Antrag auf RKI-Sonderforschungsmittel für das Jahr 2020 Projektstrang B (Kooperation: MF1, FG18, MF2, FG34)

Beitrag zur Optimierung der HIV-Inzidenzstudie von Transmissionsclusteranalysen

Projekt

Bestimmung der Infektionsdauer von HIV Neudiagnosen mit Hilfe von NGS

1. Angaben zum Antragstellenden

Antragsteller: **Max von Kleist (MF1), Kirsten Hanke (FG18), Karolin Meixenberger (FG18), Norbert Bannert (FG18)**
Kooperationspartner: Barbara Gunsenheimer-Bartmeyer (FG34), Viviane Bremer (FG34), Uwe Koppe (FG34), Bernhard Renard (MF1), Andrea Thürmer (MF2)
Arbeitsschwerpunkt: Bioinformatik Methodenentwicklung (MF1), Retroviren, molekulare HIV-Surveillance, HIV-Diagnostik und HIV-Pathogenese (FG18), HIV-Epidemiologie (FG34), Next Generation Sequencing-Verfahren (MF2)

2. Angaben über das Forschungsvorhaben

Zusammenfassung

Die exakte Datierung von HIV-Infektionen und die Erkennung von akuten HIV Ausbrüchen ist entscheidend für das Verständnis der Virusbiologie, seiner Epidemiologie und insbesondere für die epidemiologische Kontrolle und Eindämmung. Die Befragung von neudiagnostizierten Personen liefert allerdings in den meisten Fällen unzureichende Erkenntnisse, da eine sexuelle HIV-Übertragung mit einem relativ geringen Übertragungsrisiko pro Exposition verbunden ist und die Zeit zwischen Infektion und dem Auftreten erster Symptome mitunter Jahre betragen kann. Molekularbiologisch lassen sich HIV-Proben nur sehr unzuverlässig hinsichtlich der Zeit seit der Infektion (time since infection, TSI) datieren. Zudem zeigen immunbiologische Assays in der Rezenzbestimmung bestimmter Subtypen hohe Fehlerraten. Darüber hinaus ist einer der beiden bisher am RKI eingesetzten Rezenz-Assays seit März 2019 nicht mehr kommerziell erhältlich. Es wird also dringend ein alternatives Verfahren zur Rezenzbestimmung benötigt.

HIV ist ein extrem schnell mutierendes Virus, dessen Übertragung dadurch gekennzeichnet ist, dass in den allermeisten Fällen ein einziges Virus für die Neuinfektion verantwortlich ist. Folglich nimmt im Laufe einer Neuinfektion durch zufällige Mutation und exponentielles Wachstum die virale Diversität stetig zu. Dieser Aspekt kann genutzt werden, um die TSI einer Infektion zu bestimmen: Ein vielversprechende Ansatz zur Bestimmung der TSI basiert daher auf der bioinformatischen Analyse der genetischen Diversität von HIV in NGS-Patientenerstproben und der ‚Rückrechnung' des Infektionszeitpunktes.

In dem beantragten Projekt beabsichtigen wir, ausgehend von eigenen Vorarbeiten und kürzlich publizierten Methoden, eine bioinformatische Methode für die Bestimmung der TSI zu entwickeln, zu validieren und am RKI zu etablieren. Die Modellentwicklung und -validierung erfolgt anhand von Proben der HIV-1 Serokonverterstudie, deren Infektionsdatum bekannt ist. Bei erfolgreicher Validierung bietet das Verfahren die Möglichkeit, Transmissionscluster-Analysen zu optimieren sowie die serologische Rezenztestung von HIV Neudiagnosen, die im Rahmen der Inzidenzstudie durchgeführt wird, zu verbessern oder ganz zu ersetzen und somit Kosten einzusparen.

Hintergrund und Zielsetzung

Nach Angaben von UNAIDS sind aktuell etwa 37 Mio. Menschen mit HIV infiziert, die meisten davon (64%) leben in der Sub-Sahara Region. Im Jahr 2016 infizierten sich 1,8 Mio. Menschen mit dem Virus und 1,0 Mio. Infizierte starben an AIDS [1]. Seit 20 Jahren nimmt die Gesamtzahl der HIV-Neuinfektionen ab und hat sich seitdem beinahe halbiert. In Deutschland leben derzeit geschätzt ca. 86.100 HIV-Positive [2]. Basierend auf aktuellen Schätzungen hat sich jedoch zwischen dem Jahr 2000 und 2006 die Zahl der Neuinfektionen mehr als verdoppelt und verharrt seitdem bei ungefähr 3000 Fällen pro Jahr [3]. Neben der Inzidenzschätzung, wird die Rezenz (Infektion innerhalb der letzten 6 Monate) für alle HIV Neudiagnosen derzeit auch serologisch bestimmt. Allerdings sind diese Daten mit Vorsicht zu genießen, weil eine Inzidenz aus diesen Daten mit den derzeitigen Methoden aufgrund hoher „Falsch Rezenz Raten“ (FRR), schwieriger Nennerdefinitionen und niedriger Fallzahlen nur unzureichend bestimmt werden kann [4]. Die genaue Bestimmung der Anzahl der rezenten Infektionen kann jedoch von entscheidender Bedeutung für die Bewertung neuartiger Präventionsmethoden sein, wie Beispielsweise der Präexpositionsprophylaxe (PrEP), da diese auch mit einem Anstieg von HIV-Testungen einhergeht (= Anstieg von HIV Diagnosen).

