein misfolding. To ascertain whether the lack of integration was related to the absence of the region that ensures the function of the CLA motif, we replaced NQNQ (non-functional CLA motif) by NKNK (functional CLA motif), obtaining chimeras O206/NTD-M/NKNK and O206/CCD1-M/NKNK (Figure 2C). We also inserted the sequence NKNK in wt IN O206 (O206/NKNK) to verify that this insertion did not

affect the functionality of the enzyme. As shown in Figure 2D, neither integration nor reverse transcription were affected in this mutant. Integration was fully restored for the chimera containing the NTD M, while it remained undetectable for O206/CCD1-M/NKNK (Figure 2D). Therefore, the default of chimera O206/NTD-M/NQNQ appears related to the lack of the region that exerts the function of the CLA motif, whereas for O206/CCD1-M/NKNK the loss of integration was unrelated to the functionality ensured by the CLA motif (Figure 2D; Figure S1B). These results indicate that the NTD of isolate O206 can complement the absence of a functional CLA motif. Furthermore, the high similarity between the NTD of O206 and the consensus sequence O (only one substitution, K46R), suggests that this is likely the case for integrases of group O in general.

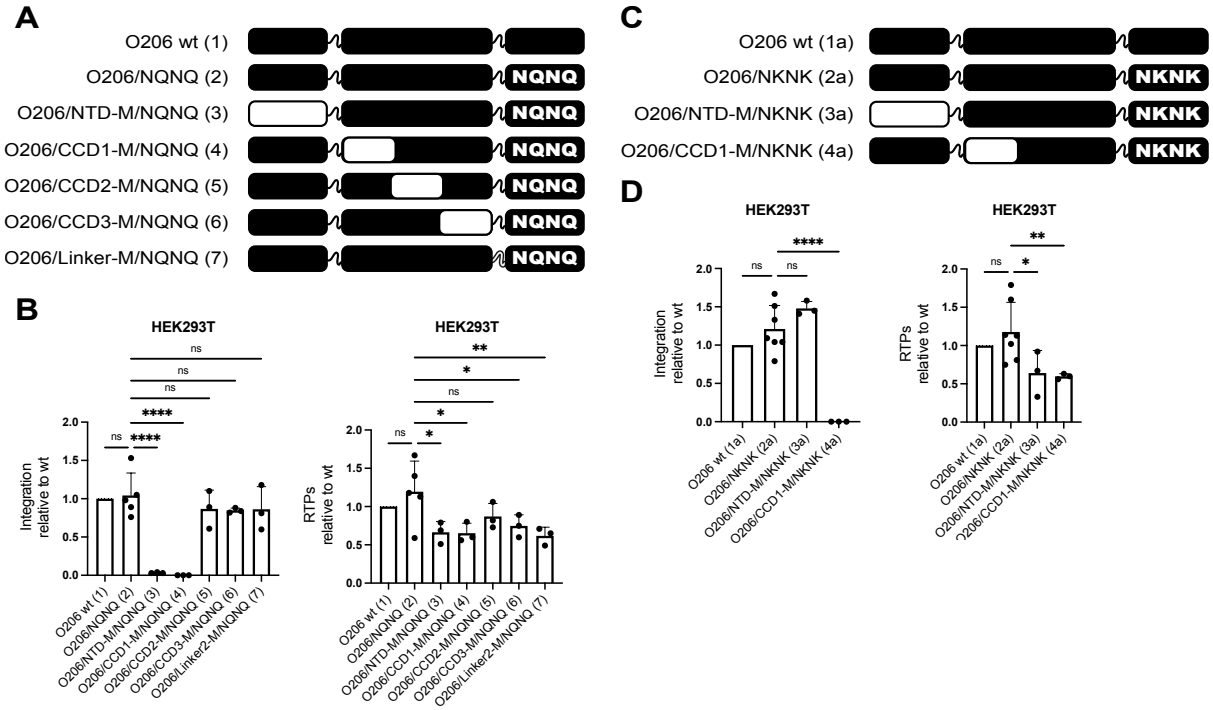


Figure 2. The NTD of isolates O complements the function of the CLA motif of isolate M. A Schematic representation of the chimeras with the NQNQ sequence in the CLA motif positions and of IN O wt, as reference at the top of the drawing. Color code is black for isolates O and white for isolates M. **B** Normalized levels of integration (left graph) and amount of RTPs (right graph), relative to the wt IN, for the chimeras shown in panel A (n=5 for O206 wt and O206/NQNQ, n=3 for the remaining samples). **C** Schematic representation of the mutants used to discern whether the loss of functionality of the two chimeras shown in panel B is related to the functionality of the CLA motif. **D** Normalized levels of integration (left graph) and amount of RTPs (right graph), relative to the wt IN, for the chimeras shown in panel C (n=7 for O206 wt and O206/NKNK, n=3 for the remaining samples). Data are shown as the average ± SD. ****p ≤ 0.0001. **p ≤ 0.01. *p ≤ 0.05. ns, not significant (one-way ANOVA with Tukey's multiple comparisons correction).

Identification and characterization of the N-terminal O group (NOG) motif

The consensus sequences of the NTDs M and O differ for 10 residues (Figure 3A). According to the score of the BLOSUM62 matrix (Henikoff and Henikoff, 1992), the replacement of four of these residues (Q7, G27, P41, H44, highlighted by a star in Figure 3A and highly conserved in group O as shown in Figure 3B) introduces more drastic changes in the properties of the protein than the substitution of the other residues. To test if the four residues $Q_7G_{27}P_{41}H_{44}$ of the NTD O are the ones allowing for the complementation of the functionality ensured by the CLA motif, we inserted them in the NTD of the IN M that harbors, in the CLA positions, the consensus sequence of isolates O (IN M/QGPH/NQKK, Figure 3C). This double mutant recovered an integration efficiency from 25% of IN M/NQKK to 100% of wt IN M, both in HEK293T and Jurkat cells (Figure 3D). The same results were obtained by replacing the whole NTD M by the NTD O (IN M/NTD-O/NQKK in Figure 3D). For both cell types, the replacement of the QGPH, also led to an improvement of reverse transcription by an approximately two-fold factor (Figure 3E) while no differences were observed in the efficiency of Pr55Gag processing for all constructions compared to the wt (Figure S3C). We refer collectively to the amino acids $Q_7G_{27}P_{41}H_{44}$ as the NOG motif, for "N-terminal O group" motif.

To understand if the NOG and the CLA motifs have an additive effect on the efficiency of integration, we then inserted the NOG motif in wt IN M (IN M/QGPH, Figure 3F). If this is the case, this IN should give levels of integration higher than wt IN M (because of the presence of the NOG motif). No improvement was instead observed for integration, nor for reverse transcription or Pr55Gag processing neither in HEK293T nor in Jurkat cells (Figures 3G and 3H; Figure S1D). All the efficiencies remained comparable between wt IN M and M/QGPH, indicating that no additive effect of the NOG and the CLA motifs.

Tracing the phylogenetic origins of the NOG and CLA motifs

To understand how these functional differences between IN M and O could have emerged, we analyzed the NOG and the CLA motifs positions in the simian viruses assumed to be the ancestors of HIV-1 M and O: SIVcpz*Ptt* and SIVgor, respectively (Figure 4A). The sequence QGPH, is highly conserved in group O and in SIVgor (Figure 4A) and it is also found in the isolate supposed to be the closest to the founder of HIV-1 O, SIVgor BQID2 (D'Arc et al., 2015) (Figure 4A). These observations strongly suggest that this motif has been selected in the great ape virus and, after transmission to humans, it has remained unaltered. In SIVcpz*Ptt* the NOG positions are highly conserved (KNDQ) and they are identical and also highly

conserved in HIV-1 M where, according to our data though, it is conserved for reasons unrelated to the functions of the QGPH sequence of group O (Figure 4A).

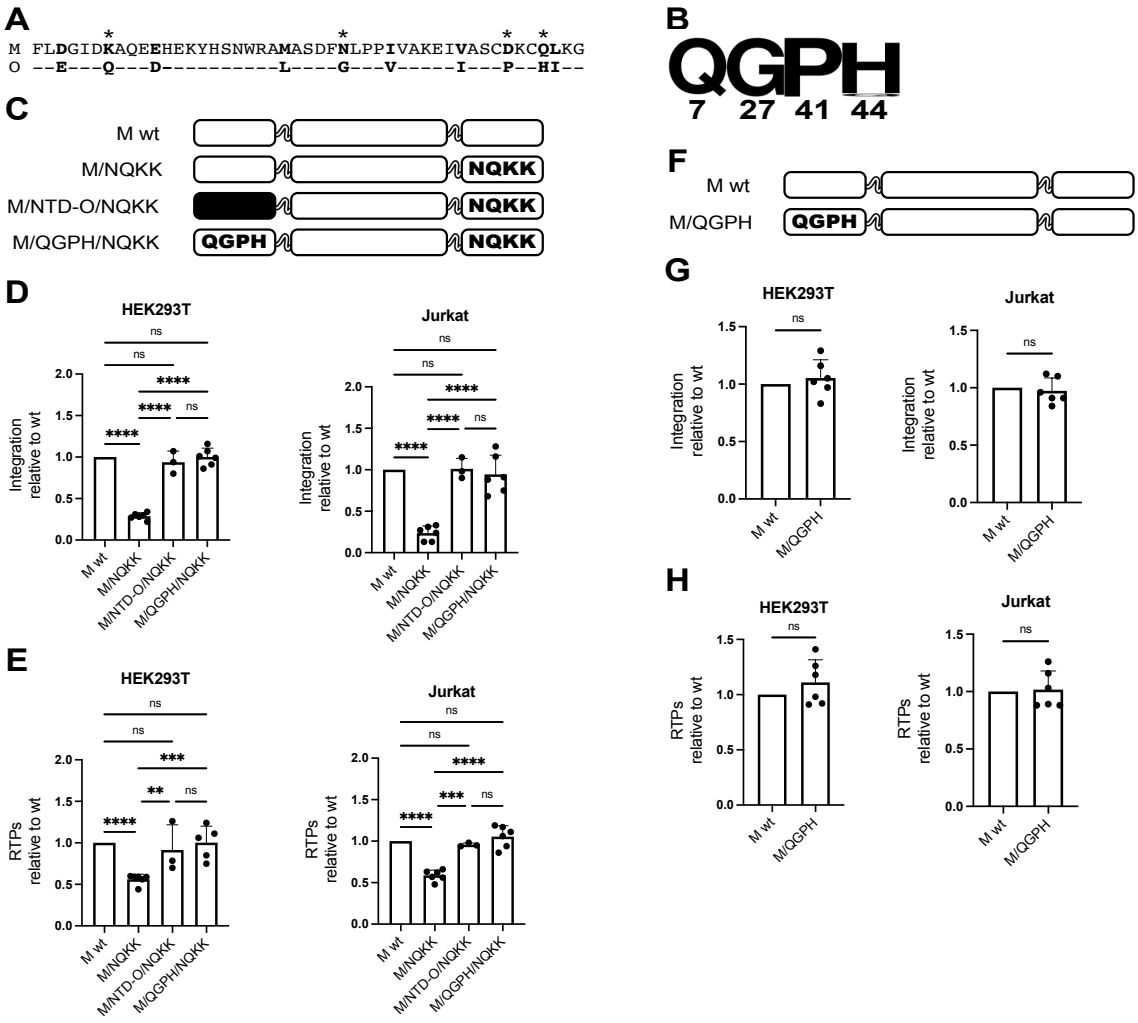


Figure 3. Identification and characterization of the N-terminal O group (NOG) motif. A Alignment of the amino acid consensus sequences of the NTD of IN M (top row) and IN O (bottom row). Unchanged amino acids in IN O with respect to IN M are indicated by a dash. Positions differing in the two sequences are in bold. Residues whose replacement gives a BLOSUM62 matrix score difference ≤ 1 are highlighted by a star. **B** Sequence conservation logo for positions 7, 14, 41 and 44 of IN O. **C** Schematic representation of the mutant IN used to evaluate the function of the NOG motif. White for isolate M and black for isolate O120. When mutated with respect to the sequence of the wt, the amino acids of the NOG or of the CLA motifs are shown in capital letters. **D** Normalized levels of integration relative to the wt IN, for the chimeras shown in panel C (n=3 for M/NTD-O/NQKK; n=6 for all the remaining samples). **E** Amounts of RTPs, relative to the wt IN, for the chimeras shown in panel C (n=3 for M/NTD-O/NQKK; n=6 for all the remaining samples). **F** Schematic representation of IN M/QGPH. (G and H) Normalized levels of integration (panel **G**) and amount of RTPs (panel **H**), relative to the wt IN, for IN M/QGPH (n=6 for all the samples). Data are shown as the average ± SD. ****p ≤ 0.0001. ***p ≤ 0.001. **p ≤ 0.01. ns, not significant (one-way ANOVA with Tukey's multiple comparisons correction for panels D and E. Two-tailed, unpaired Student's t-test for panels G and H).

Concerning the CLA positions, in SIVgor the consensus sequence is NTKK while in HIV-1 O it has been mutated to N/KQKK (Figure 4A). This could reflect adaptation to the new host, but NQKK is also the sequence of the isolate BQID2, the one assumed as the closest SIVgor to HIV-1 O (Figure 4B). Therefore, it is also possible that this sequence was transmitted to humans directly from a minor variant of the simian virus that crossed the species barrier. In the case of SIVcpz*Ptt,* although the residues found in the CLA positions are mostly K and N and therefore the same of the essential CLA motif in HIV-1 M, no conservation emerges, apart for K273 (Figure 4A). To evaluate the possibility that the NKNK sequence was, nevertheless, present in the isolate that was transferred to human and was conserved ever since, we compared the sequences found in the two isolates of SIVcpz*Ptt* that have been identified as the closest to HIV-1 M, isolates SIVcpz*Ptt* MB897 and SIVcpz*Ptt* LB715 (Heuverswyn et al., 2007). In neither case, the sequence was NKNK (Figure 4B). These observations suggest that a specific selection for this motif occurred after transmission to the human host.

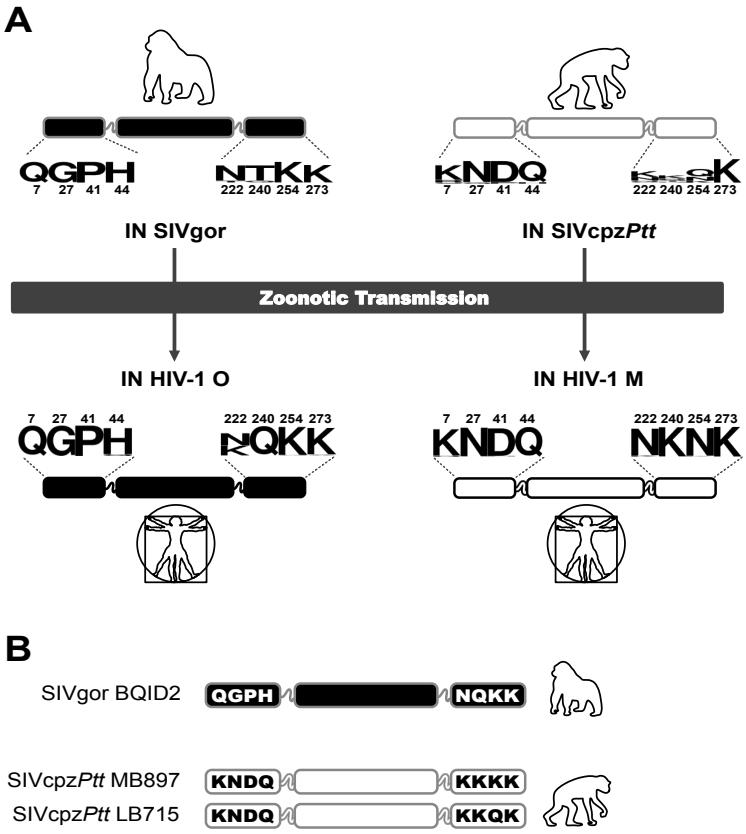


Figure 4. Tracing the phylogenetic origins of the NOG and CLA. A Sequence conservation logos of the NOG and CLA motifs in IN of groups M and O and in their ancestor viruses (SIVgor and SIVcpz*Ptt*). **B** Sequences of the isolates of SIVgor and SIVcpz*Ptt* supposed to be the closest to the isolates that were transmitted to human.

Nevertheless, we wanted to understand if a virus with an IN from SIVcpz*Ptt* but with the NKNK sequence in the CLA positions could have been infectious in human cells. To address this issue, we first produced viral particles in which we replaced the RT and IN of HIV-1 M by those of isolate SIVcpz*Ptt* MB897, hereafter simply referred to as "SIVcpz*Ptt*", that has the KKKK sequence in the CLA positions (Figure 5A). The goal was to verify that the chimeric nature of these viruses (Gag and protease from HIV-1 M; RT and IN from SIV) was not an obstacle for infection. The chimeric particles were fully processed by the protease (Figure S1E) and integration levels were twice those obtained with wt IN M (Figure 5B).

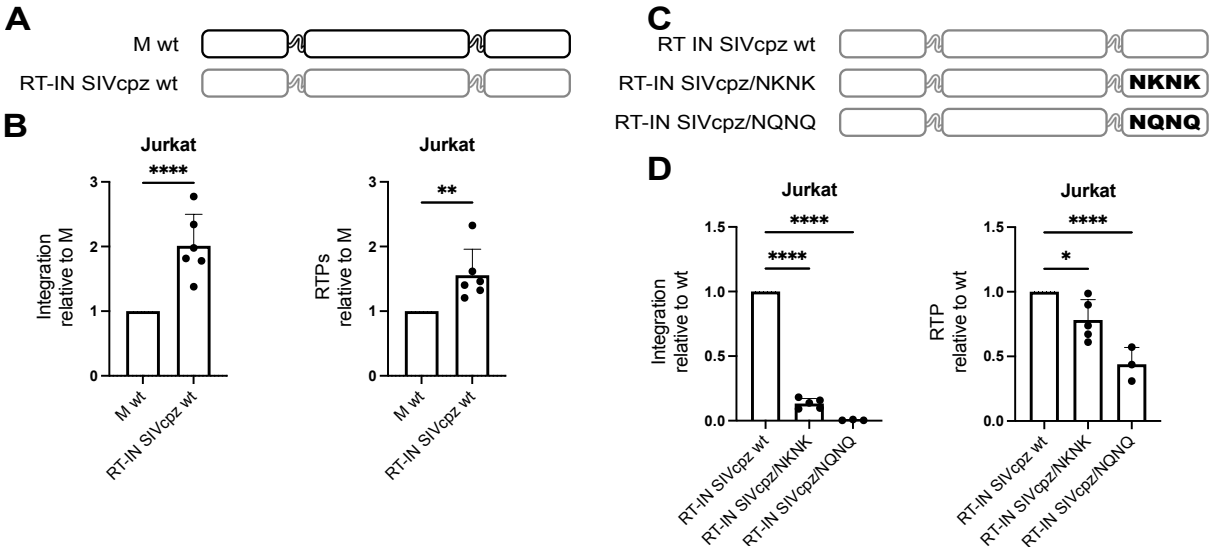


Figure 5. SIVcpzPtt CLA motif is important for integration. A Schematic representation of wt IN M and IN SIVcpz*Ptt*. **B** Normalized levels of integration (left graph) and amounts of RTPs (right graph) relative to isolate M for the IN shown in panel A (n=6). **C** Schematic representation of IN SIVcpz*Ptt* wt and the two mutants for the CLA motif positions tested for integration and reverse transcription in panel D. When mutated with respect to the sequence of the wt, the amino acids of the NOG or of the CLA motifs are shown in capital letters. **D** Normalized levels of integration (left graph) and amounts of RTPs (right graph) relative to SIVcpz*Ptt* wt (n=6 for MB897 wt; n=5 for SIVcpz*Ptt*/NKNK; n=3 for SIVcpz*Ptt*/NQNQ). Data are shown as the average ± SD. ****p ≤ 0.0001. **p ≤ 0.01. *p ≤ 0.05. (Two-tailed, unpaired Student's t-test for panel B. One-way ANOVA with Tukey's multiple comparisons correction for panel D).

Reverse transcription was also increased with respect to wt IN M (Figure 5B). In conclusion, the chimeric nature of the virus, did not affect its functionality. We therefore proceeded to replace the KKKK sequence in the CLA positions of the SIV integrase by NKNK (Figure 5C). This change was sufficient to reduce integration to around 10% with respect to wt IN SIVcpz*Ptt* (Figure 5D), while reverse transcription and Pr55Gag processing were slightly (Figure 5D) or completely unaltered (Figure S1F) respectively. This result markedly differs from what had been observed for IN M, for which the two sequences (KKKK and NKNK) yielded comparable levels of integration (Kanja et al., 2020). The replacement of the amino acids in

the CLA positions by NQNQ, condition that abolished integration in IN M, caused a drop of integration to undetectable levels, as well as a significant decrease in reverse transcription levels (Figure 5D). Pr55Gag processing levels, instead, were still unaltered (Figure S1F). Altogether, these results indicate that in SIVcpz*Ptt*, the sequence in the CLA positions is crucial to determine the levels of integration, like for IN M. In contrast with IN M, though, in SIVcpz*Ptt* the NKNK sequence did not ensure high levels of integration.

The fate of the reverse transcription products in the presence of the NOG motif

To understand by which means, in IN M, the NOG motif compensates for the lower efficiency of integration triggered by mutations in the CLA motif, we characterized the RTPs produced with the different mutants. We quantified the proportion of RTPs that were processed at their 3' end (removal, by the IN, of the terminal GT dinucleotide sequences at each 3’ end of the RTP), a prerequisite essential for integration. Replacing the NKNK sequence by NQKK in IN M (Figure 6A) reduced 3' processing by a two-fold factor both in HEK293T and Jurkat cells (Figure 6B). The replacement of the NOG motif residues of isolates M by QGPH (Figure 6A) rescued the defect, partially in HEK293T, and totally in Jurkat cells for which 3' processing was comparable to that observed for wt IN M (Figure 6B).

Finally, we looked for possible differences, with the various IN mutants used, into the process of dismantling of the capsid, since this could alter the levels of RTPs available for integration, even if equal amounts of RTPs were measured in the cell. For instance, premature uncoating can lead to the dissociation of IN from the RTPs, while closed capsid prevents the RTPs from interacting with the genome of the infected cell (Eschbach et al., 2020; Forshey et al., 2002; Stremlau et al., 2006). To this end, we used the EURT assay approach (da Silva Santos et al., 2016), in which the stability of the capsid is measured through the expression of a reporter gene carried by the VLP. The coding sequence is carried by an RNA (EUrep-RNA) that cannot be reverse transcribed but can be translated, leading to the synthesis of the firefly luciferase (Figure 6C). The experiment can be carried out copackaging with the EUrep-RNA another RNA that can be reverse transcribed (in our case "SGP" RNA, Figure 6C). In this case the luciferase signal provided by heterozygous EUrep/SGP viruses will evaluate the stability of the capsid in the presence of reverse transcription, that is the condition relevant for the present study. If only EUrep-RNA is used (Figure 6D), the assay will measure the stability of the capsid in the absence of reverse transcription.

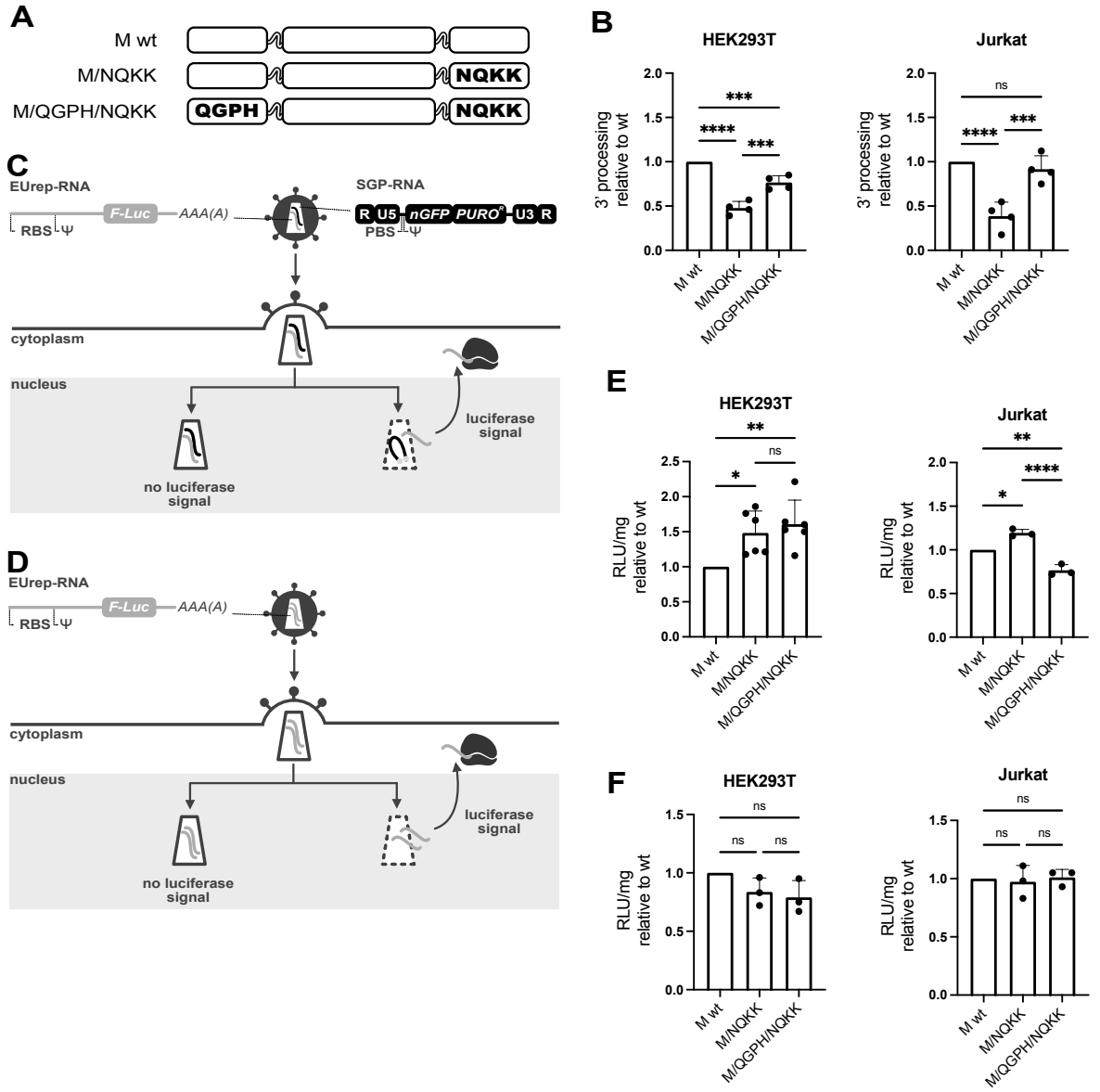


Figure 6. The fate of the reverse transcription products in the presence of the NOG motif. A Schematic representation of the IN used to evaluate the effect of the NOG motif on 3' processing. When mutated with respect to the sequence of the wt, the amino acids of the NOG or of the CLA motifs are shown in capital letters. **B** Efficiency of 3' processing, relative to the wt IN, for the mutants shown in panel A, in HEK293T and in Jurkat cells (n=4 for all samples). **C** Outline of the EURT assay, with reverse transcription, adapted from ref. 35. The two types of RNA that are co-packaged in the viral particles are shown with their essential functional features. Ψ: packaging sequence, RBS: ribosome binding site, F-luc: firefly luciferase coding sequence, AAA(A): polyA sequence, R, U5 and U3: elements of HIV-1 LTRs, PBS, HIV-1 primer binding site. The SGP-RNA also has a poly-A tale, but it is not shown for clarity, not being relevant for this experimental setting. **D** Outline of the EURT assay, without reverse transcription. **E** Luciferase expression, with reverse transcription happening inside the capsid, relative to the wt IN, for the mutants shown in panel A, in HEK293T (n=6) and in Jurkat cells (n=3). **F** Luciferase expression, without reverse transcription happening inside the capsid, relative to the wt IN, for the mutants shown in panel A, in HEK293T (n=3) and in Jurkat cells (n=3). Data are shown as average ± SD. ****p ≤ 0.0001. ***p ≤ 0.001. **p ≤ 0.01. *p ≤ 0.05. ns, not significant (one-way ANOVA with Tukey's multiple comparisons correction).

To study the stability in the presence of reverse transcription, VLPs are produced by transfection of cells that express equimolar amounts of two types of RNAs. Since the packaging and the dimerization signals are the same in the two RNAs, the resulting viral population is expected to be constituted by 50% heterozygous EUrep/SGP virions, 25% EUrep/EUrep and 25% SGP/SGP homozygous virions. While SGP/SGP viruses will not give any luciferase signal, homozygous EUrep/EUrep RNAs will interfere with the signal provided by the heterozygous EUrep/SGP particles. For this reason, the experiment was also performed in the absence of reverse transcription, to evaluate the contribution of homozygous EUrep/EUrep virions to the results obtained in the presence of reverse transcription and take this into account for the interpretation of the results.

The experiments were run using wt IN M, IN M/NQKK and IN M/QGPH/NQKK (Figure 6A) in HEK293T and in Jurkat cells. In the presence of reverse transcription (Figure 6E) the replacement of NKNK by NQKK in the CLA motif led to a modest increase in the expression of the luciferase, indicating that the mutant NQKK triggers a slight decrease of the stability of the capsid. The addition of the NOG motif had no effect in HEK293T cells (Figure 6E) while it markedly increased the stability of the capsid in Jurkat cells that became even more stable than what observed with wt IN M (Figure 6E). In the absence of reverse transcription, instead, no change in the stability of the capsid was observed among the different mutants and cells tested (Figure 6F). Therefore, the specific changes in the stability of the capsid observed in the experiment performed in the presence of reverse transcription are due to the heterozygous virions and are thus related to the ongoing reverse transcription in the viral particles.

DISCUSSION

In this work, we document that integrases of HIV-1 groups M and O have developed two phylogenetic-group specific functional motifs that can cross-complement each other. One motif (CLA) is located in the CTD of the protein of group M, the other (NOG) in the NTD of isolates of group O. This observation highlights for the first time that, depending on the phylogenetic sequence considered, two different domains of the same HIV-1 protein carry out functions that can mutually complement each other during the infectious cycle.

We previously showed that, when at least two K are present among the four amino acids that constitute the CLA motif, the positions of the individual residues can be permutated without affecting integration in eight of the ten possible combinations (Kanja et al., 2020). In the two other cases, integration was significantly reduced with the most marked decrease (to around

25% of the wt IN M) observed with the sequence $N_{222}Q_{240}K_{254}K_{273}$. Despite this, NQKK constitutes the consensus sequence of HIV-1 O, raising the question of how it could have been selected. We find here that IN O has a motif in its NTD (NOG, $Q_7G_{24}P_{41}H_{44}$) that allows to bypass the need for the CLA motif, ultimately yielding levels of integration comparable to IN M. Indeed, when in IN M, both motifs of IN O are present (M/QGPH/NQKK) the levels of integration are brought back to those of wt IN M. This is achieved by increasing the amount of reverse transcription products (RTPs) and favoring their integration by improving 3' processing. In Jurkat cells these effects were concomitant to an increase of the stability of the capsid, that could potentially favor both processes by increasing the residence time of the nucleic acids within the capsid core (Eschbach et al., 2020; Forshey et al., 2002; Stremlau et al., 2006). Dismantling of the viral capsid is a central step in the control of infectivity (Toccafondi et al., 2021). An implication of the IN in ensuring the optimal stability of the viral core by favoring the interaction between the capsid protein and cyclophilin A had been previously described (Briones et al., 2010). Reverse transcription favors dismantling of the capsid core *in vitro* and the generation of full-length RTPs has been proposed to be the main motor promoting its disassembly (Christensen et al., 2020; Rankovic et al., 2017). Our results indicate that, from the complementary standpoint, an increased stability of the capsid can favor reverse transcription.

How could such different domains have converged to ensure functions as similar as to be mutually interchangeable? The simplest explanation is that they are involved in the same mechanistic step of the infectious cycle, likely through an essential interaction with the same molecule. We showed that the three first residues of the motif ($N_{222}K_{240}N_{254}$) form a positively charged surface, absent in the case of the $N_{222}Q_{240}K_{254}$ sequence (Kanja et al., 2020). This surface was proposed to interact, possibly with the contribution of the additional K_{273}, with a repetitive, negatively charged partner (Kanja et al., 2020) (Figure 7A) as the backbone of DNA or RNA molecules. In IN O, the presence of the NOG motif is predicted to induce the formation of an alternative positively charged surface, shown in Figures 7B and 7C, which could drive the interaction to involve preferentially the NTD (Figure 7D).

The CTD of IN binds the viral RNA and altering this interaction results in dislocation of the gRNA outside the capsid, severely affecting reverse transcription (Elliott et al., 2020; Kessl et al., 2016). In this work, we evaluated the efficiency of integration of a given mutant by normalizing integration for RTPs generated with the same mutant. If the gRNA were the interacting partner of the CLA and of the NOG motifs, a decrease in reverse transcription would be observed first of all, and, after normalization, no defect in integration would be observed. This is not what we observe, though, suggesting that the function of the integrase

we highlight is a different one. This does not rule out the possibility that the interacting partner for these motifs is the gRNA, but it suggests that, if this is the case, this interaction concerns a different step than the one discussed above. The role of the interaction with the nucleic acids could mediate the control of the stability of the capsid in relation to the progression of reverse transcription, since these parameters are all modified with our mutants.

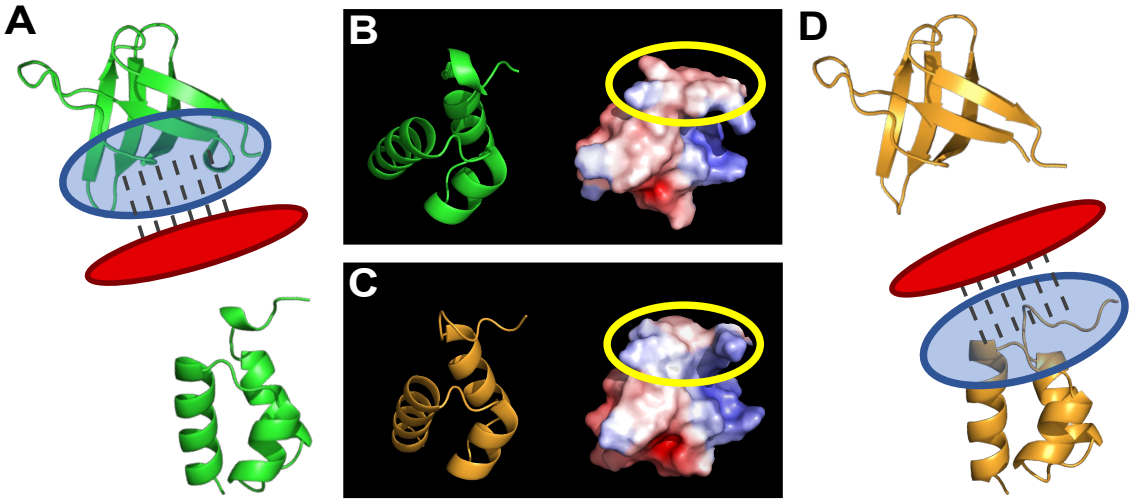


Figure 7. Model for the complementation of the CLA and NOG motifs. A CTD and NTD of IN M, PDB 6PUT (Passos et al., 2020), are shown in green, facing each other. This distribution could happen between two different IN being part of the same intasome. In this case, the CLA motif, located in the CTD, is forming a positively charged surface (in blue) that is interacting with a negatively charged partner (in red). **B**, **C** NTD M (in green) and O (in magenta) with their cartoon representation on the left, showing their orientation, and with their surface electrostatic potential on the right. The positively charged surface position exposed when the NOG motif is present is highlighted by a yellow circle. **D** CTD and NTD of IN O (obtained with AlphaFold2), are shown in magenta, facing each other. Here, the NOG motif is inducing the formation of a positively charged surface (in blue) in the NTD that will interact with a negatively charged partner (in red).

A possible scenario for the emergence of these two motifs with interchangeable functionalities is that the NOG motif was generated in the simian virus infecting gorillas (or in its ancestors) where it is highly conserved, at least within the limits of an analysis carried out on only 8 sequences available for SIVgor. The emergence of the CLA motif, instead, appears to date after transmission of the virus to the human host, since no conservation of sequence is found in this region in SIVcpz*Ptt*, although an overall trend for the presence of N and K residues (which are those that compose the consensus motif in HIV-1 M) is found. The fact that none of the combinations of these amino acids was selected in the simian virus, which instead was what occurred after transfer to humans, could indicate that the function exerted by this motif was not required in the simian cells. Otherwise, even if required, the selective pressure for the CLA motif was not as strong as it appears to be in human, allowing the co-

existence of multiple functional sequences. In both cases it is tempting to speculate that the emergence of the NKNK motif was part of the process of viral adaptation to the new host.

At this regard, the sequence NKNK in the CLA positions is found in two isolates (SIVcpz*Ptt* EK505 and Marilyn) across the population of SIVcpz*Ptt*, indicating that, *per se*, its presence is compatible with replication of simian viruses in simian cells. Therefore, NKNK could have been selected as a consensus already in the simian virus. The fact that this was not the case, instead, suggests that the advantage conferred by this motif, with respect to the other sequences in apes cells, was not as important as it is in the human ones. Thus, at least two paths could have led to the establishment of the NKNK motif in HIV-1 M: the transfer of the motif to humans by a virus that already contained it and its subsequent positive selection, or the generation of the NKNK sequence from a virus that did not initially contain the motif, followed by positive selection. When we inserted the NKNK motif in the RT-IN coding region of the SIVcpz*Ptt* and generated a chimeric HIV-1 carrying RT-IN of SIVcpz*Ptt*, integration was around 10% of that observed with its wt sequence (KKKK) (Figure 5D) in Jurkat cells. This result has been obtained using a chimerical HIV-1/SIV virus, which provides the advantage of focusing specifically on the effect of the IN and RT sequences of ape origin. However, this also raises the issue of whether the low integration levels observed could be due to the chimerical nature of the viral particles themselves. This possibility, though, appears unlikely since the same chimeric viral particles gave high levels of integration when carrying the wt sequence of SIVcpz*Ptt* IN (Figure 5B). Altogether, these results rather support the view that, once the simian virus has been transferred to humans, both sequences (RT and IN) have undergone a stepwise adaptation process to the new host that finally generated the genetic context in which the NKNK sequence in the CLA positions became optimal. The genetic flexibility that we described for the CLA locus, with several permutated sequences retaining integration ability (Kanja et al., 2020), could constitute what remains of the swarm of sequences generated by genetic drift and from which selection for the successful NKNK sequence occurred.

Noteworthy in its CLA motif positions group N also carries the NKNK sequence (Figure S2) and, also in this case, it appears from the analysis of the 11 sequences available for this virus, to be conserved in this group. The closest phylogenetically isolate of SIVcpz*Ptt* to group N is EK505, which, as mentioned above, carries the NKNK sequence. It seems therefore that for both cases of transmission of SIVcpz*Ptt* to human that resulted in the establishment of human infectious viruses (HIV-1 M and N), transmission was likely followed by selection of the NKNK sequence in the CLA motif positions, either by fixation or adaptation, as previously discussed.

This issue raised the question of why selection for the same motif did not emerge also after transfer of SIVgor to humans. In principle, the presence in this virus of an already functional motif (the NOG one) did not exclude the addition of a second motif (like the CLA one) that could have conferred a further advantage to the virus carrying both. However, here, when we tested this possibility by replacing in isolate O, the NQKK sequence by NKNK, we did not observe any increase in integration, providing a potential answer to the question. These results, together with the observation that group O does not need the CLA motif for integration, instead, are evocative of the presence of dominant epistasis of the positions that constitute the NOG motif over those making the CLA motif.

Dominant epistasis, relieving selective pressure from the CLA motif, would have allowed this region of IN O to develop, potentially, new accessory functions. Indeed, HIV-1 integrase is a multifunctional protein that, logically, acquired its diverse functions, and optimized those already acquired, progressively during evolution. Increasing evidence supports the notion that in multifunctional proteins, the initial steps toward the establishment of a new function are undertaken by genetic drift before selection for the new function is applied (Aharoni et al., 2005; Chothia et al., 2003; O'Brien and Herschlag, 1999). Intra-patient expanding HIV populations are characterized by extensive genetic drift, driven by neutral selection (Maldarelli et al., 2013), thereby creating favorable conditions for the generation of new functionalities in its proteins (Bloom et al., 2007). The presence in the CLA positions of SIVcpz*Ptt* of the same type of amino acids that would then have generated the NKNK motif in HIV-1 M, but still without selection for a consensus sequence, could constitute a snapshot of such early phases of genetic drift in the process of generation of what will then become an essential motif for integration in HIV-1 M.

In conclusion, this work sheds light on crucial aspects of the process of evolution of two phylogenetic-group specific motifs of the integrase of HIV-1, from their simian ancestors across the barrier of the zoonotic transmission to humans. By deciphering how optimization of integration is achieved in these two cases, this work contributes to improve our understanding of the rules governing viral evolution and the evolution of multifunctional viral proteins in the context of human infections.

ACKNOWLEDGMENTS

We are grateful to P. Gouet for helpful discussion. This work was supported by grants from Sidaction and the ANRS (grant ECTZ72120). E.T. was recipient of a three-years doctoral

fellowship from the ANRS and then a one-year doctoral fellowship from Sidaction (FJC-12935).

REFERENCES

Aharoni, A., Gaidukov, L., Khersonsky, O., Gould, S.M.Q., Roodveldt, C., and Tawfik, D.S. (2005). The "evolvability" of promiscuous protein functions. Nature Genetics *37*, 73–76. https://doi.org/10.1038/ng1482.

Bego, M.G., Cong, L., Mack, K., Kirchhoff, F., and Cohen, É.A. (2016). Differential Control of BST2 Restriction and Plasmacytoid Dendritic Cell Antiviral Response by Antagonists Encoded by HIV-1 Group M and O Strains. Journal of Virology *90*, 10236–10246. https://doi.org/10.1128/jvi.01131-16.

Bloom, J.D., Romero, P.A., Lu, Z., and Arnold, F.H. (2007). Neutral genetic drift can alter promiscuous protein functions, potentially aiding functional evolution. Biology Direct *2*, 17. https://doi.org/10.1186/1745-6150-2-17.

Briones, M.S., Dobard, C.W., and Chow, S.A. (2010). Role of Human Immunodeficiency Virus Type 1 Integrase in Uncoating of the Viral Core. Journal of Virology *84*, 5181–5190. https://doi.org/10.1128/jvi.02382-09.

Busschots, K., Vercammen, J., Emiliani, S., Benarous, R., Engelborghs, Y., Christ, F., and Debyser, Z. (2005). The Interaction of LEDGF/p75 with Integrase Is Lentivirus-specific and Promotes DNA Binding. Journal of Biological Chemistry *280*, 17841–17847. https://doi.org/10.1074/jbc.M411681200.

Chothia, C., Gough, J., Vogel, C., and Teichmann, S.A. (2003). Evolution of the protein repertoire. Science (1979) *300*, 1701–1703. https://doi.org/10.1126/science.1085371.

Christensen, D.E., Ganser-Pornillos, B.K., Johnson, J.S., Pornillos, O., and Sundquist, W.I. (2020). Reconstitution and visualization of HIV-1 capsid-dependent replication and integration in vitro. Science (1979) *370*. https://doi.org/10.1126/science.abc8420.

Crooks, G.E., Hon, G., Chandonia, J.-M., and Brenner, S.E. (2004). WebLogo: A Sequence Logo Generator: Figure 1. Genome Research 14, 1188–1190. https://doi.org/10.1101/gr.849004.

D'Arc, M., Ayouba, A., Esteban, A., Learn, G.H., Boué, V., Liegeois, F., Etienne, L., Tagg, N., Leendertz, F.H., Boesch, C., et al. (2015). Origin of the HIV-1 group O epidemic in western lowland gorillas. Proc Natl Acad Sci U S A *112*, E1343–E1352. https://doi.org/10.1073/pnas.1502022112.

Eijkelenboom, A.P.A.M., van den Ent, F.M.I., Vos, A., Doreleijers, J.F., Hård, K., Tullius, T.D., Plasterk, R.H.A., Kaptein, R., and Boelens, R. (1997). The solution structure of the amino-terminal HHCC domain of HIV-2 integrase: a three-helix bundle stabilized by zinc. Current Biology *7*, 739–746. https://doi.org/10.1016/S0960-9822(06)00332-0.

Elliott, J., Eschbach, J.E., Koneru, P.C., Li, W., Puray-Chavez, M., Townsend, D., Lawson, D., Engelman, A.N., Kvaratskhelia, M., and Kutluay, S.B. (2020). Integrase-RNA interactions underscore the critical role of integrase in HIV-1 virion morphogenesis. Elife *9*, 1–56. https://doi.org/10.7554/ELIFE.54311.

Engelman, A., and Craigie, R. (1992). Identification of conserved amino acid residues critical for human immunodeficiency virus type 1 integrase function in vitro. Journal of Virology *66*, 6361–6369. https://doi.org/10.1128/jvi.66.11.6361-6369.1992.

Engelman, A.N., and Kvaratskhelia, M. (2022). Multimodal Functionalities of HIV-1 Integrase. Viruses *14*, 926. https://doi.org/10.3390/v14050926.

Engelman, A., Bushman, F.D., and Craigie, R. (1993). Identification of discrete functional domains of HIV-1 integrase and their organization within an active multimeric complex. EMBO Journal *12*, 3269–3275. https://doi.org/10.1002/j.1460-2075.1993.tb05996.x.

Engelman, A., Hickman, A.B., and Craigie, R. (1994). The core and carboxyl-terminal domains of the integrase protein of human immunodeficiency virus type 1 each contribute to nonspecific DNA binding. Journal of Virology *68*, 5911–5917. https://doi.org/10.1128/jvi.68.9.5911-5917.1994.

Eschbach, J.E., Elliott, J.L., Li, W., Zadrozny, K.K., Davis, K., Mohammed, S.J., Lawson, D.Q., Pornillos, O., Engelman, A.N., and Kutluay, S.B. (2020). Capsid Lattice Destabilization Leads to Premature Loss of the Viral Genome and Integrase Enzyme during HIV-1 Infection. Journal of Virology *95*. https://doi.org/10.1128/JVI.00984-20.

Forshey, B.M., von Schwedler, U., Sundquist, W.I., and Aiken, C. (2002). Formation of a Human Immunodeficiency Virus Type 1 Core of Optimal Stability Is Crucial for Viral Replication. Journal of Virology *76*, 5667–5677. https://doi.org/10.1128/jvi.76.11.5667-5677.2002.

Gao, F., Bailes, E., Robertson, D.L., Chen, Y., Rodenburg, C.M., Michael, S.F., Cummins, L.B., Arthur, L.O., Peeters, M., Shaw, G.M., et al. (1999). Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. Nature *397*, 436–441. https://doi.org/10.1038/17130.

van Gent, D.C., Vink, C., Groeneger, A.A., and Plasterk, R.H. (1993). Complementation between HIV integrase proteins mutated in different domains. The EMBO Journal *12*, 3261–3267. https://doi.org/10.1002/j.1460-2075.1993.tb05995.x.

Henikoff, S., and Henikoff, J.G. (1992). Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A *89*, 10915–10919. https://doi.org/10.1073/pnas.89.22.10915.

Heuverswyn, F. van, Li, Y., Bailes, E., Neel, C., Lafay, B., Keele, B.F., Shaw, K.S., Takehisa, J., Kraus, M.H., Loul, S., et al. (2007). Genetic diversity and phylogeographic clustering of SIVcpzPtt in wild chimpanzees in Cameroon. Virology *368*, 155–171. https://doi.org/10.1016/j.virol.2007.06.018.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589. https://doi.org/10.1038/s41586-021-03819-2.

Kanja, M., Cappy, P., Levy, N., Oladosu, O., Schmidt, S., Rossolillo, P., Winter, F., Gasser, R., Moog, C., Ruff, M., et al. (2020). NKNK: a New Essential Motif in the C-Terminal Domain of HIV-1 Group M Integrases. Journal of Virology *94*, 1–23. https://doi.org/10.1128/JVI.01035-20.

Keele, B.F., van Heuverswyn, F., Li, Y., Bailes, E., Takehisa, J., Santiago, M.L., Bibollet-Ruche, F., Chen, Y., Wain, L. v., Liegeois, F., et al. (2006). Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. Science (1979) *313*, 523–526. https://doi.org/10.1126/science.1126531.

Kessl, J.J., Kutluay, S.B., Townsend, D., Rebensburg, S., Slaughter, A., Larue, R.C., Shkriabai, N., Bakouche, N., Fuchs, J.R., Bieniasz, P.D., et al. (2016). HIV-1 Integrase Binds the Viral RNA Genome and Is Essential during Virion Morphogenesis. Cell *166*, 1257-1268.e12. https://doi.org/10.1016/j.cell.2016.07.044.