Aufgrund sehr hoher Viruslasten kurz nach HIV-Infektion und der Unkenntnis über die eigene Infektion wird HIV meist kurz nach der Infektion übertragen [5]. Unsere eigenen Vorarbeiten haben gezeigt, dass über 50% der HIV-Infektionen bereits nach weniger als 26 Wochen = 180 Tagen weitergegeben werden [6]. Daher ist es aus epidemiologischer Sicht besonders wichtig einen akuten Ausbruch als Solchen zu erkennen, um gegebenenfalls Zielgruppenfokussierte Kontroll- und Präventionsmaßnahmen ergreifen zu können. Zudem ist die Surveillance von HIV Neuinfektionen extrem wertvoll, um das Ausmaß und den Bedarf, aber auch die Effektivität von HIV- Präventionsinterventionen (Rückgang/Zunahme von *rezenten* Infektionen) zu beurteilen [1, 7]. Aus diesem Grund wurden in der Vergangenheit verschiedene serologische Assays entwickelt, um unterscheiden zu können, ob eine Infektion neu erworben wurde oder ob es sich um eine etablierte Infektion handelt [8-14]. Zwei dieser Assays wurden ausführlich am RKI getestet, validiert und eingesetzt. Zu diesen Assays gehört der Sedia Bioscience BED HIV-1 Capture enzyme immunoassay (BED), der 2008 am RKI implementiert wurde. Dieser Test ermöglicht es aufgrund des typischen Anstiegs von HIV-spezifischen IgG-Antikörpern persistierende von frischen Infektionen zu unterscheiden [15]. Hierzu wird ein spezielles Peptid verwendet, welches die immundominante Region des gp41 von drei verschiedenen HIV-1 Subtypen (Subtyp B, D und CRF01_AE, früher Subtyp E genannt) enthält. In einem ELISA wird der relative Anteil an HIV-1-spezifischen IgG-Antikörpern unter den gesamten IgG-Antikörpern im Serum bestimmt. Der BED-ELISA war lange Zeit der einzige kommerzielle, für epidemiologische Studien zugelassene Test für HIV-1 Inzidenz und wurde weltweit verwendet [16-18]. Seit November 2018 ist dieser Assay jedoch nicht mehr kommerziell erhältlich. Zudem zeigte sich aufgrund der Selektivität für spezielle Subtypen, dass nicht-BED-Subtypen, vor allem Subtyp A, häufig nicht richtig klassifiziert werden.

Aus diesen Gründen, wurde ein weiterer Assay, der Biorad Genscreen TM HIV-1/2 Aviditäts-Assay (BRA) am RKI etabliert [19]. Avidität beschreibt die spezifische Bindungsstärke zwischen einem Antikörper und seinem Antigen. Die Bindungsstärke der HIV-spezifischen Antikörper nimmt in den ersten 1-2 Jahren der HIV-Infektion zu. Mit Hilfe des BRA wird die Avidität der HIV-spezifischen Antikörper zu den Antigenen aus dem Testkit (IgM, gp120, gp160) mit und ohne Zugabe eines denaturierenden, chaotropen Salzes (DEA-Diethylamin) gemessen. Die gebildeten Antigen-Antikörper-Komplexe dissoziieren in Gegenwart von Diethylamin, sofern sie nur durch eine geringe Avidität gekennzeichnet sind. Stärker avide Bindungen, als Folge einer bereits länger bestehenden Infektion, lassen sich dadurch nicht beeinflussen und bleiben bestehen. Aus diesem unterschiedlichen Dissoziationsverhalten kann ein Aviditätsindex bestimmt werden und somit inzidente von persistierenden Infektionen unterschieden werden.

Die Vergleichbarkeit der Ergebnisse der beiden Assays war mit 87,3% gut. Für Subtyp B zeigte der BRA mit 9% und 18% eine leicht erhöhte ‚false recent rate' (FRR) und ‚false long-term rate' (FLTR) gegenüber dem BED mit 7% und 14%. Dennoch gab es auffällige Unterschiede in der Klassifizierung innerhalb der Nicht-B-Subtypen. Im BED wurden vor allem Subtyp A und CRF02AG tendenziell häufiger als falsch rezent fehlklassifiziert (31% im BED vs 9% im BRA).

Trotz dieser Optimierungen haben sämtliche serologische Assays Schwächen, da vor allem für Proben im Grenzbereich der Rezenz bereits kleinste Unterschiede im Handling zu einer unterschiedlichen Klassifizierung führen können. Ein naheliegender Ansatz um die Genauigkeit der Rezenzbestimmung zu verbessern, liegt darin mehrere Methoden und Parameter zu kombinieren [4, 20]. Allerdings hat z.B. die Hinzunahme von weiteren Parametern, wie der CD4 Zellzahl oder der Viruslast, zu keiner signifikanten Verbesserung der Präzision der Assays geführt, da diese Parameter großen Schwankungen im Patienten unterliegen können [19]. Andere Mess-Parameter, wie beispielsweise Konzentrationen von antiviralen Medikamenten im Blut, liegen standardmäßig nicht vor. Außerdem lässt die serologische Testung keine graduelle Vorhersage zu, also wie lange genau ein Patient bereits infiziert sein könnte.

Neben serologischen Assays, die im Wesentlichen die Anpassung des humanen Wirts an die Infektion abbilden, könnten sequenzbasierte Verfahren, die die Veränderung des Virus abbilden, eine völlig unabhängige und ergänzende Informationsquelle darstellen. Mehrere kürzlich publizierte Studien haben gezeigt, dass die Sensitivität der Rezenzbestimmung anhand von NGS-Erstproben mit der von immunologischen Assays (e.g. BED) vergleichbar ist [21-23].

Die Vorhersagekraft von sequenzbasierten Verfahren ist in der HIV-Biologie begründet: So ist eine HIV-Neuinfektion in den meisten Fällen durch einen extremen Flaschenhals gekennzeichnet. Die Infektion wird in den allermeisten Fällen, und insbesondere bei sexueller Übertragung, von einem einzigen Virus etabliert [24-26]. Dies bedeutet, dass zum Zeitpunkt der Infektion die virale „Quasi-Species" völlig homogen ist und keine intra-Patienten Diversität vorhanden ist, vgl. Abb. 1B. In der Folge expandiert das Virus exponentiell bevor es nach einigen Wochen seinen „set-point" erreicht (Abb. 1A), wobei weiterhin etwa 10 Billionen Viren pro Tag generiert werden [27]. Während der Virusreplikation kommt es zu zufälligen Mutationen im viralen Genom. Diese entstehen hauptsächlich durch die Ungenauigkeit der viralen reversen Transkriptase [28], sowie durch RNA-editierende Enzyme der infizierten Wirtszellen [29, 30]. Die Mutationsrate beträgt dabei in etwa 2.3×10^{-5} pro Base und reverses Transkriptionsevent [28], was bei einer Genomlänge von etwa 10^5 Nukleotiden bedeutet, dass pro infizierter Zelle im Durchschnitt ca. zwei Mutationen im viralen Genom entstehen. Der weitaus größte Teil der Viruspopulation, der tagtäglich in einer infizierten Person entsteht ist demnach mutiert. Davon ist jedoch ein Großteil nicht lebensfähig, so dass sich nur ein winziger Bruchteil der 10 Billionen Viren vermehrt und damit im NGS detektierbar ist. Diese überlebensfähigen Viren tragen Mutationen, die entweder fitness-neutral sind oder einen selektiven Vorteil haben [31, 32].

Interessant für die Bestimmung des Zeitraums, der seit der Infektion vergangen ist, sind die (fast) fitness-neutralen Mutationen. Da diese völlig zufällig entstehen, kommt es zu einem Anstieg der Diversität über die Zeit (Abb. 1B), abhängig von der Anzahl der Replikationsereignisse und der Mutationswahrscheinlichkeit, bis schließlich ein Maximum erreicht ist, bei dem alle möglichen neutralen Mutationen entstanden sind. Dieser Anstieg der Diversität wird auch als „molekulare Uhr" bezeichnet und ist ein zentraler Bestandteil vieler Phylogeniemethoden. Anders hingegen verhält es sich bei Mutationen die selektiv von Nachteil (sie werden ersetzt), oder von Vorteil sind (sie ersetzen den ursprünglichen Genotyp). Bei selektierten Mutationen kommt es zu einer transienten Divergenzzunahme mit anschließender Abnahme. Daher sind Divergenzmaße bezogen auf selektierte Mutationen weniger tauglich für eine Rezenzbestimmung. Beispiele für selektierte Mutationen sind alle HLA-relevanten Positionen und wirtspezifischen Anpassungen im viralen Genom. Bezogen auf das Virus betrifft das insbesondere die Oberflächenproteine (*Env*). Fitness-

neutral hingegen sind oft die Drittpositionen eines Codons, weil sie die Proteinsequenz nicht verändern. Allerdings ist auch hier Vorsicht geboten, da das Virusgenom auch verschiedene RNA-assoziierte regulatorische Funktionen im viralen Lebenszyklus besitzt, so dass einige synonyme Mutationen nicht fitness-neutral sind. Beispiele hierfür sind der 5'UTR von HIV, inklusive einiger *Gag*-codierender Positionen [33-35]. Es ist daher wichtig zunächst zu bestimmen, welche Positionen (annähernd) fitness-neutral sind.

Zusammenfassend enthält die Untersuchung der Divergenz von diesen fitness-neutralen Mutationen zwei wichtige Informationen: (i) Wie wahrscheinlich ist es, dass die betrachtete Position zufällig mutiert und (ii) Wie viele Replikationen haben stattgefunden?

Ersteres ist hinreichend bekannt [28] und auch für den zweiten Punkt gibt es Indikatoren: die Anzahl der Replikationsereignisse ist annähernd proportional zu dem über die Zeit integrierten Produkt aus Infektionsdauer und Populationsgröße. Die Populationsgröße ist zudem für die Dauer nach der akuten Infektion (ab den ‚set-point') konstant, siehe Abbildung 1A.

Bisher entwickelte bioinformatische Methoden zur Infektionszeitbestimmung funktionieren wie folgt:

I) Bei der ersten Klasse von Methoden wird anhand der jeweiligen Patienten NGS-Probe ein Diversitätsmaß bestimmt (z.B. Nei-Li Divergenz). Anschließend wird anhand aller Patientenproben mit bekanntem Infektionsdatum eine lineare Regression durchgeführt, bei der der Anstieg der Divergenz mit der Infektionsdauer in Bezug gebracht wird. Nun wird der Regressionsparameter auf Daten mit unbekannten Infektionsdatum angewendet: Ist die Divergenz ober- oder unterhalb eines kritischen Werts wird die Probe als rezent, bzw. nicht rezent eingestuft. Diese Verfahren funktionieren also dann, wenn der Anstieg an Diversität zwischen Individuen annähernd identisch ist [22]. Dies ist jedoch nur dann der Fall, wenn die Infektionen jeweils auf ein einziges Virus zurückzuführen sind und bei unterschiedlichen Individuen identische „set-point" Viruslasten erreicht werden. Beides kann sehr starken inter-individuellen Variationen unterliegen (siehe Abbildung 1A), weswegen diese Verfahren verbessert werden können. Zudem ist die Wahl einer linearen Regression mit kritischem Wert eher ungeeignet unter Betrachtung der Divergenzdynamik (Abb. 1B) und aus technischer Sicht gegenüber maschinellen Lernverfahren. Trotz dieser Unzulänglichkeiten wurde in aktuellen Studien eine Sensitivität und Spezifität berichtet, die dem BED-Assay entspricht [21, 22].