Kluge, S.F., Mack, K., Iyer, S.S., Pujol, F.M., Heigele, A., Learn, G.H., Usmani, S.M., Sauter, D., Joas, S., Hotter, D., et al. (2014). Nef Proteins of Epidemic HIV-1 Group O Strains Antagonize Human Tetherin. Cell Host & Microbe *16*, 639–650. https://doi.org/10.1016/j.chom.2014.10.002.

Korber, B., Muldoon, M., Theiler, J., Gao, F., Gupta, R., Lapedes, A., Hahn, B.H., Wolinsky, S., and Bhattacharya, T. (2000). Timing the Ancestor of the HIV-1 Pandemic Strains. Science (1979) *288*, 1789–1796. https://doi.org/10.1126/science.288.5472.1789.

Kulkosky, J., Jones, K.S., Katz, R.A., Mack, J.P., and Skalka, A.M. (1992). Residues critical for retroviral integrative recombination in a region that is highly conserved among retroviral/retrotransposon integrases and bacterial insertion sequence transposases. Molecular and Cellular Biology *12*, 2331–2338. https://doi.org/10.1128/mcb.12.5.2331-2338.1992.

Lemey, P., Pybus, O.G., Rambaut, A., Drummond, A.J., Robertson, D.L., Roques, P., Worobey, M., and Vandamme, A.M. (2004). The molecular population genetics of HIV-1 group O. Genetics *167*, 1059–1068. https://doi.org/10.1534/genetics.104.026666.

Leoz, M., Feyertag, F., Kfutwah, A., Mauclère, P., Lachenal, G., Damond, F., de Oliveira, F., Lemée, V., Simon, F., Robertson, D.L., et al. (2015). The Two-Phase Emergence of Non Pandemic HIV-1 Group O in Cameroon. PLoS Pathogens *11*, 1–13. https://doi.org/10.1371/journal.ppat.1005029.

Maldarelli, F., Kearney, M., Palmer, S., Stephens, R., Mican, J., Polis, M.A., Davey, R.T., Kovacs, J., Shao, W., Rock-Kress, D., et al. (2013). HIV Populations Are Large and Accumulate High Genetic Diversity in a Nonlinear Fashion. Journal of Virology *87*, 10313–10323. https://doi.org/10.1128/JVI.01225-12.

Mourez, T., Simon, F., and Plantiera, J.C. (2013). Non-M variants of human immunodeficiency virus type. Clinical Microbiology Reviews *26*, 448–461. https://doi.org/10.1128/CMR.00012-13.

Naldini, L., Blömer, U., Gallay, P., Ory, D., Mulligan, R., Gage, F.H., Verma, I.M., and Trono, D. (1996). In vivo gene delivery and stable transduction of nondividing cells by a lentiviral vector. Science (1979) 272, 263–267. https://doi.org/10.1126/science.272.5259.263.

O'Brien, P.J., and Herschlag, D. (1999). Catalytic promiscuity and the evolution of new enzymatic activities. Chemistry and Biology *6*. https://doi.org/10.1016/S1074-5521(99)80033-7.

Passos, D.O., Li, M., Yang, R., Rebensburg, S. v., Ghirlando, R., Jeon, Y., Shkriabai, N., Kvaratskhelia, M., Craigie, R., and Lyumkis, D. (2017). Cryo-EM structures and atomic model of the HIV-1 strand transfer complex intasome. Science (1979) *355*, 89–92. https://doi.org/10.1126/science.aah5163.

Passos, D.O., Li, M., Jóźwik, I.K., Zhao, X.Z., Santos-Martins, D., Yang, R., Smith, S.J., Jeon, Y., Forli, S., Hughes, S.H., et al. (2020). Structural basis for strand-transfer inhibitor binding to HIV intasomes. Science (1979) *367*, 810–814. https://doi.org/10.1126/science.aay8015.

Peeters, M., Gueye, A., Mboup, S., Bibollet-Ruche, F., Ezaka, E., Mulanga, C., Ouedrago, R., Regine, Gandji., Mpele, P., Dibanga, G., et al. (1997). Geographical distribution of HIV-1 group O viruses in Africa. AIDS *11*, 493–498. .

Plantier, J.C., Leoz, M., Dickerson, J.E., de Oliveira, F., Cordonnier, F., Lemée, V., Damond, F., Robertson, D.L., and Simon, F. (2009). A new human immunodeficiency virus derived from gorillas. Nature Medicine *15*, 871–872. https://doi.org/10.1038/nm.2016.

Rankovic, S., Varadarajan, J., Ramalho, R., Aiken, C., and Rousso, I. (2017). Reverse Transcription Mechanically Initiates HIV-1 Capsid Disassembly. Journal of Virology *91*, 1–14. https://doi.org/10.1128/jvi.00289-17.

Santoro, M.M., and Perno, C.F. (2013). HIV-1 Genetic Variability and Clinical Implications. ISRN Microbiology *2013*, 1–20. https://doi.org/10.1155/2013/481314.

da Silva Santos, C., Tartour, K., and Cimarelli, A. (2016). A Novel Entry/Uncoating Assay Reveals the Presence of at Least Two Species of Viral Capsids During Synchronized HIV-1 Infection. PLoS Pathogens *12*. https://doi.org/10.1371/journal.ppat.1005897.

Schneider, T.D., and Stephens, R.M. (1990). Sequence logos: a new way to display consensus sequences. Nucleic Acids Research 18, 6097–6100.

Stremlau, M., Perron, M., Lee, M., Li, Y., Song, B., Javanbakht, H., Diaz-Griffero, F., Anderson, D.J., Sundquist, W.I., and Sodroski, J. (2006). Specific recognition and accelerated uncoating of retroviral capsids by the TRIM5 restriction factor. Proceedings of the National Academy of Sciences *103*, 5514–5519. https://doi.org/10.1073/pnas.0509996103.

Toccafondi, E., Lener, D., and Negroni, M. (2021). HIV-1 Capsid Core: A Bullet to the Heart of the Target Cell. Frontiers in Microbiology *12*, 1–17. https://doi.org/10.3389/fmicb.2021.652486.

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., et al. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Research 50, D439–D444. https://doi.org/10.1093/nar/gkab1061.

Wilkinson, T.A., Januszyk, K., Phillips, M.L., Tekeste, S.S., Zhang, M., Miller, J.T., le Grice, S.F.J., Clubb, R.T., and Chow, S.A. (2009). Identifying and Characterizing a Functional HIV-1 Reverse Transcriptase-binding Site on Integrase. Journal of Biological Chemistry *284*, 7931–7939. https://doi.org/10.1074/jbc.M806241200.

Zheng, R., Jenkins, T.M., and Craigie, R. (1996). Zinc folds the N-terminal domain of HIV-1 integrase, promotes multimerization, and enhances catalytic activity. Proceedings of the National Academy of Sciences *93*, 13659–13664. https://doi.org/10.1073/pnas.93.24.13659.

Zhu, K., Dobard, C., and Chow, S.A. (2004). Requirement for Integrase during Reverse Transcription of Human Immunodeficiency Virus Type 1 and the Effect of Cysteine Mutations of Integrase on Its Interactions with Reverse Transcriptase. Journal of Virology *78*, 5045–5055. https://doi.org/10.1128/jvi.78.10.5045-5055.2004.

SUPPLEMENTARY MATERIAL

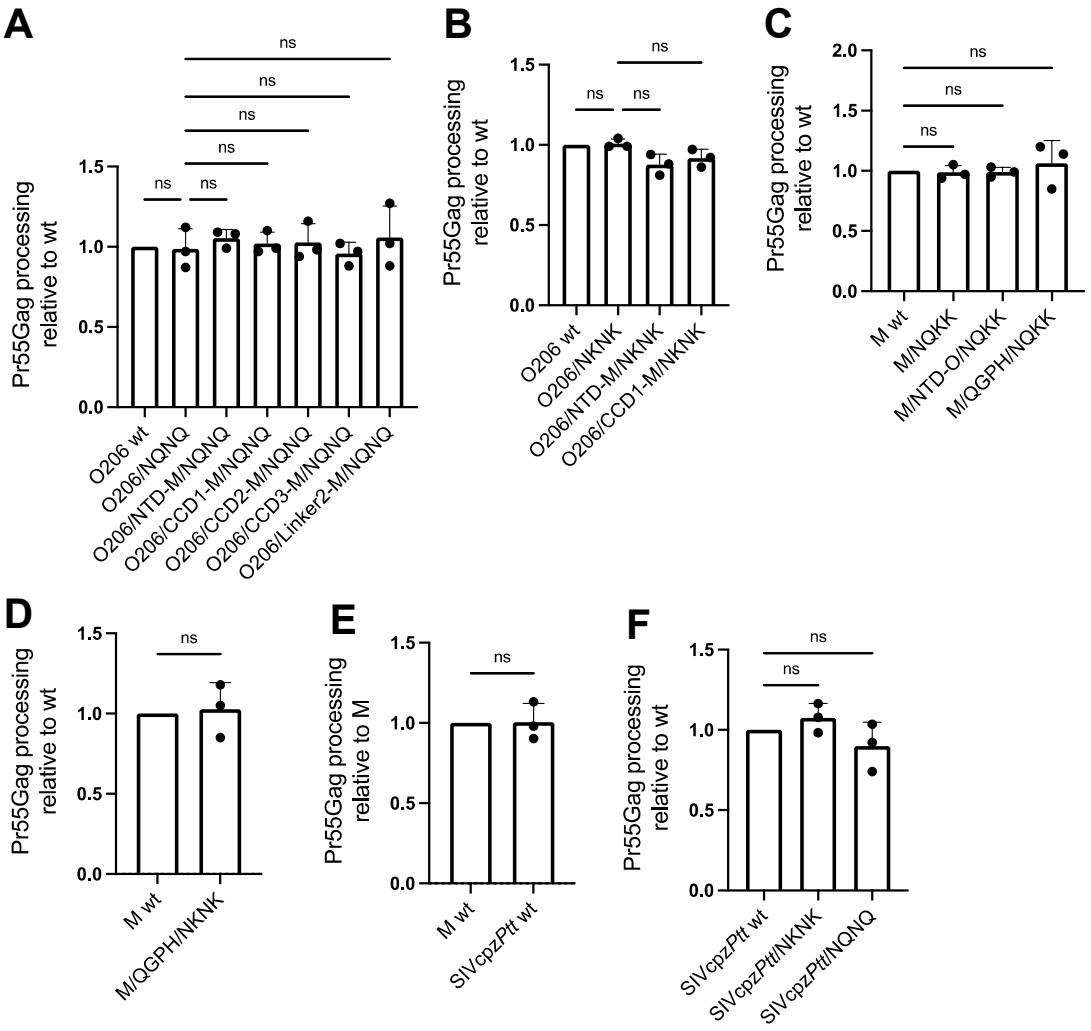


Figure S1. Pr55Gag processing of IN tested in this work was not affected. A Results for Pr55Gag processing for the constructions shown in figure 2A. Pr55Gag is not affected for all the constructions tested (n=3). **B** Results for Pr55Gag processing for the constructions shown in figure 2C. Pr55Gag is not affected for all the constructions tested (n=3). **C** Results for Pr55Gag processing for the constructions shown in figure 3C. Pr55Gag is not affected for all the constructions tested (n=3). **D** Results for Pr55Gag processing for the constructions shown in figure 3F. Pr55Gag is not affected for all the constructions tested (n=3). Data are shown as the average ± SD. ns, not significant (one-way ANOVA with Tukey's multiple comparisons correction).

Figure S2. Group N CLA motif has the same amino acidic sequence found in group M. A Sequence conservation logo of the CLA motif in isolates of group N. Obtained with WebLogo by aligning 11 sequences.

Target	Oligos/probe	Sequence	Fluorophore
Late RTPs	U5Psi fwd	GTGACTCTGGTAACTAGAGA	
	U5Psi probe	CGCTTTCAAGTCCCTGTTCGGG	FAM
	U5psi rev	GAGAGCTCCTCTCCTTTC	
Genomic DNA (ACTB)	IDT Assay ID: Hs.PT.56a.40703009.g		HEX
2LTR circles	2LTR fwd	CCCTTTTAGTCAGTGTGGAA	
	2LTR probe	TTCACTCCCAACGAAGACAAGATATCCTT	FAM
	2LTR rev	GTAGCCTTGTGTGTGGTAGA	
Perfect junctions	2LTR PJ fwd	TGTGGAAAAATCTCTAGCAGTAC	
	2LTR probe	TTCACTCCCAACGAAGACAAGATATCCTT	FAM
	2LTR rev	GTAGCCTTGTGTGTGGTAGA	

Table S1. Oligos used for qPCR.

RESULTS NOT INCLUDED IN THE ARTICLE

The NOG and the CLA motifs are mostly conserved within HIV-1 and SIV phylogenetic groups

The phylogenetic analyses presented in the manuscript are strictly focused on HIV-1 groups M and O and their ancestor simian viruses, infecting chimpanzees *Pan troglodytes troglodytes* and gorillas. In the manuscript, we show that, with one exception, the NOG and CLA positions are highly conserved within HIV-1 groups M and O, as well as in SIVcpz*Ptt* and SIVgor, even if the consensus sequences identified are not necessarily the same in the four cases. In one case, though (the CLA motif of SIVcpz*Ptt*) no conservation of sequence could be found. We also observed that the sequence of the eleven isolates of HIV-1 group N was identical to the consensus sequence of HIV-1 M, both in the CLA and in the NOG motifs (Figure 23). Similarly, the NOG motif of group P (three isolates) had the same sequence of SIVgor and HIV-1 group O (Figure 23). For the CLA motif, while HIV-1 O and SIVgor differ for the second amino acid of the motif, HIV-1 P matched the consensus of the simian viruses (NTKK) instead of that of HIV-1 O (NQKK) (Figure 23). Interpretations and discussions on HIV-1 P are obviously limited by the small number of samples available. Considering these limitations, data from groups N and P have not been included in the manuscript.

However, we reasoned that besides the cases mentioned above (included in the manuscript submitted for publication), a broader analysis of these motifs, including other SIV, will improve our understanding of their role in evolution of human and simian viruses in the context of cross-species transmission. We therefore extended our analyses to three other simian viruses: SIVrcm, SIVmnd-2 and SIVcpz*Pts*. SIVrcm and SIVmnd-2 are equally considered to be the ancestors viruses of the 3' *pol* portion of SIVcpz*Ptt*, therefore including the IN ORF. The sequences available for SIVrcm and SIVmnd-2 for the NOG and the CLA positions differ in their number and are specified for each position in their conservation logos (Figure 23). At the CLA positions, distinct levels of conservation are found among the four positions in both viruses. In SIVrcm, the first and the last positions are conserved, while the middle ones show no conservation (Figure 23), while SIVmnd-2 shows a fully conserved sequence among the 4 isolates available. The two central positions of both viruses are occupied by amino acids (E, G) significantly different compared to those found in the CLA motif of group M. However, it is interesting to note that, for both viruses, the first amino acid is identical to the one found in group M CLA motif, while the last, is the same one (K) for SIVmnd-2, and the other positively charged amino acid (R) for SIVrcm (Figure 23). Interestingly, the laboratory showed that, in IN

M, the two K of the motif could be replaced by R without affecting the efficiency of integration (Figure 20). It is therefore justified to expect that, in the context of SIVrcm, the R plays the same role of the corresponding K in the other viruses. Noteworthy, SIVrcm is the only lentivirus, among the ones studied so far, that does not carry a K in position 273. At SIVrcm NOG positions a relatively conserved sequence, with amino acids similar to those found in both M and O, is present. In SIVmnd-2 the last two NOG positions are highly conserved whereas the first two are less conserved, but yet occupied by recurrent amino acids, and overall, the sequence shows similarities with both HIV-1 group M and O motifs.

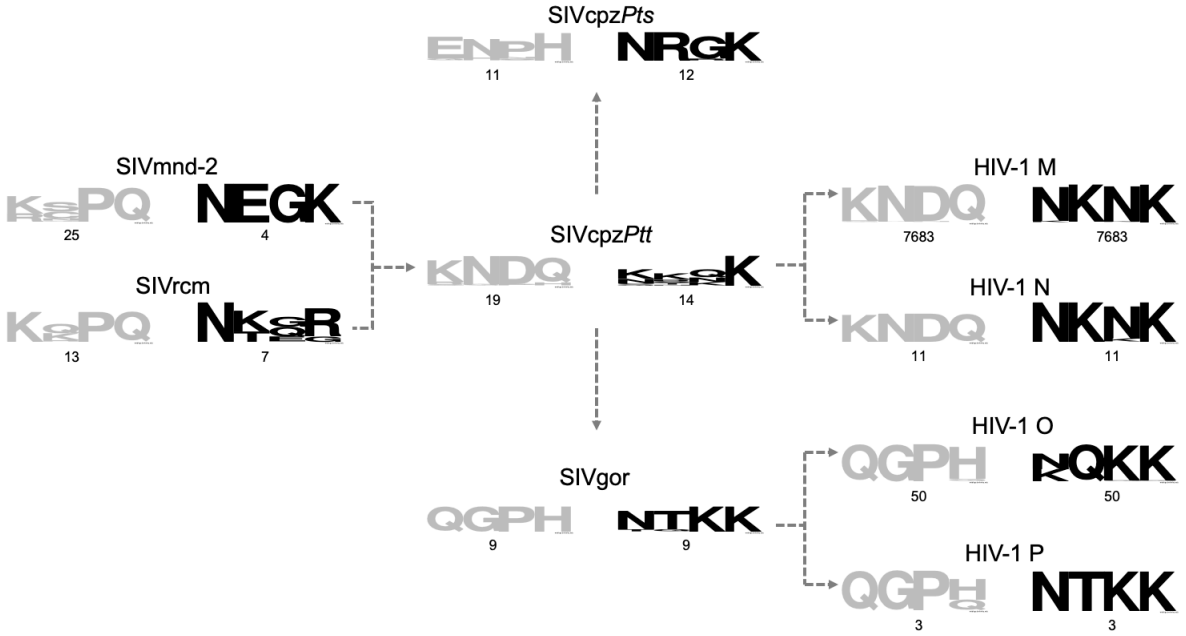


Figure 23. CLA and NOG motif phylogenetic history. Conservation logos for NOG positions (7, 27, 41, 44) and for CLA positions (222, 240, 254, 273) are shown in grey and black respectively. Under each logo is indicated the number of sequences used to obtain it. Dashed arrows represent the possible zoonotic transmissions that originated each virus (or the *pol* region of the genome as it is the case of SIVcpz*Ptt*).

Finally, we wanted to check the sequences of the other SIV infecting chimpanzees, SIVcpz*Pts*. No transmission to humans from this virus has been described so far. This could be due to trivial factors as different ways of living of these animals and, in general, a more limited promiscuity with humans. However, this could also suggest that the cross-species barrier to cross is higher than the one separating *Pan troglodytes troglodytes* from humans. At the CLA positions, the conserved sequence NRGK is deduced from the 5 sequences available (Figure 23). Except for G254, once again the amino acids found in these positions are constituted by amidic polar and positively charged residues. At the NOG positions, a sequence fairly conserved is present, constituted by amino acids found in other NOG motifs analyzed (N, P and H), but, unique for this virus, we also find an acidic residue (E) occupying

the first position where usually an amidic or a positively charged amino acid is present (Figure 23).

The amino acids found at the two motifs (NOG and CLA) are globally conserved in most the phylogenetic groups considered, and it is indicative of a positive selection for their conservation within each clade. The exceptions constituted by the CLA motifs of SIVcpz*Ptt* and SIVrcm raise the issue of why selective pressure should not be effective in these hosts. It is tempting to propose that dominant intragenic epistasis (from another domain of the protein) could have released the functional constraints that forced sequence conservation in this region, as we have proposed to be the case for the *nog* locus over the *cla* one in group O. If this were the case, the nature of the epistatic interaction between the two domains, should be different in the variants carried by these two apes viruses, specifically. It would be interesting and possible at the same time, to test this hypothesis experimentally, the moment a sufficient number of these isolates will be sampled and characterized.

At the same time, though, if it is true that the sequences were conserved within a phylogenetic group, it is also true that they were rarely the same when comparing the phylogenetic groups between them. This can indicate that, in the various hosts, the requirements were so different as to lead to the sequence divergences that are now observed in these two motifs among the various types of viruses. However, this can also be due to coevolution as an indirect consequence of the adaptation of the virus to the new host.

Indeed, if a given sequence (X) is selected in a virus in response to a host-specific requirement, after cross-species transmission three possible scenarios might occur: (1) it might become dispensable (if no such requirement exists in the new host); (2) it might be strictly preserved (if he same requirement exists in the new host); (3) it might be adapted to new requirement (if a similar but not identical requirement is present in the new host). In case 1, with time X will be likely heavily mutated, potentially acquiring new functions, becoming too different from the original X to be identified as a related sequence or it might be even deleted. In case 3, depending on how different the requirements in the new host are, the phylogenetic relationship between the two sequences can or cannot still be inferred once optimization is completed. Whichever of the three cases mentioned above one considers, though, the rest of the genome (i.e. outside locus *x*) will almost certainly undergo adaptation in response to various physiological differences between the two hosts, eventually leading to the generation of a new virus specific for the new host. During this process of adaptation, mutations at the site X could be selected because of coevolution requirements needed to preserve infectivity in response to mutations occurring outside X itself. This will lead to a

modification of X, also in the case 2, where its strict preservation was initially expected. Therefore, distinguishing between mutations directly related to the functionality of a motif from those that constitute “side effects” of coevolution, is important for identifying host specific requirements that allow or hamper transmission to a new host. The set of sequence identified in this work in the CLA and NOG motifs and their functional relationship provide, in our opinion, a precious material for such studies.

Indeed, based on our results the NOG motif of SIVgor and HIV-1 groups O and P seems to have found the same functional requirements in both the gorilla and the human host, and its sequence consequently remained unchanged. Furthermore, the coevolution that led to the adaptation in the new host did not affect this sequence. On the contrary the CLA motif of SIVcpz*Ptt* and HIV-1 M was most likely affected by coevolution, as our results of the SIV IN carrying the NKNK sequence at its CLA positions are showing. This mutant, in fact, although carrying the optimal group M CLA motif sequence, was not able to replicate in human cells, suggesting that further adaptation in, or even outside, the IN was required to select the NKNK sequence in group M. Further investigations are needed in order to understand whether the functional requirements in chimpanzee and in human were the same or not for the CLA locus. Indeed, although no conservation is found at the CLA positions of SIVcpz*Ptt*, amino acids with the same characteristics (positive and amidic), if not exactly the same amino acids, are recurrently found among its isolates, suggesting that selective pressure was present for these residues.

The O CLA motif when inserted in IN M increases the levels of 2LTRc

Among the different evaluations employed to better characterize the phenotype of the mutants tested in this work, 2LTRc levels were also used. 2LTRc are usually used as a marker of nuclear import of the RTPs, since they can only be formed by cellular nuclear enzymes, as well as a marker of integration efficiency. In fact, since 2LTRc and the integration product are formed starting from the same substrate (the linear proviral DNA), an increase of 2LTRc amounts when integration is impaired is expected to be observed. This trend was indeed found when 2LTRc levels of the M/NQKK mutant (25% of integration relative to M wt) were measured, being 2-fold higher than those of the wt, either in HEK293T or Jurkat cells (Figure 24A). In the same way, the mutant M/QGPH (integration 100% relative to M wt) showed to have similar levels of 2LTRc than the wt, as expected from its integration levels (Figure 24A). Strikingly, the double mutant M/QGPH/NQKK (integration 100% relative to M wt) also had

significantly higher levels of 2LTRc in both HEK293T and Jurkat cells (Figure 24A), despite the observed rescue in the integration efficiency.

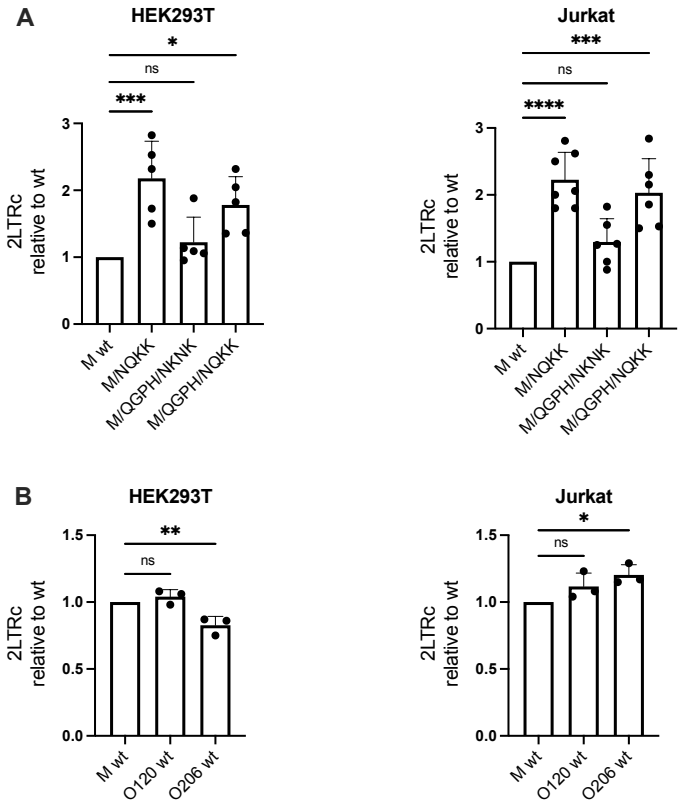


Figure 24. 2LTRc levels of IN M and O. A 2LTRc levels, relative to M wt, for the mutants carrying the O CLA and NOG sequences in HEK293T (on the left, n=5) and in Jurkat (on the right, M wt and M/NQKK n=7, M/QGPH/NKNK and M/QGPH/NQKK n=6). **B** 2LTRc levels, relative to M wt, for O120 wt and O206 wt in HEK293T (on the left, n=3) and Jurkat (on the right, n=3). Data are shown as average ± SD. ****p ≤ 0.0001. ***p ≤ 0.001. **p ≤ 0.01. *p ≤ 0.05. ns, not significant (one-way ANOVA with Tukey's multiple comparisons correction).

Since the mutant has both functional motifs (NOG and CLA) of group O, this result raised the question of whether having higher levels of 2LTRc could be a signature feature of group O. Therefore, the O isolate carrying the O consensus sequences in both motifs (QGPH and NQKK) used in this work, O120, was tested for its 2LTRc level against M wt in HEK293T and Jurkat cells. As a control to check whether this could be due to the NQKK sequence or to a general feature of the group, the same experiment was conducted in parallel with isolate O206 (carrying the same NOG motif sequence, but the KQKQ sequence in the CLA positions). In both cell types, isolate O120 showed to have 2LTRc levels similar to those of M wt (Figure 24B), while isolate O206 appears to be slightly below or above the M one, in HEK293T and Jurkat cells respectively (Figure 24B). Overall, this result confuted the hypothesis that group O could have higher 2LTRc levels *per se*, suggesting that the mutant M/QGPH/NQKK is producing more 2LTRc for other reasons.

As stated above, 2LTRc can be used as markers of RTPs nuclear import. Another possible explanation for the observed increase of 2LTRc could indeed be an alteration of the RTPs nuclear import, which, for example, could be more efficient for the mutant M/QGPH/NQKK. To assess this hypothesis, transduced cells were fractioned at the same time point at which 2LTRc are measured (24-hpt), and the amount of RTPs was measured in the nuclear and the cytoplasmic fractions. The experiment was performed with the M wt as well as the same mutants tested above (M/NQKK, M/QGPH, M/QGPH/NQKK) on HEK293T and Jurkat cells. In both cells, the RTPs were 100% localized in the nuclear fraction at 24-hpt for all the IN evaluated (Figure 25), suggesting that the nuclear import of the RTPs is equal for all of them. However, the limit of our fractionation method is that we cannot distinguish between what has really been imported and, therefore, present in the nucleoplasm and what is blocked at the nuclear membrane, therefore it is not possible to completely exclude the possibility that, indeed, the nuclear import could be responsible for the alteration in the 2LTRc levels observed with the mutant M/QGPH/NQKK.

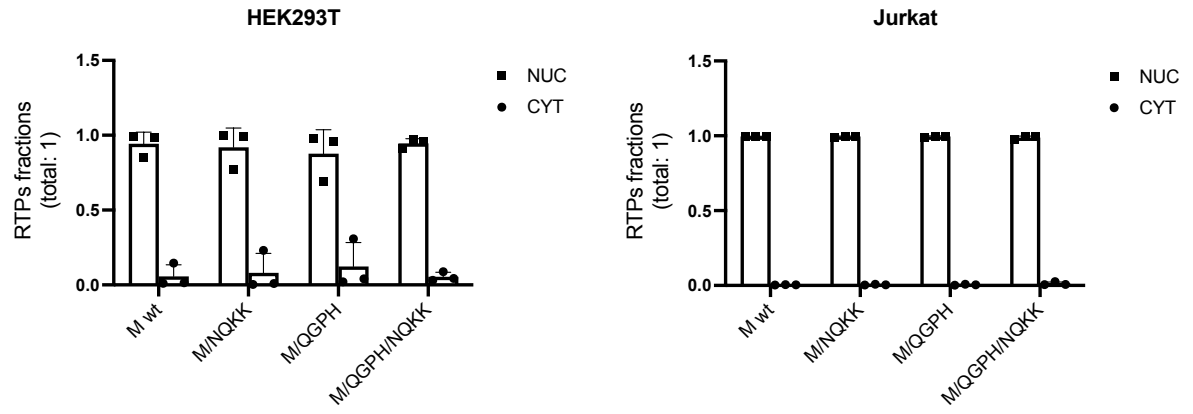


Figure 25. RTPs levels in cytoplasmic and nuclear fractions. RTPs fraction levels in the nuclear (square) and cytoplasmic (circle) fractions are shown for M wt, M/NQKK, M/QGPH, M/QGPH/NQKK (n=3). Data are shown as average ± SD.

OBJECTIVE 2: THE ROLE OF THE AMIDIC AMINO ACIDS IN THE CLA MOTIF

My PhD work dealt, essentially, with the characterization of a motif recently identified by the hosting laboratory. As mentioned above, this motif is highly conserved in IN M and constituted by the four non-contiguous amino acids NKNK. The laboratory showed that the four positions are functionally linked, which led to define this as a motif, because of the possibility of permutating the amino acids in the four positions, generally retaining functionality (Kanja 2020). The work performed subsequently was more oriented on the study of the role of the K than on that of the N of the motif. This was essentially justified by the observation that, in the process of inferring the influence of the number of K residues present in the motif on integration, it was shown that the mutant containing four K (KKKK) was as functional as the wt enzyme (Kanja et al., 2020). However, since this mutant is deprived of amidic residues, this result also indicated that the importance of the K is, somehow, dominant on the need for an amidic residue in the motif. In contrast, the NQNQ mutant, which has four amidic residues but no lysines, has no detectable integration levels, indicating that, if the absence of amidic residues can be bypassed by the presence of additional that of lysines, the reverse is not true.

The analysis of the importance of the amidic residues in the motif was performed, by replacing the N residues (present in the NKNK motif of IN M) by two Q that have physical-chemical properties very similar to asparagine, being both polar and carrying an amidic group in their lateral chain. In this case the integration efficiency was unaltered with respect to the wt enzyme (Kanja et al., 2020). In contrast, abolishing the polar nature of the amino acid, by replacing both N by L, same length of the lateral chain but no polarity, abolished integration, indicating that polarity was a crucial feature in these positions (Figure 20). However, polarity alone was not sufficient, since when the N were replaced by two T (polar and with a lateral chain of similar size of asparagines, but carrying an OH group instead of the amidic one) integration dropped to barely detectable levels (in the 2-5 % range with respect to wt IN) (Figure 20), indicating that the nature of the group carried by the lateral chain is also important (Kanja et al., 2020).

A theoretical estimate of the contribution of each type of defect (as a decrease of nuclear import of the reverse transcription product, for example) to the overall efficiency of integration was performed. This was done by assuming that if a decrease of 30%, for instance, was observed in nuclear import (to stick to the example given above) with a mutant with respect to wt IN, if that was the unique defect, one would expect to observe a decrease of integration with respect to the use of the wt protein, of 30%. By following this approach, the lab reached

the conclusion that the defects observed in nuclear import and in 3' processing could account for the whole decrease in integration observed with each of the mutants of the K of the motif. This was published as Table 1 in the Kanja's paper. The same analysis was performed for the LKLK, and TKTK mutants, but, as the paper focused on the role of the lysine of the motif, it was not inserted in the publication. Interestingly this analysis indicated that, in contrast to what observed for the mutants of the K, in this case the defects observed in nuclear import and in 3' processing were not sufficient to account for the amplitude of the decrease in integration observed with these mutants. This is manifest from Table 1 where the same values published in Kanja et al. are reported, side by side with those of the mutants of the N. By comparing, for each mutant, the values of the first line with those of the last, one compares the observed values for the efficiency of integration with those expected if 3' processing and nuclear import were the only defects in integration found with that given mutant. For the three mutants of the K residues the values are (observed efficiency with respect to wt IN *vs* expected frequency with respect to wt IN): 0.03 vs 0.09, 0.24 vs 0.26, and 0.00 vs 0.05, witnessing a remarkable similarity between observed and expected results. In sharp contrast, the mutants where the N were replaced by L or T showed values of observed integration that did not match with the expected values: 0.01 vs 0.27 and 0.05 vs 0.47, for the L and the T mutants, respectively. These results suggested that additional defects were present with these mutants.

			K mutants			**N mutants**	
	wt	D116A	NQNK	NKNQ	NQNQ	LKLK	TKTK
Observed integration relative to wt	**1.00**	**0.00**	**0.03**	**0.24**	**0.00**	**0.01**	**0.05**
Nuclear import relative to D116A	1.00	1.00	0.31	0.35	0.33	0.67	0.53
3' processing relative to wt	1.00	0.00	0.28	0.74	0.15	0.41	0.89
Expected integration relative to wt	**1.00**	**0.00**	**0.09**	**0.26**	**0.05**	**0.27**	**0.47**

Table 1. Observed and expected integration for the K and N mutants of the CLA motif. In the table observed and expected levels of integration (lines highlighted in grey) for the K and T mutants, as well as for the IN wt and the catalytically inactive IN (D116A) are shown. The observed integration values were obtained from our experimental method (see Methods). The nuclear import and 3' processing levels were calculated as explained in the Kanja et al. paper. The expected integration values were obtained by multiplying the nuclear import values for the 3' processing values. Adapted from Kanja et al., 2020.

In the previous work (Kanja et al., 2020) several steps of the replication cycle were evaluated for these mutants. During my thesis we focused on one of the few pre-integration parameters that we did not take into account before: the choice of the integration sites. Indeed, in our experimental system, we evaluate the efficiency of integration on the basis of the expression of a transgene inserted in the proviral DNA. If the choice of the integration sites were affected in the mutants, this could influence our readout, possibly introducing a discrepancy between

expected and observed results. Therefore, we started looking at the integration sites with the TKTK mutant, since integration with the LKLK mutant was so low that it would not have been possible to obtain material for these analyses. The analysis of the integration sites was carried out in collaboration with Dr. Marina Lusic, at the University of Heidelberg, where I spent one month to start the project.

The CLA motif amidic amino acids are influencing the choice of the integration sites

To understand if the TKTK mutant leads to a change of the integration sites, transduction of Jurkat cells with viral particles carrying either the IN M wt or TKTK was performed. A library enriched in integration sites of the genomic DNA was then prepared, sequenced, and analyzed.

	M wt	M/TKTK
Total sites	4674	1375
Intragenic	3401 (72%)	779 (56%)
Intergenic	1273 (28%)	596 (44%)

Table 2. Integration sites retrieved for IN M wt and IN M/TKTK. In the table the total number of integration sites retrieved for IN wt and the TKTK mutant is shown, along the fractions found in intra- or intergenic positions.

From the cells infected with IN wt a total of 4,674 sites were retrieved, while for the TKTK mutant the rescued sites were 1,375 (Table 2). Strikingly, the percentage of intra- and intergenic integration sites was different between the two integrases, with the wt preferentially targeting the intragenic sites (72%), while the TKTK mutant is integrating in these regions only 56% of the times (Table 2). Additionally, when looking at the expression levels of the targeted genes, it was clear that the mutant is integrating with a higher frequency in low-expression and/or silenced genes compared to wt IN (Figure 26). These first results support the hypothesis for which the discrepancy observed between the expected levels of integrations and the real ones, might be caused by a non-expression of the reporter genes, and consequently a non-detection of the integration event.

To further investigate this redirection of the integration sites, an analysis of the chromatin profile around them was performed. Results showed that the TKTK mutant integration sites are generally less-correlated, compared to the wt, with open chromatin markers that are typically associated to HIV-1 integration sites (Wang et al., 2007; Roth et al., 2011;

Kvaratskhelia et al., 2014; Sowd et al., 2016). The TKTK mutant integrates less in regions with markers H3K4me1 and H3K27Ac, which are super-enhancer signatures, typically abundant in SPADs (Bedwell 2021, Singh 2022). H3K36me3 modification is normally associated with gene bodies and transcription elongation and is recognized by the PWWP domain of LEDGF/p75. However, LEDGF/p75 is known to interact with the viral IN through its NTD and CCD domains, therefore a possible perturbation of this binding caused by the two-point mutations present in the CTD of the TKTK mutant is unlikely to happen. Conversely, an increase of integration sites correlating with repressive chromatin marks was observed for the TKTK mutant compared to the wt. Indeed, the study of the chromatin marker landscape shows that the mutant redirects, at least partially, the integration sites towards non-actively transcribed regions (Figure 27B), that are not normally targets for integration by HIV-1 normally (Wang et al., 2007; Roth et al., 2011; Sowd et al., 2016).

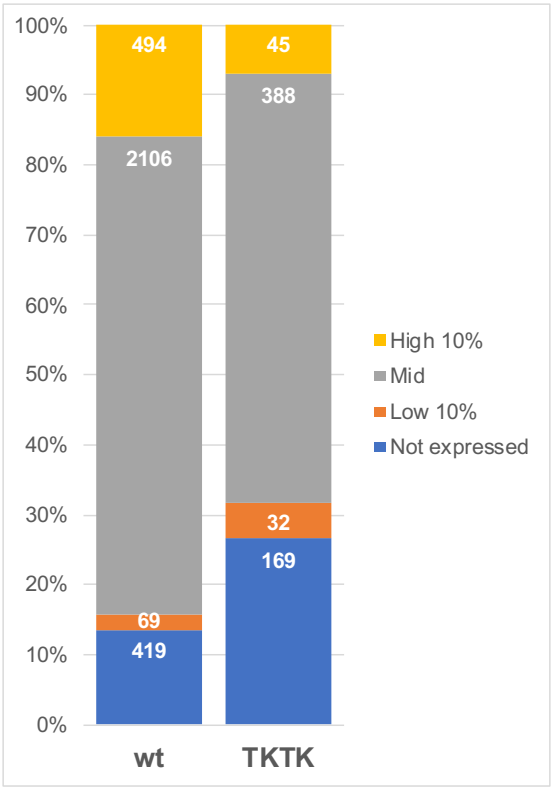


Figure 26. Relative expression levels of IN wt and the TKTK mutant integration sites. The relative expression is color coded as shown on the right. Integration sites expression levels of the IN wt and the TKTK mutant are shown as the percentage of the respective number of total sites. The absolute number of each category is shown in white.

Overall, these preliminary results suggest that the choice of the integrations sites of the TKTK mutant is indeed redirected compared to the wt enzyme. Interestingly, since the LEDGF/p75 binding site on IN does not involve the CTD, the mechanism responsible for this phenotype seems to be LEDGF/p75-independent, suggesting that it could constitute an alternative pathway to the classical one.

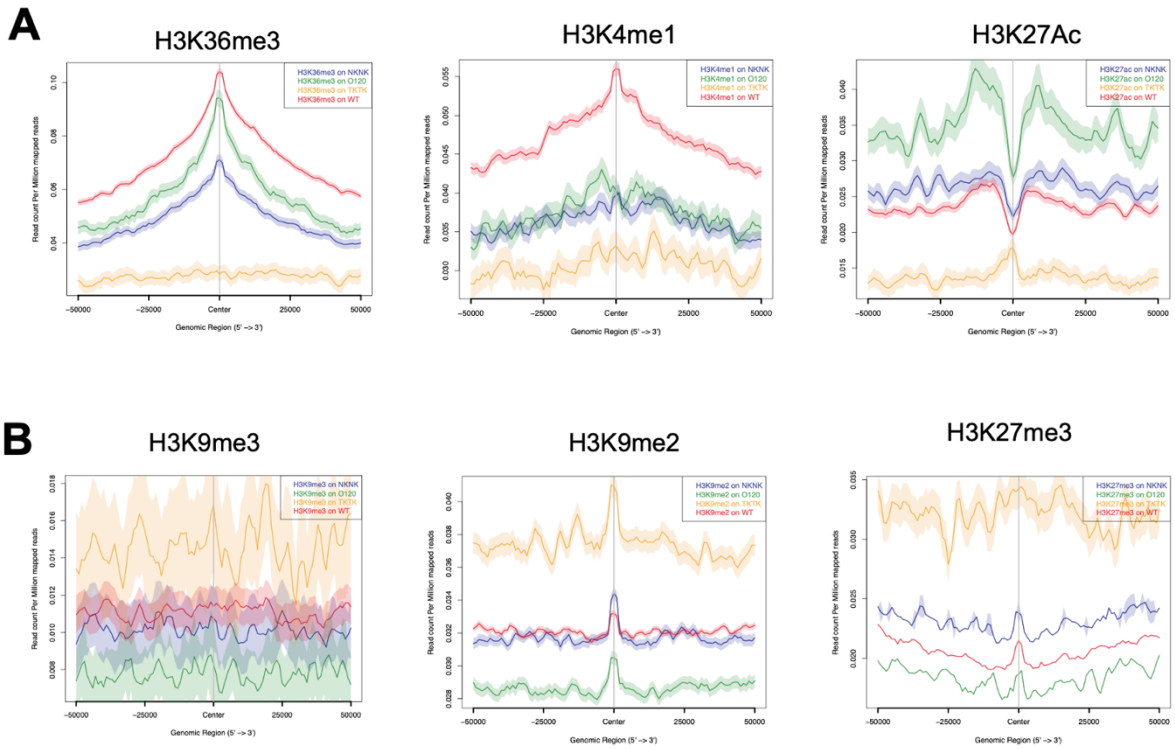


Figure 27. Signature chromatin features of IN wt and IN TKTK mutants integration sites. B, **C** Distribution of markers of open chromatin (panel B) and repressive chromatin (panel C) for the integration sites of IN wt (in blue) and the TKTK mutant (in yellow).

DISCUSSION AND PERSPECTIVES

HIV-1 groups M and O originated from two independent zoonotic transmissions from SIVcpz*Ptt* and SIVgor respectively (Gao et al., 1999; Heuverswyn et al., 2007; Keele et al., 2008; D'Arc et al., 2015). This resulted in considerable levels of genetic diversity between the two viruses, but, beside this, no significant functional difference that could contribute to explain their notably different epidemiology is known.

HIV-1 integrase is a multifunctional protein as well as one of the key viral enzymes. It is one of the most conserved viral proteins among all HIV-1 groups and subtypes (Li et al., 2015; Nagata et al., 2017). Indeed, the protein is exposed to evolutionary constraints in order to maintain its catalytic and non-catalytic functions while also being the partner of viral and host proteins. Accordingly, integrases of M and O differ for just 16% of their amino acids. They share the same domain organization, the same functional motifs (e.g., HHCC, DDE) conserved among retroviruses, and the same functionalities. On these bases, although no structure of IN O is available, it is expected that the overall fold would be similar to IN M. In support of this, most IN O/M chimeras generated in this work, are functional. Indeed, among the five chimeras obtained by replacing a region of IN O with the homologous one of IN M, only two of them (O/NTD-M/NQNQ and O/CCD1-M/NQNQ) resulted to be non-functional (Article Figure 2). Though, the insertion of the NKNK motif rescued the integration level of the NTD chimera but not that of the chimera of the first part of the CCD (Article Figure 2). These results indicated that the chimera of the NTD was non-functional because of the lack of functional motifs (carrying NQNQ at its CLA positions, and KNDQ at its NOG ones) and suggested that the loss of functionality of the CCD-1 chimera was related to other reasons, as it could be an incorrect folding. Overall, the results showing that the majority of the chimera were still functional, suggest a high level of structural identity between IN M and O. Knowing this, it is interesting to observe how the same enzyme found two different and independent evolutionary pathways to converge to the expression of the same phenotype in two phylogenetically distant context such as HIV-1 group M and group O.

In this work, along with the identification of the motifs and their characterization, we proposed that their ability to functionally complement each other relies on the binding of the same partner thanks to the two motifs (the CLA motif in M and the NOG motif in O). More specifically, as it was previously proposed by Kanja and colleagues, with a negatively charged partner through a positively charged surface. How is the NOG motif contributing to the formation of such a positively charged surface without containing any positively charged amino acids itself? We found that Q7, G27, P41, and H44 forming the NOG motif were sufficient to exert the functionality, indicating that no other amino acid present in the NTD is necessary. However, this does not exclude the possibility that less than these four amino

acids are actually needed. Based on their localization, as well as the biochemical and structural properties, amino acids P41 and H44 might be sufficient to observe the same phenotype as with the four amino acids. P41, especially, could play a key role. By looking at the IN O predicted structure obtained with AlphaFold, we localized a positively charged surface in the C-ter part of the NTD, right before the linker that connects the NTD to the CTD. This positive surface is formed by K42, H44 and K46. These three amino acids lateral chains are located on the same geometrical plane forming a positive flat surface (Figure 28). Interestingly, K42 and K46 are present also on IN M NTD, but the surface positive charge appears weaker than in O. We deduced, from the comparison of the structure M and the prediction of the structure O, that the reason for this is to be found in the presence, in NTD O, of the P and the H at positions 41 and 44 respectively, that are instead occupied in group M by a D and a Q. The presence of the H44 is locally intensifying the positive charge, contributing to it with its lateral chain. The role of the P is indirect, but perhaps more important.

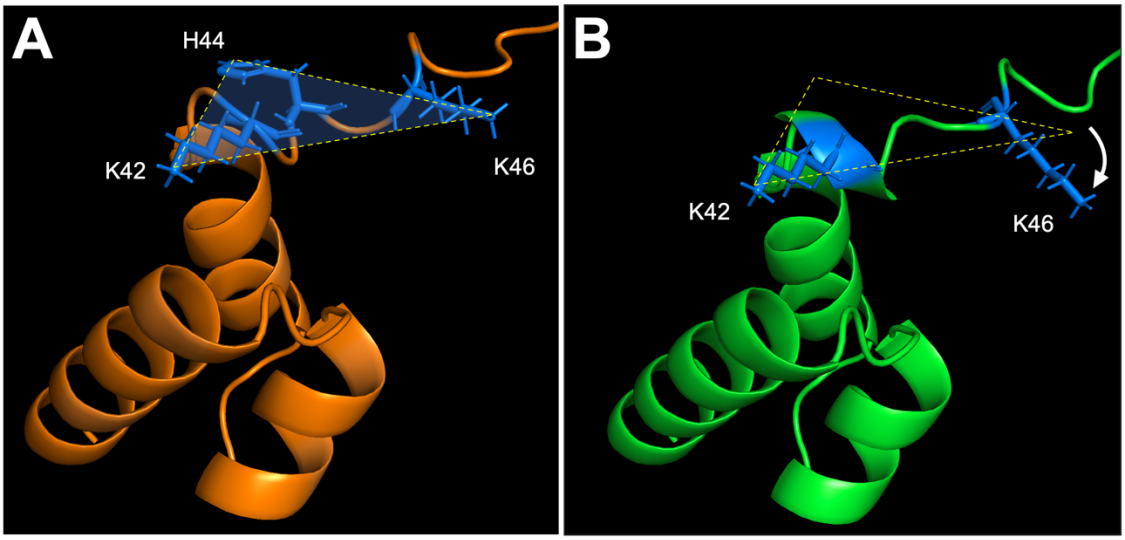


Figure 28. The positive surface of NTD O. A Structure of IN O NTD (AlphaFold model) is shown in orange. Lateral chains of aa K42, H44, and K46 are shown in blue. A triangle with yellow border and blue transparent fill highlights the positive plan formed by the three amino acids. **B** Structure of IN M NTD (PDB 6PUT) is shown. Lateral chains of aa K42 and K46 are shown in blue. The same triangle formed by the three positive aa in the NTD O is reported here, with an arrow highlighting the displacement out of the positive surface of K46.