II) Bei der zweiten Klasse von Methoden wird gezählt wie häufig die paarweise Hammingdistanz aller Reads aus einer Patientenprobe unter einem Schwellwert liegt (genome similarity index, GSI) [36]. Dieses Divergenzmaß wird in einer Logit-Regression für die Klassifizierung in rezent/nicht rezent benutzt, wobei die inter-patienten Variabilität statistisch modelliert wird. Auch wenn dieses Verfahren dem Erstgenannten methodisch überlegen ist, müsste es an das am RKI verwendete Illumina-NGS Verfahren angepasst werden, da es derzeit ausschließlich für die Pyrosequenzierung (Roche 454; lange reads) optimiert ist, die am RKI nicht mehr durchgeführt wird [36]. Es gilt zu testen, ob der Informationsgehalt (bzgl. des Hammingdistanzmaßes) bei den relativ kurzen NGS-Reads ausreichend ist.

Ein Nachteil aller Methoden ist, dass sie weder immunologische Assays, noch die gemessene Viruslast im Patienten oder Unsicherheiten in den ‚gemessenen' Infektionszeitpunkten miteinbeziehen. Beträgt beispielsweise das Intervall zwischen der letzten Negativ- und der ersten Positivtestung eines Individuums 12 Monate, so wäre der Fehler im berechneten Infektionsdatum (der Mittelwert zwischen Negativ und Positivtestung) in einem Intervall von ±[0-6] Monaten anzusiedeln. Derzeitig verwendete Methoden laufen daher Gefahr an falsch kategorisierten Proben kalibriert zu sein [22, 23, 36].

Es gilt daher an Serokonverter-Daten mit exakt eingrenzbarem Infektionszeitpunkt, oder unter Einbeziehung der Unsicherheiten zu untersuchen, wie gut die oben genannten Methoden die Infektionsdauer bestimmen können. Zweitens gilt es zu klären wie diese Methoden verbessert, bzw. angepasst werden können, z.B. unter Einbeziehung weiterer Parameter (e.g. Viruslast) und Messwerte (e.g. BRA). Nach umfangreicher Prüfung gilt es letztendlich die entwickelten Methoden am RKI zu etablieren.

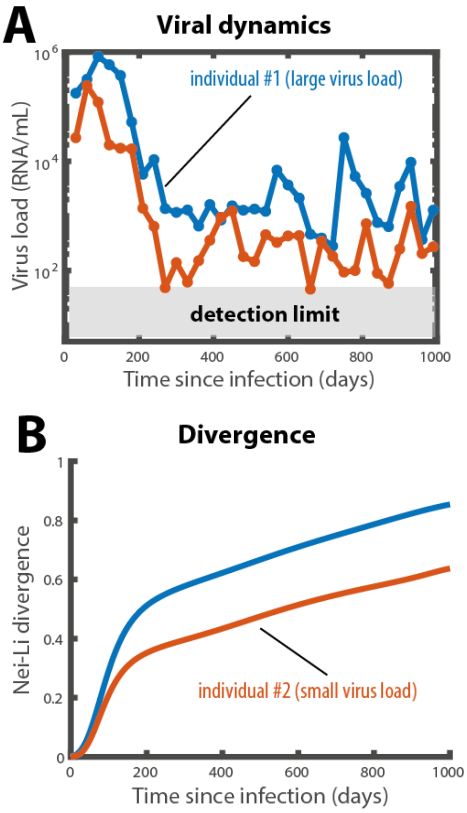

Abbildung 1: Simulierte Virusdynamik nach einem Infektionsevent und daraus resultierende Nei-Li Divergenz in der viralen Quasi-Spezies in zwei unbehandelten Individuen. **A**: Viruslastdynamik in zwei repräsentativen Individuen. Nach einem schnellen Anstieg der Viruslast erreicht dieser einen Höchstwert („peak viral load“) in den ersten 2-6 Wochen nach Infektion, um danach auf einen viralen „set-point“ abzufallen, wo die Viruslast über Jahre auf einem ähnlichen Niveau verbleibt [27]. Während die ***qualitative*** Dynamik (initialer Anstieg, Höchstwert, „set-point“) recht ähnlich in unterschiedlichen Individuen verläuft, kann sich die ***quantitative*** Dynamik stark unterscheiden: In unserem Beispiel hat Individuum #1 (blaue Linie) eine generell höhere Viruslast als Individuum #2 (rote Linie). **B**: Die Gesamtanzahl der Mutationsevents (und damit die Geschwindigkeit der Diversifikation) ist ein Resultat der viralen Dynamik. Zunächst diversifiziert das Virus schnell, was mit den hohen Viruslasten um den Höchstwert („peak viral load“) einhergeht. Danach nimmt die Diversifizierungsgeschwindigkeit ab, sobald das Virus den viralen „set-point“ erreicht. Die Unterschiede in der Virusdynamik von Individuum #1 und Individuum #2 sind in der Diversifizierungsgeschwindigkeit gespiegelt. Simulationen wurden durch Lösen der Viruslast-korrigierten Quasi-Spezies Gleichungen und der anschließenden Berechnung der Nei-Li Divergenz durchgeführt [37].

Einpassung in den Arbeitsschwerpunkt der OEs

FG18 & FG34: Das Vorhaben baut auf den laufenden Projekten zur molekularen Surveillance von HIV-Neudiagnosen (MolSurv_HIV) sowie der HIV-1 Serokonverterstudie auf, erweitert diese und wird es erlauben große NGS-Datensätze zu gewinnen, die für das anvisierte Projekt und darüber hinaus von großem Wert sind. Die genannten Studien gehören zu den Kernaufgaben der Fachgebiete 18 und 34 im Bereich der molekularen und klinischen HIV-Surveillance. Die Arbeiten sind von hoher Public Health Relevanz, da sie Informationen über den Zeitpunkt der Infektion bei einer HIV Neudiagnose liefern (siehe auch Doktorandenantrag zu HIV-Spätdiagnosen). Zudem wird die Erkennung von HIV-Ausbrüchen entscheidend optimiert. Das Vorhaben trägt maßgeblich zum besseren Verständnis der HIV-Übertragung bei: So verbessert die genaue Bestimmung des Infektionszeitpunks entscheidend die Erkennung von Transmissionsclustern und die Inferenz von Transmissionsketten. Dadurch können Risikopersonen und -gruppen besser identifiziert werden und entsprechende Public Health Maßnahmen eingeleitet werden.

MF1: Die Kernaufgabe von MF1 ist die bioinformatische Methodenentwicklungen für Public Health Anwendungen und die anschließende Bereitstellung dieser Methoden für Routineanwendungen am RKI. Das beschriebene Vorhaben sieht die Entwicklung, Validierung und Etablierung bioinformatischer Methoden für die HIV-Rezenzbestimmung vor. Wir erwarten auf dieser Schnittstelle zwischen Bioinformatik, Virologie und Epidemiologie einen sehr großen Synergieeffekt zu schaffen, was auf exzellente Weise in den Aufgabenbereich von MF1 passt. Die neu (Mai 2019) am RKI angesiedelte Forschungsgruppe von Dr. Max von Kleist kann dabei auf einen großen Erfahrungsschatz in der Zusammenarbeit mit FG18/FG34 aufbauen, siehe gemeinsame Vorarbeiten.

MF2: Die Durchführung des Projektes erfordert den Einsatz von NGS-Verfahren und die Expertise der Mitarbeiter des Fachgebietes MF2. Moderne NGS-Verfahren sind einer der Schwerpunkte von MF2.
Gemeinsame Vorarbeiten: Die Antragsteller können auf eine langjährige und äußerst erfolgreiche Zusammenarbeit aufbauen, die seinerzeit als Kooperation zwischen der FU (von Kleist) und dem RKI (FG18/FG34) begann und in mehreren gemeinsamen Projekten im Bereich Inzidenzschätzung [38], Schätzung von Transmissiondynamiken [6, 38, 39] und evolutionären Analysen [39, 40] mündete. Diese Vorarbeiten sind entscheidend für den Erfolg dieses interdisziplinären Vorhabens.

Arbeitsprogramm/vorgesehene Methoden

Das Arbeitsprogramm gliedert sich in vier Arbeitspakete (AP), wobei sich das erste Arbeitspaket mit der Generierung von NGS-Daten für ein Evaluationspanel der HIV-1 Serokonverterstudie befasst, das Zweite mit der Datenanalyse, Paket drei mit der Methodenentwicklung und -Validierung und das letzte Arbeitspaket mit der Anwendung/Etablierung der Methoden mit NGS-Daten aus der molekularen Surveillance von HIV.

Arbeitspaket 1: NGS Sequenzierung des Serokonverter Evaluationspanels

a) **Probenauswahl** (FG18, Epidemiologische Daten FG34): Für die Evaluation der Methode wird auf asservierte Plasmen von Studienpatienten der HIV-1 Serokonverterstudie zurückgegriffen. Die HIV-1 Serokonverterstudie ist eine seit 1997 durchgeführte, prospektive, multizentrische, deutschlandweite Langzeitbeobachtungsstudie, in die Patienten mit bekanntem HIV-1 Infektionszeitpunkt eingeschlossen werden. Bei „akuten“ Serokonvertern (SK) liegt ein reaktiver ELISA oder ein positiver NAT (Nachweis HIV-RNA) bei gleichzeitig unvollständigem Immunoblot vor. Für diese Studienpatienten wird das Datum des ersten reaktiven Tests als Infektionsdatum definiert. Bei „dokumentierten“ SK sind die Daten des letzten negativen und des ersten positiven HIV-Tests bekannt und dokumentiert. Für diese Studienpatienten wird das Infektionsdatum als arithmetisches Mittel zwischen den beiden Test-Zeitpunkten berechnet. Von allen Studienpatienten wird bei Einschluss in die Studie und danach jährlich eine Blutprobe genommen und es werden soziodemographische und klinische Daten erhoben, wie z.B. Geschlecht, Transmissionsweg, Nationalität, Land der Infektion, CD4-Zellzahl und Viruslast. Die Studienpatienten unterzeichnen hierfür eine Einverständniserklärung. Das Studienprotokoll wurde 2005 durch das Ethik-Komitee der Charité genehmigt (EA2/105/05) und 2013 bestätigt [41, 42].

Um die Genauigkeit der berechneten Infektionsdauer zu gewährleisten, werden für die hier geplante Analyse ausschließlich Erst- und Verlaufsproben von akuten Serokonvertern ausgewählt. Um eine entsprechende Vergleichbarkeit der Sequenziertiefe aller Proben sicher zu stellen, werden nur Proben mit Viruslasten ≥10.000 Kopien/ml eingesetzt. Unter Berücksichtigung dieser Kriterien stehen 1144 Proben von therapienaiven Studienpatienten mit Infektionsdauern von 0 bis 2180 Tagen (6 Jahre) zur Verfügung, aus denen das Evaluationspanel zusammengestellt werden soll. Bei der Auswahl soll auf die Verteilung der Infektionsdauern und der HIV-1 Subtypen geachtet werden. Außerdem werden nur Proben verwendet, die phylogenetisch nicht miteinander verwandt sind um ein potentielles Bias durch Clustereffekte zu vermeiden. Von 485 sind zudem bereits serologische Rezenz-Daten (BED und/oder BRA) vorhanden.