While the C-ter part of IN M NTD is more structured, forming a small alpha-helix, the presence of the proline in the NTD O is disrupting this secondary structure resulting in the rearrangements of the amino acids that follow, most notably the 42-44-46 triad (Figure 28). Indeed, thanks to its structural characteristics, P41 is creating a turn in the NTD of IN O which results in the alignment on the same plane of the three amino acids (K42, H44 and K46), therefore forming the positive surface, and in the exposition toward the solvent side of the latter. Several experiments need to be conducted to validate (1) whether it is true that the

P41 and H44 (or even the P41 alone) are sufficient to obtain the phenotype we observe and (2) if the positive surface formed by K42, H44 and K46 (or just K42 and K46) is essential for having a functional integrase. (1) By following the same strategy employed in our work, we would start from the mutant M/NQKK. This mutant carries the O CLA consensus sequences, and it has a default of integration (Article Figure 3) that allows to screen for a gain of function when mutating the NOG positions, knowing that the M/QGPH/NQKK mutant has wt levels of integration (Article Figure 3). First, a mutant where only the two last positions of the NOG are mutated (M/PH/NQKK, with positions 7 and 27 occupied by the original M sequence, K7 and N27) can be created and assessed for its integration levels. If our hypothesis is correct (P41 and H44 being the most important aa of the motif), we would expect to see the same restoration of integration levels as when the four amino acids are replaced (from 25% to 100% relative to the wt). Otherwise, if integration efficiency is not recovered, it would mean that also these two amino acids (Q7 and G27) are important to form the positive surface, or that they might contribute in other way which we would further investigate. Similarly, to check the role of the P41 alone, the mutant M/P/NQKK can be assessed for integration. If integration would be at wt levels also in this case, this would suggest that the two lysines are sufficient to form the positive surface, as well as confirming the crucial role of the P in relocating them. (2) Similarly, IN mutants for the three amino acids contributing to the positive surface (K42A, H44Q, K46A), of for only the two lysins (K42A and K46A), could be prepared in a M/QGPH/NQKK background (or with the M/PH/NQKK or M/P/NQKK if the results will confirm our hypothesis), allowing to evaluate the role of these mutations through a loss of functionality. Mutants will be assessed for integration to check, first of all, if the positive surface is essential and, then, whether the contribution of the H44 is significant or not. Removing the K from the O positive surface is a similar strategy to the one that highlighted the essential role of the K in the CLA motif (M/NQNQ mutant) (Kanja et al., 2020).

No matter the mechanism through which the interaction with the partner occurs, this partner must be identified. Indeed, the most likely explanation for the two functional motifs to be able to complement for each other is that they are interacting with the same partner. What can this partner be? As observed by Kanja and coworkers, the fact that the positive charges can be permutated among the four positions of the CLA motif suggest that the possible partner might have a repetitive pattern of negative charges. This feature matches with the charges of the backbone of a nucleic acid. Several studies showed how the positive charged amino acids present in IN CTD are binding to the viral RNA (Kessl et al., 2016; Elliott et al., 2020). This binding appears to be essential for the correct assembly/morphogenesis of the capsid core and the encapsidation of the genomic RNA inside it. Mutations of these amino acids

cause the formation of "eccentric" viral particles with the genomic RNA located outside the capsid core (that could be intact or affected in its assembly), leading to an unsuccessful infection since the vRNA would be quickly released into the cytoplasm of the infected cell and degraded, causing an early block of the viral life cycle. Although it is true that we did not check the binding affinity of our mutants to the vRNA, we believe that the defaults we observed do not derive from the formation of an eccentric particle. First of all, thanks to the results obtained with the EURT method, we can exclude the possibility that our mutants are affecting the formation of a regular capsid core. In the EURT experiments conducted in the absence of the SGP-RNA, indeed, we observed the same uncoating kinetic for the IN wt and the mutants. These results indicate that no abnormal capsid cores were present. Moreover, if the default in reverse transcription of mutant M/NQKK was caused by a default in the encapsidation of vRNA, by normalizing the integration products for the RTPs we would see no default for integration, which is not the case, meaning that the default is specific for integration. While the possibility that the vRNA is the motifs binding partner is not to exclude, it is clear that it would not be limited to this specific IN functionality during viral morphogenesis. Most of the effects we observed when the motifs were mutated were connected to steps involving either the vRNA, vDNA or the genomic DNA. In fact, both the CTD and the NTD have been shown to bind the vDNA (Vink et al., 1993; Engelman et al., 1994; Puras Lutzke et al., 1994; Hare et al., 2010, 2012; Maertens et al., 2010). Therefore, the option that our unknown partner could be one or more of these molecules could partially explain our results.

Then, of course, the partner could be a protein, either a viral protein or one of host origin. IN is known to interact with several viral and host proteins (van Maele et al., 2006; Raghavendra et al., 2010) and among them there could be the alleged partner of the functional motifs we identified. The interactant might also be a yet unknown partner of the integrase. If we want to speculate what this protein might be, based on our results, a few features can be predicted. Of course, it has to be a protein with a negatively charged domain and, preferentially, with a repetitive negative pattern. It could be an RNA/DNA binding protein (therefore with also a positively charged domain) that might help and/or facilitate IN in taking contact with the vRNA/vDNA. Also, based on the observed effect on the reverse transcription step, and assuming that the reverse transcription is completed inside the capsid core, the partner protein might be encapsidated during the formation of new viral particles.

The two scenarios, one where the binding partner is a nucleic acid and one where it is a protein, are non-mutually exclusive and it can be, for example, that some of the defaults we are observing are coming from the possible perturbations of the IN-DNA/RNA binding, while

other can derive from the loss of interaction with the alleged protein partner. Have a better understanding of the mechanism behind the motifs and identifying the partner/s constitute a crucial point for the continuation of this work. Among the different experiments that can be performed to try to shed light on this point, there are a few that could be a good starting point. On one hand, IN could be purified, in either its full-length form or just the domains we are interested in and tested for its ability to bind nucleic acids. Moreover, the motifs could be wt or mutated. On the other hand, IP could be performed on IN wt vs IN mutants, to try to identify the possible protein partner.

How could such a difference between the two groups have emerged in the same protein? By analyzing the sequence conservation at the CLA and NOG positions of different simian viruses we traced back the evolutionary history of the two motifs. The NOG motif sequence, QGPH, appears for the first time in the SIVgor (Figure 23), where it is present in all the analyzed sequences (although their relatively small number). We therefore supposed that this motif and its functionality were established in SIVgor, and that, once the virus transferred to humans, it did not need adaptation and was, therefore, fixed in HIV-1 group O and P (Figure 23). However, we did not check whether or not the NOG motif might have the same essential role in SIVgor or in group P as it has in group O. While a SIVgor chimeric construction might be more difficult to test, because of the phylogenetic distance between gorillas and humans, an IN from group P could be tested in the same chimeric context used to test SIVcpz*Ptt* IN. Although not responding directly to the question whether this motif was already functional and essential in SIVgor, results obtained from group P IN might help to understand the evolutionary scenario. In fact, if comparable results to those observed for IN O will be found, this would support the hypothesis that the motif functionality was indeed already present in SIVgor and that it was "passively" inherited by these two groups.

The CLA motif as we know it, NKNK, contrary to the NOG, appears for the first time in humans, in the two groups derived from SIVcpz*Ptt*, M and N (Figure 23). In both groups high conservation is present at these positions. The fact that the same conserved amino acidic sequence is present in both groups, despite the lack of conservation present at the same positions in SIVcpz*Ptt,* is noteworthy, especially since the path to arrive to this sequence seems to be different for each group. In fact, while group N is most closely related to a SIVcpz*Ptt* isolate carrying the NKNK sequence (EK505), group M closest isolates do not carry the CLA motif sequence (MB897 KKKK, LB715 KKQK) (Article Figure 4). Of course, it cannot be excluded that group M might have originated from an isolate carrying the NKNK sequence, but the fact that the motif is present and conserved in both groups (especially in group M considering the number of isolates existing in this group) makes it hard to believe that this

motif was just passively inherited. Interestingly, although the CLA motif sequence is found only in HIV-1 M and N groups, amino acids with similar characteristic, if not exactly the same ones, are often found at these positions throughout all the HIV and SIV taken in consideration. Surprisingly, the virus where the least conservation is found at the CLA positions is the precursor of the two groups carrying the NKNK sequence, the SIVcpz*Ptt* (Figure 23). It is, for example, the only time a conserved N is not present at position 222. Nevertheless, by analyzing every single sequence of SIVcpz*Ptt* isolates is clear that, although not conserved, these positions were subjected to some kind of selective pressure, allowing only certain amino acids to occupy these positions (N/Q and K/R are always present with only one exception) (Figure 29).

	222	240	254	273
MB897	K	K	K	K
MB66	K	K	Q	K
LB715	K	K	Q	K
LB7	K	K	Q	K
DP943	N	R	K	K
MT145	Q	K	Q	K
Marilyn	N	K	N	K
EK505	N	K	N	K
CAM3	N	R	R	K
CAM5	N	R	N	K
CAM13	K	T	Q	K
CAM155	K	N	Q	K

Figure 29. Frequency of amidic and positive amino acids at the CLA positions of SIVcpz*Ptt*. The amino acidic positions 222, 240, 254, and 273 are shown for SIVcpz*Ptt* isolates. Color code is blue for positive amino acids (K, R) and orange for amidic amino acids (N, Q). The only amino acids not fitting in any of these two categories (T) is shown in black.

As mentioned above, we hypothesize that of the four amino acids composing the NOG motif, only the latter two, if not only the P, might be sufficient to explain the observed phenotype. While this is definitely a hypothesis that we need to confirm, it is intriguing to speculate that what might have caused this need to adapt the amino acids present at the CLA positions in SIVcpz*Ptt* is, indeed, the lack of the P41. The P is present in both SIVrcm and SIVmnd2, the two possible ancestors of the *pol* portion coding for IN of SIVcpz*Ptt*, but not in SIVcpz*Ptt* IN (Figure 23). Furthermore, the two lysins (K42 and K46) we believe are responsible to form the positive surface in the NTD, are also present in SIVrcm and SIVmnd2, as well as being highly conserved in each virus analyzed in this work. Therefore, the NTD of these viruses, thanks to the presence of the P41, K42 and K46 could have the same organization of NTD from group O and, therefore, the same functionality. However, when after cross-species transmission to chimpanzee the P was lost, probably as an adaptation to the new host, the K were no longer forming the positive platform, as is it is the case for the NTD of IN M. Therefore, with no

optimized NOG motif sequence and consequently no positive surface, the integrase of SIVcpz*Ptt* had to adapt its CLA motif locus, and the lack of conservation found at these positions might reflect this evolution time frame. This hypothesis would suggest that a crosstalk between the NTD and the CTD of the integrase exists. Linkage constraint between different regions of the integrase have been previously showed, highlighting, among them, a link between the NTD and the CTD (Meixenberger et al., 2017). Indeed, our results also support a link between the two domains. In group O, a phenomenon of epistasis is observed of the *nog* locus over the *cla* one. Epistasis is described to be important in the evolution of viruses and is not new to HIV-1 (da Silva et al., 2010; Martínez et al., 2011). When the phenotype of a given mutation is under epistatic control, it means that the same mutation can have opposite effects on different genetic backgrounds (Storz, 2016). This is the case for the CLA motif mutation "NQNQ" that abolishes completely integration in the group M context, while, thanks to the epistatic effect of the NOG motif over it, it has no effect in group O (Article Figure 1). Epistasis can also explain the lack of evolution in O CLA motif, in fact, if the NOG motif was already present in the ancestor sequence, as it is the case for SIVgor, its epistatic effect could have affected the mutational pathway of its CLA motif, by making the beneficial mutations (as for example NKNK) less or not at all required. This seems to be confirmed by the fact that the insertion of the optimized NKNK sequence in the CLA motif positions of IN O does not lead to a better fitness of the protein. Similarly, the narrowing of the possible evolution pathways to a better phenotype and therefore the enhancement of the repeatability of the possible substitutions, as observed in the CLA motif of SIVcpz*Ptt*, might be a consequence of epistasis.

However, the perfect conditions for the NKNK sequence to be fixed arose only after cross-species transmission to humans, where a strong positive selective pressure was exerted at these positions, selecting the NKNK sequence. This is supported by the observation that a SIVcpz*Ptt* integrase mutated to carry the NKNK sequence, showed to be less functional than its wt (KKKK), in human cells. These results, while giving us an idea of what could have happened when an isolate of SIV was transferred to humans, do not really give us information about the role of the CLA motif in the SIVcpz*Ptt*. This, because not only the SIV IN was tested in a HIV-1 context, but also in human cells. To have a better understanding of the role of the CLA motif in SIV we wanted to test the same SIV chimeric constructions in EB176 (JC), which is an EBV transformed lymphoblastoid cell line from the peripheral lymphocytes of a male chimpanzee. While we were able to find and obtain this cell line, we encountered severe obstacles in culturing them and they died before we could have the chance to perform any

experiment on them. This part of the project, therefore, represents another starting point for further explorations in the future.

While we successfully identified the NOG motif and characterized its functionality, by showing that it is ensuring good levels of integration by increasing the substrate for integration (3' processed full RTPs), and, more indirectly, optimizing the uncoating step, a result we cannot yet explain is the one we obtained when 2LTRc levels were measured. Indeed, we found that the double mutant carrying the consensus sequences for both O motifs (M/QGPH/NQKK) has significantly higher level of 2LTRc relative to the wt (Figure 24). While the similar results obtained with mutant M/NQKK can be explained by the default in integration of this mutant, meaning that more linear vDNA was left as potential substrate for circularization by the nuclear enzymes, for the M/QGPH/NQKK, which has wt level of integration, this result is hard to understand. We also ruled out the possibility that this might be a group O feature by measuring 2LTRc levels of IN O120 and IN O206, which appear to be non-significantly different from the M ones (Figure 24). One possibility was that the nuclear import of this mutant could be more efficient, however when RTPs were measured in the cytoplasmic and nuclear fractions, we observed that 100% of them was located in the nuclear fractions for all the samples tested. Although it is true that with our protocol, we are not able to discriminate between what is at the nuclear membrane and what is in the nucleoplasm, it is hard to believe that a significant portion of the 100% could be associated with the nuclear membrane, especially at 24-hpt. Another possibility could be that the nuclear import difference could be observed at earlier time points than the one we used (24-hpt, the same at which 2LTRc are detected). Indeed, we tried to conduct the same experiments at 6- and 8-hpt (results not shown), but it was too early to detect late RTPs and early RTPs were too divergent among repetitions to draw any conclusion. Maybe an optimization of the protocol could solve this problem and finally shed light on the nuclear import kinetic of this mutant. A simpler explanation could be that the M/QGPH/NQKK double mutant is somehow influencing, either in a direct or non-direct way, the NHEJ machinery in charge of the formation of 2LTRc. While interactions between components of the NHEJ as Ku80 and Ku70 and the PIC (Li et al., 2001), and a direct interaction between IN and Ku70 (Knyazhanskaya et al., 2019) are known, they are believed to be essential for the chromosomal DNA repair after integration and not related to the formation of 2LTRc. Perhaps, the interaction of our mutant with these factors is affected in a way that is shifting the normal pathway to another one that leads to the formation of more 2LTRc. Nevertheless, it was shown that 2LTRc, especially when accumulated, as it is the case for the double mutant M/QGPH/NQKK, can serve as an alternative substrate for integration (Richetta et al., 2019). Therefore, while the mechanism behind the higher 2LTRc

level of this mutant are yet to be explained, they can participate to the recovery of the integration levels of this mutant.

One of the main limits of our model is that it does not explain the role of the N in the CLA motif, focusing more on the importance of the positive surface formed by the Ks. While it is true that it was previously shown that the role of Ks at the CLA positions seem to be more relevant than the one of Ns, as observed when four K are present at the CLA positions (M/KKKK 100% integration of the wt) (Figure 21), on the other hand, it was also shown how mutating the N in non-polar or polar but not amidic amino acids, led to a total loss of the integration efficiency (M/LKLK 0% integration of the wt, M/TKTK 5% integration of the wt) (Figure 21). Therefore, although dispensable when replaced by a positively charged amino acids, the amidic feature of N side chain appears to be essential in the other cases. Moreover, by analyzing the conservation of the CLA positions, it was clear that a strong positive selective pressure to have a highly conserved N in position 222 is present in all the lentiviruses analyzed. The only exception was found in SIVcpz*Ptt* where a K is more often found and in general for the CLA motif no conservation is present, yet polar amidic amino acids (N, Q) are often recurrent (Figure 29).

The high level of conservation of N222 across HIV-1 and the SIV analyzed in this work, along the loss of integration observed when N222 and N254 of IN M were replaced by L or T, point out an important role of the amidic amino acids present in the CLA motif. Furthermore, when trying to understand if the integration default could be explained by the impaired pre-integration steps analyzed, Kanja and colleagues found that there was indeed a correlation between the two when the K were mutated in the CLA motif. The expected levels of integration when taking in account the 2LTRc accumulation (as a marker of nuclear import) and the 3' processing default were perfectly correlating with the integration default observed. However, this was not the case for the N mutants. Both, the LKLK and TKTK, had levels of expected integration significantly higher than the observed ones, raising the question whether other pre-integration steps were affected (Table 1). Indeed, we found that the TKTK mutant affects the choice of integration sites. In particular, its integration sites are more often found to be in intergenic positions compared to wt (Table 2), while the intragenic sites are less frequent and more frequently found in non-transcriptionally active genes compared to the wt (Figure 26). These results might explain the gap between the expected and observed levels of integrations for the TKTK mutant. The observed integration efficiency, indeed, is measured thanks to the presence of reporter genes in the VLPs modified viral genome ($PURO^R$). Therefore, the detection of our integration events is limited to those that are located in transcriptionally active regions. The expected levels of integration for the TKTK mutant can

therefore be explained because a significant part of the integration events occurs in regions where the reporter genes cannot be transcribed. Indeed, by comparing the integration events involving transcribed regions, which are the only ones we can detect with our system, of IN wt (2,669 integration events) and IN/TKTK (465 integration events), and knowing that the nominal MOI used for IN/TKTK was half of the wt, we can estimate an integration efficiency of the TKTK mutant of about 30% of the wt, which better correlates with the expected value found in Table 1.

Directing of integration sites towards interior regions of gene bodies and actively transcribed regions are explained, by the current models, by the interaction of the CA and the IN with cellular factors CPSF6 and LEDGF/p75. While CPSF6 has the role of directing the PIC towards internal SPADs regions, LEDGF/p75 is mainly targeting the PIC in the interior regions of gene bodies. LEDGF/p75 binds the viral IN, through its IBD, at the CCD and at the NTD. Our mutant has two-point mutations in the CTD (the only domain not involved in the binding of this cellular factor) reducing the possibility that the binding with LEDGF/p75 could be affected. This suggest that the shift towards intergenic regions might be caused by other mechanisms, as, for example, the perturbation of the binding with another cellular factor or a direct role of the IN in tethering the chromatin. Indeed, depletion of LEDGF/p75 does not abolish the preference for the gene bodies for integration, suggesting that other factors might be involved in this. Another known protein with a similar role to LEDGF/p75 is HDGFL2. However, its role in our results could also be excluded, since this protein shares the same protein organization of LEDGF/p75 and is probably binding to the IN in the same way. To check whether the TKTK mutant binding to LEDGF/p75 is not perturbed, several methods can be employed. First, their binding might be directly checked via coIP of IN (wt and TKTK mutant) and LEDGF/p75 (wt and ΔIBD). Second, the integration sites map obtained with the TKTK mutant could be confronted with a map of integration sites obtained with cells infected in presence of LEDGINs, to check whether a correlation between the two distribution is present or not.

Moreover, the TKTK mutant shows to have different preferential choices for chromatin with given properties albeit with an inversed trend compared to the wt (Figure 27). Indeed, its integration sites are less associated with signatures of the open chromatin, like H3K4me1, H3K27Ac and H3K36me3 (Wang et al., 2007; Roth et al., 2011; Kvaratskhelia et al., 2014; Sowd et al., 2016). H3K4me1 and H3K27Ac, in particular, are super-enhancer signatures that constitute preferential HIV-1 target sites. Although it now seems that this preference is a consequence of the fact that SPADs, one of the favored targets for integration, have an abundance of super-enhancer regions (Bedwell et al., 2021; Singh et al., 2022). H3K36me3,

instead, is associated with gene bodies and is induced by transcription elongation. Furthermore, it is the modification preferentially recognized by the PWWP domain of LEDGF/p75. As mentioned above, however, the TKTK mutant phenotype is likely not dependent on its inability to bind to LEDGF/p75. A recent study highlighted how HIV-1 integrations sites appear to correlate more strongly with H3K36me3 associated genes, rather than chromatin-bound LEDGF/p75 associated genes (Singh et al., 2022). This observation suggest that the H3K36me3 could be the target of HIV-1 integration even in a LEDGF/p75 non-correlated way. The existence of this kind of mechanism could explain how the TKTK mutant orients integration toward sites less associated with the H3K36me3 marker, despite its binding to LEDGF/p75 should not perturbed. Conversely, the TKTK mutans integrates more frequently in sites with chromatin marks such as H3K9me2/3 and H3K27me3, associated with heterochromatin and transcriptional repression, which are normally avoided by the wt (Wang et al., 2007; Roth et al., 2011; Sowd et al., 2016).

A direct role of HIV-1 IN, and in particular its CTD, in interacting directly with chromatin and histone tails has been previously shown (Lesbats et al., 2011; Benleulmi et al., 2017; Matysiak et al., 2017; Mauro et al., 2019; Rocchi et al., 2022). Mutations in this domain, indeed, affected the insertion site selection based on chromatin density (Demeulemeester et al., 2014; Benleulmi et al., 2017). This chromatin binding was shown to be LEDGF/p75 independent, as it is happening also in its absence. However, the presence of LEDGF/p75 enhances the affinity of IN to bind chromatin and redirects the targeting (Lapaillerie et al., 2021). Overall, these data suggest that the IN might have a chromatin-binding functionality that is, at least partially, directly participating to the choice of integration site.

A recent study showed a redirection of integration sites caused by a single point mutation in the IN CTD, K258R (Winans et al., 2022). This mutation causes a 10-fold shift, compared to wt IN, of integration sites into centromeric alpha satellite repeat sequences. To understand the mechanism responsible for this redirection, the authors of the study performed an immunoprecipitation of IN wt and IN K258R to identify the interacting host proteins by mass spectrometry. An enhanced binding of IN K258R was found for two factors involved in mitotic chromosome condensation (NCAPD3 and SMC4) and for several components of the catalytic core of the protein phosphatase I (PPI) complex (Winans et al., 2022), which are involved in heterochromatin formation and regulation at the centromere (Samoshkin et al., 2009; Leonard et al., 2015; de Castro et al., 2017; Wang et al., 2017). However, when co-IP, followed by western blot, was performed on IN wt or IN K258R and their identified binding partner NCAPD3, SMC4, and two PPI complex components, overexpressed in cells, no higher binding affinity was found for IN K258R in comparison to wt IN for these cellular factors

(Winans et al., 2022). Therefore, the mechanism behind this shift was not elucidated, but the authors assume that LEDGF/p75 is likely not involved in it, since the point mutation of their mutant was located in the CTD, therefore mapped outside the binding site of LEGDF/p75 on IN (Winans et al., 2022).

The mechanism by which the TKTK mutant retargets integration also needs to be elucidated. Similarly to what Winans and colleagues did, we should clarify how much of the observed phenotype (shift in the integration sites) is due to a direct role of the CTD or to the potential disruption (caused by the two point mutations) of the interaction with a protein partner with a role similar to LEDGF/p75 (if not LEDGF/p75 itself). Identifying this/these cellular factor/s would represent one of the crucial steps for continuing this project.

If the initial phases of the mechanism of targeting of the insertion site are known, the later ones, among which the contact with the host chromatin are not well characterized. At this regard, our results, along the same lines of those found in the literature, are going into a direction where the IN CTD (directly or through the interaction with cellular factors) appears to be responsible for the choice of integration sites, and, in particular, for the tethering of the chromatin.

CONCLUSIONS

Through the main project of my doctoral thesis, that was focused on the investigation of the role of the CLA motif in group O, we were able to highlight the existence of two group-specific functional motifs in groups M and O integrases that complement each other functionally: group M CLA motif ($N_{222}K_{240}N_{254}K_{273}$), located in the CTD, and group O NOG motif ($Q_{7}G_{27}P_{41}H_{44}$), located in the NTD. This result is important in the light of understanding the epidemiological success of group M over group O. After cross-species transmission to human, the genetic features of each group are what, more than other factors, probably determined the replication success of the isolates. The fact that group M optimized and selected its CLA motif as an adaptation to the new host could represent one of the several features that conferred to this group a replication advantage in the human host. Understanding these mechanisms is important to better comprehend the story behind important threats to human health such as the pandemic of HIV-1 group M and to be able to fight it with optimized tools.

A secondary project was carried out in collaboration with Dr. Marina Lusic and her team at the University of Heidelberg, focusing on understanding the role of the amidic amino acids present in the CLA motif. The investigation on the N of the motif led to discover a new potentially LEDGF/p75-indipendent relevant role of the IN in the choice of the integration sites. Altogether, although limited to some preliminary results, this part of my PhD work successfully demonstrated the role of the IN CTD in the choice of integration sites. Understanding the mechanism behind this phenotype could open to interesting new perspectives in lentiviral vector-based therapy. Clearly the TKTK mutant has a severe integration default and it is integrating in non-transcriptionally active regions, making it difficult to imagine its employment for therapy. However, elucidating the mechanism for which HIV-1 IN choose its integration sites, could allow to "hack" it to target specific genomic regions in the human genome, which could drastically reduce the possibility of inducing insertional mutagenesis (and potentially oncogenesis) in the target cells.

Finally, during my thesis I worked on a third project, whose objective was to understand the role of the CLA motif in the mechanism of uncoating of the viral core capsid. Although this project did not reach the level required to foresee the publication of the results obtained, it allowed me to gain relevant technical and theoretical knowledge on the subject. While I was working on it, a new wave of papers highlighting nuclear entry of intact, or almost intact, capsid cores came out. Hence, when the first lockdown caused by the Covid pandemic came, I took profit of the stop imposed to my experimental activity, to exploit my knowledge on the uncoating step to write a review on the subject, focusing on the recent breakthroughs described in the literature. The review was published in Frontiers in Microbiology and is co-

signed with Dr. Daniela Lener and Dr. Matteo Negroni and can be found as Annex 1. A second aspect for which this project resulted to be instrumental for my PhD work was that it allowed me to transfer the skills I had acquired (e.g., EURT assay) to the latest development of the project that led to the Research Article presented in the results of this thesis.

BIBLIOGRAPHY

Abecasis, A. B., Vandamme, A.-M., and Lemey, P. (2009). Quantifying Differences in the Tempo of Human Immunodeficiency Virus Type 1 Subtype Evolution. *Journal of Virology* 83, 12917–12924. doi: 10.1128/jvi.01022-09.

Accola, M. A., Höglund, S., and Göttlinger, H. G. (1998). A Putative α-Helical Structure Which Overlaps the Capsid-p2 Boundary in the Human Immunodeficiency Virus Type 1 Gag Precursor Is Crucial for Viral Particle Assembly. *Journal of Virology* 72, 2072–2078. doi: 10.1128/jvi.72.3.2072-2078.1998.

Achuthan, V., Perreira, J. M., Sowd, G. A., Puray-Chavez, M., McDougall, W. M., Paulucci-Holthauzen, A., et al. (2018). Capsid-CPSF6 Interaction Licenses Nuclear HIV-1 Trafficking to Sites of Viral DNA Integration. *Cell Host and Microbe* 24, 392-404.e8. doi: 10.1016/j.chom.2018.08.002.

Aghokeng, A. F., Ayouba, A., Mpoudi-Ngole, E., Loul, S., Liegeois, F., Delaporte, E., et al. (2010). Extensive survey on the prevalence and genetic diversity of SIVs in primate bushmeat provides insights into risks for potential new cross-species transmissions. *Infection, Genetics and Evolution* 10, 386–396. doi: 10.1016/j.meegid.2009.04.014.

Aldovini, A., and Young, R. A. (1990). Mutations of RNA and protein sequences involved in human immunodeficiency virus type 1 packaging result in production of noninfectious virus. *Journal of Virology* 64, 1920–1926. doi: 10.1128/jvi.64.5.1920-1926.1990.

Alkhatib, G., Combadiere, C., Broder, C. C., Feng, Y., Kennedy, P. E., Murphy, P. M., et al. (1996). CC CKR5: A RANTES, MIP-1α, MIP-1β Receptor as a Fusion Cofactor for Macrophage-Tropic HIV-1. *Science (1979)* 272, 1955–1958. doi: 10.1126/science.272.5270.1955.

Anderson, E. M., and Maldarelli, F. (2018). The role of integration and clonal expansion in HIV infection: live long and prosper. *Retrovirology* 15, 71. doi: 10.1186/s12977-018-0448-8.

Andrake, M. D., and Skalka, A. M. (1996). Retroviral integrase, putting the pieces together. *Journal of Biological Chemistry* 271, 19633–19636. doi: 10.1074/jbc.271.33.19633.

Anisenko, A. N., Knyazhanskaya, E. S., Zalevsky, A. O., Agapkina, J. Y., Sizov, A. I., Zatsepin, T. S., et al. (2017). Characterization of HIV-1 integrase interaction with human Ku70 protein and initial implications for drug targeting. *Scientific Reports* 7, 1–14. doi: 10.1038/s41598-017-05659-5.

Anstett, K., Brenner, B., Mesplede, T., and Wainberg, M. A. (2017). HIV drug resistance against strand transfer integrase inhibitors. *Retrovirology* 14, 1–16. doi: 10.1186/s12977-017-0360-7.

Arhel, N. J., and Kirchhoff, F. (2009). Implications of Nef: Host cell interactions in viral persistence and progression to AIDS. *Current Topics in Microbiology and Immunology* 339, 147–175. doi: 10.1007/978-3-642-02175-6_8.

Arien, K., and Verhasselt, B. (2008). HIV Nef: Role in Pathogenesis and Viral Fitness. *Current HIV Research* 6, 200–208. doi: 10.2174/157016208784325001.

Ariumi, Y., Serhan, F., Turelli, P., Telenti, A., and Trono, D. (2006). The integrase interactor 1 (INI1) proteins facilitate Tat-mediated human immunodeficiency virus type 1 transcription. *Retrovirology* 3, 1–6. doi: 10.1186/1742-4690-3-47.

Auxilien, S., Keith, G., le Grice, S. F. J., and Darlix, J. L. (1999). Role of post-transcriptional modifications of primer tRNA(Lys,3) in the fidelity and efficacy of plus strand DNA transfer during HIV-1 reverse transcription. *Journal of Biological Chemistry* 274, 4412–4420. doi: 10.1074/jbc.274.7.4412.

Ayouba, A., Souquières, S., Njinku, B., Martin, P. M. v., Müller-Trutwin, M. C., Roques, P., et al. (2000). HIV-1 group N among HIV-1-seropositive individuals in Cameroon. *AIDS* 14, 2623–2625. doi: 10.1097/00002030-200011100-00033.

Badou, A., Bennasser, Y., Moreau, M., Leclerc, C., Benkirane, M., and Bahraoui, E. (2000). Tat Protein of Human Immunodeficiency Virus Type 1 Induces Interleukin-10 in Human Peripheral Blood Monocytes: Implication of Protein Kinase C-Dependent Pathway. *Journal of Virology* 74, 10551–10562. doi: 10.1128/jvi.74.22.10551-10562.2000.

Bailes, E. (2003). Hybrid Origin of SIV in Chimpanzees. *Science (1979)* 300, 1713–1713. doi: 10.1126/science.1080657.

Baird, H. A., Galetto, R., Gao, Y., Simon-Loriere, E., Abreha, M., Archer, J., et al. (2006a). Sequence determinants of breakpoint location during HIV-1 intersubtype recombination. *Nucleic Acids Research* 34, 5203–5216. doi: 10.1093/nar/gkl669.

Baird, H. A., Gao, Y., Galetto, R., Lalonde, M., Anthony, R. M., Giacomoni, V., et al. (2006b). Influence of sequence identity and unique breakpoints on the frequency of intersubtype HIV-1 recombination. *Retrovirology* 3, 1–17. doi: 10.1186/1742-4690-3-91.

Balakrishnan, M., Yant, S. R., Tsai, L., O'Sullivan, C., Bam, R. A., Tsai, A., et al. (2013). Non-Catalytic Site HIV-1 Integrase Inhibitors Disrupt Core Maturation and Induce a Reverse Transcription Block in Target Cells. *PLoS ONE* 8. doi: 10.1371/journal.pone.0074163.

Ballandras-Colas, A., Brown, M., Cook, N. J., Dewdney, T. G., Demeler, B., Cherepanov, P., et al. (2016). Cryo-EM reveals a novel octameric integrase structure for betaretroviral intasome function. *Nature* 530, 358–361. doi: 10.1038/nature16955.

Ballandras-Colas, A., Maskell, D. P., Serrao, E., Locke, J., Swuec, P., Jónsson, S. R., et al. (2017). A supramolecular assembly mediates lentiviral DNA integration. *Science (1979)* 355, 93–95. doi: 10.1126/science.aah7002.

Balliet, J. W., Kolson, D. L., Eiger, G., Kim, F. M., McGann, K. A., Srinivasan, A., et al. (1994). Distinct Effects in Primary Macrophages and Lymphocytes of the Human Immunodeficiency Virus Type 1 Accessory Genes vpr, vpu, and nef: Mutational Analysis of a Primary HIV-1 Isolate. *Virology* 200, 623–631. doi: 10.1006/viro.1994.1225.

Barré-Sinoussi, F., Chermann, J. C., Rey, F., Nugeyre, M. T., Chamaret, S., Gruest, J., et al. (1983). Isolation of a T-Lymphotropic Retrovirus from a Patient at Risk for Acquired Immune Deficiency Syndrome (AIDS). *Science (1979)* 220, 868–871. doi: 10.1126/science.6189183.

Bbosa, N., Kaleebu, P., and Ssemwanga, D. (2019). HIV subtype diversity worldwide. *Current Opinion in HIV and AIDS* 14, 153–160. doi: 10.1097/COH.0000000000000534.

Bedwell, G. J., and Engelman, A. N. (2020). Factors that mold the nuclear landscape of HIV-1 integration. *Nucleic Acids Research* 49, 1–15. doi: 10.1093/nar/gkaa1207.

Bedwell, G. J., Jang, S., Li, W., Singh, P. K., and Engelman, A. N. (2021). rigrag: High-resolution mapping of genic targeting preferences during HIV-1 integration in vitro and in vivo. *Nucleic Acids Research* 49, 7330–7346. doi: 10.1093/nar/gkab514.

Bego, M. G., Cong, L., Mack, K., Kirchhoff, F., and Cohen, É. A. (2016). Differential Control of BST2 Restriction and Plasmacytoid Dendritic Cell Antiviral Response by Antagonists Encoded by HIV-1 Group M and O Strains. *Journal of Virology* 90, 10236–10246. doi: 10.1128/jvi.01131-16.

Bejarano, D. A., Peng, K., Laketa, V., Börner, K., Jost, K. L., Lucic, B., et al. (2019). HIV-1 nuclear import in macrophages is regulated by CPSF6-capsid interactions at the nuclear pore complex. *Elife* 8, 1–31. doi: 10.7554/eLife.41800.

Bell, S. M., and Bedford, T. (2017). Modern-day SIV viral diversity generated by extensive recombination and cross-species transmission. *PLOS Pathogens* 13, e1006466. doi: 10.1371/journal.ppat.1006466.

Benleulmi, M. S., Matysiak, J., Robert, X., Miskey, C., Mauro, E., Lapaillerie, D., et al. (2017). Modulation of the functional association between the HIV-1 intasome and the nucleosome by histone amino-terminal tails. *Retrovirology* 14, 1–16. doi: 10.1186/s12977-017-0378-x.

Berkhout, B., and Jeang, K. T. (1989). Trans Activation of Human Immunodeficiency Virus Type 1 Is Sequence Specific for Both the Single-Stranded Bulge and Loop of the Trans-Acting-Responsive Hairpin: a Quantitative Analysis. *Journal of Virology* 63, 5501–5504. doi: 10.1128/jvi.63.12.5501-5504.1989.

Bernstein, H. B., Tucker, S. P., Hunter, E., Schutzbach, J. S., and Compans, R. W. (1994). Human immunodeficiency virus type 1 envelope glycoprotein is modified by O-linked oligosaccharides. *Journal of Virology* 68, 463–468. doi: 10.1128/jvi.68.1.463-468.1994.

Berry, N., Aaby, P., Tedder, R., Whittle, H., Jaffar, S., Schim van der Loeff, M., et al. (2002). Low level viremia and high CD4% predict normal survival in a cohort of HIV type-2-infected villagers. *AIDS Research and Human Retroviruses* 18, 1167–1173. doi: 10.1089/08892220260387904.

Berthoux, L., Sebastian, S., Muesing, M. A., and Luban, J. (2007). The role of lysine 186 in HIV-1 integrase multimerization. *Virology* 364, 227–236. doi: 10.1016/j.virol.2007.02.029.

Bhattacharya, A., Alam, S. L., Fricke, T., Zadrozny, K., Sedzicki, J., Taylor, A. B., et al. (2014). Structural basis of HIV-1 capsid recognition by PF74 and CPSF6. *Proc Natl Acad Sci U S A* 111, 18625–18630. doi: 10.1073/pnas.1419945112.

Bikandou, B., Ndoundou-Nkodia, M. Y., Niama, F. R., Ekwalanga, M., Obengui, O., Taty-Taty, R., et al. (2004). Genetic subtyping of gag and env regions of HIV type 1 isolates in Republic of Congo. *AIDS Research and Human Retroviruses* 20, 1005–1009. doi: 10.1089/0889222042222737.

Bishop, K. N., Verma, M., Kim, E. Y., Wolinsky, S. M., and Malim, M. H. (2008). APOBEC3G inhibits elongation of HIV-1 reverse transcripts. *PLoS Pathogens* 4, 13–20. doi: 10.1371/journal.ppat.1000231.

Bouyac-Bertoia, M., Dvorin, J. D., Fouchier, R. A. M., Jenkins, Y., Meyer, B. E., Wu, L. I., et al. (2001). HIV-1 Infection Requires a Functional Integrase NLS. *Molecular Cell* 7, 1025–1035. doi: 10.1016/S1097-2765(01)00240-4.

Bowerman, B., Brown, P. O., Bishop, J. M., and Varmus, H. E. (1989). A nucleoprotein complex mediates the integration of retroviral DNA. *Genes Dev* 3, 469–478. doi: 10.1101/gad.3.4.469.

Brenchley, J. M., and Douek, D. C. (2008). HIV infection and the gastrointestinal immune system. *Mucosal Immunology* 1, 23–30. doi: 10.1038/mi.2007.1.

Brenchley, J. M., Schacker, T. W., Ruff, L. E., Price, D. A., Taylor, J. H., Beilman, G. J., et al. (2004). CD4+ T cell depletion during all stages of HIV disease occurs predominantly in the gastrointestinal tract. *Journal of Experimental Medicine* 200, 749–759. doi: 10.1084/jem.20040874.

Briones, M. S., Dobard, C. W., and Chow, S. A. (2010). Role of Human Immunodeficiency Virus Type 1 Integrase in Uncoating of the Viral Core. *Journal of Virology* 84, 5181–5190. doi: 10.1128/jvi.02382-09.

Brown, P. O., Bowerman, B., Varmus, H. E., and Bishop, J. M. (1989). Retroviral integration: Structure of the initial covalent product and its precursor, and a role for the viral IN protein. *Proc Natl Acad Sci U S A* 86, 2525–2529. doi: 10.1073/pnas.86.8.2525.

Brussel, A., and Sonigo, P. (2004). Evidence for Gene Expression by Unintegrated Human Immunodeficiency Virus Type 1 DNA Species. *Journal of Virology* 78, 11263–11271. doi: 10.1128/jvi.78.20.11263-11271.2004.

Buffone, C. C. C., Martinez-Lopez, A., Fricke, T., Opp, S., Severgnini, M., Cifola, I., et al. (2018). Nup153 Unlocks the Nuclear Pore Complex for HIV-1 Nuclear Translocation in Nondividing Cells. *Journal of Virology* 92, 648–666. doi: 10.1128/JVI.00648-18.

Bukrinskaya, A. (2007). HIV-1 matrix protein: A mysterious regulator of the viral life cycle. *Virus Research* 124, 1–11. doi: 10.1016/j.virusres.2006.07.001.

Burdick, R. C., Li, C., Munshi, M., Rawson, J. M. O., Nagashima, K., Hu, W.-S., et al. (2020). HIV-1 uncoats in the nucleus near sites of integration. *Proceedings of the National Academy of Sciences* 0, 201920631. doi: 10.1073/pnas.1920631117.

Burke, C. J., Sanyal, G., Bruner, M. W., Ryan, J. A., LaFemina, R. L., Robbins, H. L., et al. (1992). Structural implications of spectroscopic characterization of a putative zinc finger peptide from HIV-1 integrase. *Journal of Biological Chemistry* 267, 9639–9644. doi: 10.1016/s0021-9258(19)50138-7.

Bushman, F. D., Engelman, A., Palmer, I., Wingfield, P., and Craigie, R. (1993). Domains of the integrase protein of human immunodeficiency virus type 1 responsible for polynucleotidyl transfer and zinc binding. *Proc Natl Acad Sci U S A* 90, 3428–3432. doi: 10.1073/pnas.90.8.3428.

Cappy, P., Moisan, A., de Oliveira, F., Plantier, J. C., and Negroni, M. (2017). HIV-1 sequences in the epidemic suggest an alternative pathway for the generation of the Long Terminal Repeats. *Scientific Reports* 7. doi: 10.1038/s41598-017-14135-z.

Cara, A., and Klotman, M. E. (2006). Retroviral E-DNA: persistence and gene expression in nondividing immune cells. *Journal of Leukocyte Biology* 80, 1013–1017. doi: 10.1189/jlb.0306151.

Cardozo, E. F., Andrade, A., Mellors, J. W., Kuritzkes, D. R., Perelson, A. S., and Ribeiro, R. M. (2017). Treatment with integrase inhibitor suggests a new interpretation of HIV RNA decay curves that reveals a subset of cells with slow integration. *PLoS Pathogens* 13, 1–18. doi: 10.1371/journal.ppat.1006478.

Carr, J. K., Wolfe, N. D., Torimiro, J. N., Tamoufe, U., Mpoudi-Ngole, E., Eyzaguirre, L., et al. (2010). HIV-1 recombinants with multiple parental strains in low-prevalence, remote regions of Cameroon: Evolutionary relics? *Retrovirology* 7, 1–8. doi: 10.1186/1742-4690-7-39.

Chan, C. N., Trinité, B., Lee, C. S., Mahajan, S., Anand, A., Wodarz, D., et al. (2016). HIV-1 latency and virus production from unintegrated genomes following direct infection of resting CD4 T cells. *Retrovirology* 13, 1–22. doi: 10.1186/s12977-015-0234-9.

Charneau, P., Borman, A. M., Quillent, C., Guétard, D., Chamaret, S., Cohen, J., et al. (1994). Isolation and Envelope Sequence of a Highly Divergent HIV-1 Isolate: Definition of a New HIV-1 Group. *Virology* 205, 247–253. doi: 10.1006/viro.1994.1640.

Chen, A., Weber, I. T., Harrison, R. W., and Leis, J. (2006). Identification of amino acids in HIV-1 and avian sarcoma virus integrase subsites required for specific recognition of the long terminal repeat ends. *Journal of Biological Chemistry* 281, 4173–4182. doi: 10.1074/jbc.M510628200.

Chen, D., Wang, M., Zhou, S., and Zhou, Q. (2002). HIV-1 Tat targets microtubules to induce apoptosis, a process promoted by the pro-apoptotic Bcl-2 relative Bim. *EMBO Journal* 21, 6801–6810. doi: 10.1093/emboj/cdf683.

Chen, H., and Engelman, A. (1998). The barrier-to-autointegration protein is a host factor for HIV type 1 integration. *Proc Natl Acad Sci U S A* 95, 15270–15274. doi: 10.1073/pnas.95.26.15270.

Chen, H., Wei, S. Q., and Engelman, A. (1999). Multiple integrase functions are required to form the native structure of the human immunodeficiency virus type I intasome. *Journal of Biological Chemistry* 274, 17358–17364. doi: 10.1074/jbc.274.24.17358.

Chen, J. C. H., Krucinski, J., Miercke, L. J. W., Finer-Moore, J. S., Tang, A. H., Leavitt, A. D., et al. (2000). Crystal structure of the HIV-1 integrase catalytic core and C-terminal domains: A model for viral DNA binding. *Proc Natl Acad Sci U S A* 97, 8233–8238. doi: 10.1073/pnas.150220297.

Chen, J., Nikolaitchik, O., Singh, J., Wright, A., Bencsics, C. E., Coffin, J. M., et al. (2009). High efficiency of HIV-1 genomic RNA packaging and heterozygote formation revealed by single virion analysis. *Proc Natl Acad Sci U S A* 106, 13535–13540. doi: 10.1073/pnas.0906822106.

Chen, Y., and Belmont, A. S. (2019). Genome organization around nuclear speckles. *Current Opinion in Genetics and Development* 55, 91–99. doi: 10.1016/j.gde.2019.06.008.

Chen, Y., Zhang, Y., Wang, Y., Zhang, L., Brinkman, E. K., Adam, S. A., et al. (2018). Mapping 3D genome organization relative to nuclear compartments using TSA-Seq as a cytological ruler. *Journal of Cell Biology* 217, 4025–4048. doi: 10.1083/jcb.201807108.

Chen, Z., Luckay, A., Sodora, D. L., Telfer, P., Reed, P., Gettie, A., et al. (1997). Human immunodeficiency virus type 2 (HIV-2) seroprevalence and characterization of a distinct HIV-2 genetic subtype from the natural range of simian immunodeficiency virus-infected sooty mangabeys. *Journal of Virology* 71, 3953–3960. doi: 10.1128/jvi.71.5.3953-3960.1997.

Cherepanov, P., Ambrosio, A. L. B., Rahman, S., Ellenberger, T., and Engelman, A. (2005a). Structural basis for the recognition between HIV-1 integrase and transcriptional coactivator p75. *Proc Natl Acad Sci U S A* 102, 17308–17313. doi: 10.1073/pnas.0506924102.

Cherepanov, P., Devroe, E., Silver, P. A., and Engelman, A. (2004). Identification of an evolutionarily conserved domain in human lens epithelium-derived growth factor/transcriptional co-activator p75 (LEDGF/p75) that binds HIV-1 integrase. *Journal of Biological Chemistry* 279, 48883–48892. doi: 10.1074/jbc.M406307200.

Cherepanov, P., Maertens, G., Proost, P., Devreese, B., van Beeumen, J., Engelborghs, Y., et al. (2003). HIV-1 integrase forms stable tetramers and associates with LEDGF/p75 protein in human cells. *Journal of Biological Chemistry* 278, 372–381. doi: 10.1074/jbc.M209278200.

Cherepanov, P., Sun, Z. Y. J., Rahman, S., Maertens, G., Wagner, G., and Engelman, A. (2005b). Solution structure of the HIV-1 integrase-binding domain in LEDGF/p75. *Nature Structural and Molecular Biology* 12, 526–532. doi: 10.1038/nsmb937.

Chin, C. R., Perreira, J. M., Savidis, G., Portmann, J. M., Aker, A. M., Feeley, E. M., et al. (2015). Direct Visualization of HIV-1 Replication Intermediates Shows that Capsid and CPSF6 Modulate HIV-1 Intra-nuclear Invasion and Integration. *Cell Reports* 13, 1717–1731. doi: 10.1016/j.celrep.2015.10.036.

Chiu, Y. L., and Greene, W. C. (2009). APOBEC3G: An intracellular centurion. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364, 689–703. doi: 10.1098/rstb.2008.0193.

Christ, F., Shaw, S., Demeulemeester, J., Desimmie, B. A., Marchan, A., Butler, S., et al. (2012). Small-molecule inhibitors of the LEDGF/p75 binding site of integrase block HIV replication and modulate integrase multimerization. *Antimicrobial Agents and Chemotherapy* 56, 4365–4374. doi: 10.1128/AAC.00717-12.

Christ, F., Voet, A., Marchand, A., Nicolet, S., Desimmie, B. A., Marchand, D., et al. (2010). Rational design of small-molecule inhibitors of the LEDGF/p75-integrase interaction and HIV replication. *Nature Chemical Biology* 6, 442–448. doi: 10.1038/nchembio.370.

Cihlar, T., and Fordyce, M. (2016). Current status and prospects of HIV treatment. *Current Opinion in Virology* 18, 50–56. doi: 10.1016/j.coviro.2016.03.004.

Ciuffi, A., Llano, M., Poeschla, E., Hoffmann, C., Leipzig, J., Shinn, P., et al. (2005). A role for LEDGF/p75 in targeting HIV DNA integration. *Nature Medicine* 11, 1287–1289. doi: 10.1038/nm1329.

Clavel, F., and Orenstein, J. M. (1990). A mutant of human immunodeficiency virus with reduced RNA packaging and abnormal particle morphology. *Journal of Virology* 64, 5230–5234. doi: 10.1128/jvi.64.10.5230-5234.1990.

Coffin, J. (1979). Replication, and Recombination of Retrovirus Genomes: Some Unifying Hypotheses. *J Gen Virol* 41, 1–26.

Coffin, J., and Swanstrom, R. (2013). HIV pathogenesis: Dynamics and genetics of viral populations and infected cells. *Cold Spring Harbor Perspectives in Medicine* 3. doi: 10.1101/cshperspect.a012526.