b) **Probenbearbeitung** (FG18, Sequenzierung MF2): Aus den ausgewählten Proben wird Virus-RNA isoliert und anschließend mit drei Subtyp-generischen PCRs amplifiziert, sodass die relevantesten Genombereiche (*gag*, *pol*, *env*) abgebildet werden. Anschließend werden die drei PCR-Fragmente aufgereinigt, quantifiziert und equimolar gepooled. Die Präparation für NGS und die Sequenzierung auf dem Illumina MiSeq (2x300bp) erfolgt durch MF2. Für die Auswertung der Varianten und ihrer Anteile pro Position wird eine Sequenziertiefe von 10.000 reads benötigt. Dabei wird auf eine in FG18 etablierte NGS-Pipeline zurückgegriffen (e.g. clipping, alignment, quality trimming). Anhand der finalen BAM/SAM Files können wir mit in der AG von Kleist entwickelten Tools (sam2counts [43]) die relativen Häufigkeiten jedes Nukleotids

an jeder Position im viralen Genom berechnen und zudem Unsicherheiten, ähnlich der in [33, 34] entwickelten Methoden berechnen. Dies eliminiert insbesondere Sequenzier- und PCR-Artefakte. Im nächsten Schritt werden rekursiv alle Positionen entfernt, die resistenz-assoziiert sind oder mit adaptiven Immunantworten in Verbindung gebracht wurden[44]. Die Entfernung ersterer Positionen vermeidet Artefakte konvergenter Evolution im Falle von übertragener Resistenz, während die Entfernung von HLA-assoziierten Positionen Artefakte eliminiert, die durch individuelle Immunantworten entstehen. Zudem wird bei allen Sequenzierungen eine Positivkontrolle mit bekannter Sequenz mitgeführt, um die RT- und Sequenzierfehlerrate zu berechnen und anschließend mathematisch berücksichtigen zu können.

Arbeitspaket 2: Bioinformatische Analyse zu Evolution und Informationsgehalt

Als nächstes testen wir verschiedene Divergenzmaße. Dabei ist es von Vorteil, lediglich solche Positionen zu betrachten, die entweder positiv selektiert werden oder die selektiv neutral sind. Die Beschränkung auf diese Positionen erhöht den Informationsgehalt des später anzuwendenden Divergenzmaßes. Als nächsten Schritt wird die Fitness-Neutralität aller beobachteten Mutationen getested [45], positionsspezifische Entropiemaße bestimmt und diese Größen für verschiedene genomische Regionen verglichen, um später Empfehlungen für die Eignung verschiedener Abschnitte für die Rezenzbestimmung geben zu können. Darauf aufbauend untersuchen wir den Verlauf unterschiedlicher Divergenzmaße nach Serokonversion, wie z.B. Entropie, Nei-Li Divergenz oder Genome similarity index, werden diese hinsichtlich unserer Erwartungen (Abbildung 1B) bewerten und die Vorgehensweise gegebenenfalls weiter anpassen. Zudem werden wir die in der Gruppe etablierten „direct coupling“ analysen (DCA) [43] verwenden, um co-evolvierende Positionen zu identifizieren. Letztere, z.B. der “mittlere coupling Index“ könnten ein Maß der HIV-Kompartimentalisierung sein, welches sich erst nach längerer Infektionsdauer ausprägt. Diese Untersuchungen werden wir verwenden, um verschiedener Genomabschnitte und -Codonpositionen hinsichtlich ihres Informationsgehalts über die Dauer einer Infektion zu bewerten.

Zusätzlich zu den Serokonverter-Daten existieren frei verfügbare NGS-Datensätze mit bekanntem Infektionsdatum: z.B. aus einer longitudinalen, schwedischen Studie [22, 46], einer cross-sektionalen schweizerischen Studie [21] und einer kanadischen Studie [47]. Falls notwendig, werden wir diese Daten zur Vervollständigung unseres Datensatzes miteinbeziehen, um verschiedene Metriken bezüglich Ihres Informationsgehalts zu testen und um unsere Methoden zu validieren.

Arbeitspaket 3: Bioinformatische Methodenentwicklung und Validierung

Um deren Vorhersagekraft an Serokonverterdaten zu überprüfen werden wir die bestehenden Verfahren, wie in den entsprechenden Artikeln [21-23, 36] beschrieben, implementieren und auf die verschiedenen Nukleotidpositionen in *gag*, *pol*, *env*, bzw. *gag-pol*, *pol-env* oder *gag-pol-env* anwenden, sofern diese einen Informationgehalt haben und fitness-neutral sind (siehe AP2). Wir werden die Methoden sowohl bezüglich ihrer qualitativen Vorhersagekraft (Klassifizierung: rezent/nicht-rezent), als auch ihrer quantitativen Vorhersagekraft (Schätzung des Infektionszeitpunktes) bewerten und überprüfen, ob die Hinzunahme von am RKI eingesetzten serologischen Methoden (BRA) die Klassifizierung verbessern kann.

Weiterhin, und wie in Abb. 1 verdeutlicht, werden wir eine eigene Methode entwickeln und testen, die zusätzlich die verfügbaren Viruslastmessungen miteinbezieht und von der wir uns erhoffen, dass insbesondere die Datierung im Grenzbereich zwischen rezent und nicht-rezent deutlich verbessert wird. Diese zu entwickelnde Methode wird die Unsicherheit der Klassifizierung bestimmen, so dass konfirmierende Assays (e.g. BRA) bei unklarer Klassifizierung mit herangezogen werden können.