Cohen, E. A., Dehni, G., Sodroski, J. G., and Haseltine, W. A. (1990). Human immunodeficiency virus vpr product is a virion-associated regulatory protein. *Journal of Virology* 64, 3097–3099. doi: 10.1128/jvi.64.6.3097-3099.1990.

Cohen, M. S., Shaw, G. M., McMichael, A. J., and Haynes, B. F. (2011). Acute HIV-1 Infection. *New England Journal of Medicine* 364, 1943–1954. doi: 10.1056/NEJMra1011874.

Connor, R. I., Chen, B. K., Choe, S., and Landau, N. R. (1995). Vpr is required for efficient replication of human immunodeficiency virus type-1 in mononuclear phagocytes. *Virology* 206, 935–944. doi: 10.1006/viro.1995.1016.

Conticello, S. G., Thomas, C. J. F., Petersen-Mahrt, S. K., and Neuberger, M. S. (2005). Evolution of the AID/APOBEC family of polynucleotide (deoxy)cytidine deaminases. *Molecular Biology and Evolution* 22, 367–377. doi: 10.1093/molbev/msi026.

Cook, N. J., Li, W., Berta, D., Badaoui, M., Ballandras-Colas, A., Nans, A., et al. (2020). Structural basis of second-generation HIV integrase inhibitor action and viral resistance. *Science (1979)* 367, 806–810. doi: 10.1126/science.aay4919.

Cooper, D. A., Steigbigel, R. T., Gatell, J. M., Rockstroh, J. K., Katlama, C., Yeni, P., et al. (2008). Subgroup and Resistance Analyses of Raltegravir for Resistant HIV-1 Infection. *New England Journal of Medicine* 359, 355–365. doi: 10.1056/nejmoa0708978.

Crooks, G. E., Hon, G., Chandonia, J.-M., and Brenner, S. E. (2004). WebLogo: A Sequence Logo Generator: Figure 1. *Genome Research* 14, 1188–1190. doi: 10.1101/gr.849004.

Cujec, T. P., Cho, H., Maldonado, E., Meyer, J., Reinberg, D., and Peterlin, B. M. (1997). The human immunodeficiency virus transactivator Tat interacts with the RNA polymerase II holoenzyme. *Molecular and Cellular Biology* 17, 1817–1823. doi: 10.1128/mcb.17.4.1817.

da Silva, J., Coetzer, M., Nedellec, R., Pastore, C., and Mosier, D. E. (2010). Fitness epistasis and constraints on adaptation in a human immunodeficiency virus type 1 protein region. *Genetics* 185, 293–303. doi: 10.1534/genetics.109.112458.

da Silva Santos, C., Tartour, K., and Cimarelli, A. (2016). A Novel Entry/Uncoating Assay Reveals the Presence of at Least Two Species of Viral Capsids During Synchronized HIV-1 Infection. *PLoS Pathogens* 12. doi: 10.1371/journal.ppat.1005897.

Damond, F., Apetrei, C., Robertson, D. L., Souquière, S., Leprêtre, A., Matheron, S., et al. (2001). Variability of human immunodeficiency virus type 2 (HIV-2) infecting patients living in France. *Virology* 280, 19–30. doi: 10.1006/viro.2000.0685.

Dar, M. J., Monel, B., Krishnan, L., Shun, M.-C., di Nunzio, F., Helland, D. E., et al. (2009). Biochemical and virological analysis of the 18-residue C-terminal tail of HIV-1 integrase. *Retrovirology* 6, 94. doi: 10.1186/1742-4690-6-94.

D'Arc, M., Ayouba, A., Esteban, A., Learn, G. H., Boué, V., Liegeois, F., et al. (2015). Origin of the HIV-1 group O epidemic in western lowland gorillas. *Proc Natl Acad Sci U S A* 112, E1343–E1352. doi: 10.1073/pnas.1502022112.

Das, A. T., Klaver, B., and Berkhout, B. (1999). A Hairpin Structure in the R Region of the Human Immunodeficiency Virus Type 1 RNA Genome Is Instrumental in Polyadenylation Site Selection. *Journal of Virology* 73, 81–91. doi: 10.1128/jvi.73.1.81-91.1999.

Das, A. T., Koken, S. E. C., Oude Essink, B. B., van Wamel, J. L. B., and Berkhout, B. (1994). Human immunodeficiency virus uses tRNALys,3 as primer for reverse transcription in HeLa-CD4+ cells. *FEBS Letters* 341, 49–53. doi: 10.1016/0014-5793(94)80238-6.

David Pauza, C. (1990). Two bases are deleted from the termini of HIV-1 linear DNA during integrative recombination. *Virology* 179, 886–889. doi: 10.1016/0042-6822(90)90161-J.

de Castro, I. J., Budzak, J., di Giacinto, M. L., Ligammari, L., Gokhan, E., Spanos, C., et al. (2017). Repo-Man/PP1 regulates heterochromatin formation in interphase. *Nature Communications* 8. doi: 10.1038/ncomms14048.

de Houwer, S., Demeulemeester, J., Thys, W., Rocha, S., Dirix, L., Gijsbers, R., et al. (2014). The HIV-1 integrase mutant R263A/K264A is 2-fold defective for TRN-SR2 binding and viral nuclear import. *Journal of Biological Chemistry* 289, 25351–25361. doi: 10.1074/jbc.M113.533281.

de Oliveira, F., Cappy, P., Lemée, V., Moisan, A., Pronier, C., Bocket, L., et al. (2018). Detection of numerous HIV-1/MO recombinants in France. *AIDS* 32, 1289–1299. doi: 10.1097/QAD.0000000000001814.

de Oliveira, F., Mourez, T., Vessiere, A., Ngoupo, P. A., Alessandri-Gradt, E., Simon, F., et al. (2017). Multiple HIV-1/M + HIV-1/O dual infections and new HIV-1/MO inter-group recombinant forms detected in Cameroon. *Retrovirology* 14, 1–11. doi: 10.1186/s12977-016-0324-3.

de Sousa, J. D., Alvarez, C., Vandamme, A. M., and Müller, V. (2012). Enhanced heterosexual transmission hypothesis for the origin of pandemic HIV-1. *Viruses* 4, 1950–1983. doi: 10.3390/v4101950.

DeAnda, F., Hightower, K. E., Nolte, R. T., Hattori, K., Yoshinaga, T., Kawasuji, T., et al. (2013). Dolutegravir Interactions with HIV-1 Integrase-DNA: Structural Rationale for Drug Resistance and Dissociation Kinetics. *PLoS ONE* 8, 1–12. doi: 10.1371/journal.pone.0077448.

Demeulemeester, J., Vets, S., Schrijvers, R., Madlala, P., de Maeyer, M., de Rijck, J., et al. (2014). HIV-1 integrase variants retarget viral integration and are associated with disease progression in a chronic infection cohort. *Cell Host and Microbe* 16, 651–662. doi: 10.1016/j.chom.2014.09.016.

Desimmie, B. A., Schrijvers, R., Demeulemeester, J., Borrenberghs, D., Weydert, C., Thys, W., et al. (2013). LEDGINs inhibit late stage HIV-1 replication by modulating integrase multimerization in the virions. *Retrovirology* 10, 57. doi: 10.1186/1742-4690-10-57.

Dettwiler, S., Aringhieri, C., Cardinale, S., Keller, W., and Barabino, S. M. L. (2004). Distinct sequence motifs within the 68-kDa subunit of cleavage factor Im mediate RNA binding, protein-protein interactions, and subcellular localization. *Journal of Biological Chemistry* 279, 35788–35797. doi: 10.1074/jbc.M403927200.

Devroe, E., Engelman, A., and Silver, P. A. (2003). Intracellular transport of human immunodeficiency virus type 1 integrase. *Journal of Cell Science* 116, 4401–4408. doi: 10.1242/jcs.00747.

Dharan, A., Bachmann, N., Talley, S., Zwikelmaier, V., and Campbell, E. M. (2020). Nuclear pore blockade reveals that HIV-1 completes reverse transcription and uncoating in the nucleus. *Nat Microbiol*. doi: 10.1038/s41564-020-0735-8.

Dingwall, C., Ernberg, I., Gait, M. J., Green, S. M., Heaphy, S., Karn, J., et al. (1990). HIV-1 tat protein stimulates transcription by binding to a U-rich bulge in the stem of the TAR RNA structure. *EMBO Journal* 9, 4145–4153. doi: 10.1002/j.1460-2075.1990.tb07637.x.

Dobard, C. W., Briones, M. S., and Chow, S. A. (2007). Molecular Mechanisms by Which Human Immunodeficiency Virus Type 1 Integrase Stimulates the Early Steps of Reverse Transcription. *Journal of Virology* 81, 10037–10046. doi: 10.1128/jvi.00519-07.

Dorfman, T., Mammano, F., Haseltine, W. A., and Göttlinger, H. G. (1994). Role of the matrix protein in the virion association of the human immunodeficiency virus type 1 envelope glycoprotein. *Journal of Virology* 68, 1689–1696. doi: 10.1128/jvi.68.3.1689-1696.1994.

Douglas, J. L., Viswanathan, K., McCarroll, M. N., Gustin, J. K., Früh, K., and Moses, A. v. (2009). Vpu Directs the Degradation of the Human Immunodeficiency Virus Restriction Factor BST-2/Tetherin via a βTrCP-Dependent Mechanism. *Journal of Virology* 83, 7931–7947. doi: 10.1128/jvi.00242-09.

Douglas, N. W., Munro, G. H., and Daniels, R. S. (1997). HIV/SIV glycoproteins: Structure-function relationships. *Journal of Molecular Biology* 273, 122–149. doi: 10.1006/jmbi.1997.1277.

Drelich, M., Wilhelm, R., and Mous, J. (1992). Identification of amino acid residues critical for endonuclease and integration activities of HIV-1 IN protein in Vitro. *Virology* 188, 459–468. doi: 10.1016/0042-6822(92)90499-F.

Dupont, L., Bloor, S., Williamson, J. C., Cuesta, S. M., Shah, R., Teixeira-Silva, A., et al. (2021). The SMC5/6 complex compacts and silences unintegrated HIV-1 DNA and is antagonized by Vpr. *Cell Host and Microbe* 29, 792-805.e6. doi: 10.1016/j.chom.2021.03.001.

Dyda, F., Hickman, A. B., Jenkins, T. M., Engelman, A., Craigie, R., and Davies, D. R. (1994). Crystal structure of the catalytic domain of HIV-1 integrase: Similarity to other polynucleotidyl transferases. *Science (1979)* 266, 1981–1986. doi: 10.1126/science.7801124.

Eichorst, J. P., Chen, Y., Mueller, J. D., and Mansky, L. M. (2018). Distinct pathway of human T-cell leukemia virus type 1 Gag punctum biogenesis provides new insights into enveloped virus assembly. *mBio* 9. doi: 10.1128/mBio.00758-18.

Eidahl, J. O., Crowe, B. L., North, J. A., McKee, C. J., Shkriabai, N., Feng, L., et al. (2013). Structural basis for high-affinity binding of LEDGF PWWP to mononucleosomes. *Nucleic Acids Research* 41, 3924–3936. doi: 10.1093/nar/gkt074.

Eijkelenboom, A. P. A. M., Puras Lutzke, R. A., Boelens, R., Plasterk, R. H. A., Kaptein, R., and Hård, K. (1995). The DNA-binding domain of HIV-1 integrase has an SH3-like fold. *Nature Structural & Molecular Biology* 2, 807–810. doi: 10.1038/nsb0995-807.

Eisele, E., and Siliciano, R. F. (2012). Redefining the Viral Reservoirs that Prevent HIV-1 Eradication. *Immunity* 37, 377–388. doi: 10.1016/j.immuni.2012.08.010.

El-Amine, R., Germini, D., Zakharova, V. v., Tsfasman, T., Sheval, E. v., Louzada, R. A. N., et al. (2018). HIV-1 Tat protein induces DNA damage in human peripheral blood B-lymphocytes via mitochondrial ROS production. *Redox Biology* 15, 97–108. doi: 10.1016/j.redox.2017.11.024.

Elliott, J., Eschbach, J. E., Koneru, P. C., Li, W., Puray-Chavez, M., Townsend, D., et al. (2020). Integrase-RNA interactions underscore the critical role of integrase in HIV-1 virion morphogenesis. *Elife* 9, 1–56. doi: 10.7554/ELIFE.54311.

Emiliani, S., Mousnier, A., Busschots, K., Maroun, M., van Maele, B., Tempé, D., et al. (2005). Integrase Mutants Defective for Interaction with LEDGF/p75 Are Impairedin Chromosome Tethering and HIV-1Replication. *Journal of Biological Chemistry* 280, 25517–25523. doi: 10.1074/jbc.M501378200.

Engelman, A. (1999). In vivo analysis of retroviral integrase structure and function. *Adv Virus Res* 52, 411–426. doi: 10.1016/s0065-3527(08)60309-7.

Engelman, A., Bushman, F. D., and Craigie, R. (1993). Identification of discrete functional domains of HIV-1 integrase and their organization within an active multimeric complex. *EMBO Journal* 12, 3269–3275. doi: 10.1002/j.1460-2075.1993.tb05996.x.

Engelman, A., and Craigie, R. (1992). Identification of conserved amino acid residues critical for human immunodeficiency virus type 1 integrase function in vitro. *Journal of Virology* 66, 6361–6369. doi: 10.1128/jvi.66.11.6361-6369.1992.

Engelman, A., Englund, G., Orenstein, J. M., Martin, M. A., and Craigie, R. (1995). Multiple effects of mutations in human immunodeficiency virus type 1 integrase on viral replication. *Journal of Virology* 69, 2729–2736. doi: 10.1128/jvi.69.5.2729-2736.1995.

Engelman, A., Hickman, A. B., and Craigie, R. (1994). The core and carboxyl-terminal domains of the integrase protein of human immunodeficiency virus type 1 each contribute to nonspecific DNA binding. *Journal of Virology* 68, 5911–5917. doi: 10.1128/jvi.68.9.5911-5917.1994.

Engelman, A., Liu, Y., Chen, H., Farzan, M., and Dyda, F. (1997). Structure-based mutagenesis of the catalytic domain of human immunodeficiency virus type 1 integrase. *Journal of Virology* 71, 3507–3514. doi: 10.1128/jvi.71.5.3507-3514.1997.

Engelman, A., Mizuuchi, K., and Craigie, R. (1991). HIV-1 DNA integration: Mechanism of viral DNA cleavage and DNA strand transfer. *Cell* 67, 1211–1221. doi: 10.1016/0092-8674(91)90297-C.

Engelman, A. N. (2019). Multifaceted HIV integrase functionalities and therapeutic strategies for their inhibition. *Journal of Biological Chemistry* 294, 15137–15157. doi: 10.1074/jbc.REV119.006901.

Engelman, A. N., and Cherepanov, P. (2017). Retroviral intasomes arising. *Current Opinion in Structural Biology* 47, 23–29. doi: 10.1016/j.sbi.2017.04.005.

Engelman, A. N., and Kvaratskhelia, M. (2022). Multimodal Functionalities of HIV-1 Integrase. *Viruses* 14, 926. doi: 10.3390/v14050926.

Engelman, A. N., and Maertens, G. N. (2018). *Virus-host interactions in retrovirus integration*. Elsevier Inc. doi: 10.1016/B978-0-12-811185-7.00004-2.

Engelman, A. N., and Singh, P. K. (2018). Cellular and molecular mechanisms of HIV-1 integration targeting. *Cellular and Molecular Life Sciences* 75, 2491–2507. doi: 10.1007/s00018-018-2772-5.

Erickson-Viitanen, S., Manfredi, J., Viitanen, P., Tribe, D. E., Tritch, R., Hutchison, C. A., et al. (1989). Cleavage of HIV-1 gag Polyprotein Synthesized In Vitro: Sequential Cleavage by the Viral Protease. *AIDS Research and Human Retroviruses* 5, 577–591. doi: 10.1089/aid.1989.5.577.

Espeseth, A. S., Fishel, R., Hazuda, D., Huang, Q., Xu, M., Yoder, K., et al. (2011). Sirna screening of a targeted library of DNA repair factors in HIV infection reveals a role for base excision repair in HIV integration. *PLoS ONE* 6. doi: 10.1371/journal.pone.0017612.

Esposito, D., and Craigie, R. (1998). Sequence specificity of viral end DNA binding by HIV-1 integrase reveals critical regions for protein-DNA interaction. *EMBO Journal* 17, 5832–5843. doi: 10.1093/emboj/17.19.5832.

Etienne, L., Hahn, B. H., Sharp, P. M., Matsen, F. A., and Emerman, M. (2013). Gene Loss and Adaptation to Hominids Underlie the Ancient Origin of HIV-1. *Cell Host & Microbe* 14, 85–92. doi: 10.1016/j.chom.2013.06.002.

Fan, J., Negroni, M., and Robertson, D. L. (2007). The distribution of HIV-1 recombination breakpoints. *Infection, Genetics and Evolution* 7, 717–723. doi: 10.1016/j.meegid.2007.07.012.

Faria, N. R., Rambaut, A., Suchard, M. A., Baele, G., Bedford, T., Ward, M. J., et al. (2014). The early spread and epidemic ignition of HIV-1 in human populations. *Science (1979)* 346, 56–61. doi: 10.1126/science.1256739.

Farnet, C. M., and Bushman, F. D. (1997). HIV-1 cDNA integration: Requirement of HMG I(Y) protein for function of preintegration complexes in vitro. *Cell* 88, 483–492. doi: 10.1016/S0092-8674(00)81888-7.

Farnet, C. M., and Haseltine, W. A. (1990). Integration of human immunodeficiency virus type 1 DNA in vitro. *Proc Natl Acad Sci U S A* 87, 4164–4168. doi: 10.1073/pnas.87.11.4164.

Felber, B. K., Hadzopoulou-Cladaras, M., Cladaras, C., Copeland, T., and Pavlakis, G. N. (1989). rev protein of human immunodeficiency virus type 1 affects the stability and transport of the viral mRNA. *Proceedings of the National Academy of Sciences* 86, 1495–1499. doi: 10.1073/pnas.86.5.1495.

Feng, L., Dharmarajan, V., Serrao, E., Hoyte, A., Larue, R. C., Slaughter, A., et al. (2016). The Competitive Interplay between Allosteric HIV-1 Integrase Inhibitor BI/D and LEDGF/p75 during the Early Stage of HIV-1 Replication Adversely Affects Inhibitor Potency. *ACS Chemical Biology* 11, 1313–1321. doi: 10.1021/acschembio.6b00167.

Feng, Y., Broder, C. C., Kennedy, P. E., and Berger, E. A. (1996). HIV-1 entry cofactor: Functional cDNA cloning of a seven-transmembrane, G protein-coupled receptor. *Science (1979)* 272, 872–877. doi: 10.1126/science.272.5263.872.

Fenwick, C., Amad, M., Bailey, M. D., Bethell, R., Bös, M., Bonneau, P., et al. (2014). Preclinical profile of BI 224436, a novel hiv-1 non-catalytic-site integrase inhibitor. *Antimicrobial Agents and Chemotherapy* 58, 3233–3244. doi: 10.1128/AAC.02719-13.

Ferris, A. L., Wu, X., Hughes, C. M., Stewart, C., Smith, S. J., Milne, T. A., et al. (2010). Lens epithelium-derived growth factor fusion proteins redirect HIV-1 DNA integration. *Proc Natl Acad Sci U S A* 107, 3135–3140. doi: 10.1073/pnas.0914142107.

Finzi, D., Blankson, J., Siliciano, J. D., Margolick, J. B., Chadwick, K., Pierson, T., et al. (1999). Latent infection of CD4+ T cells provides a mechanism for lifelong persistence of HIV-1, even in patients on effective combination therapy. *Nature Medicine* 5, 512–517. doi: 10.1038/8394.

Fontana, J., Jurado, K. A., Cheng, N., Ly, N. L., Fuchs, J. R., Gorelick, R. J., et al. (2015). Distribution and Redistribution of HIV-1 Nucleocapsid Protein in Immature, Mature, and Integrase-Inhibited Virions: a Role for Integrase in Maturation. *Journal of Virology* 89, 9765–9780. doi: 10.1128/jvi.01522-15.

Francis, A. C., Marin, M., Prellberg, M. J., Palermino-Rowland, K., and Melikyan, G. B. (2020a). HIV-1 Uncoating and Nuclear Import Precede the Completion of Reverse Transcription in Cell Lines and in Primary Macrophages. *Viruses* 12, 1234. doi: 10.3390/v12111234.

Francis, A. C., Marin, M., Singh, P. K., Achuthan, V., Prellberg, M. J., Palermino-Rowland, K., et al. (2020b). HIV-1 replication complexes accumulate in nuclear speckles and integrate into speckle-associated genomic domains. *Nature Communications* 11. doi: 10.1038/s41467-020-17256-8.

Franke, E. K., Yuan, H. E. H., and Luban, J. (1994). Specific incorporation of cyclophilin a into HIV-1 virions. *Nature* 372, 359–362. doi: 10.1038/372359a0.

Fujiwara, T., and Mizuuchi, K. (1988). Retroviral DNA integration: Structure of an integration intermediate. *Cell* 54, 497–504. doi: 10.1016/0092-8674(88)90071-2.

Galganski, L., Urbanek, M. O., and Krzyzosiak, W. J. (2017). Nuclear speckles: Molecular organization, biological function and role in disease. *Nucleic Acids Research* 45, 10350–10368. doi: 10.1093/nar/gkx759.

Gallay, P., Hope, T., Chin, D., and Trono, D. (1997). HIV-1 infection of nondividing cells through the recognition of integrase by the importin/karyopherin pathway. *Proc Natl Acad Sci U S A* 94, 9825–9830. doi: 10.1073/pnas.94.18.9825.

Galli, A., Kearney, M., Nikolaitchik, O. A., Yu, S., Chin, M. P. S., Maldarelli, F., et al. (2010). Patterns of Human Immunodeficiency Virus Type 1 Recombination Ex Vivo Provide Evidence for Coadaptation of Distant Sites, Resulting in Purifying Selection for Intersubtype Recombinants during Replication. *Journal of Virology* 84, 7651–7661. doi: 10.1128/JVI.00276-10.

Gallo, R. C., Sarin, P. S., Gelmann, E. P., Robert-Guroff, M., Richardson, E., Kalyanaraman, V. S., et al. (1983). Isolation of Human T-Cell Leukemia Virus in Acquired Immune Deficiency Syndrome (AIDS). *Science (1979)* 220, 865–867. doi: 10.1126/science.6601823.

Gamble, T. R., Vajdos, F. F., Yoo, S., Worthylake, D. K., Houseweart, M., Sundquist, W. I., et al. (1996). Crystal structure of human cyclophilin A bound to the amino-terminal domain of HIV-1 capsid. *Cell* 87, 1285–1294. doi: 10.1016/S0092-8674(00)81823-1.

Gao, F., Bailes, E., Robertson, D. L., Chen, Y., Rodenburg, C. M., Michael, S. F., et al. (1999). Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. *Nature* 397, 436–441. doi: 10.1038/17130.

Gao, F., Yue, L., White, A. T., Pappas, P. G., Barchue, J., Hanson, A. P., et al. (1992). Human infection by genetically diverse SIVSM-related HIV-2 in West Africa. *Nature* 358, 495–499. doi: 10.1038/358495a0.

Garrus, J. E., von Schwedler, U. K., Pornillos, O. W., Morham, S. G., Zavitz, K. H., Wang, H. E., et al. (2001). Tsg101 and the vacuolar protein sorting pathway are essential for HIV-1 budding. *Cell* 107, 55–65. doi: 10.1016/S0092-8674(01)00506-2.

Gaynor, R. (1992). Cellular Transcription Factors Involved in the Regulation of HIV-1 Gene Expression. *AIDS* 6, 347–364. doi: 10.1097/00002030-199204000-00001.

Geis, F. K., and Goff, S. P. (2019). Unintegrated HIV-1 DNAs are loaded with core and linker histones and transcriptionally silenced. *Proceedings of the National Academy of Sciences* 116, 23735–23742. doi: 10.1073/pnas.1912638116.

Geis, F. K., Sabo, Y., Chen, X., Li, Y., Lu, C., and Goff, S. P. (2022). CHAF1A/B mediate silencing of unintegrated HIV-1 DNAs early in infection. *Proc Natl Acad Sci U S A* 119. doi: 10.1073/pnas.2116735119.

Gelbart, M., and Stern, A. (2020). Site-Specific Evolutionary Rate Shifts in HIV-1 and SIV. *Viruses* 12, 8–12. doi: 10.3390/v12111312.

Gien, H., Morse, M., McCauley, M. J., Kitzrow, J. P., Musier-Forsyth, K., Gorelick, R. J., et al. (2022). HIV-1 Nucleocapsid Protein Binds Double-Stranded DNA in Multiple Modes to Regulate Compaction and Capsid Uncoating. *Viruses* 14. doi: 10.3390/v14020235.

Gijsbers, R., Ronen, K., Vets, S., Malani, N., de Rijck, J., McNeely, M., et al. (2010). LEDGF hybrids efficiently retarget lentiviral integration into heterochromatin. *Molecular Therapy* 18, 552–560. doi: 10.1038/mt.2010.36.

Goffinet, C., Allespach, I., Homann, S., Tervo, H. M., Habermann, A., Rupp, D., et al. (2009). HIV-1 Antagonism of CD317 Is Species Specific and Involves Vpu-Mediated Proteasomal Degradation of the Restriction Factor. *Cell Host and Microbe* 5, 285–297. doi: 10.1016/j.chom.2009.01.009.

Gottlieb, M. S., Schanker, M. D., Fan, P. T., Saxon, M. D., and Weisman, J. D. (1981). Centers for Disease Control (CDC). *Morb Mortal Wkly Rep* 30, 205–252.

Goujon, C., Arfi, V., Pertel, T., Luban, J., Lienard, J., Rigal, D., et al. (2008). Characterization of Simian Immunodeficiency Virus SIV SM /Human Immunodeficiency Virus Type 2 Vpx Function in Human Myeloid Cells. *Journal of Virology* 82, 12335–12345. doi: 10.1128/JVI.01181-08.

Goujon, C., Jarrosson-Wuillème, L., Bernaud, J., Rigal, D., Darlix, J. L., and Cimarelli, A. (2006). With a little help from a friend: Increasing HIV transduction of monocyte-derived dendritic cells with virion-like particles of SIVMAC. *Gene Therapy* 13, 991–994. doi: 10.1038/sj.gt.3302753.

Gray, R. R., Tatem, A. J., Lamers, S., Hou, W., Laeyendecker, O., Serwadda, D., et al. (2009). Spatial phylodynamics of HIV-1 epidemic emergence in east Africa. *AIDS* 23, F9–F17. doi: 10.1097/QAD.0b013e32832faf61.

Grobler, J. A., Stillmock, K., Hu, B., Witmer, M., Felock, P., Espeseth, A. S., et al. (2002). Diketo acid inhibitor mechanism and HIV-1 integrase: Implications for metal binding in the active site of phosphotransferase enzymes. *Proc Natl Acad Sci U S A* 99, 6661–6666. doi: 10.1073/pnas.092056199.

Gross, I., Hohenberg, H., Wilk, T., Wiegers, K., Grättinger, M., Müller, B., et al. (2000). A conformational switch controlling HIV-1 morphogenesis. *EMBO Journal* 19, 103–113. doi: 10.1093/emboj/19.1.103.

Gruber, A. R., Martin, G., Keller, W., and Zavolan, M. (2012). Cleavage factor Im is a key regulator of 3' UTR length. *RNA Biology* 9, 1405–1412. doi: 10.4161/rna.22570.

Gupta, K., Brady, T., Dyer, B. M., Malani, N., Hwang, Y., Male, F., et al. (2014). Allosteric inhibition of human immunodeficiency virus integrase: Late block during viral replication and abnormal multimerization involving specific protein domains. *Journal of Biological Chemistry* 289, 20477–20488. doi: 10.1074/jbc.M114.551119.

Gupta, K., Turkki, V., Sherrill-Mix, S., Hwang, Y., Eilers, G., Taylor, L., et al. (2016). Structural Basis for Inhibitor-Induced Aggregation of HIV Integrase. *PLoS Biology* 14, 1–24. doi: 10.1371/journal.pbio.1002584.

Gupta, R. K., and Towers, G. J. (2009). A Tail of Tetherin: How Pandemic HIV-1 Conquered the World. *Cell Host & Microbe* 6, 393–395. doi: 10.1016/j.chom.2009.11.002.

Gürtler, L. G., Hauser, P. H., Eberle, J., von Brunn, A., Knapp, S., Zekeng, L., et al. (1994). A new subtype of human immunodeficiency virus type 1 (MVP-5180) from Cameroon. *Journal of Virology* 68, 1581–1585. doi: 10.1128/jvi.68.3.1581-1585.1994.

Hahn, B. H., Shaw, G. M., de Cock, K. M., and Sharp, P. M. (2000). AIDS as a zoonosis: Scientific and public health implications. *Science (1979)* 287, 607–614. doi: 10.1126/science.287.5453.607.

Hamel, D. J., Sankalé, J. L., Eisen, G., Meloni, S. T., Mullins, C., Gueye-Ndiaye, A., et al. (2007). Twenty years of prospective molecular epidemiology in Senegal: Changes in HIV diversity. *AIDS Research and Human Retroviruses* 23, 1189–1196. doi: 10.1089/aid.2007.0037.

Hare, S., di Nunzio, F., Labeja, A., Wang, J., Engelman, A., and Cherepanov, P. (2009a). Structural basis for functional tetramerization of lentiviral integrase. *PLoS Pathogens* 5. doi: 10.1371/journal.ppat.1000515.

Hare, S., Gupta, S. S., Valkov, E., Engelman, A., and Cherepanov, P. (2010). Retroviral intasome assembly and inhibition of DNA strand transfer. *Nature* 464, 232–236. doi: 10.1038/nature08784.

Hare, S., Maertens, G. N., and Cherepanov, P. (2012). 3′-Processing and strand transfer catalysed by retroviral integrase in crystallo. *EMBO Journal* 31, 3020–3028. doi: 10.1038/emboj.2012.118.

Hare, S., Shun, M. C., Gupta, S. S., Valkov, E., Engelman, A., and Cherepanov, P. (2009b). A novel co-crystal structure affords the design of gain-of-function lentiviral integrase mutants in the presence of modified PSIP1/LEDGF/p75. *PLoS Pathogens* 5. doi: 10.1371/journal.ppat.1000259.

Hare, S., Smith, S. J., Métifiot, M., Jaxa-Chamiec, A., Pommier, Y., Hughes, S. H., et al. (2011). Structural and functional analyses of the second-generation integrase strand transfer inhibitor dolutegravir (S/GSK1349572). *Molecular Pharmacology* 80, 565–572. doi: 10.1124/mol.111.073189.

Harper, A. L., Skinner, L. M., Sudol, M., and Katzman, M. (2001). Use of Patient-Derived Human Immunodeficiency Virus Type 1 Integrases To Identify a Protein Residue That Affects Target Site Selection. *Journal of Virology* 75, 7756–7762. doi: 10.1128/jvi.75.16.7756-7762.2001.

Harris, R. S., and Dudley, J. P. (2015). APOBECs and virus restriction. *Virology* 479–480, 131–145. doi: 10.1016/j.virol.2015.03.012.

Harris, R. S., Hultquist, J. F., and Evans, D. T. (2012). The restriction factors of human immunodeficiency virus. *Journal of Biological Chemistry* 287, 40875–40883. doi: 10.1074/jbc.R112.416925.

Hasegawa, M., Kishino, H., and Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* 22, 160–174.

Hattori, N., Michaels, F., Fargnoli, K., Marcon, L., Gallo, R. C., and Franchini, G. (1990). The human immunodeficiency virus type 2 vpr gene is essential for productive infection of human macrophages. *Proc Natl Acad Sci U S A* 87, 8080–8084. doi: 10.1073/pnas.87.20.8080.

Hayouka, Z., Rosenbluh, J., Levin, A., Loya, S., Lebendiker, M., Veprintsev, D., et al. (2007). Inhibiting HIV-1 integrase by shifting its oligomerization equilibrium. *Proceedings of the National Academy of Sciences* 104, 8316–8321. doi: 10.1073/pnas.0700781104.

Hazuda, D. J., Felock, P., Witmer, M., Wolfe, A., Stillmock, K., Grobler, J. A., et al. (2000). Inhibitors of strand transfer that prevent integration and inhibit HIV-1 replication in cells. *Science (1979)* 287, 646–650. doi: 10.1126/science.287.5453.646.

He, J., Choe, S., Walker, R., di Marzio, P., Morgan, D. O., and Landau, N. R. (1995). Human immunodeficiency virus type 1 viral protein R (Vpr) arrests cells in the G2 phase of the cell cycle by inhibiting p34cdc2 activity. *Journal of Virology* 69, 6705–6711. doi: 10.1128/jvi.69.11.6705-6711.1995.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., et al. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell* 38, 576–589. doi: 10.1016/j.molcel.2010.05.004.

Hemelaar, J. (2012). The origin and diversity of the HIV-1 pandemic. *Trends in Molecular Medicine* 18, 182–192. doi: 10.1016/j.molmed.2011.12.001.

Hemelaar, J., Gouws, E., Ghys, P. D., and Osmanov, S. (2006). Global and regional distribution of HIV-1 genetic subtypes and recombinants in 2004. *Aids* 20, 13–23. doi: 10.1097/01.aids.0000247564.73009.bc.

Hemelaar, J., Gouws, E., Ghys, P. D., and Osmanov, S. (2011). Global trends in molecular epidemiology of HIV-1 during 2000–2007. *AIDS* 25, 679–689. doi: 10.1097/QAD.0b013e328342ff93.

Heuer, T. S., and Brown, P. O. (1997). Mapping features of HIV-1 integrase near selected sites on viral and target DNA molecules in an active enzyme - DNA complex by photo-cross- linking. *Biochemistry* 36, 10655–10665. doi: 10.1021/bi970782h.

Heuverswyn, F. van, Li, Y., Bailes, E., Neel, C., Lafay, B., Keele, B. F., et al. (2007). Genetic diversity and phylogeographic clustering of SIVcpzPtt in wild chimpanzees in Cameroon. *Virology* 368, 155–171. doi: 10.1016/j.virol.2007.06.018.

Hill, C. P., Worthylake, D., Bancroft, D. P., Christensen, A. M., and Sundquist, W. I. (1996). Crystal structures of the trimeric human immunodeficiency virus type 1 matrix protein: Implications for membrane association and assembly. *Proc Natl Acad Sci U S A* 93, 3099–3104. doi: 10.1073/pnas.93.7.3099.

Hill, M., Bellamy-McIntyre, A., Vella, L., Campbell, S., Marshall, J., Tachedjian, G., et al. (2006). Alteration of the Proline at Position 7 of the HIV-1 Spacer Peptide p1 Suppresses Viral Infectivity in a Strain Dependent Manner. *Current HIV Research* 5, 69–78. doi: 10.2174/157016207779316323.

Hladik, F., and McElrath, M. J. (2008). Setting the Stage-HIV Host invasion. *Nature Reviews Immunology* 8, 447–457. doi: 10.1038/nri2302.Setting.

Ho DD, Neumann AU, Perelson AS, Chen W, Leonard JM, and M, M. (1995). Rapid Turnover of Plasma Virions and CD4 Lymphocytes in HIV-1 Infection. *Nature* 373, 123–126.

Holmes, R. K., Koning, F. A., Bishop, K. N., and Malim, M. H. (2007). APOBEC3F can inhibit the accumulation of HIV-1 reverse transcription products in the absence of hypermutation: Comparisons with APOBEC3G. *Journal of Biological Chemistry* 282, 2587–2595. doi: 10.1074/jbc.M607298200.

Hrecka, K., Hao, C., Gierszewska, M., Swanson, S. K., Kesik-Brodacka, M., Srivastava, S., et al. (2011). Vpx relieves inhibition of HIV-1 infection of macrophages mediated by the SAMHD1 protein. *Nature* 474, 658–661. doi: 10.1038/nature10195.

Hu, W. S., and Temin, H. M. (1990). Genetic consequences of packaging two RNA genomes in one retroviral particle: Pseudodiploidy and high rate of genetic recombination. *Proc Natl Acad Sci U S A* 87, 1556–1560. doi: 10.1073/pnas.87.4.1556.

Iijima, S., Nitahara-Kasahara, Y., Kimata, K., Zhong Zhuang, W., Kamata, M., Isogai, M., et al. (2004). Nuclear localization of Vpr is crucial for the efficient replication of HIV-1 in primary CD4 + T cells. *Virology* 327, 249–261. doi: 10.1016/j.virol.2004.06.024.

Ishikawa, K., Janssens, W., Banor, J. S., Shinno, T., Piedade, J., Sata, T., et al. (2001). Genetic analysis of HIV type 2 from Ghana and Guinea-Bissau, West Africa. *AIDS Research and Human Retroviruses* 17, 1661–1663. doi: 10.1089/088922201753342077.

Iwabu, Y., Fujita, H., Kinomoto, M., Kaneko, K., Ishizaka, Y., Tanaka, Y., et al. (2009). HIV-1 accessory protein Vpu internalizes cell-surface BST-2/tetherin through transmembrane interactions leading to lysosomes. *Journal of Biological Chemistry* 284, 35060–35072. doi: 10.1074/jbc.M109.058305.

Izumoto, Y., Kuroda, T., Harada, H., Kishimoto, T., and Nakamura, H. (1997). Hepatoma-derived growth factor belongs to a gene family in mice showing significant homology in the amino terminus. *Biochemical and Biophysical Research Communications* 238, 26–32. doi: 10.1006/bbrc.1997.7233.

Jacks, T., Powert, M. D., Masiarzt, F. R., Luciwt, P. A., Barrt, P. J., and Varmus, H. E. (1988). Characterization of ribosomal frameshifting in HIV-1 gag-pol expression. *Nature* 331, 280–283.

Jacobo-Molina, A., Ding, J., Nanni, R. G., Clark, A. D., Lu, X., Tantillo, C., et al. (1993). Crystal structure of human immunodeficiency virus type 1 reverse transcriptase complexed with double-stranded DNA at 3.0 Å resolution shows bent DNA. *Proc Natl Acad Sci U S A* 90, 6320–6324. doi: 10.1073/pnas.90.13.6320.

Jaeger, J., Restle, T., and Steitz, T. A. (1998). The structure of HIV-1 reverse transcriptase complexed with an RNA pseudoknot inhibitor. *EMBO Journal* 17, 4535–4542. doi: 10.1093/emboj/17.15.4535.

Jayappa, K. D., Ao, Z., Yang, M., Wang, J., and Yao, X. (2011). Identification of critical motifs within HIV-1 integrase required for importin α3 interaction and viral cDNA nuclear import. *Journal of Molecular Biology* 410, 847–862. doi: 10.1016/j.jmb.2011.04.011.

Jenkins, T. M., Engelman, A., Ghirlando, R., and Craigie, R. (1996). A soluble active mutant of HIV-1 integrase: Involvement of both the core and caeboxyl-terminal domains in multimerization. *Journal of Biological Chemistry* 271, 7712–7718. doi: 10.1074/jbc.271.13.7712.

Jia, B., Serra-Moreno, R., Neidermyer, W., Rahmberg, A., Mackey, J., Fofana, I. ben, et al. (2009). Species-specific activity of SIV Nef and HIV-1 Vpu in overcoming restriction by tetherin/BST2. *PLoS Pathogens* 5. doi: 10.1371/journal.ppat.1000429.

Jochelson, K., Mothibeli, M., and Leger, J.-P. (1991). Human Immunodeficiency Virus and Migrant Labor in South Africa. *International Journal of Health Services* 21, 157–173. doi: 10.2190/11UE-L88J-46HN-HR0K.

Johnson, B. C., Métifiot, M., Ferris, A., Pommier, Y., and Hughes, S. H. (2013). A Homology Model of HIV-1 Integrase and Analysis of Mutations Designed to Test the Model. *Journal of Molecular Biology* 425, 2133–2146. doi: 10.1016/j.jmb.2013.03.027.

Jordan, A., Defechereux, P., and Verdin, E. (2001). The site of HIV-1 integration in the human genome determines basal transcriptional activity and response to Tat transactivation. *EMBO Journal* 20, 1726–1738. doi: 10.1093/emboj/20.7.1726.

Jowett, J. B. M., Hockley, D. J., Nermut, M. v., and Jones, I. M. (1993). Distinct signals in human immunodeficiency virus type 1 Pr55 necessary for RNA binding and particle formation (Journal of general virology (1992), 73 (3079-3086)). *Journal of General Virology* 74, 943.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589. doi: 10.1038/s41586-021-03819-2.

Jurado, K. A., Wang, H., Slaughter, A., Feng, L., Kessl, J. J., Koh, Y., et al. (2013). Allosteric integrase inhibitor potency is determined through the inhibition of HIV-1 particle maturation. *Proc Natl Acad Sci U S A* 110, 8690–8695. doi: 10.1073/pnas.1300703110.

Kalish, M. L., Robbins, K. E., Pieniazek, D., Schaefer, A., Nzilambi, N., Quinn, T. C., et al. (2004). Recombinant viruses and early global HIV-1 epidemic. *Emerging Infectious Diseases* 10, 1227–1234. doi: 10.3201/eid1007.030904.

Kalland, K. H., Szilvay, A. M., Brokstad, K. A., Saetrevik, W., and Haukenes, G. (1994). The human immunodeficiency virus type 1 Rev protein shuttles between the cytoplasm and nuclear compartments. *Molecular and Cellular Biology* 14, 7436–7444. doi: 10.1128/mcb.14.11.7436-7444.1994.

Kalpana, G. v., Marmon, S., Wang, W., Crabtree, G. R., and Goff, S. P. (1994). Binding and stimulation of HIV-1 integrase by a human homolog of yeast transcription factor SNF5. *Science (1979)* 266, 2002–2006. doi: 10.1126/science.7801128.

Kanja, M., Cappy, P., Levy, N., Oladosu, O., Schmidt, S., Rossolillo, P., et al. (2020). NKNK: a New Essential Motif in the C-Terminal Domain of HIV-1 Group M Integrases. *Journal of Virology* 94, 1–23. doi: 10.1128/JVI.01035-20.

Keele, B. F., Giorgi, E. E., Salazar-Gonzalez, J. F., Decker, J. M., Pham, K. T., Salazar, M. G., et al. (2008). Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proceedings of the National Academy of Sciences* 105, 7552–7557. doi: 10.1073/pnas.0802203105.

Keele, B. F., Jones, J. H., Terio, K. A., Estes, J. D., Rudicell, R. S., Wilson, M. L., et al. (2009). Increased mortality and AIDS-like immunopathology in wild chimpanzees infected with SIVcpz. *Nature* 460, 515–519. doi: 10.1038/nature08200.

Keele, B. F., van Heuverswyn, F., Li, Y., Bailes, E., Takehisa, J., Santiago, M. L., et al. (2006). Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. *Science (1979)* 313, 523–526. doi: 10.1126/science.1126531.

Kent, W. J. (2002). BLAT —The BLAST -Like Alignment Tool. *Genome Research* 12, 656–664. doi: 10.1101/gr.229202.

Kessl, J. J., Jena, N., Koh, Y., Taskent-Sezgin, H., Slaughter, A., Feng, L., et al. (2012). Multimode, cooperative mechanism of action of allosteric HIV-1 integrase inhibitors. *Journal of Biological Chemistry* 287, 16801–16811. doi: 10.1074/jbc.M112.354373.

Kessl, J. J., Kutluay, S. B., Townsend, D., Rebensburg, S., Slaughter, A., Larue, R. C., et al. (2016). HIV-1 Integrase Binds the Viral RNA Genome and Is Essential during Virion Morphogenesis. *Cell* 166, 1257-1268.e12. doi: 10.1016/j.cell.2016.07.044.

Kestier, H. W., Ringler, D. J., Mori, K., Panicali, D. L., Sehgal, P. K., Daniel, M. D., et al. (1991). Importance of the nef gene for maintenance of high virus loads and for development of AIDS. *Cell* 65, 651–662. doi: 10.1016/0092-8674(91)90097-I.

Kielian, M. (2006). Class II virus membrane fusion proteins. *Virology* 344, 38–47. doi: 10.1016/j.virol.2005.09.036.

Kirchhoff, F. (2010). Immune Evasion and Counteraction of Restriction Factors by HIV-1 and Other Primate Lentiviruses. *Cell Host & Microbe* 8, 55–67. doi: 10.1016/j.chom.2010.06.004.

Kirchhoff, F., Schindler, M., Specht, A., Arhel, N., and Münch, J. (2008). Role of Nef in primate lentiviral immunopathogenesis. *Cellular and Molecular Life Sciences* 65, 2621–2636. doi: 10.1007/s00018-008-8094-2.

Klatzmann, D., Champagne, E., Chamaret, S., Gruest, J., Guetard, D., Hercend, T., et al. (1984). T-lymphocyte T4 molecule behaves as the receptor for human retrovirus LAV. *Nature* 312, 767–768. doi: 10.1038/312767a0.

Kleiman, L. (2002). tRNALys3: The primer tRNA for reverse transcription in HIV-1. *IUBMB Life* 53, 107–114. doi: 10.1080/15216540211469.

Kluge, S. F., Mack, K., Iyer, S. S., Pujol, F. M., Heigele, A., Learn, G. H., et al. (2014). Nef Proteins of Epidemic HIV-1 Group O Strains Antagonize Human Tetherin. *Cell Host & Microbe* 16, 639–650. doi: 10.1016/j.chom.2014.10.002.

Knyazhanskaya, E., Anisenko, A., Shadrina, O., Kalinina, A., Zatsepin, T., Zalevsky, A., et al. (2019). NHEJ pathway is involved in post-integrational DNA repair due to Ku70 binding to HIV-1 integrase. *Retrovirology* 16, 1–17. doi: 10.1186/s12977-019-0492-z.

Koh, Y., Wu, X., Ferris, A. L., Matreyek, K. A., Smith, S. J., Lee, K., et al. (2013). Differential Effects of Human Immunodeficiency Virus Type 1 Capsid and Cellular Factors Nucleoporin 153 and LEDGF/p75 on the Efficiency and Specificity of Viral DNA Integration. *Journal of Virology* 87, 648–658. doi: 10.1128/jvi.01148-12.

Kohlstaedt, L. A., Wang, J., Friedman, J. M., Rice, P. A., and Steitz, T. A. (1992). "Crystal Structure at 3.5 Å Resolution of HIV-1 Reverse Transcriptase Complexed with an Inhibitor," in *Structural Insights into Gene Expression and Protein Synthesis*, 254–261. doi: 10.1142/9789811215865_0026.

Korber, B., Gaschen, B., Yusim, K., Thakallapally, R., Kesmir, C., and Detours, V. (2001). Evolutionary and immunological implications of contemporary HIV-1 variation. *British Medical Bulletin* 58, 19–42. doi: 10.1093/bmb/58.1.19.

Korber, B., Muldoon, M., Theiler, J., Gao, F., Gupta, R., Lapedes, A., et al. (2000). Timing the Ancestor of the HIV-1 Pandemic Strains. *Science (1979)* 288, 1789–1796. doi: 10.1126/science.288.5472.1789.

Kulkosky, J., Jones, K. S., Katz, R. A., Mack, J. P., and Skalka, A. M. (1992). Residues critical for retroviral integrative recombination in a region that is highly conserved among retroviral/retrotransposon integrases and bacterial insertion sequence transposases. *Molecular and Cellular Biology* 12, 2331–2338. doi: 10.1128/mcb.12.5.2331-2338.1992.

Kuritzkes, D. R. (2009). HIV-1 entry inhibitors: An overview. *Current Opinion in HIV and AIDS* 4, 82–87. doi: 10.1097/COH.0b013e328322402e.

Kuzembayeva, M., Dilley, K., Sardo, L., and Hu, W.-S. (2014). Life of psi: How full-length HIV-1 RNAs become packaged genomes in the viral particles. *Virology* 454–455, 362–370. doi: 10.1016/j.virol.2014.01.019.

Kvaratskhelia, M., Sharma, A., Larue, R. C., Serrao, E., and Engelman, A. (2014). Molecular mechanisms of retroviral integration site selection. *Nucleic Acids Research* 42, 10209–10225. doi: 10.1093/nar/gku769.

la Porte, A., Cano, J., Wu, X., Mitra, D., and Kalpana, G. v. (2016). An Essential Role of INI1/hSNF5 Chromatin Remodeling Protein in HIV-1 Posttranscriptional Events and Gag/Gag-Pol Stability. *Journal of Virology* 90, 9889–9904. doi: 10.1128/jvi.00323-16.

Lackner, A. A., Lederman, M. M., and Rodriguez, B. (2012). HIV pathogenesis: The host. *Cold Spring Harbor Perspectives in Medicine* 2, 1–23. doi: 10.1101/cshperspect.a007005.

Laguette, N., Sobhian, B., Casartelli, N., Ringeard, M., Chable-Bessia, C., Ségéral, E., et al. (2011). SAMHD1 is the dendritic- and myeloid-cell-specific HIV-1 restriction factor counteracted by Vpx. *Nature* 474, 654–657. doi: 10.1038/nature10117.