Abschließend werden wir für das beste Verfahren die subtypenspezifische Vorhersagequalität evaluieren.

Arbeitspaket 4: Anwendung auf die MolSurv_HIV-Daten

Im Rahmen der Molekularen Surveillance von HIV (MolSurv_HIV) werden routinemäßig 1200-1500 Proben/Jahr von HIV-Neudiagnosen aus ganz Deutschland untersucht und sequenziert. Im Gegensatz zu den Patientenproben aus der Serokonverterstudie ist der Infektionszeitpunkt unbekannt und kann auch nicht nachträglich geschätzt werden. Für alle eingehenden Proben wird deshalb ein serologischer Rezenzassay durchgeführt (ca. 2200 Proben/Jahr) um Neuinfektionen von Langzeitinfektionen unterscheiden zu können. Aufgrund der oben genannten Limitationen dieser Assays und des mit ihnen verbundenen Materials- und Personalaufwands, wäre eine Umstellung der Rezenztestung auf ein Sequenz-basiertes Verfahren eine erhebliche Aufwands- und Kostenerleichterung. Aus diesem Grund soll in einer umfassenden Analyse der in Arbeitspaket 3 erstellte Algorithmus überprüft und mit den BED bzw. BRA-Ergebnissen verglichen bzw. kombiniert werden. Das kann sowohl für aktuelle Daten als auch retrospektiv für bereits vorhandene Sequenzen erfolgen.

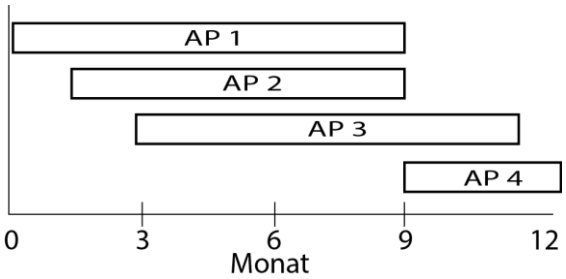

Abbildung 2: Zeitplan für das anvisierte Projekt. Arbeitspakete (AP) sind für einen nahtlosen Übertrag von Daten, Methoden und Erkenntnissen angelegt. Um Raum für konzeptionelle Planung zu schaffen, ist es daher geplant, dass Arbeitspakete sich zeitlich überschneiden.

Zeitplan

Die Durchführung des Projektes ist für den Zeitraum Januar bis Dezember 2020 vorgesehen. Die fortlaufende zeitliche Durchführung der verschiedenen Arbeitspakete für diesen Zeitraum ist in Abbildung 2 dargestellt.

Geplante Fortführung nach Ende der Förderung durch Sonderforschungsmittel

Bei dem Vorhaben handelt es sich um ein Projekt, das die Rezenzbestimmung und Ausbruchsdetektion verbessern soll. Insbesondere werden dazu bestehende Methoden getestet und optimiert. Das Ziel ist es, sequenzbasierte Verfahren am RKI als zusätzliche oder sogar alleinige Informationsquelle zu etablieren. Wir rechnen damit, dass wir im Rahmen der Förderung die Methodenevaluierung, und –optimierung abgeschlossen haben, so dass wir den Mehrwert für deren routinemäßigen Einsatz am RKI einschätzen können. Im Falle eines Mehrwertes könnte die optimierte Methode standartmäßig am RKI in der verstetigten Inzidenzstudie (InzSurv_HIV) und in der molekularen Surveillance von HIV Neudiagnosen (MolSurv_HIV) eingesetzt werden, ohne dass es weiterer Entwicklung bedarf. Das dargelegte Konzept sieht deshalb keine Fortführung der Methodenentwicklung nach Ende der Förderperiode vor.

Risikoabschätzung

In Anbetracht des etablierten Sequenzierprotokolls am RKI besteht ein geringes Risiko, dass die NGS Sequenzierungen von zu geringer Qualität für spätere Auswertungen sind, so dass insgesamt eine Bewertung bestehender Methoden „under-powered“ ist: zum Beispiel weil die PCR-Qualität und Sequenziertiefe unzureichend sind, eine gleichmäßige Probenabdeckung bezüglich der Infektionsdauern nicht gewährleistet ist, oder die NGS-Sequenzen zu spät zur Verfügung stehen. In diesem Fall würden wir, wie in Arbeitspaket AP2 beschrieben, öffentlich verfügbare Datensätze hinzunehmen.

Obwohl mehrere kürzlich erschienenen Veröffentlichungen einen enormen Mehrwert von NGS-basierten Rezenzschätzungen suggerieren [21], besteht ein Restrisiko, dass unsere Analysen ergeben, dass die Renzenzschätzung mit existierenden Methoden nicht verbessert werden kann. Für diesen Fall haben wir eine recht flexible Vorgehensweise in AP3 gewählt, die unter anderem die Methodenoptimierung und -entwicklung vorsieht. Im sehr unwahrscheinlichen Falle eines negativen Resultats „Analyse der NGS-

Sequenzen eines neudiagnostizierten Virus verbessert die Rezenzschätzung nicht“ würde dennoch ein großer Erkenntnisgewinn erreicht werden. Zudem können die generierten NGS Sequenzen für weitere Fragestellungen herangezogen werden. Somit wäre letztlich in jedem Fall ein Mehrwert generiert worden.