Lahouassa, H., Daddacha, W., Hofmann, H., Ayinde, D., Logue, E. C., Dragin, L., et al. (2012). SAMHD1 restricts the replication of human immunodeficiency virus type 1 by depleting the intracellular pool of deoxynucleoside triphosphates. *Nature Immunology* 13, 223–228. doi: 10.1038/ni.2236.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* 10. doi: 10.1186/gb-2009-10-3-r25.

Lanman, J., Lam, T. K. T., Barnes, S., Sakalian, M., Emmett, M. R., Marshall, A. G., et al. (2003). Identification of novel interactions in HIV-1 capsid protein assembly by high-resolution mass spectrometry. *Journal of Molecular Biology* 325, 759–772. doi: 10.1016/S0022-2836(02)01245-7.

Lapaillerie, D., Lelandais, B., Mauro, E., Lagadec, F., Tumiotto, C., Miskey, C., et al. (2021). Modulation of the intrinsic chromatin binding property of HIV-1 integrase by LEDGF/p75. *Nucleic Acids Research* 49, 11241–11256. doi: 10.1093/nar/gkab886.

le Rouzic, E., and Benichou, S. (2005). The Vpr protein from HIV-1: Distinct roles along the viral life cycle. *Retrovirology* 2, 1–14. doi: 10.1186/1742-4690-2-11.

le Rouzic, E., Bonnard, D., Chasset, S., Bruneau, J. M., Chevreuil, F., le Strat, F., et al. (2013). Dual inhibition of HIV-1 replication by integrase-LEDGF allosteric inhibitors is predominant at the post-integration stage. *Retrovirology* 10. doi: 10.1186/1742-4690-10-144.

le Tortorec, A., and Neil, S. J. D. (2009). Antagonism to and Intracellular Sequestration of Human Tetherin by the Human Immunodeficiency Virus Type 2 Envelope Glycoprotein. *Journal of Virology* 83, 11966–11978. doi: 10.1128/jvi.01515-09.

Leavitt, A. D., Robles, G., Alesandro, N., and Varmus, H. E. (1996). Human immunodeficiency virus type 1 integrase mutants retain in vitro integrase activity yet fail to integrate viral DNA efficiently during infection. *Journal of Virology* 70, 721–728. doi: 10.1128/jvi.70.2.721-728.1996.

Lee, K., Mulky, A., Yuen, W., Martin, T. D., Meyerson, N. R., Choi, L., et al. (2012). HIV-1 Capsid-Targeting Domain of Cleavage and Polyadenylation Specificity Factor 6. *Journal of Virology* 86, 3851–3860. doi: 10.1128/jvi.06607-11.

Lee, M. S., and Craigie, R. (1998). A previously unidentified host protein protects retroviral DNA from autointegration. *Proc Natl Acad Sci U S A* 95, 1528–1533. doi: 10.1073/pnas.95.4.1528.

Lee, S. P., Xiao, J., Knutson, J. R., Lewis, M. S., and Han, M. K. (1997). Zn2+ promotes the self-association of human immunodeficiency virus type-1 integrase in vitro. *Biochemistry* 36, 173–180. doi: 10.1021/bi961849o.

Leitner, T., Dazza, M. C., Ekwalanga, M., Apetrei, C., and Saragosti, S. (2007). Sequence diversity among chimpanzee simian immunodeficiency viruses (SIVcpz) suggests that SIVcpzPts was derived from SIVcpzPtt through additional recombination events. *AIDS Research and Human Retroviruses* 23, 1114–1118. doi: 10.1089/aid.2007.0071.

Lemey, P., Pybus, O. G., Rambaut, A., Drummond, A. J., Robertson, D. L., Roques, P., et al. (2004). The molecular population genetics of HIV-1 group O. *Genetics* 167, 1059–1068. doi: 10.1534/genetics.104.026666.

Lemey, P., Rambaut, A., and Pybus, O. (2006). HIV evolutionary dynamics within and among hosts. *AIDS Reviews* 8, 125–40.

Lenassi, M., Cagney, G., Liao, M., Vaupotič, T., Bartholomeeusen, K., Cheng, Y., et al. (2010). HIV Nef is Secreted in Exosomes and Triggers Apoptosis in Bystander CD4+ T Cells. *Traffic* 11, 110–122. doi: 10.1111/j.1600-0854.2009.01006.x.

Leonard, J., Sen, N., Torres, R., Sutani, T., Jarmuz, A., Shirahige, K., et al. (2015). Condensin Relocalization from Centromeres to Chromosome Arms Promotes Top2 Recruitment during Anaphase. *Cell Reports* 13, 2336–2344. doi: 10.1016/j.celrep.2015.11.041.

Leoz, M., Feyertag, F., Kfutwah, A., Mauclère, P., Lachenal, G., Damond, F., et al. (2015). The Two-Phase Emergence of Non Pandemic HIV-1 Group O in Cameroon. *PLoS Pathogens* 11, 1–13. doi: 10.1371/journal.ppat.1005029.

LeRoy, G., Oksuz, O., Descostes, N., Aoi, Y., Ganai, R. A., Kara, H. O., et al. (2019). LEDGF and HDGF2 relieve the nucleosome-induced barrier to transcription in differentiated cells. *Science Advances* 5, 1–13. doi: 10.1126/sciadv.aay3068.

Lesbats, P., Botbol, Y., Chevereau, G., Vaillant, C., Calmels, C., Arneodo, A., et al. (2011). Functional coupling between HIV-1 integrase and the SWI/SNF chromatin remodeling complex for efficient in vitro integration into stable nucleosomes. *PLoS Pathogens* 7. doi: 10.1371/journal.ppat.1001280.

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

Lewinski, M. K., Yamashita, M., Emerman, M., Ciuffi, A., Marshall, H., Crawford, G., et al. (2006). Retroviral DNA integration: Viral and cellular determinants of target-site selection. *PLoS Pathogens* 2, 0611–0622. doi: 10.1371/journal.ppat.0020060.

Li, G., Piampongsant, S., Faria, R. N., Voet, A., Pineda-Peña, A. C., Khouri, R., et al. (2015). An integrated map of HIV genome-wide variation from a population perspective. *Retrovirology* 12. doi: 10.1186/s12977-015-0148-6.

Li, L., Olvera, J. M., Yoder, K. E., Mitchell, R. S., Butler, S. L., Lieber, M., et al. (2001). Role of the non-homologous DNA end joining pathway in the early steps of retroviral infection. *The EMBO Journal* 20, 3272–3281. doi: 10.1093/emboj/20.12.3272.

Li, M., Chen, X., Wang, H., Jurado, K. A., Engelman, A. N., and Craigie, R. (2020a). A Peptide Derived from Lens Epithelium–Derived Growth Factor Stimulates HIV-1 DNA Integration and Facilitates Intasome Structural Studies. *Journal of Molecular Biology* 432, 2055–2066. doi: 10.1016/j.jmb.2020.01.040.

Li, W., Singh, P. K., Sowd, G. A., Bedwell, G. J., Jang, S., Achuthan, V., et al. (2020b). CPSF6-Dependent Targeting of Speckle-Associated Domains Distinguishes Primate from Nonprimate Lentiviral Integration. *mBio* 11, 1–20. doi: 10.1128/mBio.02254-20.

Li, Y., Ndjango, J.-B., Learn, G. H., Ramirez, M. A., Keele, B. F., Bibollet-Ruche, F., et al. (2012). Eastern Chimpanzees, but Not Bonobos, Represent a Simian Immunodeficiency Virus Reservoir. *Journal of Virology* 86, 10776–10791. doi: 10.1128/jvi.01498-12.

Lihua J Zhu, Claude Gazin, Nathan D Lawson, Hervé Pagès, Simon M Lin, David S Lapointe, et al. (2010). ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics* 11, 237–247. Available at: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-237.

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 and Microbe* 11, 194–204. doi: 10.1016/j.chom.2012.01.004.

Lim, S. Y., Rogers, T., Chan, T., Whitney, J. B., Kim, J., Sodroski, J., et al. (2010). TRIM5α modulates immunodeficiency virus control in rhesus monkeys. *PLoS Pathogens* 6, 1–11. doi: 10.1371/journal.ppat.1000738.

Lin, C.-W., and Engelman, A. (2003). The Barrier-to-Autointegration Factor Is a Component of Functional Human Immunodeficiency Virus Type 1 Preintegration Complexes. *Journal of Virology* 77, 5030–5036. doi: 10.1128/jvi.77.8.5030-5036.2003.

Llano, M., Delgado, S., Vanegas, M., and Poeschla, E. M. (2004). Lens epithelium-derived growth factor/p75 prevents proteasomal degradation of HIV-1 integrase. *Journal of Biological Chemistry* 279, 55570–55577. doi: 10.1074/jbc.M408508200.

Llano, M., Saenz, D. T., Meehan, A., Wongthida, P., Peretz, M., Walker, W. H., et al. (2006). An essential role for LEDGF/p75 in HIV integration. *Science (1979)* 314, 461–464. doi: 10.1126/science.1132319.

Lodi, P. J., Ernst, J. A., Kuszewski, J., Hickman, A. B., Engelman, A., Craigie, R., et al. (1995). Solution Structure of the DNA Binding Domain of HIV-1 Integrase. *Biochemistry* 34, 9826–9833. doi: 10.1021/bi00031a002.

Louis, J. M., Hashed, N. T., Parris, K. D., Kimmel, A. R., and Jerina, D. M. (1994). Kinetics and mechanism of autoprocessing of human immunodeficiency virus type 1 protease from an analog of the Gag-Pol polyprotein. *Proc Natl Acad Sci U S A* 91, 7970–7974. doi: 10.1073/pnas.91.17.7970.

Loussert-Ajaka, I., Chaix, M. L., Korber, B., Letourneur, F., Gomas, E., Allen, E., et al. (1995). Variability of human immunodeficiency virus type 1 group O strains isolated from Cameroonian patients living in France. *Journal of Virology* 69, 5640–5649. doi: 10.1128/jvi.69.9.5640-5649.1995.

Lu, R., Limón, A., Ghory, H. Z., and Engelman, A. (2005). Genetic Analyses of DNA-Binding Mutants in the Catalytic Core Domain of Human Immunodeficiency Virus Type 1 Integrase. *Journal of Virology* 79, 2493–2505. doi: 10.1128/jvi.79.4.2493-2505.2005.

Lusic, M., and Siliciano, R. F. (2017). Nuclear landscape of HIV-1 infection and integration. *Nature Reviews Microbiology* 15, 69–82. doi: 10.1038/nrmicro.2016.162.

Lutzke, R. A. P., and Plasterk, R. H. A. (1998). Structure-Based Mutational Analysis of the C-Terminal DNA-Binding Domain of Human Immunodeficiency Virus Type 1 Integrase: Critical Residues for Protein Oligomerization and DNA Binding. *Journal of Virology* 72, 4841–4848. doi: 10.1128/jvi.72.6.4841-4848.1998.

Machida, S., Depierre, D., Chen, H. C., Thenin-Houssier, S., Petitjean, G., Doyen, C. M., et al. (2020). Exploring histone loading on HIV DNA reveals a dynamic nucleosome positioning between unintegrated and integrated viral genome. *Proc Natl Acad Sci U S A* 117, 6822–6830. doi: 10.1073/pnas.1913754117.

Madison, M. K., Lawson, D. Q., Elliott, J., Ozantürk, A. N., Koneru, P. C., Townsend, D., et al. (2017). Allosteric HIV-1 Integrase Inhibitors Lead to Premature Degradation of the Viral RNA Genome and Integrase in Target Cells. *Journal of Virology* 91, 1–22. doi: 10.1128/JVI.00821-17.

Maertens, G. N., Engelman, A. N., and Cherepanov, P. (2022). Structure and function of retroviral integrase. *Nature Reviews Microbiology* 20, 20–34. doi: 10.1038/s41579-021-00586-9.

Maertens, G. N., Hare, S., and Cherepanov, P. (2010). The mechanism of retroviral integration from X-ray structures of its key intermediates. *Nature* 468, 326–329. doi: 10.1038/nature09517.

Maillot, B., Lévy, N., Eiler, S., Crucifix, C., Granger, F., Richert, L., et al. (2013). Structural and Functional Role of INI1 and LEDGF in the HIV-1 Preintegration Complex. *PLoS ONE* 8. doi: 10.1371/journal.pone.0060734.

Maldarelli, F., Wu, X., Su, L., Simonetti, F. R., Shao, W., Hill, S., et al. (2014). Specific HIV integration sites are linked to clonal expansion and persistence of infected cells. *Science (1979)* 345, 179–183. doi: 10.1126/science.1254194.

Malim, M. H. (2009). APOBEC proteins and intrinsic resistance to HIV-1 infection. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364, 675–687. doi: 10.1098/rstb.2008.0185.

Malim, M. H., and Emerman, M. (2008). HIV-1 Accessory Proteins-Ensuring Viral Survival in a Hostile Environment. *Cell Host and Microbe* 3, 388–398. doi: 10.1016/j.chom.2008.04.008.

Mangeat, B., Gers-Huber, G., Lehmann, M., Zufferey, M., Luban, J., and Piguet, V. (2009). HIV-1 Vpu neutralizes the antiviral factor tetherin/BST-2 by binding it and directing its beta-TrCP2-dependent degradation. *PLoS Pathogens* 5. doi: 10.1371/journal.ppat.1000574.

Mangeat, B., Turelli, P., Caron, G., Friedli, M., Perrin, L., and Trono, D. (2003). Broad antiretroviral defence by human APOBEC3G through lethal editing of nascent reverse transcripts. *Nature* 424, 99–103. doi: 10.1038/nature01709.

Mansky, L. M. (2002). HIV mutagenesis and the evolution of antiretroviral drug resistance. *Drug Resistance Updates* 5, 219–223. doi: 10.1016/S1368-7646(02)00118-8.

Mariani, R., Chen, D., Schröfelbauer, B., Navarro, F., König, R., Bollman, B., et al. (2003). Species-specific exclusion of APOBEC3G from HIV-1 virions by Vif. *Cell* 114, 21–31. doi: 10.1016/S0092-8674(03)00515-4.

Marini, B., Kertesz-Farkas, A., Ali, H., Lucic, B., Lisek, K., Manganaro, L., et al. (2015). Nuclear architecture dictates HIV-1 integration site selection. *Nature* 521, 227–231. doi: 10.1038/nature14226.

Marshall, H. M., Ronen, K., Berry, C., Llano, M., Sutherland, H., Saenz, D., et al. (2007). Role of PSIP 1/LEDGF/p75 in lentiviral infectivity and integration targeting. *PLoS ONE* 2. doi: 10.1371/journal.pone.0001340.

Martin, G., Gruber, A. R., Keller, W., and Zavolan, M. (2012). Genome-wide Analysis of Pre-mRNA 3' End Processing Reveals a Decisive Role of Human Cleavage Factor I in the Regulation of 3' UTR Length. *Cell Reports* 1, 753–763. doi: 10.1016/j.celrep.2012.05.003.

Martin, J. A., and Wang, Z. (2011). Next-generation transcriptome assembly. *Nature Reviews Genetics* 12, 671–682. doi: 10.1038/nrg3068.

Martínez, J. P., Bocharov, G., Ignatovich, A., Reiter, J., Dittmar, M. T., Wain-Hobson, S., et al. (2011). Fitness ranking of individual mutants drives patterns of epistatic interactions in HIV-1. *PLoS ONE* 6. doi: 10.1371/journal.pone.0018375.

Martin-Serrano, J., Zang, T., and Bieniasz, P. D. (2001). HIV-1 and Ebola virus encode small peptide motifs that recruit Tsg101 to sites of particle assembly to facilitate egress. *Nature Medicine* 7, 1313–1319. doi: 10.1038/nm1201-1313.

Masuda, T., Planelles, V., Krogstad, P., and Chen, I. S. (1995). Genetic analysis of human immunodeficiency virus type 1 integrase and the U3 att site: unusual phenotype of mutants in the zinc finger-like domain. *Journal of Virology* 69, 6687–6696. doi: 10.1128/jvi.69.11.6687-6696.1995.

Mathew, S., Nguyen, M., Wu, X., Pal, A., Shah, V. B., Prasad, V. R., et al. (2013). INI1/hSNF5-interaction defective HIV-1 IN mutants exhibit impaired particle morphology, reverse transcription and integration in vivo. *Retrovirology* 10, 1–20. doi: 10.1186/1742-4690-10-66.

Matysiak, J., Lesbats, P., Mauro, E., Lapaillerie, D., Dupuy, J. W., Lopez, A. P., et al. (2017). Modulation of chromatin structure by the FACT histone chaperone complex regulates HIV-1 integration. *Retrovirology* 14, 1–20. doi: 10.1186/s12977-017-0363-4.

Mauro, E., Lesbats, P., Lapaillerie, D., Chaignepain, S., Maillot, B., Oladosu, O., et al. (2019). Human H4 tail stimulates HIV-1 integration through binding to the carboxy-terminal domain of integrase. *Nucleic Acids Research* 47, 3607–3618. doi: 10.1093/nar/gkz091.

McDaniel, Y. Z., Wang, D., Love, R. P., Adolph, M. B., Mohammadzadeh, N., Chelico, L., et al. (2020). Deamination hotspots among APOBEC3 family members are defined by both target site sequence context and ssDNA secondary structure. *Nucleic Acids Research* 48, 1353–1371. doi: 10.1093/NAR/GKZ1164.

McKee, C. J., Kessl, J. J., Shkriabai, N., Dar, M. J., Engelman, A., and Kvaratskhelia, M. (2008). Dynamic modulation of HIV-1 integrase structure and function by cellular lens epithelium-derived growth factor (LEDGF) protein. *Journal of Biological Chemistry* 283, 31802–31812. doi: 10.1074/jbc.M805843200.

McNatt, M. W., Zang, T., Hatziioannou, T., Bartlett, M., Fofana, I. ben, Johnson, W. E., et al. (2009). Species-specific activity of HIV-1 Vpu and positive selection of tetherin transmembrane domain variants. *PLoS Pathogens* 5, 9–12. doi: 10.1371/journal.ppat.1000300.

Meissner, M. E., Talledge, N., and Mansky, L. M. (2022). Molecular Biology and Diversification of Human Retroviruses. *Frontiers in Virology* 2, 1–18. doi: 10.3389/fviro.2022.872599.

Meixenberger, K., Yousef, K. P., Smith, M. R., Somogyi, S., Fiedler, S., Bartmeyer, B., et al. (2017). Molecular evolution of HIV-1 integrase during the 20 years prior to the first approval of integrase inhibitors. *Virology Journal* 14, 223. doi: 10.1186/s12985-017-0887-1.

Meltzer, B., Dabbagh, D., Guo, J., Kashanchi, F., Tyagi, M., and Wu, Y. (2018). Tat controls transcriptional persistence of unintegrated HIV genome in primary human macrophages. *Virology* 518, 241–252. doi: 10.1016/j.virol.2018.03.006.

Menéndez-Arias, L. (2008). Mechanisms of resistance to nucleoside analogue inhibitors of HIV-1 reverse transcriptase. *Virus Research* 134, 124–146. doi: 10.1016/j.virusres.2007.12.015.

Meng, X., Zhao, G., Yufenyuy, E., Ke, D., Ning, J., DeLucia, M., et al. (2012). Protease Cleavage Leads to Formation of Mature Trimer Interface in HIV-1 Capsid. *PLoS Pathogens* 8. doi: 10.1371/journal.ppat.1002886.

Meyer, B. E., and Malim, M. H. (1994). The HIV-1 Rev trans-activator shuttles between the nucleus and the cytoplasm. *Genes and Development* 8, 1538–1547. doi: 10.1101/gad.8.13.1538.

Miller, M. D., Farnet, C. M., and Bushman, F. D. (1997). Human immunodeficiency virus type 1 preintegration complexes: studies of organization and composition. *J Virol* 71, 5382–5390. doi: 10.1128/jvi.71.7.5382-5390.1997.

Miller, M. D., Wang, B., and Bushman, F. D. (1995). Human immunodeficiency virus type 1 preintegration complexes containing discontinuous plus strands are competent to integrate in vitro. *Journal of Virology* 69, 3938–3944. doi: 10.1128/jvi.69.6.3938-3944.1995.

Miyagi, E., Opi, S., Takeuchi, H., Khan, M., Goila-Gaur, R., Kao, S., et al. (2007). Enzymatically Active APOBEC3G Is Required for Efficient Inhibition of Human Immunodeficiency Virus Type 1. *Journal of Virology* 81, 13346–13353. doi: 10.1128/jvi.01361-07.

Mohammed, K. D., Topper, M. B., and Muesing, M. A. (2011). Sequential Deletion of the Integrase (Gag-Pol) Carboxyl Terminus Reveals Distinct Phenotypic Classes of Defective HIV-1. *Journal of Virology* 85, 4654–4666. doi: 10.1128/jvi.02374-10.

Mourez, T., Simon, F., and Plantiera, J. C. (2013). Non-M variants of human immunodeficiency virus type. *Clinical Microbiology Reviews* 26, 448–461. doi: 10.1128/CMR.00012-13.

Mousnier, A., Kubat, N., Massias-Simon, A., Ségéral, E., Rain, J. C., Benarous, R., et al. (2007). Von Hippel-Lindau binding protein 1-mediated degradation of integrase affects HIV-1 gene expression at a postintegration step. *Proc Natl Acad Sci U S A* 104, 13615–13620. doi: 10.1073/pnas.0705162104.

Mulanga-Kabeya, C., Nzilambi, N., Edidi, B., Minlangu, M., Tshimpaka, T., Kambembo, L., et al. (1998). Evidence of stable HIV seroprevalences in selected populations in the Democratic Republic of the Congo. *AIDS* 12, 905–910. doi: 10.1097/00002030-199808000-00013.

Mulder, L. C. F., and Muesing, M. A. (2000). Degradation of HIV-1 integrase by the N-end rule pathway. *Journal of Biological Chemistry* 275, 29749–29753. doi: 10.1074/jbc.M004670200.

Munir, S., Thierry, S., Subra, F., Deprez, E., and Delelis, O. (2013). Quantitative analysis of the time-course of viral DNA forms during the HIV-1 life cycle. *Retrovirology* 10, 1–18. doi: 10.1186/1742-4690-10-87.

Nagata, S., Imai, J., Makino, G., Tomita, M., and Kanai, A. (2017). Evolutionary Analysis of HIV-1 Pol Proteins Reveals Representative Residues for Viral Subtype Differentiation. *Frontiers in Microbiology* 8, 1–10. doi: 10.3389/fmicb.2017.02151.

Naldini, L., Blömer, U., Gallay, P., Ory, D., Mulligan, R., Gage, F. H., et al. (1996). In vivo gene delivery and stable transduction of nondividing cells by a lentiviral vector. *Science (1979)* 272, 263–267. doi: 10.1126/science.272.5259.263.

Neel, C. C., Etienne, L., Li, Y., Takehisa, J., Rudicell, R. S., Ndong Bass, I., et al. (2010). Molecular Epidemiology of Simian Immunodeficiency Virus Infection in Wild-Living Gorillas. *Journal of Virology* 84, 1464–1476. doi: 10.1128/JVI.02129-09.

Negroni, M., and Buc, H. (2000). Copy-choice recombination by reverse transcriptases: Reshuffling of genetic markers mediated by RNA chaperones. *Proc Natl Acad Sci U S A* 97, 6385–6390. doi: 10.1073/pnas.120520497.

Neil, S. J. D., Sandrin, V., Sundquist, W. I., and Bieniasz, P. D. (2007). An Interferon-α-Induced Tethering Mechanism Inhibits HIV-1 and Ebola Virus Particle Release but Is Counteracted by the HIV-1 Vpu Protein. *Cell Host and Microbe* 2, 193–203. doi: 10.1016/j.chom.2007.08.001.

Neil, S. J. D., Zang, T., and Bieniasz, P. D. (2008). Tetherin inhibits retrovirus release and is antagonized by HIV-1 Vpu. *Nature* 451, 425–430. doi: 10.1038/nature06553.

Niama, F. R., Toure-Kane, C., Vidal, N., Obengui, P., Bikandou, B., Ndoundou Nkodia, M. Y., et al. (2006). HIV-1 subtypes and recombinants in the Republic of Congo. *Infection, Genetics and Evolution* 6, 337–343. doi: 10.1016/j.meegid.2005.12.001.

Nishihara, T. (1995). Feeding ecology of western lowland gorillas in the Nouabalé-Ndoki National Park, Congo. *Primates* 36, 151–168. doi: 10.1007/BF02381342.

Nishizawa, Y., Usukura, J., Singh, D. P., Chylack, L. T., and Shinohara, T. (2001). Spatial and temporal dynamics of two alternatively spliced regulatory factors, lens epithelium-derived growth factor (LEDGF/p75) and p52, in the nucleus. *Cell and Tissue Research* 305, 107–114. doi: 10.1007/s004410100398.

Nzilambi, N., de Cock, K. M., Forthal, D. N., Francis, H., Ryder, R. W., Malebe, I., et al. (1988). The Prevalence of Infection with Human Immunodeficiency Virus over a 10-Year Period in Rural Zaire. *New England Journal of Medicine* 318, 276–279. doi: 10.1056/NEJM198802043180503.

Okada, A., and Iwatani, Y. (2016). APOBEC3G-mediated G-to-A hypermutation of the HIV-1 genome: The missing link in antiviral molecular mechanisms. *Frontiers in Microbiology* 7, 1–8. doi: 10.3389/fmicb.2016.02027.

Pace, M. J., Graf, E. H., Agosto, L. M., Mexas, A. M., Male, F., Brady, T., et al. (2012). Directly infected resting CD4+T cells can produce HIV Gag without spreading infection in a model of HIV latency. *PLoS Pathogens* 8, 15. doi: 10.1371/journal.ppat.1002818.

Pandrea, I., Apetrei, C., Robertson, D. L., Onanga, R., Gao, F., Makuwa, M., et al. (2002). Analysis of partial pol and env sequences indicates a high prevalence of HIV type 1 recombinant strains circulating in Gabon. *AIDS Research and Human Retroviruses* 18, 1103–1116. doi: 10.1089/088922202320567842.

Passos, D. O., Li, M., Craigie, R., and Lyumkis, D. (2021). *Retroviral integrase: Structure, mechanism, and inhibition*. 1st ed. Elsevier Inc. doi: 10.1016/bs.enz.2021.06.007.

Passos, D. O., Li, M., Jóźwik, I. K., Zhao, X. Z., Santos-Martins, D., Yang, R., et al. (2020). Structural basis for strand-transfer inhibitor binding to HIV intasomes. *Science (1979)* 367, 810–814. doi: 10.1126/science.aay8015.

Passos, D. O., Li, M., Yang, R., Rebensburg, S. v., Ghirlando, R., Jeon, Y., et al. (2017). Cryo-EM structures and atomic model of the HIV-1 strand transfer complex intasome. *Science (1979)* 355, 89–92. doi: 10.1126/science.aah5163.

Paxton, W., Connor, R. I., and Landau, N. R. (1993). Incorporation of Vpr into human immunodeficiency virus type 1 virions: requirement for the p6 region of gag and mutational analysis. *Journal of Virology* 67, 7229–7237. doi: 10.1128/jvi.67.12.7229-7237.1993.

Peeters, M., Gueye, A., Mboup, S., Bibollet-Ruche, F., Ezaka, E., Mulanga, C., et al. (1997). Geographical distribution of HIV-1 group O viruses in Africa. *AIDS* 11, 493–498.

Peeters, M., Liegeois, F., Torimiro, N., Bourgeois, A., Mpoudi, E., Vergne, L., et al. (1999). Characterization of a Highly Replicative Intergroup M/O Human Immunodeficiency Virus Type 1 Recombinant Isolated from a Cameroonian Patient. *Journal of Virology* 73, 7368–7375. doi: 10.1128/jvi.73.9.7368-7375.1999.

Pettit, S. C., Lindquist, J. N., Kaplan, A. H., and Swanstrom, R. (2005). Processing sites in the human immunodeficiency virus type 1 (HIV-1) Gag-Pro-Pol precursor are cleaved by the viral protease at different rates. *Retrovirology* 2, 12–16. doi: 10.1186/1742-4690-2-66.

Pettit, S. C., Moody, M. D., Wehbie, R. S., Kaplan, A. H., Nantermet, P. v, Klein, C. A., et al. (1994). The p2 domain of human immunodeficiency virus type 1 Gag regulates sequential proteolytic processing and is required to produce fully infectious virions. *Journal of Virology* 68, 8017–8027. doi: 10.1128/jvi.68.12.8017-8027.1994.

Pieniazek, D., Ellenberger, D., Janini, L. M., Ramos, A. C., Nkengasong, J., Sassan-Morokro, M., et al. (1999). Predominance of human immunodeficiency virus type 2 subtype B in Abidjan, Ivory Coast. *AIDS Research and Human Retroviruses* 15, 603–608. doi: 10.1089/088922299311132.

Planelles, V., and Benichou, S. (2009). Vpr and its interactions with cellular proteins. *Current Topics in Microbiology and Immunology* 339, 177–200. doi: 10.1007/978-3-642-02175-6_9.

Plantier, J. C., Leoz, M., Dickerson, J. E., de Oliveira, F., Cordonnier, F., Lemée, V., et al. (2009). A new human immunodeficiency virus derived from gorillas. *Nature Medicine* 15, 871–872. doi: 10.1038/nm.2016.

Popov, S., Rexach, M., Ratner, L., Blobel, G., and Bukrinsky, M. (1998). Viral protein R regulates docking of the HIV-1 preintegration complex to the nuclear pore complex. *Journal of Biological Chemistry* 273, 13347–13352. doi: 10.1074/jbc.273.21.13347.

Popper, S. J., Sarr, A. D., Guèye-Ndiaye, A., Mboup, S., Essex, M. E., and Kanki, P. J. (2000). Low Plasma Human Immunodeficiency Virus Type 2 Viral Load Is Independent of Proviral Load: Low Virus Production In Vivo. *Journal of Virology* 74, 1554–1557. doi: 10.1128/jvi.74.3.1554-1557.2000.

Pornillos, O., and Ganser-Pornillos, B. K. (2019). Maturation of retroviruses. *Current Opinion in Virology* 36, 47–55. doi: 10.1016/j.coviro.2019.05.004.

Pradeepa, M. M., Sutherland, H. G., Ule, J., Grimes, G. R., and Bickmore, W. A. (2012). Psip1/Ledgf p52 binds methylated histone H3K36 and splicing factors and contributes to the regulation of alternative splicing. *PLoS Genetics* 8. doi: 10.1371/journal.pgen.1002717.

Preston, B. D., Poiesz, B. J., and Loeb, L. A. (1988). Fidelity of HIV-1 reverse transcriptase. *Science (1979)* 242, 1168–1171. doi: 10.1126/science.2460924.

Price, A. J., Fletcher, A. J., Schaller, T., Elliott, T., Lee, K. E., KewalRamani, V. N., et al. (2012). CPSF6 Defines a Conserved Capsid Interface that Modulates HIV-1 Replication. *PLoS Pathogens* 8. doi: 10.1371/journal.ppat.1002896.

Price, A. J., Jacques, D. A., McEwan, W. A., Fletcher, A. J., Essig, S., Chin, J. W., et al. (2014). Host Cofactors and Pharmacologic Ligands Share an Essential Interface in HIV-1 Capsid That Is Lost upon Disassembly. *PLoS Pathogens* 10. doi: 10.1371/journal.ppat.1004459.

Puras Lutzke, R. A., Vink, C., and Plasterk, R. H. A. (1994). Characterization of the minimal DNA-binding domain of the HIV integrase protein. *Nucleic Acids Research* 22, 4125–4131. doi: 10.1093/nar/22.20.4125.

Qin, S., and Min, J. (2014). Structure and function of the nucleosome-binding PWWP domain. *Trends in Biochemical Sciences* 39, 536–547. doi: 10.1016/j.tibs.2014.09.001.

Qu, K., Ke, Z., Zila, V., Anders-Össwein, M., Glass, B., Mücksch, F., et al. (2021). Maturation of the matrix and viral membrane of HIV-1. *Science (1979)* 373, 700–704. doi: 10.1126/science.abe6821.

Quashie, P. K., Mesplède, T., Han, Y.-S., Oliveira, M., Singhroy, D. N., Fujiwara, T., et al. (2012). Characterization of the R263K Mutation in HIV-1 Integrase That Confers Low-Level Resistance to the Second-Generation Integrase Strand Transfer Inhibitor Dolutegravir. *Journal of Virology* 86, 2696–2705. doi: 10.1128/jvi.06591-11.

Quillent, C., Borman, A. M., Paulous, S., Dauguet, C., and Clavel, F. (1996). Extensive regions of pol are required for efficient human immunodeficiency virus polyprotein processing and particle maturation. *Virology* 219, 29–36. doi: 10.1006/viro.1996.0219.

Quinn, T. C. (1994). Population migration and the spread of types 1 and 2 human immunodeficiency viruses. *Proceedings of the National Academy of Sciences* 91, 2407–2414. doi: 10.1073/pnas.91.7.2407.

Quiñones-Mateu, M. E., Albright, J. L., Mas, A., Soriano, V., and Arts, E. J. (1998). Analysis of pol Gene Heterogeneity, Viral Quasispecies, and Drug Resistance in Individuals Infected with Group O Strains of Human Immunodeficiency Virus Type 1. *Journal of Virology* 72, 9002–9015. doi: 10.1128/jvi.72.11.9002-9015.1998.

Raghavendra, N. K., Shkriabai, N., Graham, R. L. J., Hess, S., Kvaratskhelia, M., and Wu, L. (2010). Identification of host proteins associated with HIV-1 preintegration complexes isolated from infected CD4+cells. *Retrovirology* 7. doi: 10.1186/1742-4690-7-66.

Rahman, S., Lu, R., Vandegraaff, N., Cherepanov, P., and Engelman, A. (2007). Structure-based mutagenesis of the integrase-LEDGF/p75 interface uncouples a strict correlation between in vitro protein binding and HIV-1 fitness. *Virology* 357, 79–90. doi: 10.1016/j.virol.2006.08.011.

Raja, A., Venturi, M., Kwong, P., and Sodroski, J. (2003). CD4 Binding Site Antibodies Inhibit Human Immunodeficiency Virus gp120 Envelope Glycoprotein Interaction with CCR5. *Journal of Virology* 77, 713–718. doi: 10.1128/jvi.77.1.713-718.2003.

Rambaut, A., Robertson, D. L., Pybus, O. G., Peeters, M., and Holmes, E. C. (2001). Phylogeny and the origin of HIV-1. *Nature* 410, 1047–1048. doi: 10.1038/35074179.

Ramírez, F., Ryan, D. P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A. S., et al. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Research* 44, W160–W165. doi: 10.1093/NAR/GKW257.

Rasheedi, S., Shun, M. C., Serrao, E., Sowd, G. A., Qian, J., Hao, C., et al. (2016). The Cleavage and polyadenylation specificity factor 6 (CPSF6) subunit of the capsid-recruited pre-messenger RNA cleavage factor I (CFIm) complex mediates HIV-1 integration into genes. *Journal of Biological Chemistry* 291, 11809–11819. doi: 10.1074/jbc.M116.721647.

Rayfield, M. A., Sullivan, P., Bandea, C. I., Britvan, L., Otten, R. A., Pau, C. P., et al. (1996). HIV-1 group O virus identified for the first time in the United States. *Emerg Infect Dis* 2, 209–212. doi: 10.3201/eid0203.960307.

Resnick, R., Omer, C. A., and Faras, A. J. (1984). Involvement of retrovirus reverse transcriptase-associated RNase H in the initiation of strong-stop (+) DNA synthesis and the generation of the long terminal repeat. *Journal of Virology* 51, 813–821. doi: 10.1128/jvi.51.3.813-821.1984.

Rhim, H., Park, J., and Morrow, C. D. (1991). Deletions in the tRNA(Lys) primer-binding site of human immunodeficiency virus type 1 identify essential regions for reverse transcription. *Journal of Virology* 65, 4555–4564. doi: 10.1128/jvi.65.9.4555-4564.1991.

Rice, A. P. (2017). The HIV-1 Tat Protein: Mechanism of Action and Target for HIV-1 Cure Strategies. *Current Pharmaceutical Design* 23, 4098–4102. doi: 10.2174/1381612823666170704130635.

Richetta, C., Thierry, S., Thierry, E., Lesbats, P., Lapaillerie, D., Munir, S., et al. (2019). Two-long terminal repeat (LTR) DNA circles are a substrate for HIV-1 integrase. *Journal of Biological Chemistry* 294, 8286–8295. doi: 10.1074/jbc.RA118.006755.

Roberts, J. D., Bebenek, K., and Kunkel, T. A. (1988). The accuracy of reverse transcriptase from HIV-1. *Science (1979)* 242, 1171–1173. doi: 10.1126/science.2460925.

Robertson, D. L., Anderson, J. P., Bradac, J. A., Carr, J. K., Foley, B., Funkhouser, R. K., et al. (2000). HIV-1 Nomenclature Proposal. *Science (1979)* 288, 55–55. doi: 10.1126/science.288.5463.55d.

Rocchi, C., Gouet, P., Parissi, V., and Fiorini, F. (2022). The C-Terminal Domain of HIV-1 Integrase: A Swiss Army Knife for the Virus? *Viruses* 14, 1397. doi: 10.3390/v14071397.

Rong, L., Zhang, J., Lu, J., Pan, Q., Lorgeoux, R.-P., Aloysius, C., et al. (2009). The Transmembrane Domain of BST-2 Determines Its Sensitivity to Down-Modulation by Human Immunodeficiency Virus Type 1 Vpu. *Journal of Virology* 83, 7536–7546. doi: 10.1128/jvi.00620-09.

Roques, P., Robertson, D. L., Souquière, S., Apetrei, C., Nerrienet, E., Barré-Sinoussi, F., et al. (2004). Phylogenetic characteristics of three new HIV-1 N strains and implications for the origin of group N. *AIDS* 18, 1371–1381. doi: 10.1097/01.aids.0000125990.86904.28.

Roques, P., Robertson, D. L., Souquière, S., Damond, F., Ayouba, A., Farfara, I., et al. (2002). Phylogenetic analysis of 49 newly derived HIV-1 group O strains: High viral diversity but no group M-like subtype structure. *Virology* 302, 259–273. doi: 10.1006/viro.2002.1430.

Rose, K. M., Hirsch, V. M., and Bouamr, F. (2020). Budding of a retrovirus: Some assemblies required. *Viruses* 12. doi: 10.3390/v12101188.

Roth, M. J., Schwartzberg, P. L., and Goff, S. P. (1989). Structure of the termini of DNA intermediates in the integration of retroviral DNA: Dependence on IN function and terminal DNA sequence. *Cell* 58, 47–54. doi: 10.1016/0092-8674(89)90401-7.

Roth, S. L., Malani, N., and Bushman, F. D. (2011). Gammaretroviral Integration into Nucleosomal Target DNA In Vivo . *Journal of Virology* 85, 7393–7401. doi: 10.1128/jvi.00635-11.

Rousseau, C. M., Learn, G. H., Bhattacharya, T., Nickle, D. C., Heckerman, D., Chetty, S., et al. (2007). Extensive Intrasubtype Recombination in South African Human Immunodeficiency Virus Type 1 Subtype C Infections. *Journal of Virology* 81, 4492–4500. doi: 10.1128/jvi.02050-06.

Rowland-Jones, S. L., and Whittle, H. C. (2007). Out of Africa: What can we learn from HIV-2 about protective immunity to HIV-1? *Nature Immunology* 8, 329–331. doi: 10.1038/ni0407-329.

Rudicell, R. S., Jones, J. H., Wroblewski, E. E., Learn, G. H., Li, Y., Robertson, J. D., et al. (2010). Impact of simian immunodeficiency virus infection on chimpanzee population dynamics. *PLoS Pathogens* 6. doi: 10.1371/journal.ppat.1001116.

Rüegsegger, U., Beyer, K., and Keller, W. (1996). Purification and characterization of human cleavage factor Im involved in the 3′ end processing of messenger RNA precursors. *Journal of Biological Chemistry* 271, 6107–6113. doi: 10.1074/jbc.271.11.6107.

Rüegsegger, U., Blank, D., and Keller, W. (1998). Human pre-mRNA cleavage factor Im Is related to spliceosomal SR proteins and can be reconstituted in vitro from recombinant subunits. *Molecular Cell* 1, 243–253. doi: 10.1016/S1097-2765(00)80025-8.

Ruvolo, M. (1997). Molecular phylogeny of the hominoids; inferences from multiple independent DNA sequence data sets. *Molecular Biology and Evolution* 14, 248–265.

Saag, M. S., Gandhi, R. T., Hoy, J. F., Landovitz, R. J., Thompson, M. A., Sax, P. E., et al. (2020). Antiretroviral Drugs for Treatment and Prevention of HIV Infection in Adults. *JAMA* 324, 1651. doi: 10.1001/jama.2020.17025.

Samoshkin, A., Arnaoutov, A., Jansen, L. E. T., Ouspenski, I., Dye, L., Karpova, T., et al. (2009). Human condensin function is essential for centromeric chromatin assembly and proper sister kinetochore orientation. *PLoS ONE* 4. doi: 10.1371/journal.pone.0006831.

Santiago, M. L., Rodenburg, C. M., Kamenya, S., Bibollet-Ruche, F., Gao, F., Bailes, E., et al. (2002). SIVcpz in Wild Chimpanzees. *Science (1979)* 295, 465–465. doi: 10.1126/science.295.5554.465.

Sato, K., Misawa, N., Takeuchi, J. S., Kobayashi, T., Izumi, T., Aso, H., et al. (2018). Experimental Adaptive Evolution of Simian Immunodeficiency Virus SIVcpz to Pandemic Human Immunodeficiency Virus Type 1 by Using a Humanized Mouse Model. *J Virol* 92, 1–21. doi: 10.1128/jvi.01905-17.

Sauter, D., Schindler, M., Specht, A., Landford, W. N., Münch, J., Kim, K.-A., et al. (2009). Tetherin-Driven Adaptation of Vpu and Nef Function and the Evolution of Pandemic and Nonpandemic HIV-1 Strains. *Cell Host & Microbe* 6, 409–421. doi: 10.1016/j.chom.2009.10.004.

Sauter, D., Specht, A., and Kirchhoff, F. (2010). Tetherin: Holding on and letting go. *Cell* 141, 392–398. doi: 10.1016/j.cell.2010.04.022.

Sawyer, S. L., Emerman, M., and Malik, H. S. (2004). Ancient adaptive evolution of the primate antiviral DNA-editing enzyme APOBEC3G. *PLoS Biology* 2. doi: 10.1371/journal.pbio.0020275.

Sawyer, S. L., Wu, L. I., Emerman, M., and Malik, H. S. (2005). Positive selection of primate TRIM5α identifies a critical species-specific retroviral restriction domain. *Proc Natl Acad Sci U S A* 102, 2832–2837. doi: 10.1073/pnas.0409853102.

Schacker, T., Collier, A. C., Hughes, J., Shea, T., and Corey, L. (1996). Annals of Internal Medicine: Clinical and Epidemiologic Features of Primary HIV Infection. *Annals of Internal Medicine* 125, 257–264. doi: 10.7326/0003-4819-125-4-199608150-00001.

Schaller, T., Ocwieja, K. E., Rasaiyaah, J., Price, A. J., Brady, T. L., Roth, S. L., et al. (2011). HIV-1 capsid-cyclophilin interactions determine nuclear import pathway, integration targeting and replication efficiency. *PLoS Pathogens* 7. doi: 10.1371/journal.ppat.1002439.

Schindler, M., Münch, J., Kutsch, O., Li, H., Santiago, M. L., Bibollet-Ruche, F., et al. (2006). Nef-Mediated Suppression of T Cell Activation Was Lost in a Lentiviral Lineage that Gave Rise to HIV-1. *Cell* 125, 1055–1067. doi: 10.1016/j.cell.2006.04.033.

Schmidt, J. M., de Manuel, M., Marques-Bonet, T., Castellano, S., and Andrés, A. M. (2019). The impact of genetic adaptation on chimpanzee subspecies differentiation. doi: 10.1371/journal.pgen.1008485.

Schmökel, J., Sauter, D., Schindler, M., Leendertz, F. H., Bailes, E., Dazza, M.-C., et al. (2011). The Presence of a vpu Gene and the Lack of Nef-Mediated Downmodulation of T Cell Receptor-CD3 Are Not Always Linked in Primate Lentiviruses. *Journal of Virology* 85, 742–752. doi: 10.1128/JVI.02087-10.

Schneider, T. D., and Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research* 18, 6097–6100.

Schrijvers, R., de Rijck, J., Demeulemeester, J., Adachi, N., Vets, S., Ronen, K., et al. (2012). LEDGF/p75-independent HIV-1 replication demonstrates a role for HRP-2 and remains sensitive to inhibition by LEDGINs. *PLoS Pathog* 8. doi: 10.1371/journal.ppat.1002558.

Selyutina, A., Persaud, M., Lee, K., KewalRamani, V., and Diaz-Griffero, F. (2020). Nuclear Import of the HIV-1 Core Precedes Reverse Transcription and Uncoating. *Cell Reports* 32, 108201. doi: 10.1016/j.celrep.2020.108201.

Serrao, E., Thys, W., Demeulemeester, J., Al-Mawsawi, L. Q., Christ, F., Debyser, Z., et al. (2012). A Symmetric Region of the HIV-1 Integrase Dimerization Interface Is Essential for Viral Replication. *PLoS ONE* 7, 1–11. doi: 10.1371/journal.pone.0045177.

Shan, L., Deng, K., Gao, H., Xing, S., Capoferri, A. A., Durand, C. M., et al. (2017). Transcriptional Reprogramming during Effector-to-Memory Transition Renders CD4+ T Cells Permissive for Latent HIV-1 Infection. *Immunity* 47, 766-775.e3. doi: 10.1016/j.immuni.2017.09.014.

Sharma, A., Slaughter, A., Jena, N., Feng, L., Kessl, J. J., Fadel, H. J., et al. (2014). A New Class of Multimerization Selective Inhibitors of HIV-1 Integrase. *PLoS Pathogens* 10. doi: 10.1371/journal.ppat.1004171.

Sharp, P. M., and Hahn, B. H. (2011). Origins of HIV and the AIDS pandemic. *Cold Spring Harbor Perspectives in Medicine* 1. doi: 10.1101/cshperspect.a006841.

Shun, M. C., Raghavendra, N. K., Vandegraaff, N., Daigle, J. E., Hughes, S., Kellam, P., et al. (2007). LEDGF/p75 functions downstream from preintegration complex formation to effect gene-specific HIV-1 integration. *Genes and Development* 21, 1767–1778. doi: 10.1101/gad.1565107.

Silvers, R. M., Smith, J. A., Schowalter, M., Litwin, S., Liang, Z., Geary, K., et al. (2010). Modification of integration site preferences of an HIV-1-based vector by expression of a novel synthetic protein. *Human Gene Therapy* 21, 337–349. doi: 10.1089/hum.2009.134.

Simon, F., Mauclère, P., Roques, P., Loussert-Ajaka, I., Müller-Trutwin, M. C., Saragosti, S., et al. (1998). Identification of a new human immunodeficiency virus type 1 distinct from group M and group O. *Nature Medicine* 4, 1032–1037. doi: 10.1038/2017.

Simon-Loriere, E., and Holmes, E. C. (2011). Why do RNA viruses recombine? *Nature Reviews Microbiology* 9, 617–626. doi: 10.1038/nrmicro2614.

Singh, P. K., Bedwell, G. J., and Engelman, A. N. (2022). Spatial and Genomic Correlates of HIV-1 Integration Site Targeting. *Cells* 11, 655. doi: 10.3390/cells11040655.

Singh, P. K., Plumb, M. R., Ferris, A. L., Iben, J. R., Wu, X., Fadel, H. J., et al. (2015). LEDGF/p75 interacts with mRNA splicing factors and targets HIV-1 integration to highly spliced genes. *Genes and Development* 29, 2287–2297. doi: 10.1101/gad.267609.115.

Skoko, D., Li, M., Huang, Y., Mizuuchi, M., Cai, M., Bradley, C. M., et al. (2009). Barrier-to-autointegration factor (BAF) condenses DNA by looping. *Proc Natl Acad Sci U S A* 106, 16610–16615. doi: 10.1073/pnas.0909077106.

Sloan, R. D., Donahue, D. A., Kuhl, B. D., Bar-Magen, T., and Wainberg, M. A. (2010). Expression of Nef from unintegrated HIV-1 DNA downregulates cell surface CXCR4 and CCR5 on T-lymphocytes. *Retrovirology* 7, 1–10. doi: 10.1186/1742-4690-7-44.

Sloan, R. D., and Wainberg, M. A. (2011). The role of unintegrated DNA in HIV infection. *Retrovirology* 8, 1–15. doi: 10.1186/1742-4690-8-52.

Smyth, R. P., Davenport, M. P., and Mak, J. (2012). The origin of genetic diversity in HIV-1. *Virus Research* 169, 415–429. doi: 10.1016/j.virusres.2012.06.015.

Soper, A., Kimura, I., Nagaoka, S., Konno, Y., Yamamoto, K., Koyanagi, Y., et al. (2018). Type I interferon responses by HIV-1 infection: Association with disease progression and control. *Frontiers in Immunology* 8, 1–11. doi: 10.3389/fimmu.2017.01823.

Sorin, M., Cano, J., Das, S., Mathew, S., Wu, X., Davies, K. P., et al. (2009). Recruitment of a SAP18-HDAC1 complex into HIV-1 virions and its requirement for viral replication. *PLoS Pathogens* 5. doi: 10.1371/journal.ppat.1000463.

Sorin, M., Yung, E., Wu, X., and Kalpana, G. v. (2006). HIV-1 replication in cell lines harboring INI1/hSNF5 mutations. *Retrovirology* 3, 1–15. doi: 10.1186/1742-4690-3-56.

Sowd, G. A., Serrao, E., Wang, H., Wang, W., Fadel, H. J., Poeschla, E. M., et al. (2016). A critical role for alternative polyadenylation factor CPSF6 in targeting HIV-1 integration to transcriptionally active chromatin. *Proc Natl Acad Sci U S A* 113, E1054–E1063. doi: 10.1073/pnas.1524213113.

Spearman, P., Wang, J. J., vander Heyden, N., and Ratner, L. (1994). Identification of human immunodeficiency virus type 1 Gag protein domains essential to membrane binding and particle assembly. *Journal of Virology* 68, 3232–3242. doi: 10.1128/jvi.68.5.3232-3242.1994.

Spector, D. L., and Lamond, A. I. (2011). Nuclear speckles. *Cold Spring Harbor Perspectives in Biology* 3, 1–12. doi: 10.1101/cshperspect.a000646.

Stanford, C. B., and Nkurunungi, J. B. (2003). Behavioral ecology of sympatric chimpanzees and gorillas in Bwindi Impenetrable National Park, Uganda: Diet. *International Journal of Primatology* 24, 901–918. doi: 10.1023/A.1024689008159.

Stopak, K., de Noronha, C., Yonemoto, W., and Greene, W. C. (2003). HIV-1 Vif blocks the antiviral activity of APOBEC3G by impairing both its translation and intracellular stability. *Molecular Cell* 12, 591–601. doi: 10.1016/S1097-2765(03)00353-8.

Storz, J. F. (2016). Causes of molecular convergence and parallelism in protein evolution. *Nature Reviews Genetics* 17, 239–250. doi: 10.1038/nrg.2016.11.

Strack, B., Calistri, A., Craig, S., Popova, E., and Göttlinger, H. G. (2003). AIP1/ALIX is a binding partner for HIV-1 p6 and EIAV p9 functioning in virus budding. *Cell* 114, 689–699. doi: 10.1016/S0092-8674(03)00653-6.

Sundquist, W. I., Krausslich, H.-G., Kra, H., and Krausslich, H.-G. (2012). HIV-1 Assembly, Budding, and Maturation. *Cold Spring Harbor Perspectives in Medicine* 2, a006924–a006924. doi: 10.1101/cshperspect.a006924.

Suo, Z., and Johnson, K. A. (1997). Effect of RNA secondary structure on the kinetics of DNA synthesis catalyzed by HIV-1 reverse transcriptase. *Biochemistry* 36, 12459–12467. doi: 10.1021/bi971217h.

Switzer, W. M., Parekh, B., Shanmugam, V., Bhullar, V., Phillips, S., Ely, J. J., et al. (2005). The epidemiology of simian immunodeficiency virus infection in a large number of wild- and captive-born chimpanzees: Evidence for a recent introduction following chimpanzee divergence. *AIDS Research and Human Retroviruses* 21, 335–342. doi: 10.1089/aid.2005.21.335.

Takehisa, J., Kraus, M. H., Ayouba, A., Bailes, E., van Heuverswyn, F., Decker, J. M., et al. (2009). Origin and Biology of Simian Immunodeficiency Virus in Wild-Living Western Gorillas. *Journal of Virology* 83, 1635–1648. doi: 10.1128/jvi.02311-08.

Taube, R., Fujinaga, K., Wimmer, J., Barboric, M., and Peterlin, B. M. (1999). Tat transactivation: A model for the regulation of eukaryotic transcriptional elongation. *Virology* 264, 245–253. doi: 10.1006/viro.1999.9944.

Tebit, D. M., Lobritz, M., Lalonde, M., Immonen, T., Singh, K., Sarafianos, S., et al. (2010). Divergent Evolution in Reverse Transcriptase (RT) of HIV-1 Group O and M Lineages: Impact on Structure, Fitness, and Sensitivity to Nonnucleoside RT Inhibitors. *Journal of Virology* 84, 9817–9830. doi: 10.1128/JVI.00991-10.

Tekeste, S. S., Wilkinson, T. A., Weiner, E. M., Xu, X., Miller, J. T., le Grice, S. F. J., et al. (2015). Interaction between Reverse Transcriptase and Integrase Is Required for Reverse Transcription during HIV-1 Replication. *Journal of Virology* 89, 12058–12069. doi: 10.1128/jvi.01471-15.

Thierry, S., Munir, S., Thierry, E., Subra, F., Leh, H., Zamborlini, A., et al. (2015). Integrase inhibitor reversal dynamics indicate unintegrated HIV-1 dna initiate de novo integration. *Retrovirology* 12, 1–12. doi: 10.1186/s12977-015-0153-9.

Thoulouze, M. I., Sol-Foulon, N., Blanchet, F., Dautry-Varsat, A., Schwartz, O., and Alcover, A. (2006). Human Immunodeficiency Virus Type-1 Infection Impairs the Formation of the Immunological Synapse. *Immunity* 24, 547–561. doi: 10.1016/j.immuni.2006.02.016.

Tisne, C. (2005). Structural Bases of the Annealing of Primer Lys tRNA to the HIV-1 Viral RNA. *Current HIV Research* 3, 147–156. doi: 10.2174/1570162053506919.

Toccafondi, E., Lener, D., and Negroni, M. (2021). HIV-1 Capsid Core: A Bullet to the Heart of the Target Cell. *Frontiers in Microbiology* 12, 1–17. doi: 10.3389/fmicb.2021.652486.

Tsiang, M., Jones, G. S., Hung, M., Samuel, D., Novikov, N., Mukund, S., et al. (2011). Dithiothreitol causes HIV-1 integrase dimer dissociation while agents interacting with the integrase dimer interface promote dimer formation. *Biochemistry* 50, 1567–1581. doi: 10.1021/bi101504w.

Tsiang, M., Jones, G. S., Niedziela-Majka, A., Kan, E., Lansdon, E. B., Huang, W., et al. (2012). New class of HIV-1 integrase (IN) inhibitors with a dual mode of action. *Journal of Biological Chemistry* 287, 21189–21203. doi: 10.1074/jbc.M112.347534.

Turlure, F., Maertens, G., Rahman, S., Cherepanov, P., and Engelman, A. (2006). A tripartite DNA-binding element, comprised of the nuclear localization signal and two AT-hook motifs, mediates the association of LEDGF/p75 with chromatin in vivo. *Nucleic Acids Research* 34, 1653–1665. doi: 10.1093/nar/gkl052.

Vallari, A., Holzmayer, V., Harris, B., Yamaguchi, J., Ngansop, C., Makamche, F., et al. (2011). Confirmation of Putative HIV-1 Group P in Cameroon. *Journal of Virology* 85, 1403–1407. doi: 10.1128/jvi.02005-10.

van Damme, N., Goff, D., Katsura, C., Jorgenson, R. L., Mitchell, R., Johnson, M. C., et al. (2008). The Interferon-Induced Protein BST-2 Restricts HIV-1 Release and Is Downregulated from the Cell Surface by the Viral Vpu Protein. *Cell Host and Microbe* 3, 245–252. doi: 10.1016/j.chom.2008.03.001.

van Damme, N., and Guatelli, J. (2008). HIV-1 Vpu inhibits accumulation of the envelope glycoprotein within clathrin-coated, Gag-containing endosomes. *Cellular Microbiology* 10, 1040–1057. doi: 10.1111/j.1462-5822.2007.01101.x.

van de Perre, P., Lepage, P., Kestelyn, P., Hekker, A. C., Rouvroy, D., Bogaerts, J., et al. (1984). Acquired Immunodeficiency Syndrome in Rwanda. *The Lancet* 324, 62–65. doi: 10.1016/S0140-6736(84)90240-X.

van der Loeff, M. F. S., Awasana, A. A., Sarge-Njie, R., van der Sande, M., Jaye, A., Sabally, S., et al. (2006). Sixteen years of HIV surveillance in a West African research clinic reveals divergent epidemic trends of HIV-1 and HIV-2. *International Journal of Epidemiology* 35, 1322–1328. doi: 10.1093/ije/dyl037.

van der Loeff, M. F. S., Larke, N., Kaye, S., Berry, N., Ariyoshi, K., Alabi, A., et al. (2010). Undetectable plasma viral load predicts normal survival in HIV-2-infected people in a West African village. *Retrovirology* 7, 2–11. doi: 10.1186/1742-4690-7-46.

van Gent, D. C., Groeneger, A. A. M. O., and Plasterk, R. H. A. (1992). Mutational analysis of the integrase protein of human immunodeficiency virus type 2. *Proc Natl Acad Sci U S A* 89, 9598–9602. doi: 10.1073/pnas.89.20.9598.

van Gent, D. C., Vink, C., Groeneger, A. A., and Plasterk, R. H. (1993). Complementation between HIV integrase proteins mutated in different domains. *The EMBO Journal* 12, 3261–3267. doi: 10.1002/j.1460-2075.1993.tb05995.x.

van Maele, B., Busschots, K., Vandekerckhove, L., Christ, F., and Debyser, Z. (2006). Cellular co-factors of HIV-1 integration. *Trends in Biochemical Sciences* 31, 98–105. doi: 10.1016/j.tibs.2005.12.002.

van Nuland, R., van Schaik, F. M. A., Simonis, M., van Heesch, S., Cuppen, E., Boelens, R., et al. (2013). Nucleosomal DNA binding drives the recognition of H3K36-methylated nucleosomes by the PSIP1-PWWP domain. *Epigenetics and Chromatin* 6, 1–12. doi: 10.1186/1756-8935-6-12.

van Wamel, J. L. B., and Berkhout, B. (1998). The first strand transfer during HIV-1 reverse transcription can occur either intramolecularly or intermolecularly. *Virology* 244, 245–251. doi: 10.1006/viro.1998.9096.

Vandekerckhove, L., Christ, F., van Maele, B., de Rijck, J., Gijsbers, R., van den Haute, C., et al. (2006). Transient and Stable Knockdown of the Integrase Cofactor LEDGF/p75 Reveals Its Role in the Replication Cycle of Human Immunodeficiency Virus. *Journal of Virology* 80, 1886–1896. doi: 10.1128/jvi.80.4.1886-1896.2006.

vanden Haesevelde, M. M., Peeters, M., Jannes, G., Jannens, W., van der Groen, G., Sharp, P. M., et al. (1996). Sequence Analysis of a Highly Divergent HIV-1-Related Lentivirus Isolated from a Wild Captured Chimpanzee. *Virology* 221, 346–350. doi: 10.1006/viro.1996.0384.

Vangroenweghe, D. (2001). The earliest cases of human immunodeficiency virus type 1 group M in Congo-Kinshasa, Rwanda and Burundi and the origin of acquired immune deficiency syndrome. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 356, 923–925. doi: 10.1098/rstb.2001.0876.

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., et al. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research* 50, D439–D444. doi: 10.1093/nar/gkab1061.

Veazey, R. S., DeMaria, M., Chalifoux, L. v., Shvetz, D. E., Pauley, D. R., Knight, H. L., et al. (1998). Gastrointestinal Tract as a Major Site of CD4 + T Cell Depletion and Viral Replication in SIV Infection. *Science (1979)* 280, 427–431. doi: 10.1126/science.280.5362.427.

Vidal, N., Peeters, M., Mulanga-Kabeya, C., Nzilambi, N., Robertson, D., Ilunga, W., et al. (2000). Unprecedented Degree of Human Immunodeficiency Virus Type 1 (HIV-1) Group M Genetic Diversity in the Democratic Republic of Congo Suggests that the HIV-1 Pandemic Originated in Central Africa. *Journal of Virology* 74, 10498–10507. doi: 10.1128/jvi.74.22.10498-10507.2000.

Vincent, K. A., York-Higgins, D., Quiroga, M., and Brown, P. O. (1990). Host sequences flanking the HIV provirus. *Nucleic Acids Research* 18, 6045–6047. doi: 10.1093/nar/18.20.6045.

Vink, C., Groeneger, A. A. M. oude, and Plasterk, R. H. A. (1993). Identification of the catalytic and DNA-binding region of the human immunodeficiency virus type I integrase protein. *Nucleic Acids Research* 21, 1419–1425. doi: 10.1093/nar/21.6.1419.

Vink, C., Groenink, M., Elgersma, Y., Fouchier, R. A., Tersmette, M., and Plasterk, R. H. (1990). Analysis of the junctions between human immunodeficiency virus type 1 proviral DNA and human DNA. *Journal of Virology* 64, 5626–5627. doi: 10.1128/jvi.64.11.5626-5627.1990.

Visseaux, B., Damond, F., Matheron, S., Descamps, D., and Charpentier, C. (2016). Hiv-2 molecular epidemiology. *Infection, Genetics and Evolution* 46, 233–240. doi: 10.1016/j.meegid.2016.08.010.

Vodicka, M. A., Koepp, D. M., Silver, P. A., and Emerman, M. (1998). HIV-1 Vpr interacts with the nuclear transport pathway to promote macrophage infection. *Genes and Development* 12, 175–185. doi: 10.1101/gad.12.2.175.

Vranckx, L. S., Demeulemeester, J., Saleh, S., Boll, A., Vansant, G., Schrijvers, R., et al. (2016). LEDGIN-mediated Inhibition of Integrase–LEDGF/p75 Interaction Reduces Reactivation of Residual Latent HIV. *EBioMedicine* 8, 248–264. doi: 10.1016/j.ebiom.2016.04.039.

Wain, L. v., Bailes, E., Bibollet-Ruche, F., Decker, J. M., Keele, B. F., van Heuverswyn, F., et al. (2007). Adaptation of HIV-1 to Its Human Host. *Molecular Biology and Evolution* 24, 1853–1860. doi: 10.1093/molbev/msm110.

Wang, G. P., Ciuffi, A., Leipzig, J., Berry, C. C., and Bushman, F. D. (2007). HIV integration site selection: Analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. *Genome Research* 17, 1186–1194. doi: 10.1101/gr.6286907.

Wang, H., Farnung, L., Dienemann, C., and Cramer, P. (2020). Structure of H3K36-methylated nucleosome–PWWP complex reveals multivalent cross-gyre binding. *Nature Structural & Molecular Biology* 27, 8–13. doi: 10.1038/s41594-019-0345-4.

Wang, H., Jurado, K. A., Wu, X., Shun, M. C., Li, X., Ferris, A. L., et al. (2012). HRP2 determines the efficiency and specificity of HIV-1 integration in LEDGF/p75 knockout cells but does not contribute to the antiviral activity of a potent LEDGF/p75-binding site integrase inhibitor. *Nucleic Acids Research* 40, 11518–11530. doi: 10.1093/nar/gks913.

Wang, J., Blevins, T., Podicheti, R., Haag, J. R., Tan, E. H., Wang, F., et al. (2017). Mutation of Arabidopsis SMC4 identifies condensin as a corepressor of pericentromeric transposons and conditionally expressed genes. *Genes and Development* 31, 1601–1614. doi: 10.1101/gad.301499.117.

Wang, J., Smerdon, S. J., Jäger, J., Kohlstaedt, L. A., Rice, P. A., Friedman, J. M., et al. (1994). Structural basis of asymmetry in the human immunodeficiency virus type 1 reverse transcriptase heterodimer. *Proc Natl Acad Sci U S A* 91, 7242–7246. doi: 10.1073/pnas.91.15.7242.

Weiss, R. A. (2002). HIV Receptors and Cellular Tropism. *IUBMB Life (International Union of Biochemistry and Molecular Biology: Life)* 53, 201–205. doi: 10.1080/15216540212652.

Wertheim, J. O., and Worobey, M. (2009). Dating the age of the SIV lineages that gave rise to HIV-1 and HIV-2. *PLoS Computational Biology* 5. doi: 10.1371/journal.pcbi.1000377.

Wilkinson, T. A., Januszyk, K., Phillips, M. L., Tekeste, S. S., Zhang, M., Miller, J. T., et al. (2009). Identifying and Characterizing a Functional HIV-1 Reverse Transcriptase-binding Site on Integrase. *Journal of Biological Chemistry* 284, 7931–7939. doi: 10.1074/jbc.M806241200.

Willey, R. L., Maldarelli, F., Martin, M. A., and Strebel, K. (1992). Human immunodeficiency virus type 1 Vpu protein regulates the formation of intracellular gp160-CD4 complexes. *Journal of Virology* 66, 226–234. doi: 10.1128/jvi.66.1.226-234.1992.

Winans, S., and Goff, S. P. (2020). Mutations altering acetylated residues in the CTD of HIV-1 integrase cause defects in proviral transcription at early times after integration of viral DNA. *PLOS Pathogens* 16, e1009147. doi: 10.1371/journal.ppat.1009147.

Winans, S., Yu, H. J., de los Santos, K., Wang, G. Z., KewalRamani, V. N., and Goff, S. P. (2022). A point mutation in HIV-1 integrase redirects proviral integration into centromeric repeats. *Nature Communications* 13, 1474. doi: 10.1038/s41467-022-29097-8.

Wiskerchen, M., and Muesing, M. A. (1995). Human immunodeficiency virus type 1 integrase: effects of mutations on viral ability to integrate, direct viral gene expression from unintegrated viral DNA templates, and sustain viral propagation in primary cells. *Journal of Virology* 69, 376–386. doi: 10.1128/jvi.69.1.376-386.1995.

Woo, J., Robertson, D. L., and Lovell, S. C. (2014). Constraints from protein structure and intra-molecular coevolution influence the fitness of HIV-1 recombinants. *Virology* 454–455, 34–39. doi: 10.1016/j.virol.2014.01.029.

Worobey, M., Gemmel, M., Teuwen, D. E., Haselkorn, T., Kunstman, K., Bunce, M., et al. (2008). Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature* 455, 661–664. doi: 10.1038/nature07390.

Wu, W., Blumberg, B. M., Fay, P. J., and Bambara, R. A. (1995). Strand transfer mediated by human immunodeficiency virus reverse transcriptase in vitro is promoted by pausing and results in misincorporation. *Journal of Biological Chemistry* 270, 325–332. doi: 10.1074/jbc.270.1.325.

Wu, X., Liu, H., Xiao, H., Conway, J. A., Hehl, E., Kalpana, G. v., et al. (1999). Human Immunodeficiency Virus Type 1 Integrase Protein Promotes Reverse Transcription through Specific Interactions with the Nucleoprotein Reverse Transcription Complex. *Journal of Virology* 73, 2126–2135. doi: 10.1128/jvi.73.3.2126-2135.1999.

Wu, Y. (2004). HIV-1 gene expression: Lessons from provirus and non-integrated DNA. *Retrovirology* 1, 1–10. doi: 10.1186/1742-4690-1-13.

Yamaguchi, J., McArthur, C., Vallari, A., Sthreshley, L., Cloherty, G. A., Berg, M. G., et al. (2019). Complete genome sequence of CG-0018a-01 establishes HIV-1 subtype L. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 1. doi: 10.1097/qai.0000000000002246.

Yamaguchi, J., Vallari, A. S., Swanson, P., Bodelle, P., Kaptué, L., Ngansop, C., et al. (2002). Evaluation of HIV type 1 group O isolates: Identification of five phylogenetic clusters. *AIDS Research and Human Retroviruses* 18, 269–282. doi: 10.1089/088922202753472847.

Yoder, K. E., Espeseth, A., Wang, X. hong, Fang, Q., Russo, M. T., Lloyd, R. S., et al. (2011). The base excision repair pathway is required for efficient lentivirus integration. *PLoS ONE* 6. doi: 10.1371/journal.pone.0017862.

Yu, G., Wang, L. G., Han, Y., and He, Q. Y. (2012). ClusterProfiler: An R package for comparing biological themes among gene clusters. *OMICS A Journal of Integrative Biology* 16, 284–287. doi: 10.1089/omi.2011.0118.

Yu, X. F., Matsuda, Z., Yu, Q. C., Lee, T. H., and Essex, M. (1995). Role of the C terminus Gag protein in human immunodeficiency virus type 1 virion assembly and maturation. *Journal of General Virology* 76, 3171–3179. doi: 10.1099/0022-1317-76-12-3171.

Yu, X. F., Yu, Q. C., Essex, M., and Lee, T. H. (1991). The vpx gene of simian immunodeficiency virus facilitates efficient viral replication in fresh lymphocytes and macrophage. *Journal of Virology* 65, 5088–5091. doi: 10.1128/jvi.65.9.5088-5091.1991.

Yung, E., Sorin, M., Pal, A., Craig, E., Morozov, A., Delattre, O., et al. (2001). Inhibition of HIV-1 virion production by a transdominant mutant of integrase interactor 1. *Nature Medicine* 7, 920–926. doi: 10.1038/90959.

Zaitseva, L., Cherepanov, P., Leyens, L., Wilson, S. J., Rasaiyaah, J., and Fassati, A. (2009). HIV-1 exploits importin 7 to maximize nuclear import of its DNA genome. *Retrovirology* 6, 1–18. doi: 10.1186/1742-4690-6-11.

Zhang, F., Wilson, S. J., Landford, W. C., Virgen, B., Gregory, D., Johnson, M. C., et al. (2009). Nef Proteins from Simian Immunodeficiency Viruses Are Tetherin Antagonists. *Cell Host & Microbe* 6, 54–67. doi: 10.1016/j.chom.2009.05.008.

Zhang, J., and Crumpacker, C. (2022). HIV UTR, LTR, and Epigenetic Immunity. *Viruses* 14, 1084. doi: 10.3390/v14051084.

Zhang, L., Huang, Y., He, T., Cao, Y., and Ho, D. D. (1996). HIV-1 subtype and second-receptor use. *Nature* 383, 768–768. doi: 10.1038/383768a0.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., et al. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biology* 9. doi: 10.1186/gb-2008-9-9-r137.

Zhao, X. Z., Smith, S. J., Maskell, D. P., Metifiot, M., Pye, V. E., Fesen, K., et al. (2016). HIV-1 Integrase Strand Transfer Inhibitors with Reduced Susceptibility to Drug Resistant Mutant Integrases. *ACS Chemical Biology* 11, 1074–1081. doi: 10.1021/acschembio.5b00948.

Zhao, X. Z., Smith, S. J., Maskell, D. P., Métifiot, M., Pye, V. E., Fesen, K., et al. (2017). Structure-Guided Optimization of HIV Integrase Strand Transfer Inhibitors. *Journal of Medicinal Chemistry* 60, 7315–7332. doi: 10.1021/acs.jmedchem.7b00596.

Zheng, R., Jenkins, T. M., and Craigie, R. (1996). Zinc folds the N-terminal domain of HIV-1 integrase, promotes multimerization, and enhances catalytic activity. *Proceedings of the National Academy of Sciences* 93, 13659–13664. doi: 10.1073/pnas.93.24.13659.

Zhu, P., Liu, J., Bess, J., Chertova, E., Lifson, J. D., Grisé, H., et al. (2006). Distribution and three-dimensional structure of AIDS virus envelope spikes. *Nature* 441, 847–852. doi: 10.1038/nature04817.

Zhu, T., Korber, B. T., Nahmias, A. J., Hooper, E., Sharp, P. M., and Ho, D. D. (1998). An African HIV-1 sequence from 1959 and implications for the of the epidemic. *Nature* 391, 594–597. doi: 10.1038/35400.

Zhu, Y., Pe'ery, T., Peng, J., Ramanathan, Y., Marshall, N., Marshall, T., et al. (1997). Transcription elongation factor P-TEFb is required for HIV-1 Tat transactivation in vitro. *Genes and Development* 11, 2622–2632. doi: 10.1101/gad.11.20.2622.

Zhu, Y., Wang, G. Z., Cingöz, O., and Goff, S. P. (2018). NP220 mediates silencing of unintegrated retroviral DNA. *Nature* 564, 278–282. doi: 10.1038/s41586-018-0750-6.

ANNEXES

ANNEX I: HIV-1 CAPSID CORE: A BULLET TO THE HEART OF THE TARGET CELL

Review Article

Toccafondi E, Lener D and Negroni M

frontiers in Microbiology

REVIEW
published: 01 April 2021
doi: 10.3389/fmicb.2021.652486

HIV-1 Capsid Core: A Bullet to the Heart of the Target Cell

Elenia Toccafondi, Daniela Lener and Matteo Negroni**

CNRS, Architecture et Réactivité de l'ARN, UPR 9002, Université de Strasbourg, Strasbourg, France

The first step of the intracellular phase of retroviral infection is the release of the viral capsid core in the cytoplasm. This structure contains the viral genetic material that will be reverse transcribed and integrated into the genome of infected cells. Up to recent times, the role of the capsid core was considered essentially to protect this genetic material during the earlier phases of this process. However, increasing evidence demonstrates that the permanence inside the cell of the capsid as an intact, or almost intact, structure is longer than thought. This suggests its involvement in more aspects of the infectious cycle than previously foreseen, particularly in the steps of viral genomic material translocation into the nucleus and in the phases preceding integration. During the trip across the infected cell, many host factors are brought to interact with the capsid, some possessing antiviral properties, others, serving as viral cofactors. All these interactions rely on the properties of the unique component of the capsid core, the capsid protein CA. Likely, the drawback of ensuring these multiple functions is the extreme genetic fragility that has been shown to characterize this protein. Here, we recapitulate the busy agenda of an HIV-1 capsid in the infectious process, in particular in the light of the most recent findings.

Keywords: HIV-1, capsid, uncoating, reverse transcription, cellular cofactors, restriction factors, genetic fragility, nuclear transport

OPEN ACCESS

Edited by:
Gkikas Magiorkinis, National and Kapodistrian University of Athens, Greece

Reviewed by:
Tara Patricia Hurst, Birmingham City University, United Kingdom
Hiroaki Takeuchi, Tokyo Medical and Dental University, Japan

****Correspondence:***
Daniela Lener
d.lener@ibmc-cnrs.unistra.fr
Matteo Negroni
m.negroni@ibmc-cnrs.unistra.fr

Specialty section:
This article was submitted to Virology, a section of the journal Frontiers in Microbiology

Received: *12 January 2021*
Accepted: *15 March 2021*
Published: *01 April 2021*

Citation:
Toccafondi E, Lener D and Negroni M (2021) HIV-1 Capsid Core: A Bullet to the Heart of the Target Cell. Front. Microbiol. 12:652486. doi: 10.3389/fmicb.2021.652486

INTRODUCTION

Retroviral infection begins with the fusion of the viral and cell membranes, carried out by the viral envelope proteins (Coffin et al., 1997). This causes the entry in the cytoplasm of the viral capsid core (also simply referred here as the core), a shell constituted by approximately 1,500 copies of the capsid protein CA. The capsid core contains the viral genomic RNA (gRNA) and protects it from cellular sensors of innate immunity and antiviral factors. The infectious cycle requires the reverse transcription of the gRNA to convert it into double-stranded DNA. The capsid core favors this step by providing a confined environment where the concentration of the viral components is high. At the moment of integration, though, the genetic material must have been released from the core, in order to interact with, and integrate into, the chromosomes. When and how the protective shell is dismantled is still not clear. According to the earliest models, disassembling of the core occurred soon after its entry into the cytoplasm (Bukrinsky et al., 1993; Miller et al., 1997; Fassati and Goff, 2001). This view has been challenged recently by an increasing number of observations that support the idea that capsid cores remain intact or almost intact, long after their entry into the cell, and even once in the nucleus (Burdick et al., 2020; Dharan et al., 2020; Selyutina et al., 2020b). This implies that the core constitutes a protective shell all along the trip from entry to

almost the occurrence of integration. This review focuses on these aspects of viral infection: how and where the capsid core is dismantled in the light of the latest observations and which cellular factors, including those that control its stability, it comes across during its longer than expected presence in the newly infected cell.

STRUCTURAL BASES DETERMINING THE STABILITY OF THE CAPSID CORE

The capsid core is generated by the proteolytic processing of the Gag and Gag-Pol precursors that must free the CA protein. In the immature budding particle, these precursors assemble with each other to form the immature Gag lattice, a spherical protein shell located immediately underneath the lipidic envelope of the particle (Briggs et al., 2009). This structure is constituted by a vast majority of Gag precursors that include, from the N to the C terminus, the matrix (MA), the capsid (CA), the spacer peptide 1 (SP1), the nucleocapsid (NC), the spacer peptide 2 (SP2), and peptide 6 (p6) domains (Henderson et al., 1992; **Figure 1A**). Present in the lattice (at a ratio of approximately 1:20 with respect to Gag) are some molecules of Gag-Pol precursors, that contain MA, CA, SP1, and NC fused to the protease (PR), the reverse transcriptase (RT), and the integrase (IN) domains (Jacks et al., 1987; Reil et al., 1993; **Figure 1A**).

The structure of CA has been determined for the free protein, showing an organization in two globular domains (the N-terminal, NTD, and the C-terminal, CTD, domains) connected by a flexible linker (**Figure 1B**). The NTD is composed of seven alpha-helices and a beta-hairpin on the amino-terminal side while the CTD is composed of four alpha-helices (Gamble et al., 1996, 1997; Gitti et al., 1996). This structural arrangement has then been confirmed also for the CA domain in the Gag precursor (Tang et al., 2002; Schur et al., 2016; Wagner et al., 2016b). In the immature Gag lattice, MA points toward the exterior of the viral particle and, proceeding toward the interior, are present the NTD and CTD of CA and the SP1 domain, respectively (**Figure 1C**). Each of these domains multimerizes forming hexamers (Wright et al., 2007; Briggs et al., 2009; Schur et al., 2015, 2016). The interaction among CTDs of CA, stabilized by the six-helix bundles formed by SP1, is responsible for the formation of the immature Gag lattice, while the NTD of CA is not strictly required for assembly and it rather has the role of spacing the hexamers within the Gag lattice (Accola et al., 2000; Wright et al., 2007; Briggs et al., 2009; Bharat et al., 2012; Schur et al., 2016; Wagner et al., 2016b; **Figure 1D**).

Multimerization, which occurs soon after budding, activates the viral protease, embedded in the Gag-Pol precursor. Once activated, the PR proceeds to an ordered sequence of cuts that cleave the Gag and Gag-Pol precursors into their individual components (Pettit et al., 1994, 2005). For CA, the first cleavage occurs at the junction between MA and CA. Subsequently, SP1 undergoes a conformational switch that allows the cleavage of the CA-SP1 junction releasing the free CA protein (Pettit et al., 2005). Once released, CA dissociates from the hexamers of the Gag lattice and spontaneously re-assemble to reform hexamers and form pentamers. The arrangement of CA NTD and CTD in the hexamers of mature capsid is different from that of the hexamers of the lattice. The orientation is inverted, with the NTDs that point toward the center of the structure and, by interacting with each other, stabilize the structure of the hexamer. The CTDs, in contrast, are located toward the exterior, in a radial disposition, and are involved in inter-hexamers interactions, holding together the capsid core (Ganser-Pornillos et al., 2007; Byeon et al., 2009; Pornillos et al., 2009; Zhao et al., 2013; Mattei et al., 2016; **Figure 1E**). Approximately 250 hexamers are involved, together with 12 pentamers, in the formation of the fullerene cone structure, 120 nm long and 60 nm wide (Ganser et al., 1999; Li et al., 2000; De Marco et al., 2010; Zhao et al., 2013; **Figure 2A**). Even for a given virus, the fullerene cones can vary in number of CA molecules, shape, and positioning of the 12 pentamers. This variability makes this structure highly pleiomorphic, which endows it with a certain conformational flexibility, an important feature for a viral component that has a central role in the interaction with several factors both of viral and of cellular origin (Ganser-Pornillos et al., 2004; Mattei et al., 2016). Pentamers are highly similar to hexamers in their structure, although the pocket between the CA domains in hexamers that, as discussed below, interacts with host factors, is unfolded in pentamers (**Figure 2B**). It is therefore expected that this interaction, if still occurring, is modified in the case of the pentamers. Also, the interactions between the monomers are slightly different in pentamers (Ganser et al., 1999; Cardone et al., 2009; Pornillos et al., 2011; Mattei et al., 2016). A detailed knowledge of the interactions established between CA monomers is important since several cellular components specifically recognize only the multimerized form of the protein, implying that the interactions between CA monomers generate functional elements *per se*.

TURNING CELL PROTEINS INTO VIRAL COFACTORS

The infectious cycle is strictly intertwined with the cell components. The viral proteins, indeed, interact with various cell proteins that can act as antiviral factors or as viral cofactors. Among these, some have been shown to interact directly with the capsid. They include the cyclophilin A (CypA) (Luban et al., 1993), the cleavage and polyadenylation specificity factor 6 (F6) (Lee et al., 2010), two proteins that are part of the nuclear pore complex (NPC) (Nup358 and Nup153), and the transportin 3 (TNPO3) (Brass et al., 2008; König et al., 2008; **Table 1**).

The first intracellular protein to be described to interact with HIV-1 CA was CypA that was identified through a two-hybrid screening of a human cDNA library of proteins interacting with Gag (Luban et al., 1993). Importantly, ever since this observation, the interaction with CypA has been shown not to be specific for HIV-1 but to be common among lentiviruses, for which it has been documented to exist for millions of years (Katzourakis et al., 2007; Gilbert et al., 2009; Goldstone et al., 2010; Malfavon-Borja et al., 2013; Mu et al., 2014). CypA is a peptidylprolyl isomerase that is incorporated in the viral particle *via* an interaction with G221 and P222 of Gag (G89 and P90 in mature capsid), and it

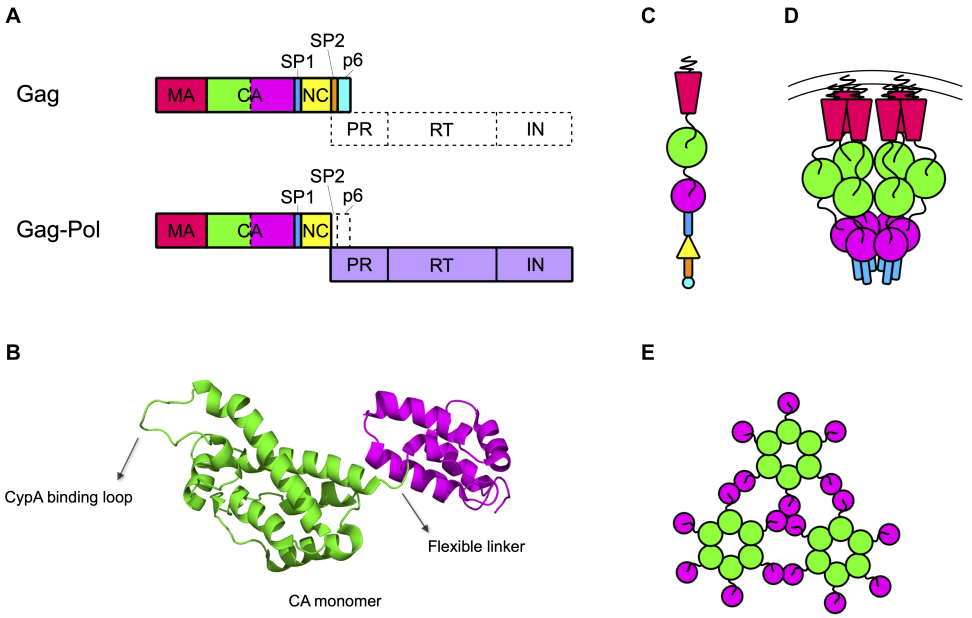


FIGURE 1 | Capsid forms throughout the HIV life cycle. **(A)** Gag and Gag-Pol precursors simplified structures. Gag precursor includes the matrix protein (MA), the capsid (CA, depicted with the NTD in green and the CTD in magenta), the spacer peptide 1 (SP1), the nucleocapsid (NC), the spacer peptide 2 (SP2), and the peptide 6 (p6). A frameshift during translation allows the production of Gag-Pol precursor, with a ratio of 1:20 with respect to the Gag precursor. In this structure the NC is fused to the protease (PR), the reverse transcriptase (RT), and the integrase (IN) domains. **(B)** Structure of CA monomer. CA is composed of two domains connected by a flexible linker: the NTD (in green), formed by a beta-hairpin and seven alpha-helices, and the CTD (in magenta), formed by four alpha-helices. The CypA binding loop in the NTD is indicated. PDB ID: 6WAP (Lu et al., 2020). **(C)** Schematic structure of the Gag precursor composed from top to bottom of MA, CA-NTD, CA-CTD, SP1, NC, SP2, and p6. **(D)** Schematic structure of a hexamer in the immature lattice, after the first proteolytic cleavage, which occurs between SP1 and NC. The MA are attached to the membrane through their myristoylated domain. Proceeding toward the center of the viral particle there are three hexameric structures composed by the CA-NTDs, CA-CTDs, and SP1. **(E)** Schematic top view of the mature capsid lattice where CA monomers are arranged in hexamers and are connected to each other through the NTDs, while the CTDs are involved in the interactions between hexamers.

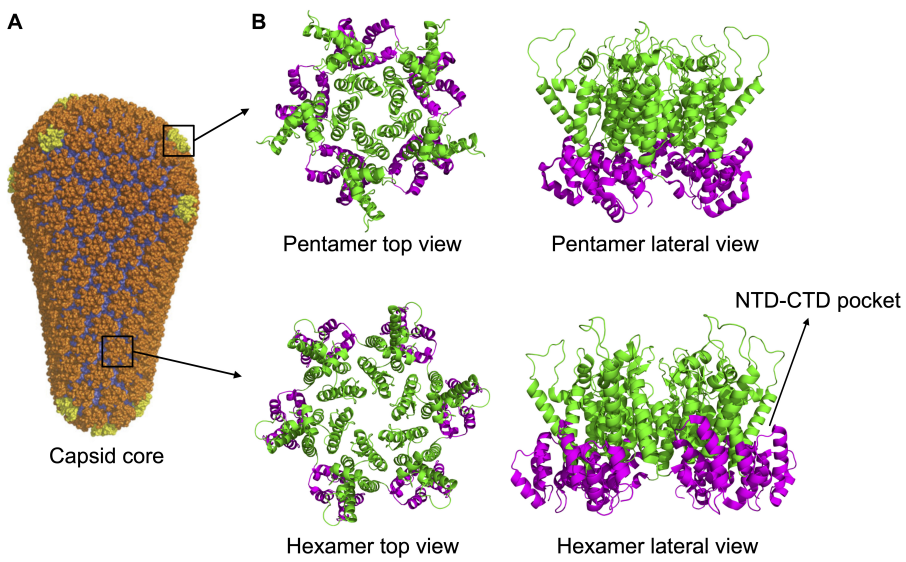


FIGURE 2 | Capsid core structure. **(A)** The mature capsid core has the shape of a fullerene cone, formed by 125 hexamers (in orange) and 12 pentamers (in yellow). Image republished with permission of Nature Publishing Group (Pornillos et al., 2011). **(B)** Top and lateral view of pentameric and hexameric capsid assemblies. In both structures, the NTDs (in green) are forming the inner ring while the CTDs (in magenta) are forming the external ring. The pocket present in the hexamer, at the NTD-CTD interface (involved in the interaction with host factors, see main text) is indicated. The pocket is absent in the pentamer. PDB IDs: 5MCX, 5MCY (Mattei et al., 2016).

TABLE 1 | Host factors interacting with the viral capsid.

Host factor	Gene	Biological role[a]	Role in HIV-1 Infection	Interaction with the capsid
Bicaudal D2 Protein	*BICD2*	Links the dynein motor complex to its cargos.	• Promotes the trafficking of viral cores toward the nucleus (Dharan et al., 2017).	Interacts with the assembled core through its C-terminal domain (Dharan et al., 2017; Carnes et al., 2018).
Cleavage and Polyadenylation Specificity Factor 6	*CPSF6*	One of the four subunits of the cleavage factor Im (CFIm), required for 3′-end RNA cleavage and polyadenylation processing.	• Participates in the nuclear import of the RTC/PIC complex (Chin et al., 2015; Burdick et al., 2020). • Involved in the choice of the integration sites (Chin et al., 2015; Rasheedi et al., 2016; Sowd et al., 2016; Achuthan et al., 2018; Francis and Melikyan, 2018; Bejarano et al., 2019).	Binds the hexameric form of CA in the nucleus at the NTD-CTD pocket (Lee et al., 2012; Price et al., 2012, 2014; Bhattacharya et al., 2014).
Cyclophilin A	*PPIA*	Cytoplasmatic peptidylprolyl *cis-trans* isomerase involved in proteins folding.	• Helps to maintain the stability of the capsid core (Li et al., 2009; Setiawan et al., 2016). • Involved in the choice of the nuclear import pathway (Schaller et al., 2011). • Protection from host restriction factors like TRIM5 (Kim et al., 2019; Selyutina et al., 2020a; Yu et al., 2020).	Binds to the capsid core in the cytoplasm by recognizing a conserved loop present in the NTD of CA (Franke et al., 1994; Gamble et al., 1996).
Extracellular Signal-Regulated Kinase 2	*MAPK1*	Serine/threonine-protein kinase part of the MAP kinase signal transduction pathway.	• Indirectly involved in promoting the uncoating step since its phosphorylation substrate is then recognized by Pin1 (Misumi et al., 2010; Dochi et al., 2014).	Phosphorylates the Ser16 of CA (Dochi et al., 2014).
Fasciculation and Elongation Protein Zeta 1	*FEZ1*	Kinesin-1 adaptor protein participating in the transport of cargos along microtubules.	• Promotes trafficking of the capsid core toward the nucleus (Malikov et al., 2015; Huang et al., 2019).	Binds the core at the hexamer pore (Huang et al., 2019).
Maternal Embryonic Leucine Zipper Kinase	*MELK*	Serine/threonine-protein kinase involved in many cellular pathways.	• Promotes viral uncoating (Takeuchi et al., 2017).	Phosphorylates the Ser149 of CA (Takeuchi et al., 2017).
MX Dynamin Like GTPase B	*MX2*	Interferon-induced dynamin-like GTPase protein located in the peripheric region of the nucleus.	• Blocks viral nuclear entry (Dicks et al., 2018; Kane et al., 2018). • Reduces integration efficiency (Kane et al., 2013; Liu et al., 2013; Matreyek et al., 2014).	Interacts with a negatively charged surface of CA (Smaga et al., 2019).
Non-POU Domain Containing Octamer Binding	*NONO*	RNA-binding protein with various roles in the nucleus including transcriptional regulation and RNA splicing.	• Restricts infection by activation of the immune response, *via* cGAS, after recognition of CA (Lahaye et al., 2018).	Binds to CA associated with the RTC/PIC complexes in the nucleus (Gao et al., 2013; Lahaye et al., 2013, 2018).
Nucleoporin 153	*NUP153*	NPC protein located in the nuclear basket of the complex with a role in the nucleocytoplasmic transport of proteins and mRNAs.	• Participates in the nuclear import of the viral complex (König et al., 2008; Matreyek and Engelman, 2011; Di Nunzio et al., 2012, 2013). • Directly or indirectly involved in the choice of the integration site (Koh et al., 2013; Marini et al., 2015).	It interacts with the multimeric form of CA at the NTD-CTD pocket at the same binding site of CPSF6 (Buffone et al., 2018; Bejarano et al., 2019).
Nucleoporin 358	*RANBP2*	RAN-binding protein located on the cytoplasmatic filaments of the NPC that promotes the nuclear import of large cargos.	• Favors the nuclear import of the viral complex (Schaller et al., 2011; Di Nunzio et al., 2012; Meehan et al., 2014; Dharan et al., 2016; Burdick et al., 2017). • Promotes uncoating of the capsid core at the NPC (Bichel et al., 2013).	Binds to the NTD domain of CA *via* a cyclophilin-homology domain as it approaches the NPC (Schaller et al., 2011).
Peptidylprolyl *Cis/Trans* Isomerase, NIMA-Interacting 1	*PIN1*	Peptidyl-prolyl *cis/trans* isomerase that specifically binds to phosphorylated ser/thr-pro motifs.	• Participates in the uncoating step (Misumi et al., 2010).	Recognizes the phosphorylated Ser16 of CA (Misumi et al., 2010).
Transportin 1	*TNPO1*	Involved in nuclear protein import as a receptor for nuclear localization signal.	• Involved in keeping the correct stability of the capsid core (Fernandez et al., 2019). • Helps the viral nuclear import (Fernandez et al., 2019).	Binds to the CypA binding-loop (Fernandez et al., 2019).

(Continued)

TABLE 1 | Continued

Host factor	Gene	Biological role[a]	Role in HIV-1 Infection	Interaction with the capsid
Transportin 3	*TNPO3*	Beta-karyopherin protein involved in the nuclear import of serine/arginine-rich (SR) proteins.	• Participates in the nuclear import step (Christ et al., 2008; Logue et al., 2011). • Involved in post-nuclear entry steps (Valle-Casuso et al., 2012; Shah et al., 2013). • Favors infection by participating in the nuclear localization of CPSF6 (De Iaco et al., 2013; Fricke et al., 2013).	Even if TNPO3 is also found in the cytoplasm, it most likely interacts with CA in the nucleus (Valle-Casuso et al., 2012; Shah et al., 2013).
Tripartite Motif Containing 5	*TRIM5*	Member of the tripartite protein family (TRIM) located in the cytoplasm of the cell where it autoassembles in cytoplasmic bodies.	• Affects the stability of the capsid core by either reducing it (Stremlau et al., 2006; Roa et al., 2012) or increasing it (Lu et al., 2015; Quinn et al., 2018). • Induces CA degradation *via* the proteasome (Lukic et al., 2011; Danielson et al., 2012; Kutluay et al., 2013) and/or the autophagy pathway (O'Connor et al., 2010; Mandell et al., 2014; Keown et al., 2018).	Forms a net around the intact capsid core in the cytoplasm by binding near or at the CypA binding site on CA (Quinn et al., 2018; Kim et al., 2019; Selyutina et al., 2020a; Yu et al., 2020).

[a]*Adapted from RefSeq.*

is found with a stoichiometry of approximately 1:10 (CypA:Gag) (Franke et al., 1994; Braaten et al., 1996b). Despite the fact that CypA is packaged in the viral particle from the infected cell, which could suggest that it plays a role at the level of the producer cells, it has been shown that it is the interaction between CA and the CypA molecules present in the target cells to be the major determinant for the effect exerted by CypA on HIV-1 infection (Hatziioannou et al., 2005). CypA interacts with the capsid core in two different ways. On one hand, the active site interacts with G89 and P90 of the P_{85}VHAGPIAP$_{93}$ loop (Gamble et al., 1996; **Figure 1B**) and, due to its isomerase activity, could destabilize the core (Braaten et al., 1996a,b; Bosco et al., 2002; Ylinen et al., 2009). On the other hand, other parts of the protein contact the hexamer interface and, bridging hexamers, likely stabilize the capsid core (Liu et al., 2016; Ni et al., 2020). Indeed, the effect of CypA on infection is to alter the stability of the capsid core, albeit the results are rather controversial since, depending on the cell type, it has been shown either to increase or to decrease it (Li et al., 2009; Setiawan et al., 2016). However, since mutating the CypA binding site on CA or the use of cyclosporin A (CsA), a drug that competes with the CA for CypA binding, both severely interfere with HIV infectivity (Franke et al., 1994; Braaten et al., 1996b) it appears that the virus relies on the interaction with this cellular cofactor to reach the optimal stability of the core. Another role of CypA during infection is to avoid the recognition by the tripartite motif (TRIM) containing protein TRIM5 of the capsid core either by inducing a conformational change through its isomerase activity or by steric hindrance (Kim et al., 2019; Ni et al., 2020; Selyutina et al., 2020a; Yu et al., 2020). Finally, the interaction between CA and CypA also appears to regulate the pathway of nuclear import of the reverse transcription and/or pre-integration complexes (RTC/PIC) that differs, according to whether CypA interacts with CA or not (Schaller et al., 2011).

Many cytoplasmic factors interact with the capsid core, on its way to the nucleus. Bicaudal D2 protein (BICD2) and the fasciculation and elongation protein zeta 1 (FEZ1) are two dynein adaptor proteins, required for HIV-1 infection, that interact with HIV-1 assembled multimeric cores (Malikov et al., 2015; Dharan et al., 2017; Carnes et al., 2018; Huang et al., 2019). Their depletion results in impaired cytoplasmic trafficking, uncoating, and nuclear import (Dharan et al., 2017; Huang et al., 2019). Uncoating has also been shown to be influenced by other host factors, as Pin1, MELK, ERK2, and TRN-1. Pin1 is a peptidyl-prolyl isomerase that facilitates HIV-1 core disassembly by interacting with the phosphorylated Ser16-Thr17 motif (Misumi et al., 2010). Responsible for the phosphorylation of Ser16 is the extracellular signal-regulated kinase 2 (ERK2), a cellular factor that is incorporated in the viral particle through its interaction with CA (Dochi et al., 2014). Another kinase involved in destabilizing the viral capsid, in this case through phosphorylation of Ser149, is the maternal embryonic leucine zipper kinase (MELK). The mutant where Ser149 is replaced by the phosphor-mimetic amino acid Glu undergoes premature disassembly of the capsid core and is impaired in nuclear import of the reverse transcription products (Takeuchi et al., 2017). Finally, β-karyopherin transportin 1 (TRN-1) recognizes the CypA binding site with high affinity and it can displace CypA from its association to the core. Knock out of TRN-1 leads to reduced infection and premature uncoating (Fernandez et al., 2019). Overall, the trend observed with these factors indicates that they are required in order to maintain in balance the subtle equilibrium between uncoating and retention of a closed capsid required to accomplish infection. Defects in nuclear import observed by depleting these factors appear to be a consequence of alteration of capsid uncoating rather than a direct interference with the import process.

Nuclear pore complex proteins regulate trafficking between the nucleus and the cytoplasm in eukaryotic cells (Strambio-De-Castillia et al., 2010; Labokha and Fassati, 2013). Two of these proteins are well-characterized interactants of HIV-1 CA: Nup153 and Nup358 (also known as RANPB2) (Brass et al., 2008; König et al., 2008). Nup358 is associated with filaments

that stem from the pore into the cytoplasm and it promotes the recruitment of nuclear import cargos (Hutten et al., 2009). It contains a cyclophilin-homology domain that is responsible for the interaction with CA (Schaller et al., 2011). As CypA, Nup358 has a *cis-trans* prolyl isomerization activity through which it can promote capsid core uncoating by catalyzing isomerization of CA (Bichel et al., 2013). This suggests that uncoating of the viral core could occur, at least partially, at the nuclear pore, once docked onto Nup358. Accordingly, depletion of Nup358 severely affects HIV-1 nuclear import, with a reduction of the amount of RTC/PIC docked at the NPC (Zhang et al., 2010; Schaller et al., 2011; Di Nunzio et al., 2012; Meehan et al., 2014; Dharan et al., 2016; Burdick et al., 2017). Nup153 is one of the components of the nuclear basket involved in the NPC formation and Nups recruitment (Vollmer et al., 2015). Through its C-terminal domain it binds the NTD-CTD pocket of CA (Buffone et al., 2018; Bejarano et al., 2019) and it favors its translocation into the nucleus (König et al., 2008; Matreyek and Engelman, 2011; Di Nunzio et al., 2012, 2013). Its depletion also alters the choice of the sites of integration (Koh et al., 2013; Marini et al., 2015). Since Nup153 binds CA hexamers with high affinity compared to monomeric CA (Matreyek and Engelman, 2011; Di Nunzio et al., 2012; Buffone et al., 2018), this translocation likely involves capsid cores that, if not intact, are at least partially assembled.

Another cellular protein interacting with CA is CPSF6, a pre-mRNA splicing factor, and a member of the serine/arginine-rich protein family (Rüegsegger et al., 1998). CPSF6 is part of the cleavage factor I (CFI_m), together with CPSF5 and CPSF7, but its activities related to HIV-1 do not involve the other proteins of the complex (Rasheedi et al., 2016). CPSF6 binding site on CA is bipartite as CPSF6 binds at the N-terminal region of CA monomers but also at the NTD-CTD pocket of adjacent monomer on CA hexamers (Lee et al., 2012; Price et al., 2012, 2014; Bhattacharya et al., 2014). CPSF6 was initially identified to be relevant for HIV-1 infection through the functional screening of a mouse cDNA expression library that led to the isolation of a truncated form of CPSF6 (CPSF6-358) inhibiting HIV-1 replication (Lee et al., 2010). The truncation removes in CFSP6-358 the C-terminal arginine-serine like domain (RSLD) that is required for its nuclear import by transportin 3 (TNPO3) (Jang et al., 2019). As a consequence, the two forms of CPSF6 display different localizations inside the cell, with CPSF6 being predominantly nuclear while CPSF6-358 is found exclusively in the cytoplasm (Lee et al., 2010). This difference is responsible for the antiviral effect exerted exclusively by CPSF6-358 that blocks HIV-1 infection by interacting with the capsid core in the cytoplasm and preventing nuclear import (Lee et al., 2010). The integral form of CPSF6, in contrast, favors HIV-1 infection. Its effect is dependent on the cell type considered. Indeed, CPSF6 is an important factor in primary CD4+ T cells and macrophages, where it directs integration toward euchromatin regions, and its deletion leads to an accumulation of RTC/PIC complexes at the nuclear pore and integration in chromatin regions close to the nuclear pore (Chin et al., 2015; Rasheedi et al., 2016; Sowd et al., 2016; Achuthan et al., 2018; Francis and Melikyan, 2018; Bejarano et al., 2019; Burdick et al., 2020). These effects are not observed in HeLa or HEK 293T cells (Lee et al., 2010; Kane et al., 2018; Bejarano et al., 2019). The CPSF6 binding site on CA appears to overlap the region recognized by the nuclear pore protein Nup153, important for HIV-1 nuclear import, as discussed above, implying a competition for CA binding that could favor, once imported in the nucleus, the release from Nup153 to allow CPSF6 binding and its translocation into deeper nuclear regions (Bejarano et al., 2019).

Transportin 3 is a β-karyopherin that transports serine/arginine-rich splicing factors in the nucleus (Kataoka et al., 1999; Lai et al., 2000). It binds to HIV-1 CA and its depletion affects HIV-1 infection (Christ et al., 2008; Krishnan et al., 2010; Logue et al., 2011; Zhou et al., 2011; Valle-Casuso et al., 2012; Shah et al., 2013). The role of TNPO3 in HIV-1 infection is still debated. Some studies suggest a role in nuclear import (Christ et al., 2008; Logue et al., 2011) while others rather suggest an implication in post-nuclear import, but prior to integration (Zhou et al., 2011; Valle-Casuso et al., 2012; Shah et al., 2013). However, TNPO3 is also responsible for the nuclear import of CPSF6 (De Iaco et al., 2013; Maertens et al., 2014; Jang et al., 2019) which, in HIV-1 infection, favors nuclear transport, as discussed above. It is therefore possible that the effects on HIV-1 infectivity attributed to TNPO3 are not only direct but also a consequence of the effect of TNPO3 on CPSF6 (De Iaco et al., 2013; Fricke et al., 2013). In support of this view is the observation that another effect of the depletion of TNPO3 is a change in the choice of the integration sites (Ocwieja et al., 2011), which is the same phenotype observed when depleting CPSF6.

Besides assisting various steps of the infectious process from the mechanistic standpoint, as capsid uncoating or nuclear translocation, these host factors also have a role in the escape from innate immunity. For example, infection by viruses with mutated CA that no longer interact with several of these factors (CPSF6, CypA, and Nup358), triggers an interferon-mediated antiviral response in human monocyte-derived macrophages (Rasaiyaah et al., 2013). Consequently, the capsid is subject to positive selection for maintaining the interaction with these proteins. At the same time, it is also the target of several cellular factors endowed with antiviral activity, from which it has to escape, adding a layer of selective pressure. The most well-characterized of these factors are constituted by a member of the tripartite motif-containing proteins family TRIM5 (Stremlau et al., 2004), the myxovirus resistance gene A and B (MxA and MxB) (Liu et al., 2013), and the non-POU domain-containing octamer binding protein (NONO) (Lahaye et al., 2018; **Table 1**).

ANTIVIRAL FACTORS TARGETING THE CAPSID

An important cellular antiviral factor directed against the capsid is TRIM5α. TRIM5α was isolated from rhesus macaque ($TRIM5\alpha_{rh}$) in the context of studies aimed at understanding the reasons for the inability of HIV-1 to establish productive infections in Old World monkey cell lines (Shibata et al., 1995; Hofmann et al., 1999; Besnier et al., 2002; Cowan et al., 2002). Independently, a variant of this protein (TRIMCyp), exclusive

to owl monkeys, was identified for its ability to confer the same phenotype of restriction to HIV-1 infection (Sayah et al., 2004; Stremlau et al., 2004). In both cases, the viral target was identified to be the capsid and, in particular, the assembled core rather than the monomeric form of CA (Cowan et al., 2002; Hatziioannou et al., 2004; Stremlau et al., 2006).

As members of the TRIM family, TRIM5α$_{rh}$ and TRIMCyp are composed of a N-terminal tripartite motif constituted by the RING domain, a B-box 2 domain, and a coiled-coil domain (Reymond et al., 2001). The TRIM is followed by a C-terminal domain: cyclophilin A in TRIMCyp, and the PRYSPRY in TRIM5α. These domains bind the CA protein at or near the CypA-binding domain (**Figure 3**; Quinn et al., 2018; Kim et al., 2019; Selyutina et al., 2020a; Yu et al., 2020). TRIM5α and TRIMCyp dimerize through the coiled-coil domain, which places the two B-box 2 domains at each extremity of an antiparallel dimer. The B-box 2 domain can form trimers allowing the formation of a network of hexamers. These hexamers can assemble into a hexagonal lattice around an incoming retroviral capsid core, in which the C-terminal domains interact with the capsid (Sebastian and Luban, 2005; Li et al., 2016; Wagner et al., 2016a; Quinn et al., 2018; Yu et al., 2020). If the mechanisms of binding of TRIM5 to the capsid core are well understood, by which means it restricts HIV-1 infection is still debated. Some studies suggest that the ability of the protein to form a net around the capsid is sufficient to perturb the capsid core stability and, therefore, infectivity. The net would either induce the destabilization of the capsid core, resulting in a premature and non-productive uncoating (Stremlau et al., 2006; Zhao et al., 2011; Roa et al., 2012), or increase its stability by reducing the intrinsic flexibility of the core and of the CypA-binding loop in particular (Lu et al., 2015; Quinn et al., 2018). In both cases, infectivity would be perturbed. Other works indicate alternative pathways, activated by TRIM5α, to degrade the capsid core, as the recruitment of the proteasome, thanks to the ability of TRIM5α to undergo self-ubiquitylation, thanks to the RING domain (Fletcher et al., 2018) while associated to the capsid core (Lukic et al., 2011; Danielson et al., 2012; Kutluay et al., 2013) or by inducing selective autophagy of the capsid core (O'Connor et al., 2010; Mandell et al., 2014; Keown et al., 2018). However, neither blocking the proteasome nor the pathways leading to autophagy abolishes the restriction activity of TRIM5α suggesting that several, non-exclusive, pathways are activated in response to the recognition of the viral core (Perez-Caballero et al., 2005; Anderson et al., 2006; Diaz-Griffero et al., 2006; Wu et al., 2006; Kutluay et al., 2013; Imam et al., 2016; Keown et al., 2018).

The wealth of information concerning the restricting function of TRIM5α comes primarily from studies of the rhesus monkey protein. Indeed, the human ortholog of TRIM5α does not block HIV-1 infection in cell lines (Hatziioannou et al., 2004; Stremlau et al., 2004; Yap et al., 2004), although it protects human cells from the infection by some non-human retroviruses (Hatziioannou et al., 2004; Keckesova et al., 2004; Perron et al., 2004; Yap et al., 2004). Furthermore, a stabilized form of TRIM5α, obtained by producing a fusion protein with mCherry, protects human T cells in humanized murine models of HIV-1 infection (Richardson et al., 2014). Human TRIM5α is also involved in

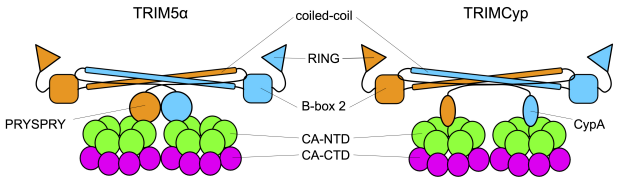


FIGURE 3 | Interaction between TRIM and the capsid. TRIM5α and TRIMCyp are represented in their dimeric form. Each monomer (in orange and in blue) is formed by the RING domain, the B-Box 2 domain, the coiled-coil domain and the C-terminal domain which is the one responsible for the interaction with the capsid core. In TRIM5α this domain is the PRYSPRY domain while in TRIMCyp is CypA.

IFNa-induced inhibition against HIV-1 infection (Kane et al., 2016; OhAinle et al., 2018; Jimenez-Guardeño et al., 2019). In fact, high levels of IFNa activate the immunoproteasome, inducing a rapid turnover of TRIM5α that, being bound to the capsid core, drives to its degradation blocking viral replication (Jimenez-Guardeño et al., 2019).

The weak restriction of HIV-1 by human TRIM5α is suggested to be due to inefficient recognition of the capsid core (Stremlau et al., 2005; Yap et al., 2005; Merindol et al., 2018). The fact that the binding site of CypA on capsid cores overlaps (at least partially) the region bound by TRIM5α could lead to competitive inhibition of binding of TRIM5α, contributing to the inefficient recognition of the core by TRIM5α (Kim et al., 2019; Selyutina et al., 2020a; Yu et al., 2020). The lack of effectiveness of the human TRIM5α protein against infection with the human variant of the virus may reflect the recent exposure of humans to this virus. Alternatively, it could be imagined that HIV possesses a yet to be defined activity that counteracts that of TRIM5α.

The human *myxovirus resistance* (*Mx*) B protein (MxB, also known as Mx2) is an important anti-HIV factor that targets the viral capsid (Goujon et al., 2013; Kane et al., 2013; Liu et al., 2013; Matreyek et al., 2014). It is a dynamin-like GTPase, a family of proteins highly conserved in all vertebrates (Verhelst et al., 2013). MxB is constituted by a globular GTPase domain, a C-terminal stalk domain, a bundle signaling element (BPE), and a non-structured N-terminal domain (Gao et al., 2011). It localizes on the cytoplasmic side of the nuclear envelope, near the NPC (King et al., 2004). This antiviral factor is effective against herpesvirus, murine cytomegalovirus (MCMV), and HIV-1 (Goujon et al., 2013; Kane et al., 2013; Liu et al., 2013; Crameri et al., 2018; Jaguva Vasudevan et al., 2018; Schilling et al., 2018). In the N-terminal domain of MxB there is a positively charged motif, the ^{11}RRR^{13} motif, that recognizes a negatively charged surface highly conserved among lentiviral capsid cores (Smaga et al., 2019). This interaction is responsible for the restriction of the infection (Goujon et al., 2015; Schulte et al., 2015) that, depending on the experimental conditions used, has been attributed either to a decrease of nuclear import of the RTC/PIC complexes by interfering with nuclear pore associated proteins (Dicks et al., 2018; Kane et al., 2018) or to a decrease of integration levels (Kane et al., 2013; Liu et al., 2013; Matreyek et al., 2014). Finally, a possible implication of MxB in the restriction response of the host

restriction factor SAMHD1 has been recently suggested although it is still not clear how this is exerted (Buffone et al., 2019).

Another host factor with anti-HIV-1 activity related to targeting the viral capsid is the non-POU domain-containing octamer-binding protein (NONO), a member of the *Drosophila behavior/human splicing* (DBHS) family. The proteins of this family are characterized by the presence of two N-terminal RNA recognition motifs (RRMs), a NonA/paraspeckle domain (NOPS), and a C-terminal coiled-coil domain (Knott et al., 2016). NONO is a nuclear protein and has both RNA- and DNA-binding properties and it is involved in the activation of the innate immune response in dendritic cells and macrophages upon HIV infections, with a more efficient response against HIV-2 than HIV-1 (Lahaye et al., 2018). In the nucleus, NONO binds CA associated with the RTC/PIC complexes, and its restriction effect is exerted through the DNA sensor cyclic GMP-AMP synthase (cGAS), which activates the innate immune response by sensing the viral double-stranded DNA (Gao et al., 2013; Lahaye et al., 2013, 2018). Without NONO, cGAS is found in the cytosol and it does not activate the immune response (Lahaye et al., 2018).

THE VIRAL UNCOATING STEP AND THE IMPORTANCE OF ITS TIMING

The timing of dismantling of the viral capsid is a crucial aspect for a successful infection since premature disassembly would expose the components of the reverse transcription complex to the antiviral responses of the host cell and it would dilute the viral components by releasing them into the cytoplasm. On the other hand, the delayed dismantling of the capsid core could affect the process of integration by sequestering the reverse transcription products. To date, not only when and where reverse transcription and dismantling of the capsid core occurs is still an open question, but it is even still debated if and how the two processes are connected. Indeed, while some works show that DNA synthesis promotes uncoating (Hulme et al., 2011, 2015; Yang et al., 2013; Cosnefroy et al., 2016; Francis et al., 2016; Mamede et al., 2017; Rankovic et al., 2017), others show that the inhibition of reverse transcription neither affects uncoating nor the nuclear import of the RTC/PIC (Lukic et al., 2014; Burdick et al., 2017; Bejarano et al., 2019; Selyutina et al., 2020b).

Answering these questions is technically challenging, though. A major difficulty comes from the intrinsic properties of the capsid cores, discussed above, that is at the origin of the generation of polymorphic capsid cores, most of which intrinsically unstable and, therefore, non-infectious (Thomas et al., 2007; Mattei et al., 2016). It is, in fact, considered that only a minority of viral particles entering the cell leads to successful infection, while the majority is constituted by defective cores that undergo proteasomal degradation. The earliest studies on the capsid were mostly based on the biochemical tracking of the intact capsid in the infected cell. These analyses, consequently, followed the fate of the capsids at the "population" level and documented a rapid dismantling of the capsid after entry into the cell. The minority of stable capsids that, according to recent data, is responsible for productive infection, was not detected.

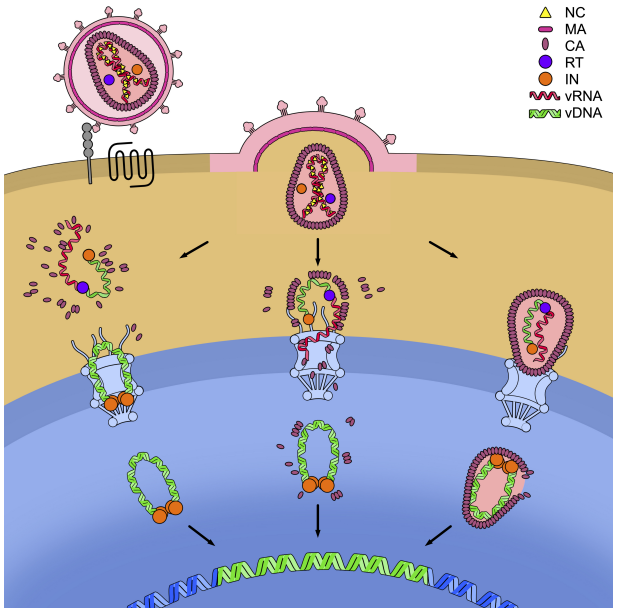


FIGURE 4 | Models for the timing of uncoating. HIV-1 enters the cell after recognition by the envelope glycoproteins of the cellular receptor CD4 (in gray) and the cellular co-receptor CXC4 or CCR5 (in black). This leads to the fusion of the cell and viral membranes and to the release of the capsid core in the cytoplasm. In the figure, the three models of uncoating covered in this review are depicted: the cytoplasmic uncoating (on the left), the uncoating at the nuclear pore complex (NPC) (in the center), and the nuclear uncoating (on the right). In each model the reverse transcription of the viral genomic RNA (vRNA) (in red) into viral DNA (vDNA) (in green) has to be completed, allowing its integration in the host genome (in blue). The reverse transcription complex (RTC) is schematically shown as the association of a molecule of reverse transcriptase (RT, in purple) to the vRNA and single-stranded vDNA. The completed vDNA forms the pre-integration complex (PIC), shown as the double-stranded vDNA bound to a tetramer of integrase (IN, in orange).

The advent of techniques that allow following, by different means, the individual capsids has permitted focusing on the minority of capsids that persist in the cell changing our view of the timing of uncoating of the particles relevant for productive infection. The different scenarios that have been depicted for the dismantling of the capsid core are recapitulated hereafter.

Cytoplasmatic Disassembly

According to the earliest models, uncoating occurs in the cytoplasm, soon after viral entry (early cytoplasmic disassembly) (Miller et al., 1997; Fassati and Goff, 2001). This model was supported by biochemical studies showing the lack of detectable CA in the cytoplasm (Bukrinsky et al., 1993; Miller et al., 1997; Fassati and Goff, 2001). However, increasing evidence showing the presence of CA and/or capsid cores in the cytoplasm of the infected cells has subsequently challenged this view (McDonald et al., 2002; Forshey et al., 2005; Shi and Aiken, 2006; Stremlau et al., 2006; Kutluay et al., 2013; Yang et al., 2013). It has thus been proposed that uncoating still occurs in the cytoplasm (Miller et al., 1997; Fassati and Goff, 2001) but delayed with respect to viral entry (late cytoplasmic disassembly) and coupled with

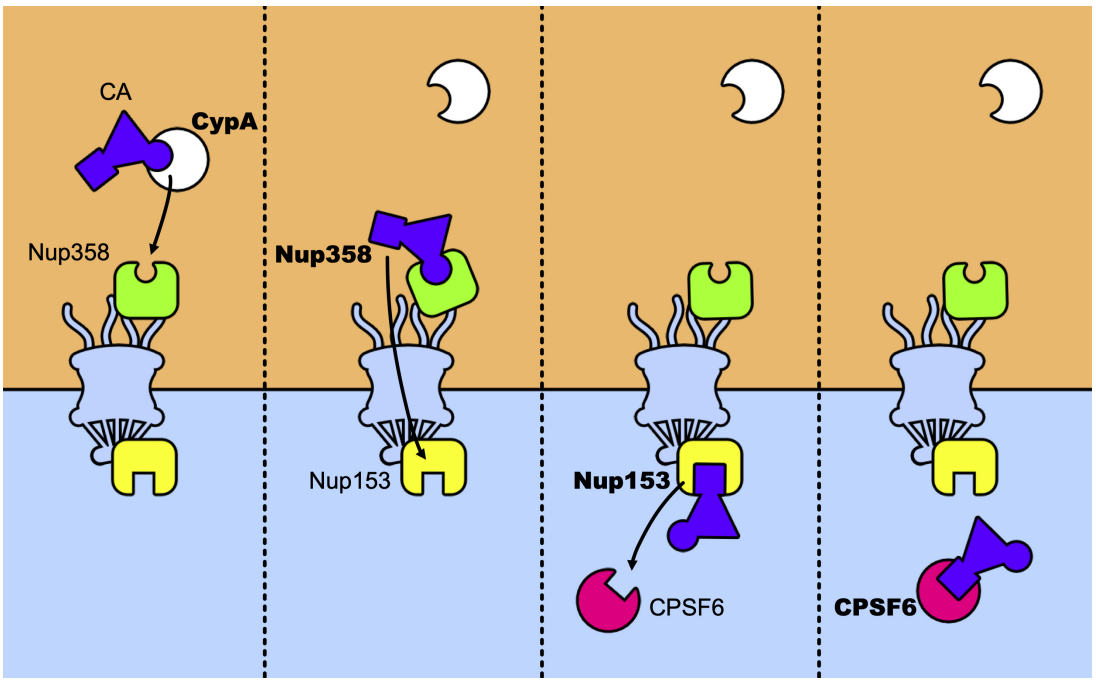


FIGURE 5 | Relay race of the capsid core in the host cell. From left to right a temporal view of how CA is passed between host factors in its trip toward the nucleus. The capsid core is schematically represented as a purple triangle with two host factors binding sites highlighted: the CypA binding-loop (the circle) and the NTD-CTD pocket (the square). The first to bind to the core is CypA, which recognizes the CypA-binding domain, located in the CA-NTD. The same binding site is recognized by Nup358 and its binding anchors the capsid core at the NPC, allowing its nuclear import. Then, Nup153 binds to the NTD-CTD pocket of the assembled capsid, which is the same recognition site of CPSF6. When CPSF6 takes the place of Nup153 on the binding site it can translocate the capsid core (intact or not) to deeper nuclear regions.

reverse transcription (Hulme et al., 2011; Cosnefroy et al., 2016). A longer presence of an assembled capsid in the cytoplasm appeared also more plausible since it accounted for the protective role of the capsid from the exposure of the viral genome to host restriction factors and to the potential activation of the IFN-mediated antiviral response (Iwasaki, 2012). To date, it is accepted that uncoating in the cytoplasm concerns a fraction of the infecting particles and that, in general, it is only partial, with capsid hexamers that remain associated with the RTC/PIC complex, where they exert important functions in late steps of the infectious cycle (see below).

Disassembly at the Nuclear Pore

As lentiviruses, unique among retroviruses, are able to infect non-replicating cells, entry into the nuclear compartment must proceed through the nuclear pore. Since the capsid core is larger than the nuclear pore, it was considered that the intact capsid could not be imported into the nucleus and, rather, it was blocked once docked at the level of the NPC (Arhel et al., 2007; Matreyek and Engelman, 2011; Schaller et al., 2011; Burdick et al., 2017; Francis and Melikyan, 2018; Francis et al., 2020; Zurnic Bönisch et al., 2020). Uncoating would then occur *in situ*, before import of the RTC/PIC could be possible. In support of this view came the measure of the time of residence of the viral complex at the nuclear pore that, for HIV-1, spans between 30 and 90 min (Burdick et al., 2017; Francis and Melikyan, 2018). Since macromolecular complexes of sizes similar to the RTC/PIC of HIV-1 have very short times of nuclear entry and a total binding time to the NPC of few milliseconds (Kelich et al., 2015), it was inferred that the longer time observed for HIV reflected the need for the capsid core to be dismantled and release the RTC/PIC. This way, the capsid core would protect the RTC/PIC from exposure to the proteasome until it has reached the proximity of the point of entry into the nucleus (Francis and Melikyan, 2018).

Nuclear Disassembly

Increasing evidence, though, supports the possibility that, despite the apparent incompatibility in terms of size, the capsid core enters the nucleus intact or almost intact, and disassembles only once inside it. It has indeed been shown that several host factors interact, at the nuclear level, with the assembled capsid rather than CA monomers (Matreyek and Engelman, 2011; Di Nunzio et al., 2012; Valle-Casuso et al., 2012; Chin et al., 2015; Buffone et al., 2018; Bejarano et al., 2019). Furthermore, assuming that uncoating is favored by reverse transcription (Hulme et al., 2011, 2015; Cosnefroy et al., 2016; Francis et al., 2016; Mamede et al., 2017; Rankovic et al., 2017), if it constitutes a requirement for nuclear import of the RTC/PIC, blocking reverse transcription would be expected to affect nuclear import. This was not the case though, while increasing evidence supports a model where reverse transcription is completed only once in the nucleus (Burdick et al., 2017, 2020; Francis and Melikyan, 2018;

Bejarano et al., 2019; Dharan et al., 2020; Francis et al., 2020; Rensen et al., 2020; Selyutina et al., 2020b). The most compelling evidence in favor of the idea that uncoating can occur in the nucleus then came from a series of recent works (Burdick et al., 2020; Dharan et al., 2020; Selyutina et al., 2020b). By labeling the capsid core with the GFP, producing a GFP-CA fusion protein, Burdick and coworkers observed that the core enters the nucleus while still intact (or almost intact), that reverse transcription is completed, and, finally, that uncoating occurs close to the integration sites approximately 1.5 h before integration (Burdick et al., 2020). In a concomitant work, Dharan et al. (2020) employed an inducible blockade of nuclear import at different time points and then evaluated the fate of the capsid cores that had entered the nucleus. In this setting, two main observations were made. One was that the completion of reverse transcription, as inferred by sensitivity to treatment with an inhibitor of reverse transcription, was posterior to nuclear import. The second observation was that, even after blocking nuclear import, the infection was susceptible to treatment with PF74. Since this compound inhibits infection through binding specifically the interface between CA monomers, these observations indicated that assembled (or partially assembled) capsid cores were present in the nuclear fraction. Finally, the observations that uncoating and reverse transcription are completed in the nucleus, have also been confirmed by the biochemical analyses of the purified cytosolic and nuclear fractions in infected cells by Selyutina et al. (2020b).

These various models of dismantling of the capsid core are not mutually exclusive and it is possible that, depending on the cell type considered, the relative predominance of one or the other scenario is found. Might this be under the form of RTC/PIC deprived of CA, of a partially dismantled or of an intact capsid core, the viral element containing the genetic material must however, be translocated across the nuclear pore of the cell (**Figure 4**).

GETTING INTO THE NUCLEUS, SOMEHOW

The main nuclear import pathway of HIV-1 appears as a relay race where the capsid core is passed from CypA to Nup358, which passes it across the nuclear pore to Nup153 that will finally pass it to CPSF6 (**Figure 5**). However, alternative pathways exist. Mutants N74D and A77V of CA, identified for their less efficient binding to CPSF6 no longer require CypA, Nup153, Nup358, and TNPO3 (Lee et al., 2010; Schaller et al., 2011; Ambrose et al., 2012; Saito et al., 2016; Buffone et al., 2018). Despite this, they retain levels of infectivity comparable to those of the wt viruses, in primary cells. This suggests that, in these cells, alternative pathways are favored by these mutations. Concomitantly, these mutations induce uncoating at the nuclear pore and shift the integration sites to perinuclear regions (Burdick et al., 2020), in line with studies that show the importance of CPSF6 for nuclear import and the choice of the integration sites (Chin et al., 2015; Rasheedi et al., 2016; Sowd et al., 2016; Achuthan et al., 2018; Francis and Melikyan, 2018; Bejarano et al., 2019). Along the same lines, blocking transport across the nuclear pore by an inducible NPC blockade (Dharan et al., 2020), neither abolished nuclear import of the capsid nor blocked infection, indicating that nuclear pores can present a heterogeneous composition of nucleoporins and that factors alternative to the canonical Nup153, Nup358, and TNPO3 can also be used by the virus to achieve integration, in accordance with previous observations (Dicks et al., 2018; Kane et al., 2018). It is tempting to speculate that the use of these alternative factors is indicative of ancestral, less efficient, pathways at the expense of which the current canonical pathways of infection have evolved. On this note, the interaction between CA and CPSF6 seems to be preserved by selective pressure *in vivo* (Henning et al., 2014; Saito et al., 2016). This shift in the nuclear entry pathway would be a consequence of the use of previously unemployed cellular cofactors that allowed to optimize various steps of the infectious cycle and to improve escape from innate immunity.

Size also matters for nuclear import. Depending on where disassembly occurs (**Figure 4**), the nature and, consequently, the size of the complex that must cross the nuclear barrier changes considerably. The intact capsid core is around 60 nm wide (Briggs et al., 2003) while the nuclear pore is no larger than 40 nm (Von Appen et al., 2015). As discussed above this incongruence has long been considered a reason to exclude the possibility that the intact capsid core can be imported into the nucleus. Recently, by using a new method of visualization of capsid cores, based on immunogold labeling, Blanco-Rodriguez et al. (2020) showed that the capsid core undergoes important structural rearrangements before, during, and after nuclear import, leading to the formation of a pearl necklace-like shape that decorates the reverse transcribed DNA. The CA molecules, present in this structure that is considerably less wide than the intact capsid, could more easily mediate nuclear import. The possibility that structural rearrangements also involve the nuclear pore counterpart has been foreseen. Indeed, the NPC displays a marked structural flexibility that can be involved in the passage of large complexes as viral capsids (Knockenhauer and Schwartz, 2016; Mahamid et al., 2016). Furthermore, recent measurements of the inner diameter of the NPC by using cryo-EM on intact infected T cells have estimated a width of the internal channel of the pore of 64 nm, thereby slightly larger than the capsid core (Zila et al., 2021). The structure of the nuclear pore was dilated rather than rearranged with respect to previous observations made on HeLa cells where the canal appeared considerably narrower (Von Appen et al., 2015). In conclusion, increasing evidence supports the view that still "structured" capsid cores do enter the nucleus, this might be due to either partial uncoating that induces higher plasticity of the capsid core, either to structural rearrangements of the nuclear pore, either both.

GENETIC FRAGILITY OF THE CAPSID: A MARK OF MULTIPLE CONSTRAINTS?

The retroviral capsid core is responsible for chaperoning the viral genetic material all along from the fusion of the viral and cellular membranes till its entry (or even after) into the nucleus.

To accomplish this, the mature CA protein must meet several structural requirements to retain its ability to multimerize in order to assemble into the capsid core (and this relying on two different types of contacts, one giving rise to CA hexamers, the other generating pentamers, as discussed above), to interact with numerous cellular factors (**Table 1**) and to escape from adaptive immunity, being a target of cytotoxic T lymphocytes (CTLs) (Troyer et al., 2009). Furthermore, as a domain of Gag and Gag-Pol precursors, it must retain structural arrangements that do not interfere with the proteolytic processing of these molecules. Altogether, these constraints can account for the extreme genetic fragility of the protein (Rihn et al., 2013).

Genetic robustness is the ability to retain functionality when mutations are introduced in the protein (Visser et al., 2003; Wagner, 2005). Two main factors contribute to determining the genetic robustness of a protein. One is the number of functions the protein has to ensure and, consequently, the number of interactants it must come into contact with, in order to carry out its functions. The other is its architectural organization. For example, the presence of intrinsically disordered regions confers genetic robustness to proteins (Brown et al., 2002; Hultqvist et al., 2017). In the case of HIV-1 CA, the high number of partners it interacts with is likely the main determinant.

Local fluctuations in the degree of fragility are observed in CA. Internal regions of the protein are less tolerant of mutations as well as helices regions in the NTD rather than in the CTD and in the interhelical loops among which, surprisingly, the loop interacting with CypA. In particular, the region with the highest fragility is the one encoding the alpha-helices present in the NTD (Manocheewa et al., 2013; Rihn et al., 2013). This region is responsible, in the assembled core, for the interaction of each monomer with each other on the internal side of the hexamer, to form the internal ring (**Figure 2B**; Li et al., 2000; Pornillos et al., 2009, 2011). In addition, NTDs interact with the CTDs of adjacent monomers on the external portion of the CA (Lanman et al., 2003, 2004; Pornillos et al., 2009). These interactions must be finely tuned since during the extracellular life of the virus they must be stable enough to maintain a closed capsid core, but once inside the target cell they must allow the progressive dismantling of the structure, with the appropriate timing, as discussed above (Forshey et al., 2002). Maintaining this delicate equilibrium can account for the fragility of these regions. Of particular interest are the epitopes recognized by CTLs that appear particularly vulnerable to the introduction of genetic polymorphisms. A similar situation is found for the external regions of the HIV-1 envelope, which are the target of heavy artillery by the immune response, in this case humoral. It has been shown that in these regions the genomic sequence has evolved in such a way as to reduce the mutation rate (Geller et al., 2015), an observation interpreted as a mechanism to limit the cost of deleterious mutations, particularly high in these regions (Simon-Loriere et al., 2009; Hamoudi et al., 2013; Gasser et al., 2016). Marked genetic fragility could therefore constitute a common signature of regions under strong immune selection. Finally, several mutations that have a positive effect on viral replication *in vitro* were not found in natural populations, suggesting the existence of additional, presently unknown, sources of selection that counterselect some positive mutants but not others (Rihn et al., 2013). Identifying these sources of selection appears an important step for understanding the molecular bases of successful viral replication *in vivo*.

The marked genetic fragility of the capsid therefore likely derives from the cumulative requirements for interacting with a plethora of cellular factors that the virus has learned to deal with, for an optimal adaptation to its host. This fragility is probably responsible for the limited capacity of the capsid to avoid the immune response of the host (Troyer et al., 2009) and encourages to design new drugs targeting this protein. Drugs from which, in strict analogy to what occurs for the immune response, it should be difficult to escape.

CONCLUDING REMARKS

The ultimate goal of a retrovirus is to reach the genetic material of the infected cell to integrate its own. To do so, the infectious cycle passes through two phases, an extracellular and an intracellular one. For each of these, a shell has been optimized. We now know that, as many vulnerable aspects of the envelope proteins are largely not accessible until the target cell has not been reached, also for the intracellular delivery of its genetic material, the virus does not leave a large window of opportunity for the host cell to sense and attack its genetic material. This, until the final destination is almost reached.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

Work in MN's laboratory is supported by Sidaction and the ANRS (Franch Agency for Rasearches on AIDS and viral hepatitis). ET is recipient of a doctoral contract from the ANRS.

REFERENCES

Accola, M. A., Strack, B., and Göttlinger, H. G. (2000). Efficient particle production by minimal Gag constructs which retain the carboxy-terminal domain of human immunodeficiency virus type 1 Capsid-p2 and a late assembly domain. *J. Virol.* 74, 5395–5402. doi: 10.1128/jvi.74.12.5395-5402.2000

Achuthan, V., Perreira, J. M., Sowd, G. A., Puray-Chavez, M., McDougall, W. M., Paulucci-Holthauzen, A., et al. (2018). Capsid-CPSF6 interaction licenses nuclear HIV-1 trafficking to sites of viral DNA integration. *Cell Host Microbe* 24, 392–404.e8. doi: 10.1016/j.chom.2018.08.002

Ambrose, Z., Lee, K., Ndjomou, J., Xu, H., Oztop, I., Matous, J., et al. (2012). Human immunodeficiency virus type 1 capsid mutation N74D alters

cyclophilin a dependence and impairs macrophage infection. *J. Virol.* 86, 4708–4714. doi: 10.1128/jvi.05887-11

Anderson, J. L., Campbell, E. M., Wu, X., Vandegraaff, N., Engelman, A., and Hope, T. J. (2006). Proteasome inhibition reveals that a functional preintegration complex intermediate can be generated during restriction by diverse TRIM5 proteins. *J. Virol.* 80, 9754–9760. doi: 10.1128/jvi.01052-06

Arhel, N. J., Souquere-Besse, S., Munier, S., Souque, P., Guadagnini, S., Rutherford, S., et al. (2007). HIV-1 DNA Flap formation promotes uncoating of the pre-integration complex at the nuclear pore. *EMBO J.* 26, 3025–3037. doi: 10.1038/sj.emboj.7601740

Bejarano, D. A., Peng, K., Laketa, V., Börner, K., Jost, K. L., Lucic, B., et al. (2019). HIV-1 nuclear import in macrophages is regulated by CPSF6-capsid interactions at the nuclear pore complex. *ELife* 8:e41800. doi: 10.7554/eLife.41800

Besnier, C., Takeuchi, Y., and Towers, G. (2002). Restriction of lentivirus in monkeys. *Proc. Natl. Acad. Sci. U.S.A.* 99, 11920–11925. doi: 10.1073/pnas.172384599

Bharat, T. A. M., Davey, N. E., Ulbrich, P., Riches, J. D., De Marco, A., Rumlova, M., et al. (2012). Structure of the immature retroviral capsid at 8Å resolution by cryo-electron microscopy. *Nature* 487, 385–389. doi: 10.1038/nature11169

Bhattacharya, A., Alam, S. L., Fricke, T., Zadrozny, K., Sedzicki, J., Taylor, A. B., et al. (2014). Structural basis of HIV-1 capsid recognition by PF74 and CPSF6. *Proc. Natl. Acad. Sci. U.S.A.* 111, 18625–18630. doi: 10.1073/pnas.1419945112

Bichel, K., Price, A. J., Schaller, T., Towers, G. J., Freund, S. M. V., and James, L. C. (2013). HIV-1 capsid undergoes coupled binding and isomerization by the nuclear pore protein NUP358. *Retrovirology* 10:81. doi: 10.1186/1742-4690-10-81

Blanco-Rodriguez, G., Gazi, A., Monel, B., Frabetti, S., Scoca, V., Mueller, F., et al. (2020). Remodeling of the core leads HIV-1 preintegration complex into the nucleus of human lymphocytes. *J. Virol.* 94, 1–20. doi: 10.1128/jvi.00135-20

Bosco, D. A., Eisenmesser, E. Z., Pochapsky, S., Sundquist, W. I., and Kern, D. (2002). Catalysis of cis/trans isomerization in native HIV-1 capsid by human cyclophilin A. *Proc. Natl. Acad. Sci. U.S.A.* 99, 5247–5252. doi: 10.1073/pnas.082100499

Braaten, D., Aberham, C., Franke, E. K., Yin, L., Phares, W., and Luban, J. (1996a). Cyclosporine A-resistant human immunodeficiency virus type 1 mutants demonstrate that Gag encodes the functional target of cyclophilin A. *J. Virol.* 70, 5170–5176. doi: 10.1128/jvi.70.8.5170-5176.1996

Braaten, D., Franke, E. K., and Luban, J. (1996b). Cyclophilin A is required for an early step in the life cycle of human immunodeficiency virus type 1 before the initiation of reverse transcription. *J. Virol.* 70, 3551–3560. doi: 10.1128/jvi.70.6.3551-3560.1996

Brass, A. L., Dykxhoorn, D. M., Benita, Y., Yan, N., Engelman, A., Xavier, R. J., et al. (2008). Identification of host proteins required for HIV infection through a functional genomic screen. *Science* 319, 921–926. doi: 10.1126/science.1152725

Briggs, J. A. G., Riches, J. D., Glass, B., Bartonova, V., Zanetti, G., and Kräusslich, H. G. (2009). Structure and assembly of immature HIV. *Proc. Natl. Acad. Sci. U.S.A.* 106, 11090–11095. doi: 10.1073/pnas.0903535106

Briggs, J. A. G., 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. doi: 10.1093/emboj/cdg143

Brown, C. J., Takayama, S., Campen, A. M., Vise, P., Marshall, T. W., Oldfield, C. J., et al. (2002). Evolutionary rate heterogeneity in proteins with long disordered regions. *J. Mol. Evol.* 55, 104–110. doi: 10.1007/s00239-001-2309-6

Buffone, C., Kutzner, J., Opp, S., Martinez-Lopez, A., Selyutina, A., Coggings, S. A., et al. (2019). The ability of SAMHD1 to block HIV-1 but not SIV requires expression of MxB. *Virology* 531, 260–268. doi: 10.1016/j.virol.2019.03.018

Buffone, C., Martinez-Lopez, A., Fricke, T., Opp, S., Severgnini, M., Cifola, I., et al. (2018). Nup153 unlocks the nuclear pore complex for HIV-1 nuclear translocation in nondividing cells. *J. Virol.* 92, 1–29. doi: 10.1128/JVI.00648-18

Bukrinsky, M. I., Sharova, N., Mcdonald, T. L., Pushkarskaya, T., Tarpley, W. G., and Stevenson, M. (1993). Association of integrase, matrix, and reverse transcriptase antigens of human immunodeficiency virus type 1 with viral nucleic acids following acute infection. *Proc. Natl. Acad. Sci. U.S.A.* 90, 6125–6129. doi: 10.1073/pnas.90.13.6125

Burdick, R. C., Delviks-Frankenberry, K. A., Chen, J., Janaka, S. K., Sastri, J., Hu, W. S., et al. (2017). Dynamics and regulation of nuclear import and nuclear movements of HIV-1 complexes. *PLoS Pathog.* 13:e1006570. doi: 10.1371/journal.ppat.1006570

Burdick, R. C., Li, C., Munshi, M., Rawson, J. M. O., Nagashima, K., Hu, W.-S., et al. (2020). HIV-1 uncoats in the nucleus near sites of integration. *Proc. Natl. Acad. Sci. U.S.A.* 117, 5486–5493. doi: 10.1073/pnas.1920631117

Byeon, I. J. L., Meng, X., Jung, J., Zhao, G., Yang, R., Ahn, J., et al. (2009). Structural convergence between Cryo-EM and NMR reveals intersubunit interactions critical for HIV-1 capsid function. *Cell* 139, 780–790. doi: 10.1016/j.cell.2009.10.010

Cardone, G., Purdy, J. G., Cheng, N., Craven, R. C., and Steven, A. C. (2009). Visualization of a missing link in retrovirus capsid assembly. *Nature* 457, 694–698. doi: 10.1038/nature07724

Carnes, S. K., Sheehan, J. H., and Aiken, C. (2018). Inhibitors of the HIV-1 capsid, a target of opportunity. *Curr. Opin. HIV AIDS* 13, 359–365. doi: 10.1097/COH.0000000000000472

Chin, C. R., Perreira, J. M., Savidis, G., Portmann, J. M., Aker, A. M., Feeley, E. M., et al. (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. doi: 10.1016/j.celrep.2015.10.036

Christ, F., Thys, W., De Rijck, J., Gijsbers, R., Albanese, A., Arosio, D., et al. (2008). Transportin-SR2 imports HIV into the nucleus. *Curr. Biol.* 18, 1192–1202. doi: 10.1016/j.cub.2008.07.079

Coffin, J. M., Hughes, S. H., and Varmus, H. E. (1997). *Retroviruses*. Cold Spring Harbor NY: Cold Spring Harbor Laboratory Press.

Cosnefroy, O., Murray, P. J., and Bishop, K. N. (2016). HIV-1 capsid uncoating initiates after the first strand transfer of reverse transcription. *Retrovirology* 13:58. doi: 10.1186/s12977-016-0292-7

Cowan, S., Hatziioannou, T., Cunningham, T., Muesing, M. A., Gottlinger, H. G., and Bieniasz, P. D. (2002). Cellular inhibitors with Fv1-like activity restrict human and simian immunodeficiency virus tropism. *Proc. Natl. Acad. Sci. U.S.A.* 99, 11914–11919. doi: 10.1073/pnas.162299499

Crameri, M., Bauer, M., Caduff, N., Walker, R., Steiner, F., Franzoso, F. D., et al. (2018). MxB is an interferon-induced restriction factor of human herpesviruses. *Nat. Commun.* 9:1980. doi: 10.1038/s41467-018-04379-2

Danielson, C. M., Cianci, G. C., and Hope, T. J. (2012). Recruitment and dynamics of proteasome association with rhTRIM5α cytoplasmic complexes during HIV-1 infection. *Traffic* 13, 1206–1217. doi: 10.1111/j.1600-0854.2012.01381.x

De Iaco, A., Santoni, F., Vannier, A., Guipponi, M., Antonarakis, S., and Luban, J. (2013). TNPO3 protects HIV-1 replication from CPSF6-mediated capsid stabilization in the host cell cytoplasm. *Retrovirology* 10:20. doi: 10.1186/1742-4690-10-20

De Marco, A., Müller, B., Glass, B., Riches, J. D., Kräusslich, H. G., and Briggs, J. A. G. (2010). Structural analysis of HIV-1 maturation using cryo- electron tomography. *PLoS Pathog.* 6:e1001215. doi: 10.1371/journal.ppat.1001215

Dharan, A., Bachmann, N., Talley, S., Zwikelmaier, V., and Campbell, E. M. (2020). Nuclear pore blockade reveals that HIV-1 completes reverse transcription and uncoating in the nucleus. *Nat. Microbiol.* 5, 1088–1095. doi: 10.1038/s41564-020-0735-8

Dharan, A., Opp, S., Abdel-Rahim, O., Keceli, S. K., Imam, S., Diaz-Griffero, F., et al. (2017). Bicaudal D2 facilitates the cytoplasmic trafficking and nuclear import of HIV-1 genomes during infection. *Proc. Natl. Acad. Sci. U.S.A.* 114, E10707–E10716. doi: 10.1073/pnas.1712033114

Dharan, A., Talley, S., Tripathi, A., Mamede, J. I., Majetschak, M., Hope, T. J., et al. (2016). KIF5B and Nup358 cooperatively mediate the nuclear import of HIV-1 during infection. *PLoS Pathog.* 12:e1005700. doi: 10.1371/journal.ppat.1005700

Di Nunzio, F., Danckaert, A., Fricke, T., Perez, P., Fernandez, J., Perret, E., et al. (2012). Human nucleoporins promote HIV-1 docking at the nuclear pore, nuclear import and integration. *PLoS One* 7:e46037. doi: 10.1371/journal.pone.0046037

Di Nunzio, F., Fricke, T., Miccio, A., Valle-Casuso, J. C., Perez, P., Souque, P., et al. (2013). Nup153 and Nup98 bind the HIV-1 core and contribute to the early steps of HIV-1 replication. *Virology* 440, 8–18. doi: 10.1016/j.virol.2013.02.008

Diaz-Griffero, F., Li, X., Javanbakht, H., Song, B., Welikala, S., Stremlau, M., et al. (2006). Rapid turnover and polyubiquitylation of the retroviral restriction factor TRIM5. *Virology* 349, 300–315. doi: 10.1016/j.virol.2005.12.040

Dicks, M. D. J., Betancor, G., Jimenez-Guardeño, J. M., Pessel-Vivares, L., Apolonia, L., Goujon, C., et al. (2018). Multiple components of the nuclear

pore complex interact with the amino-terminus of MX2 to facilitate HIV-1 restriction. *PLoS Pathog.* 14:e1007408. doi: 10.1371/journal.ppat.1007408

Dochi, T., Nakano, T., Inoue, M., Takamune, N., Shoji, S., Sano, K., et al. (2014). Phosphorylation of human immunodeficiency virus type 1 capsid protein at serine 16, required for peptidyl-prolyl isomerase-dependent uncoating, is mediated by virion-incorporated extracellular signal-regulated kinase 2. *J. Gen. Virol.* 95, 1156–1166. doi: 10.1099/vir.0.060053-0

Fassati, A., and Goff, S. P. (2001). Characterization of intracellular reverse transcription complexes of human immunodeficiency virus type 1. *J. Virol.* 75, 3626–3635. doi: 10.1128/jvi.75.8.3626-3635.2001

Fernandez, J., Machado, A. K., Lyonnais, S., Chamontin, C., Gärtner, K., Léger, T., et al. (2019). Transportin-1 binds to the HIV-1 capsid via a nuclear localization signal and triggers uncoating. *Nat. Microbiol.* 4, 1840–1850. doi: 10.1038/s41564-019-0575-6

Fletcher, A. J., Vaysburd, M., Maslen, S., Zeng, J., Skehel, J. M., Towers, G. J., et al. (2018). Trivalent RING assembly on retroviral capsids activates TRIM5 ubiquitination and innate immune signaling. *Cell Host Microbe* 24, 761–775.e6. doi: 10.1016/j.chom.2018.10.007

Forshey, B. M., Shi, J., and Aiken, C. (2005). Structural requirements for recognition of the human immunodeficiency virus type 1 core during host restriction in owl monkey cells. *J. Virol.* 79, 869–875. doi: 10.1128/jvi.79.2.869-875.2005

Forshey, B. M., Von Schwedler, U., Sundquist, W. I., and Aiken, C. (2002). Formation of a human immunodeficiency virus type 1 core of optimal stability is crucial for viral replication. *J. Virol.* 76, 5667–5677. doi: 10.1128/jvi.76.11.5667-5677.2002

Francis, A. C., Marin, M., Prellberg, M. J., Palermino-Rowland, K., and Melikyan, G. B. (2020). HIV-1 uncoating and nuclear import precede the completion of reverse transcription in cell lines and in primary macrophages. *Viruses* 12:1234. doi: 10.3390/v12111234

Francis, A. C., Marin, M., Shi, J., Aiken, C., and Melikyan, G. B. (2016). Time-Resolved imaging of single HIV-1 uncoating in vitro and in living cells. *PLoS Pathog.* 12:e1005709. doi: 10.1371/journal.ppat.1005709

Francis, A. C., and Melikyan, G. B. (2018). Single HIV-1 imaging reveals progression of infection through CA-dependent steps of docking at the nuclear pore, uncoating, and nuclear transport. *Cell Host Microbe* 23, 536–548.e6. doi: 10.1016/j.chom.2018.03.009

Franke, E. K., Yuan, H. E. H., and Luban, J. (1994). Specific incorporation of cyclophilin a into HIV-1 virions. *Nature* 372, 359–362. doi: 10.1038/372359a0

Fricke, T., Valle-Casuso, J. C., White, T. E., Brandariz-Nuñez, A., Bosche, W. J., Reszka, N., et al. (2013). The ability of TNPO3-depleted cells to inhibit HIV-1 infection requires CPSF6. *Retrovirology* 10:46. doi: 10.1186/1742-4690-10-46

Gamble, T. R., Vajdos, F. F., Yoo, S., Worthylake, D. K., Houseweart, M., Sundquist, W. I., et al. (1996). Crystal structure of human cyclophilin A bound to the amino-terminal domain of HIV-1 capsid. *Cell* 87, 1285–1294. doi: 10.1016/S0092-8674(00)81823-1

Gamble, T. R., Yoo, S., Vajdos, F. F., Von Schwedler, U. K., Worthylake, D. K., Wang, H., et al. (1997). Structure of the carboxyl-terminal dimerization domain of the HIV-1 capsid protein. *Science* 278, 849–853. doi: 10.1126/science.278.5339.849

Ganser, B. K., Li, S., Klishko, V. Y., Finch, J. T., and Sundquist, W. I. (1999). Assembly and analysis of conical models for the HIV-1 core. *Science* 283, 80–83. doi: 10.1126/science.283.5398.80

Ganser-Pornillos, B. K., Cheng, A., and Yeager, M. (2007). Structure of full-length HIV-1 CA: a model for the mature capsid lattice. *Cell* 131, 70–79. doi: 10.1016/j.cell.2007.08.018

Ganser-Pornillos, B. K., Von Schwedler, U. K., Stray, K. M., Aiken, C., and Sundquist, W. I. (2004). Assembly properties of the human immunodeficiency virus type 1 CA protein. *J. Virol.* 78, 2545–2552. doi: 10.1128/jvi.78.5.2545-2552.2004

Gao, D., Wu, Y.-T., Aroh, C., Yan, N., Sun, L., Wu, J., et al. (2013). Cyclic GMP-AMP synthase is an innate immune sensor of HIV and other retroviruses - follow up paper couple months later. *Science* 339, 786–791. doi: 10.1126/science.1229963

Gao, S., Von der Malsburg, A., Dick, A., Faelber, K., Schröder, G. F., Haller, O., et al. (2011). Structure of myxovirus resistance protein a reveals intra- and intermolecular domain interactions required for the antiviral function. *Immunity* 35, 514–525. doi: 10.1016/j.immuni.2011.07.012

Gasser, R., Hamoudi, M., Pellicciotta, M., Zhou, Z., Visdeloup, C., Colin, P., et al. (2016). Buffering deleterious polymorphisms in highly constrained parts of HIV-1 envelope by flexible regions. *Retrovirology* 13:50. doi: 10.1186/s12977-016-0285-6

Geller, R., Domingo-Calap, P., Cuevas, J. M., Rossolillo, P., Negroni, M., and Sanjuán, R. (2015). The external domains of the HIV-1 envelope are a mutational cold spot. *Nat. Commun.* 6:8571. doi: 10.1038/ncomms9571

Gilbert, C., Maxfield, D. G., Goodman, S. M., and Feschotte, C. (2009). Parallel germline infiltration of a lentivirus in two malagasy lemurs. *PLoS Genet.* 5:e1000425. doi: 10.1371/journal.pgen.1000425

Gitti, R. K., Lee, B. M., Walker, J., Summers, M. F., Yoo, S., and Sundquist, W. I. (1996). Structure of the amino-terminal core domain of the HIV-1 capsid protein. *Science* 273, 231–235. doi: 10.1126/science.273.5272.231

Goldstone, D. C., Yap, M. W., Robertson, L. E., Haire, L. F., Taylor, W. R., Katzourakis, A., et al. (2010). Structural and functional analysis of prehistoric lentiviruses uncovers an ancient molecular interface. *Cell Host Microbe* 8, 248–259. doi: 10.1016/j.chom.2010.08.006

Goujon, C., Greenbury, R. A., Papaioannou, S., Doyle, T., and Malim, M. H. (2015). A triple-arginine motif in the amino-terminal domain and oligomerization are required for HIV-1 inhibition by human MX2. *J. Virol.* 89, 4676–4680. doi: 10.1128/jvi.00169-15

Goujon, C., Moncorgé, O., Bauby, H., Doyle, T., Ward, C. C., Schaller, T., et al. (2013). Human MX2 is an interferon-induced post-entry inhibitor of HIV-1 infection. *Nature* 502, 559–562. doi: 10.1038/nature12542

Hamoudi, M., Simon-Loriere, E., Gasser, R., and Negroni, M. (2013). Genetic diversity of the highly variable V1 region interferes with human immunodeficiency virus type 1 envelope functionality. *Retrovirology* 10:114. doi: 10.1186/1742-4690-10-114

Hatziioannou, T., Perez-Caballero, D., Cowan, S., and Bieniasz, P. D. (2005). Cyclophilin interactions with incoming human immunodeficiency virus type 1 capsids with opposing effects on infectivity in human cells. *J. Virol.* 79, 176–183. doi: 10.1128/JVI.79.1.176-183.2005

Hatziioannou, T., Perez-Caballero, D., Yang, A., Cowan, S., and Bieniasz, P. D. (2004). Retrovirus resistance factors Ref1 and Lv1 are species-specific variants of TRIM5α. *Proc. Natl. Acad. Sci. U.S.A.* 101, 10774–10779. doi: 10.1073/pnas.0402361101

Henderson, L. E., Bowers, M. A., Sowder, R. C., Serabyn, S. A., Johnson, D. G., Bess, J. W., et al. (1992). Gag proteins of the highly replicative MN strain of human immunodeficiency virus type 1: posttranslational modifications, proteolytic processings, and complete amino acid sequences. *J. Virol.* 66, 1856–1865. doi: 10.1128/jvi.66.4.1856-1865.1992

Henning, M. S., Dubose, B. N., Burse, M. J., Aiken, C., and Yamashita, M. (2014). In vivo functions of CPSF6 for HIV-1 as revealed by HIV-1 capsid evolution in HLA-B27-positive subjects. *PLoS Pathog.* 10:e1003868. doi: 10.1371/journal.ppat.1003868

Hofmann, W., Schubert, D., LaBonte, J., Munson, L., Gibson, S., Scammell, J., et al. (1999). Species-specific, postentry barriers to primate immunodeficiency virus infection. *J. Virol.* 73, 10020–10028. doi: 10.1128/jvi.73.12.10020-10028.1999

Huang, P. T., Summers, B. J., Xu, C., Perilla, J. R., Malikov, V., Naghavi, M. H., et al. (2019). FEZ1 is recruited to a conserved cofactor site on capsid to promote HIV-1 trafficking. *Cell Rep.* 28, 2373–2385.e7. doi: 10.1016/j.celrep.2019.07.079

Hulme, A. E., Kelley, Z., Foley, D., and Hope, T. J. (2015). Complementary assays reveal a low level of CA associated with viral complexes in the nuclei of HIV-1-infected cells. *J. Virol.* 89, 5350–5361. doi: 10.1128/jvi.00476-15

Hulme, A. E., Perez, O., and Hope, T. J. (2011). Complementary assays reveal a relationship between HIV-1 uncoating and reverse transcription. *Proc. Natl. Acad. Sci. U.S.A.* 108, 9975–9980. doi: 10.1073/pnas.1014522108

Hultqvist, G., Åberg, E., Camilloni, C., Sundell, G. N., Andersson, E., Dogan, J., et al. (2017). Emergence and evolution of an interaction between intrinsically disordered proteins. *ELife* 6:e16059. doi: 10.7554/eLife.16059

Hutten, S., Wälde, S., Spillner, C., Hauber, J., and Kehlenbach, R. H. (2009). The nuclear pore component Nup358 promotes transportin-dependent nuclear import. *J. Cell Sci.* 122, 1100–1110. doi: 10.1242/jcs.040154

Imam, S., Talley, S., Nelson, R. S., Dharan, A., O'Connor, C., Hope, T. J., et al. (2016). TRIM5α degradation via autophagy is not required for retroviral restriction. *J. Virol.* 90, 3400–3410. doi: 10.1128/jvi.03033-15

Iwasaki, A. (2012). Innate immune recognition of HIV-1. *Immunity* 37, 389–398. doi: 10.1016/j.immuni.2012.08.011

Jacks, T., Powert, M. D., Masiarzt, F. R., Luciwt, P. A., Barrt, P. J., and Varmus, H. E. (1987). Characterization of ribosomal frameshifting in HIV-1 gag-pol expression. *Nature* 331, 280–283.

Jaguva Vasudevan, A. A., Bähr, A., Grothmann, R., Singer, A., Häussinger, D., Zimmermann, A., et al. (2018). MXB inhibits murine cytomegalovirus. *Virology* 522, 158–167. doi: 10.1016/j.virol.2018.07.017

Jang, S., Cook, N. J., Pye, V. E., Bedwell, G. J., Dudek, A. M., Singh, P. K., et al. (2019). Differential role for phosphorylation in alternative polyadenylation function versus nuclear import of SR-like protein CPSF6. *Nucleic Acids Res.* 47, 4663–4683. doi: 10.1093/nar/gkz206

Jimenez-Guardeño, J. M., Apolonia, L., Betancor, G., and Malim, M. H. (2019). Immunoproteasome activation enables human TRIM5α restriction of HIV-1. *Nat. Microbiol.* 4, 933–940. doi: 10.1038/s41564-019-0402-0

Kane, M., Rebensburg, S. v., Takata, M. A., Zang, T. M., Yamashita, M., Kvaratskhelia, M., et al. (2018). Nuclear pore heterogeneity influences HIV-1 infection and the antiviral activity of MX2. *ELife* 7:e35738. doi: 10.7554/eLife.35738

Kane, M., Yadav, S. S., Bitzegeio, J., Kutluay, S. B., Zang, T., Wilson, S. J., et al. (2013). MX2 is an interferon-induced inhibitor of HIV-1 infection. *Nature* 502, 563–566. doi: 10.1038/nature12653

Kane, M., Zang, T. M., Rihn, S. J., Zhang, F., Kueck, T., Alim, M., et al. (2016). Identification of interferon-stimulated genes with antiretroviral activity. *Cell Host Microbe* 20, 392–405. doi: 10.1016/j.chom.2016.08.005

Kataoka, N., Bachorik, J. L., and Dreyfuss, G. (1999). Transportin-SR, a nuclear import receptor for SR proteins. *J. Cell Biol.* 145, 1145–1152. doi: 10.1083/jcb.145.6.1145

Katzourakis, A., Tristem, M., Pybus, O. G., and Gifford, R. J. (2007). Discovery and analysis of the first endogenous lentivirus. *Proc. Natl. Acad. Sci. U.S.A.* 104, 6261–6265. doi: 10.1073/pnas.0700471104

Keckesova, Z., Ylinen, L. M. J., and Towers, G. J. (2004). The human and African green monkey TRIM5α genes encode Ref1 and Lv1 retroviral restriction factor activities. *Proc. Natl. Acad. Sci. U.S.A.* 101, 10780–10785. doi: 10.1073/pnas.0402474101

Kelich, J. M., Ma, J., Dong, B., Wang, Q., Chin, M., Magura, C. M., et al. (2015). Super-resolution imaging of nuclear import of adeno-associated virus in live cells. *Mol. Ther. Methods Clin. Dev.* 2:15047. doi: 10.1038/mtm.2015.47

Keown, J. R., Black, M. M., Ferron, A., Yap, M., Barnett, M. J., Grant Pearce, F., et al. (2018). A helical LC3-interacting region mediates the interaction between the retroviral restriction factor Trim5 and mammalian autophagy-related ATG8 proteins. *J. Biol. Chem.* 293, 18378–18386. doi: 10.1074/jbc.RA118.004202

Kim, K., Dauphin, A., Komurlu, S., McCauley, S. M., Yurkovetskiy, L., Carbone, C., et al. (2019). Cyclophilin A protects HIV-1 from restriction by human TRIM5α. *Nat. Microbiol.* 4, 2044–2051. doi: 10.1038/s41564-019-0592-5

King, M. C., Raposo, G., and Lemmon, M. A. (2004). Inhibition of nuclear import and cell-cycle progression by mutated forms of the dynamin-like GTPase MxB. *Proc. Natl. Acad. Sci. U.S.A.* 101, 8957–8962. doi: 10.1073/pnas.0403167101

Knockenhauer, K. E., and Schwartz, T. U. (2016). The nuclear pore complex as a flexible and dynamic gate. *Cell* 164, 1162–1171. doi: 10.1016/j.cell.2016.01.034

Knott, G. J., Bond, C. S., and Fox, A. H. (2016). The DBHS proteins SFPQ, NONO and PSPC1: a multipurpose molecular scaffold. *Nucleic Acids Res.* 44, 3989–4004. doi: 10.1093/nar/gkw271

Koh, Y., Wu, X., Ferris, A. L., Matreyek, K. A., Smith, S. J., Lee, K., et al. (2013). Differential effects of human immunodeficiency virus type 1 capsid and cellular factors nucleoporin 153 and LEDGF/p75 on the efficiency and specificity of viral DNA integration. *J. Virol.* 87, 648–658. doi: 10.1128/jvi.01148-12

König, R., Zhou, Y., Elleder, D., Diamond, T. L., Bonamy, G. M. C., Irelan, J. T., et al. (2008). Global analysis of host-pathogen interactions that regulate early-stage HIV-1 replication. *Cell* 135, 49–60. doi: 10.1016/j.cell.2008.07.032

Krishnan, L., Matreyek, K. A., Oztop, I., Lee, K., Tipper, C. H., Li, X., et al. (2010). The requirement for cellular transportin 3 (TNPO3 or TRN-SR2) during infection maps to human immunodeficiency virus type 1 capsid and not integrase. *J. Virol.* 84, 397–406. doi: 10.1128/jvi.01899-09

Kutluay, S. B., Perez-Caballero, D., and Bieniasz, P. D. (2013). Fates of retroviral core components during unrestricted and TRIM5-restricted infection. *PLoS Pathog.* 9:e1003214. doi: 10.1371/journal.ppat.1003214

Labokha, A. A., and Fassati, A. F. (2013). Viruses challenge selectivity barrier of nuclear pores. *Viruses* 5, 2410–2423. doi: 10.3390/v5102410

Lahaye, X., Gentili, M., Silvin, A., Conrad, C., Picard, L., Jouve, M., et al. (2018). NONO detects the nuclear HIV capsid to promote cGAS-mediated innate immune activation. *Cell* 175, 488–501.e22. doi: 10.1016/j.cell.2018.08.062

Lahaye, X., Satoh, T., Gentili, M., Cerboni, S., Conrad, C., Hurbain, I., et al. (2013). The capsids of HIV-1 and HIV-2 determine immune detection of the viral cDNA by the innate sensor cGAS in dendritic cells. *Immunity* 39, 1132–1142. doi: 10.1016/j.immuni.2013.11.002

Lai, M. C., Lin, R. I., Huang, S. Y., Tsai, C. W., and Tarn, W. Y. (2000). A human importin-β family protein, transportin-SR2, interacts with the phosphorylated RS domain of SR proteins. *J. Biol. Chem.* 275, 7950–7957. doi: 10.1074/jbc.275.11.7950

Lanman, J., Lam, T. K. T., Barnes, S., Sakalian, M., Emmett, M. R., Marshall, A. G., et al. (2003). Identification of novel interactions in HIV-1 capsid protein assembly by high-resolution mass spectrometry. *J. Mol. Biol.* 325, 759–772. doi: 10.1016/S0022-2836(02)01245-7

Lanman, J., Lam, T. K. T., Emmett, M. R., Marshall, A. G., Sakalian, M., and Prevelige, P. E. (2004). Key interactions in HIV-1 maturation identified by hydrogen-deuterium exchange. *Nat. Struct. Mol. Biol.* 11, 676–677. doi: 10.1038/nsmb790

Lee, K., Mulky, A., Yuen, W., Martin, T. D., Meyerson, N. R., Choi, L., et al. (2012). HIV-1 capsid-targeting domain of cleavage and polyadenylation specificity factor 6. *J. Virol.* 86, 3851–3860. doi: 10.1128/jvi.06607-11

Lee, K. E., Ambrose, Z., Martin, T. D., Oztop, I., Mulky, A., Julias, J. G., et al. (2010). Flexible use of nuclear import pathways by HIV-1. *Cell Host Microbe* 7, 221–233. doi: 10.1016/j.chom.2010.02.007

Li, S., Hill, C. P., Sundquist, W. I., and Finch, J. T. (2000). Image reconstructions of helical assemblies of the HIV-1 CA protein. *Nature* 407, 409–413. doi: 10.1038/35030177

Li, Y., Kar, A. K., and Sodroski, J. (2009). Target cell type-dependent modulation of human immunodeficiency virus type 1 capsid disassembly by cyclophilin A. *J. Virol.* 83, 10951–10962. doi: 10.1128/jvi.00682-09

Li, Y. L., Chandrasekaran, V., Carter, S. D., Woodward, C. L., Christensen, D. E., Dryden, K. A., et al. (2016). Primate TRIM5 proteins form hexagonal nets on HIV-1 capsids. *ELife* 5:e16269. doi: 10.7554/eLife.16269

Liu, C., Perilla, J. R., Ning, J., Lu, M., Hou, G., Ramalho, R., et al. (2016). Cyclophilin A stabilizes the HIV-1 capsid through a novel non-canonical binding site. *Nat. Commun.* 7:10714. doi: 10.1038/ncomms10714

Liu, Z., Pan, Q., Ding, S., Qian, J., Xu, F., Zhou, J., et al. (2013). The interferon-inducible MxB protein inhibits HIV-1 infection. *Cell Host Microbe* 14, 398–410. doi: 10.1016/j.chom.2013.08.015

Logue, E. C., Taylor, K. T., Goff, P. H., and Landau, N. R. (2011). The cargo-binding domain of transportin 3 is required for lentivirus nuclear import. *J. Virol.* 85, 12950–12961. doi: 10.1128/jvi.05384-11

Lu, M., Hou, G., Zhang, H., Suiter, C. L., Ahn, J., Byeon, I. J. L., et al. (2015). Dynamic allostery governs cyclophilin A-HIV capsid interplay. *Proc. Natl. Acad. Sci. U.S.A.* 112, 14617–14622. doi: 10.1073/pnas.1516920112

Lu, M., Russell, R. W., Bryer, A. J., Quinn, C. M., Hou, G., Zhang, H., et al. (2020). Atomic-resolution structure of HIV-1 capsid tubes by magic-angle spinning NMR. *Nat. Struct. Mol. Biol.* 27, 863–869. doi: 10.1038/s41594-020-0489-2

Luban, J., Bossolt, K. L., Franke, E. K., Kalpana, G. v., and Goff, S. P. (1993). Human immunodeficiency virus type 1 Gag protein binds to cyclophilins A and B. *Cell* 73, 1067–1078. doi: 10.1016/0092-8674(93)90637-6

Lukic, Z., Dharan, A., Fricke, T., Diaz-Griffero, F., and Campbell, E. M. (2014). HIV-1 uncoating is facilitated by dynein and kinesin 1. *J. Virol.* 88, 13613–13625. doi: 10.1128/jvi.02219-14

Lukic, Z., Hausmann, S., Sebastian, S., Rucci, J., Sastri, J., Robia, S. L., et al. (2011). TRIM5α associates with proteasomal subunits in cells while in complex with HIV-1 virions. *Retrovirology* 8:93. doi: 10.1186/1742-4690-8-93

Maertens, G. N., Cook, N. J., Wang, W., Hare, S., Gupta, S. S., Öztop, I., et al. (2014). Structural basis for nuclear import of splicing factors by human Transportin 3. *Proc. Natl. Acad. Sci. U.S.A.* 111, 2728–2733. doi: 10.1073/pnas.1320755111

Mahamid, J., Pfeffer, S., Schaffer, M., Villa, E., Danev, R., Cuellar, L. K., et al. (2016). Visualizing the molecular sociology at the HeLa cell nuclear periphery. *Science* 351, 969–972. doi: 10.1126/science.aad8857

Malfavon-Borja, R., Wu, L. I., Emerman, M., and Malik, H. S. (2013). Birth, decay, and reconstruction of an ancient TRIMCyp gene fusion in primate genomes. *Proc. Natl. Acad. Sci. U.S.A.* 110, E583-E592. doi: 10.1073/pnas.1216542110

Malikov, V., da Silva, E. S., Jovasevic, V., Bennett, G., de Souza Aranha Vieira, D. A., Schulte, B., et al. (2015). HIV-1 capsids bind and exploit the kinesin-1 adaptor FEZ1 for inward movement to the nucleus. *Nat. Commun.* 6:7660. doi: 10.1038/ncomms7660

Mamede, J. I., Cianci, G. C., Anderson, M. R., and Hope, T. J. (2017). Early cytoplasmic uncoating is associated with infectivity of HIV-1. *Proc. Natl. Acad. Sci. U.S.A.* 114, E7169–E7178. doi: 10.1073/pnas.1706245114

Mandell, M. A., Jain, A., Arko-Mensah, J., Chauhan, S., Kimura, T., Dinkins, C., et al. (2014). TRIM proteins regulate autophagy and can target autophagic substrates by direct recognition. *Dev. Cell* 30, 394–409. doi: 10.1016/j.devcel.2014.06.013

Manocheewa, S., Swain, J. V., Lanxon-Cookson, E., Rolland, M., and Mullins, J. I. (2013). Fitness costs of mutations at the HIV-1 capsid hexamerization interface. *PLoS One* 8:e66065. doi: 10.1371/journal.pone.0066065

Marini, B., Kertesz-Farkas, A., Ali, H., Lucic, B., Lisek, K., Manganaro, L., et al. (2015). Nuclear architecture dictates HIV-1 integration site selection. *Nature* 521, 227–231. doi: 10.1038/nature14226

Matreyek, K. A., and Engelman, A. (2011). The requirement for nucleoporin NUP153 during human immunodeficiency virus type 1 infection is determined by the viral capsid. *J. Virol.* 85, 7818–7827. doi: 10.1128/jvi.00325-11

Matreyek, K. A., Wang, W., Serrao, E., Singh, K. P., Levin, H. L., and Engelman, A. (2014). Host and viral determinants for MxB restriction of HIV-1 infection. *Retrovirology* 11:90. doi: 10.1186/s12977-014-0090-z

Mattei, S., Glass, B., Hagen, W. J. H., Kräusslich, H. G., and Briggs, J. A. G. (2016). The structure and flexibility of conical HIV-1 capsids determined within intact virions. *Science* 354, 1434–1437. doi: 10.1126/science.aah4972

McDonald, D., Vodicka, M. A., Lucero, G., Svitkina, T. M., Borisy, G. G., Emerman, M., et al. (2002). Visualization of the intracellular behavior of HIV in living cells. *J. Cell Biol.* 159, 441–452. doi: 10.1083/jcb.200203150

Meehan, A. M., Saenz, D. T., Guevera, R., Morrison, J. H., Peretz, M., Fadel, H. J., et al. (2014). A cyclophilin homology domain-independent role for Nup358 in HIV-1 infection. *PLoS Pathog.* 10:e1003969. doi: 10.1371/journal.ppat.1003969

Merindol, N., El-Far, M., Sylla, M., Masroori, N., Dufour, C., Li, J. X., et al. (2018). HIV-1 capsids from B27/B57+ elite controllers escape Mx2 but are targeted by TRIM5α, leading to the induction of an antiviral state. *PLoS Pathog.* 14:e1007398. doi: 10.1371/journal.ppat.1007398

Miller, M. D., Farnet, C. M., and Bushman, F. D. (1997). Human immunodeficiency virus type 1 preintegration complexes: studies of organization and composition. *J. Virol.* 71, 5382–5390. doi: 10.1128/jvi.71.7.5382-5390.1997

Misumi, S., Inoue, M., Dochi, T., Kishimoto, N., Hasegawa, N., Takamune, N., et al. (2010). Uncoating of human immunodeficiency virus type 1 requires prolyl isomerase Pin1. *J. Biol. Chem.* 285, 25185–25195. doi: 10.1074/jbc.M110.114256

Mu, D., Yang, H., Zhu, J. W., Liu, F. L., Tian, R. R., Zheng, H. Y., et al. (2014). Independent birth of a novel TRIMCyp in tupaia belangeri with a divergent function from its paralog TRIM5. *Mol. Biol. Evol.* 31, 2985–2997. doi: 10.1093/molbev/msu238

Ni, T., Gerard, S., Zhao, G., Dent, K., Ning, J., Zhou, J., et al. (2020). Intrinsic curvature of the HIV-1 CA hexamer underlies capsid topology and interaction with cyclophilin A. *Nat. Struct. Mol. Biol.* 27, 855–862. doi: 10.1038/s41594-020-0467-8

O'Connor, C., Pertel, T., Gray, S., Robia, S. L., Bakowska, J. C., Luban, J., et al. (2010). p62/Sequestosome-1 associates with and sustains the expression of retroviral restriction factor TRIM5α. *J. Virol.* 84, 5997–6006. doi: 10.1128/jvi.02412-09

Ocwieja, K. E., Brady, T. L., Ronen, K., Huegel, A., Roth, S. L., Schaller, T., et al. (2011). HIV integration targeting: a pathway involving transportin-3 and the nuclear pore protein RanBP2. *PLoS Pathog.* 7:e1001313. doi: 10.1371/journal.ppat.1001313

OhAinle, M., Helms, L., Vermeire, J., Roesch, F., Humes, D., Basom, R., et al. (2018). A virus-packageable CRISPR screen identifies host factors mediating interferon inhibition of HIV. *ELife* 7:e39823. doi: 10.7554/eLife.39823

Perez-Caballero, D., Hatziioannou, T., Zhang, F., Cowan, S., and Bieniasz, P. D. (2005). Restriction of human immunodeficiency virus type 1 by TRIM-CypA occurs with rapid kinetics and independently of cytoplasmic bodies, ubiquitin, and proteasome activity. *J. Virol.* 79, 15567–15572. doi: 10.1128/jvi.79.24.15567-15572.2005

Perron, M. J., Stremlau, M., Song, B., Ulm, W., Mulligan, R. C., and Sodroski, J. (2004). TRIM5α mediates the postentry block to N-tropic murine leukemia viruses in human cells. *Proc. Natl. Acad. Sci. U.S.A.* 101, 11827–11832. doi: 10.1073/pnas.0403364101

Pettit, S. C., Lindquist, J. N., Kaplan, A. H., and Swanstrom, R. (2005). Processing sites in the human immunodeficiency virus type 1 (HIV-1) Gag-Pro-Pol precursor are cleaved by the viral protease at different rates. *Retrovirology* 2, 12–16. doi: 10.1186/1742-4690-2-66

Pettit, S. C., Moody, M. D., Wehbie, R. S., Kaplan, A. H., Nantermet, P. v, Klein, C. A., et al. (1994). The p2 domain of human immunodeficiency virus type 1 Gag regulates sequential proteolytic processing and is required to produce fully infectious virions. *J. Virol.* 68, 8017–8027. doi: 10.1128/jvi.68.12.8017-8027.1994

Pornillos, O., Ganser-Pornillos, B. K., Kelly, B. N., Hua, Y., Whitby, F. G., Stout, C. D., et al. (2009). X-Ray structures of the hexameric building block of the HIV capsid. *Cell* 137, 1282–1292. doi: 10.1016/j.cell.2009.04.063

Pornillos, O., Ganser-Pornillos, B. K., and Yeager, M. (2011). Atomic-level modelling of the HIV capsid. *Nature* 469, 424–427. doi: 10.1038/nature09640

Price, A. J., Fletcher, A. J., Schaller, T., Elliott, T., Lee, K. E., KewalRamani, V. N., et al. (2012). CPSF6 defines a conserved capsid interface that modulates HIV-1 replication. *PLoS Pathog.* 8:e1002896. doi: 10.1371/journal.ppat.1002896

Price, A. J., Jacques, D. A., McEwan, W. A., Fletcher, A. J., Essig, S., Chin, J. W., et al. (2014). Host cofactors and pharmacologic ligands share an essential interface in HIV-1 capsid that is lost upon disassembly. *PLoS Pathog.* 10:e1004459. doi: 10.1371/journal.ppat.1004459

Quinn, C. M., Wang, M., Fritz, M. P., Runge, B., Ahn, J., Xu, C., et al. (2018). Dynamic regulation of HIV-1 capsid interaction with the restriction factor TRIM5α identified by magic-angle spinning NMR and molecular dynamics simulations. *Proc. Natl. Acad. Sci. U.S.A.* 115, 11519–11524. doi: 10.1073/pnas.1800796115

Rankovic, S., Varadarajan, J., Ramalho, R., Aiken, C., and Rousso, I. (2017). Reverse transcription mechanically initiates HIV-1 capsid disassembly. *J. Virol.* 91:e00289-17. doi: 10.1128/jvi.00289-17

Rasaiyaah, J., Tan, C. P., Fletcher, A. J., Price, A. J., Blondeau, C., Hilditch, L., et al. (2013). HIV-1 evades innate immune recognition through specific cofactor recruitment. *Nature* 503, 402–405. doi: 10.1038/nature12769

Rasheedi, S., Shun, M. C., Serrao, E., Sowd, G. A., Qian, J., Hao, C., et al. (2016). The Cleavage and polyadenylation specificity factor 6 (CPSF6) subunit of the capsid-recruited pre-messenger RNA cleavage factor I (CFIm) complex mediates HIV-1 integration into genes. *J. Biol. Chem.* 291, 11809–11819. doi: 10.1074/jbc.M116.721647

Reil, H., Kollmus, H., Weidle, U. H., and Hauser, H. (1993). A heptanucleotide sequence mediates ribosomal frameshifting in mammalian cells. *J. Virol.* 67, 5579–5584. doi: 10.1128/jvi.67.9.5579-5584.1993

Rensen, E., Mueller, F., Scoca, V., Parmar, J., Souque, P., Zimmer, C., et al. (2020). Clustering and reverse transcription of HIV-1 genomes in nuclear niches of macrophages. *EMBO J.* 40:e105247. doi: 10.1101/2020.04.12.038067

Reymond, A., Meroni, G., Fantozzi, A., Merla, G., Cairo, S., Luzi, L., et al. (2001). The tripartite motif family identifies cell compartments. *EMBO J.* 20, 2140–2151. doi: 10.1093/emboj/20.9.2140

Richardson, M. W., Guo, L., Xin, F., Yang, X., and Riley, J. L. (2014). Stabilized human TRIM5α protects human T cells from HIV-1 infection. *Mol. Ther.* 22, 1084–1095. doi: 10.1038/mt.2014.52

Rihn, S. J., Wilson, S. J., Loman, N. J., Alim, M., Bakker, S. E., Bhella, D., et al. (2013). Extreme Genetic Fragility of the HIV-1 Capsid. *PLoS Pathog.* 9:e1003461. doi: 10.1371/journal.ppat.1003461

Roa, A., Hayashi, F., Yang, Y., Lienlaf, M., Zhou, J., Shi, J., et al. (2012). RING domain mutations uncouple TRIM5 restriction of HIV-1 from inhibition of reverse transcription and acceleration of uncoating. *J. Virol.* 86, 1717–1727. doi: 10.1128/jvi.05811-11

Rüegsegger, U., Blank, D., and Keller, W. (1998). Human pre-mRNA cleavage factor Im Is related to spliceosomal SR proteins and can be reconstituted *in vitro* from recombinant subunits. *Mol. Cell* 1, 243–253. doi: 10.1016/S1097-2765(00)80025-8

Saito, A., Henning, M. S., Serrao, E., Dubose, B. N., Teng, S., Huang, J., et al. (2016). Capsid-CPSF6 interaction is dispensable for HIV-1 replication in primary cells

but is selected during virus passage in vivo. *J. Virol.* 90, 6918–6935. doi: 10.1128/jvi.00019-16

Sayah, D. M., Sokolskaja, E., Berthoux, L., and Luban, J. (2004). Cyclophilin A retrotransposition into TRIM5 explains owl monkey resistance to HIV-1. *Nature* 430, 569–573. doi: 10.1038/nature02777

Schaller, T., Ocwieja, K. E., Rasaiyaah, J., Price, A. J., Brady, T. L., Roth, S. L., et al. (2011). HIV-1 capsid-cyclophilin interactions determine nuclear import pathway, integration targeting and replication efficiency. *PLoS Pathog.* 7:e1002439. doi: 10.1371/journal.ppat.1002439

Schilling, M., Bulli, L., Weigang, S., Graf, L., Naumann, S., Patzina, C., et al. (2018). Human MxB protein is a pan-herpesvirus restriction factor. *J. Virol.* 92:e01056-18. doi: 10.1128/jvi.01056-18

Schulte, B., Buffone, C., Opp, S., Di Nunzio, F., De Souza Aranha Vieira, D. A., Brandariz-Nuñez, A., et al. (2015). Restriction of HIV-1 requires the N-terminal region of MxB as a capsid-binding Motif but Not as a nuclear localization signal. *J. Virol.* 89, 8599–8610. doi: 10.1128/jvi.00753-15

Schur, F. K. M., Hagen, W. J. H., Rumlová, M., Ruml, T., Müller, B., Kraüsslich, H. G., et al. (2015). Structure of the immature HIV-1 capsid in intact virus particles at 8.8 Å resolution. *Nature* 517, 505–508. doi: 10.1038/nature13838

Schur, F. K. M., Obr, M., Hagen, W. J. H., Wan, W., Jakobi, A. J., Kirkpatrick, J. M., et al. (2016). Assembly and maturation. *Sci. Rep.* 353, 506–508.

Sebastian, S., and Luban, J. (2005). TRIM5α selectively binds a restriction-sensitive retroviral capsid. *Retrovirology* 2:40. doi: 10.1186/1742-4690-2-40

Selyutina, A., Persaud, M., Bulnes-Ramos, A., Buffone, C., Scoca, V., Di Nunzio, F., et al. (2020a). Cyclophilin a prevents HIV-1 restriction in lymphocytes by blocking human TRIM5α binding to the viral core. *Cell Rep.* 30:678037. doi: 10.1101/678037

Selyutina, A., Persaud, M., KewalRamani, V., and Diaz-Griffero, F. (2020b). Nuclear import of the HIV-1 core precedes reverse transcription and uncoating. *bioRxiv* [Preprint]. doi: 10.1101/2020.03.31.018747

Setiawan, L. C., van Dort, K. A., Rits, M. A. N., and Kootstra, N. A. (2016). Mutations in CypA binding region of HIV-1 capsid affect capsid stability and viral replication in primary macrophages. *AIDS Res. Hum. Retroviruses.* 32, 390–398. doi: 10.1089/aid.2014.0361

Shah, V. B., Shi, J., Hout, D. R., Oztop, I., Krishnan, L., Ahn, J., et al. (2013). The host proteins transportin SR2/TNPO3 and cyclophilin a exert opposing effects on HIV-1 uncoating. *J. Virol.* 87, 422–432. doi: 10.1128/jvi.07177-11

Shi, J., and Aiken, C. (2006). Saturation of TRIM5α-mediated restriction of HIV-1 infection depends on the stability of the incoming viral capsid. *Virology* 350, 493–500. doi: 10.1016/j.virol.2006.03.013

Shibata, R., Sakai, H., Kawamura, M., Tokunaga, K., and Adachi, A. (1995). Early replication block of human immunodeficiency virus type 1 in monkey cells. *J. Gen. Virol.* 76, 2723–2730. doi: 10.1099/0022-1317-76-11-2723

Simon-Loriere, E., Galetto, R., Hamoudi, M., Archer, J., Lefeuvre, P., Martin, D. P., et al. (2009). Molecular mechanisms of recombination restriction in the envelope gene of the human immunodeficiency virus. *PLoS Pathog.* 5:e1000418. doi: 10.1371/journal.ppat.1000418

Smaga, S. S., Xu, C., Summers, B. J., Digianantonio, K. M., Perilla, J. R., and Xiong, Y. (2019). MxB restricts HIV-1 by targeting the Tri-hexamer interface of the viral capsid. *Structure* 27, 1234–1245.e5. doi: 10.1016/j.str.2019.04.015

Sowd, G. A., Serrao, E., Wang, H., Wang, W., Fadel, H. J., Poeschla, E. M., et al. (2016). A critical role for alternative polyadenylation factor CPSF6 in targeting HIV-1 integration to transcriptionally active chromatin. *Proc. Natl. Acad. Sci. U.S.A.* 113, E1054–E1063. doi: 10.1073/pnas.1524213113

Strambio-De-Castillia, C., Niepel, M., and Rout, M. P. (2010). The nuclear pore complex: bridging nuclear transport and gene regulation. *Nat. Rev. Mol. Cell Biol.* 11, 490–501. doi: 10.1038/nrm2928

Stremlau, M., Owens, C. M., Perron, M. J., Kiessling, M., Autissier, P., and Sodroski, J. (2004). The cytoplasmic body component TRIM5α restricts HIV-1 infection in Old World monkeys. *Nature* 427, 848–853. doi: 10.1038/nature02343

Stremlau, M., Perron, M., Lee, M., Li, Y., Song, B., Javanbakht, H., et al. (2006). Specific recognition and accelerated uncoating of retroviral capsids by the TRIM5 restriction factor. *Proc. Natl. Acad. Sci. U.S.A.* 103, 5514–5519. doi: 10.1073/pnas.0509996103

Stremlau, M., Perron, M., Welikala, S., and Sodroski, J. (2005). Species-specific variation in the B30.2(SPRY) domain of TRIM5α determines the potency of human immunodeficiency virus restriction. *J. Virol.* 79, 3139–3145. doi: 10.1128/jvi.79.5.3139-3145.2005

Takeuchi, H., Saito, H., Noda, T., Miyamoto, T., Yoshinaga, T., Terahara, K., et al. (2017). Phosphorylation of the HIV-1 capsid by MELK triggers uncoating to promote viral cDNA synthesis. *PLoS Pathog.* 13:e1006441. doi: 10.1371/journal.ppat.1006441

Tang, C., Ndassa, Y., and Summers, M. F. (2002). Structure of the N-terminal 283-residue fragment of the immature HIV-1 Gag polyprotein. *Nat. Struct. Biol.* 9, 537–543. doi: 10.1038/nsb806

Thomas, J. A., Ott, D. E., and Gorelick, R. J. (2007). Efficiency of human immunodeficiency virus type 1 postentry infection processes: evidence against disproportionate numbers of defective virions. *J. Virol.* 81, 4367–4370. doi: 10.1128/jvi.02357-06

Troyer, R. M., McNevin, J., Liu, Y., Zhang, S. C., Krizan, R. W., Abraha, A., et al. (2009). Variable fitness impact of HIV-1 escape mutations to cytotoxic T lymphocyte (CTL) response. *PLoS Pathog.* 5:e100365. doi: 10.1371/journal.ppat.1000365

Valle-Casuso, J. C., Di Nunzio, F., Yang, Y., Reszka, N., Lienlaf, M., Arhel, N., et al. (2012). TNPO3 is required for HIV-1 replication after nuclear import but prior to integration and binds the HIV-1 core. *J. Virol.* 86, 5931–5936. doi: 10.1128/jvi.00451-12

Verhelst, J., Hulpiau, P., and Saelens, X. (2013). Mx proteins: antiviral gatekeepers that restrain the uninvited. *Microbiol. Mol. Biol. Rev.* 77, 551–566. doi: 10.1128/mmbr.00024-13

Visser, J. A. G. M., Hermisson, J., Wagner, G. P., Meyers, L. A., Bagheri-Chaichian, H., Blanchard, J. L., et al. (2003). Perspective: evolution and detection of genetic robustness. *Evolution* 57, 1959–1972. doi: 10.1111/j.0014-3820.2003.tb00377.x

Vollmer, B., Lorenz, M., Moreno-Andrés, D., Bodenhöfer, M., De Magistris, P., Astrinidis, S. A., et al. (2015). Nup153 recruits the Nup107-160 complex to the inner nuclear membrane for interphasic nuclear pore complex assembly. *Dev. Cell* 33, 717–728. doi: 10.1016/j.devcel.2015.04.027

Von Appen, A., Kosinski, J., Sparks, L., Ori, A., DiGuilio, A. L., Vollmer, B., et al. (2015). In situ structural analysis of the human nuclear pore complex. *Nature* 526, 140–143. doi: 10.1038/nature15381

Wagner, A. (2005). Robustness, evolvability, and neutrality. *FEBS Lett.* 579, 1772–1778. doi: 10.1016/j.febslet.2005.01.063

Wagner, J. M., Roganowicz, M. D., Skorupka, K., Alam, S. L., Christensen, D., Doss, G., et al. (2016a). Mechanism of B-box 2 domain-mediated higher-order assembly of the retroviral restriction factor TRIM5α. *ELife* 5:e16309. doi: 10.7554/eLife.16309

Wagner, J. M., Zadrozny, K. K., Chrustowicz, J., Purdy, M. D., Yeager, M., Ganser-Pornillos, B. K., et al. (2016b). Crystal structure of an HIV assembly and maturation switch. *ELife* 5, e17063. doi: 10.7554/eLife.17063

Wright, E. R., Schooler, J. B., Ding, H. J., Kieffer, C., Fillmore, C., Sundquist, W. I., et al. (2007). Electron cryotomography of immature HIV-1 virions reveals the structure of the CA and SP1 Gag shells. *EMBO J.* 26, 2218–2226. doi: 10.1038/sj.emboj.7601664

Wu, X., Anderson, J. L., Campbell, E. M., Joseph, A. M., and Hope, T. J. (2006). Proteasome inhibitors uncouple rhesus TRIM5α restriction of HIV-1 reverse transcription and infection. *Proc. Natl. Acad. Sci. U.S.A.* 103, 7465–7470. doi: 10.1073/pnas.0510483103

Yang, Y., Fricke, T., and Diaz-Griffero, F. (2013). Inhibition of reverse transcriptase activity increases stability of the HIV-1 Core. *J. Virol.* 87, 683–687. doi: 10.1128/jvi.01228-12

Yap, M. W., Nisole, S., Lynch, C., and Stoye, J. P. (2004). Trim5α protein restricts both HIV-1 and murine leukemia virus. *Proc. Natl. Acad. Sci. U.S.A.* 101, 10786–10791. doi: 10.1073/pnas.0402876101

Yap, M. W., Nisole, S., and Stoye, J. P. (2005). A single amino acid change in the SPRY domain of human Trim5α Leads to HIV-1 restriction. *Curr. Biol.* 15, 73–78. doi: 10.1016/j.cub.2004.12.042

Ylinen, L. M. J., Schaller, T., Price, A., Fletcher, A. J., Noursadeghi, M., James, L. C., et al. (2009). Cyclophilin a levels dictate infection efficiency of human immunodeficiency virus type 1 capsid escape mutants A92E and G94D. *J. Virol.* 83, 2044–2047. doi: 10.1128/jvi.01876-08

Yu, A., Skorupka, K. A., Pak, A. J., Ganser-Pornillos, B. K., Pornillos, O., and Voth, G. A. (2020). TRIM5α self-assembly and compartmentalization of the HIV-1 viral capsid. *Nat. Commun.* 11:1307. doi: 10.1038/s41467-020-15106-1

Zhang, R., Mehla, R., and Chauhan, A. (2010). Perturbation of host nuclear membrane component RanBP2 impairs the nuclear import of human

immunodeficiency virus -1 preintegration complex (DNA). *PLoS One* 5:e15620. doi: 10.1371/journal.pone.0015620

Zhao, G., Ke, D., Vu, T., Ahn, J., Shah, V. B., Yang, R., et al. (2011). Rhesus TRIM5α disrupts the HIV-1 capsid at the inter-hexamer interfaces. *PLoS Pathog.* 7:e1002009. doi: 10.1371/journal.ppat.1002009

Zhao, G., Perilla, J. R., Yufenyuy, E. L., Meng, X., Chen, B., Ning, J., et al. (2013). Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics. *Nature* 497, 643–646. doi: 10.1038/nature12162

Zhou, L., Sokolskaja, E., Jolly, C., James, W., Cowley, S. A., and Fassati, A. (2011). Transportin 3 promotes a nuclear maturation step required for efficient HIV-1 integration. *PLoS Pathog.* 7:e1002194. doi: 10.1371/journal.ppat.1002194

Zila, V., Margiotta, E., Turonova, B., Müller, T. G., Zimmerli, C. E., Mattei, S., et al. (2021). Cone-shaped HIV-1 capsids are transported through intact nuclear pores. *Cell* 184, 1032–1046.e1–e8. doi: 10.1016/j.cell2021.01.025

Zurnic Bönisch, I., Dirix, L., Lemmens, V., Borrenberghs, D., De Wit, F., Vernaillen, F., et al. (2020). Capsid-Labelled HIV to investigate the role of capsid during nuclear import and integration. *J. Virol.* 94:e01024-19. doi: 10.1128/JVI.01024-19

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Université de Strasbourg

École Doctorale des Sciences de la Vie et de la Santé STRASBOURG

Elenia TOCCAFONDI

Identification des propriétés phylogénétiques spécifiques des intégrases M et O du VIH-1

Résumé

Les transmissions de virus simiens à l'homme ont donné naissance aux différents groupes de VIH-1. Nous avons récemment identifié un motif fonctionnel (CLA), dans le domaine C-terminal de l'intégrase, essentiel pour l'intégration dans le groupe M. Ici, nous avons constaté que le motif est dispensable pour les isolats du groupe O, en raison de la présence, dans leur domaine N-terminal d'un autre motif spécifique, NOG, qui est mutuellement interchangeable avec le motif CLA. Alors que le motif NOG est déjà hautement conservé chez les ancêtres simiens du groupe O, SIVgor, chez SIVcpz*Ptt*, l'ancêtre du VIH-1 M, aucun consensus pour le motif CLA n'est trouvé, suggérant que le motif a été sélectionné après transmission à l'homme. La caractérisation fonctionnelle des intégrases contenant le motif NOG retrace les voies mécanistes suivies par ces deux virus pour assurer une intégration efficace, améliorant notre compréhension de l'évolution des virus et de leurs protéines multifonctionnelles dans les infections humaines.

Mots-clés : intégrase, intégration, VIH-1, groupe M, groupe O, protéine multifonctionnelle, évolution, SIV

Summary

Transmissions of simian viruses to humans gave rise to the different groups of HIV-1. We recently identified a functional motif (CLA), in the C-terminal domain of integrase, essential for integration in group M. Here, we found that the motif is dispensable for isolates of group O, because of the presence, in their N-terminal domain of another specific motif, NOG, that is mutually interchangeable with the CLA motif. While the NOG motif is highly conserved already in the simian ancestors of group O, SIVgor, in SIVcpz*Ptt*, the ancestor of HIV-1 M, no consensus for the CLA motif is found, suggesting that the motif was selected after transmission to humans. Functional characterization of NOG-motif-containing integrases traces the mechanistic paths followed by these two viruses to ensure efficient integration, improving our understanding of the evolution of viruses and of their multifunctional proteins in human infections.

Keywords: integrase, integration, HIV-1, group M, group O, multifunctional protein, evolution, SIV