3. Zusammenarbeit mit anderen Wissenschaftlern/Wissenschaftlerinnen

Institutsintern: Das beantragte Vorhaben basiert in hohem Maße auf dem Probenmaterial der Studien MolSurv_HIV und der HIV-1 Serokonverterstudie. Beide Studien sind langjährige und erfolgreiche Kooperationsprojekte zwischen FG18 und FG34. Wir möchten darauf hinweisen, dass die Arbeiten des hier zur Förderung vorgeschlagenen Projektes kein Bestandteil der genannten Studien sind. Die Sequenzierungen der Proben in diesen Studien erfolgen unter Einbindung von MF2, das ganz wesentlich zu der Optimierung der Verfahren beiträgt, während MF1 die bioinformatischen Arbeiten durchführen wird. Wir möchten darauf hinweisen, dass das hier angestrebte Vorhaben entscheidende Zuarbeiten für das von FG34 gestellte Promotionsprojekt zu HIV-Spätdiagnosen liefern kann.

Die produktive Zusammenarbeit aller beteiligten Kooperationspartner ist durch mehrere gemeinsame Publikationen in den vergangenen vier Jahren belegt, siehe auch ‚gemeinsame Vorarbeiten‘.

Die vorgesehenen Laborarbeiten erfolgen in FG18 und sollen von einem qualifizierten wissenschaftlichen Mitarbeiter mit molekularbiologischer und serologischer Laborerfahrung und NGS Expertise ausgeführt werden. Für die bioinformatischen Analysen in MF1 steht eine Bioinformatik-Doktorandin aus der Gruppe von Dr. von Kleist, die in das Themengebiet bereits eingearbeitet ist, und in 2020 die Promotion beenden wird zur Verfügung. Die anstehenden Arbeiten werden in enger Abstimmung erfolgen.

4. Ist das Vorhaben bei einer anderen Förderorganisation eingereicht?

Eingereicht: Nein

Zur Beantragung vorgesehen: Nein

5. Beantragte Mittel

Bedarf	Einzelkosten	Anzahl	Gesamtkosten
WA, E13, 50% (FG18)[+]	30.000,-	1	30.000,-
WA, E13, 67% (MF1)[#]	30.000,-	1	30.000,-
Studentische Hilfskraft[§]	6.000,-	1	6.000,-
PCR und NGS-Material*	15.000,-	1	15.000,-
∑	-	-	**81.000,-**

[+]*Wiss. Mitarbeiter E13 (50%), 12 Monate für Laborarbeit in FG18 (Probenauswahl, -sortierung, Amplifizierung etc.)*

[#]*Wiss. Mitarbeiter E13, 67%, 9 Monate für MF1 (Bioinformatische Analysen, Methodenentwicklung, etc.)*

[§]*Die Studentische Hilfskraft (40Std/Mo) wird bei der Datenbearbeitung aushelfen.*

**Ca. 60€/Probe für Extraktion/Amplifizierung/Sequenzierung (5 Zeitabschnitte à 50 Proben pro Abschnitt) = 15,000*

Referenzen

1. UNAIDS. *Global HIV Statistics*. 2018. doi: http://www.unaids.org/sites/default/files/media_asset/UNAIDS_FactSheet_en.pdf
2. Epidemiologisches Bulletin. *HIV-Jahresbericht 2017*. doi: https://www.rki.de/DE/Content/Infekt/EpidBull/Archiv/2018/Ausgaben/47_18.pdf.
3. Epidemiologisches Bulletin. *Schätzung der Zahl der HIV-Neuinfektionen und der Gesamtzahl von Menschen mit HIV in Deutschland*. 47/2018. doi: https://www.rki.de/DE/Content/Infekt/EpidBull/Archiv/2018/Ausgaben/47_18.pdf.
4. UNAIDS/WHO. *When and how to use assays for recent infection to estimate HIV incidence at a population level*. 2011. doi: https://www.who.int/diagnostics_laboratory/hiv_incidence_may13_final.pdf.
5. Volz EM, Ionides E, Romero-Severson EO, Brandt MG, Mokotoff E, Koopman JS. *HIV-1 transmission during early infection in men who have sex with men: a phylodynamic analysis*. PLoS Med. 2013;10(12):e1001568;
6. Pouran Yousef K, Meixenberger K, Smith MR, Somogyi S, Gromoller S, Schmidt D, et al. *Inferring HIV-1 Transmission Dynamics in Germany From Recently Transmitted Viruses. Journal of acquired immune deficiency syndromes*. 2016;73(3):356-63.
7. ECDC. *HIV/AIDS Surveillance in Europe 2018*. https://ecdceuropaeu/en/publications-data/hivaids-surveillance-europe-2018-2017-data (accessed 05/12/2018). 2018.

8. Barin F, Meyer L, Lancar R, Deveau C, Gharib M, Laporte A, et al. *Development and validation of an immunoassay for identification of recent human immunodeficiency virus type 1 infections and its use on dried serum spots*. Journal of clinical microbiology. 2005;43(9):4441-7.
9. Duong YT, Qiu M, De AK, Jackson K, Dobbs T, Kim AA, et al. *Detection of recent HIV-1 infection using a new limiting-antigen avidity assay: potential for HIV-1 incidence estimates and avidity maturation studies*. PloS one. 2012;7(3):e33328.
10. Duong YT, Mavengere Y, Patel H, Moore C, Manjengwa J, Sibandze D, et al. *Poor performance of the determine HIV-1/2 Ag/Ab combo fourth-generation rapid test for detection of acute infections in a National Household Survey in Swaziland*. Journal of clinical microbiology. 2014;52(10):3743-8.
11. Janssen RS, Satten GA, Stramer SL, Rawal BD, O'Brien TR, Weiblen BJ, et al. *New testing strategy to detect early HIV-1 infection for use in incidence estimates and for clinical and prevention purposes*. JAMA : the journal of the American Medical Association. 1998;280(1):42-8.
12. Keating SM, Hanson D, Lebedeva M, Laeyendecker O, Ali-Napo NL, Owen SM, et al. *Lower-sensitivity and avidity modifications of the vitros anti-HIV 1+2 assay for detection of recent HIV infections and incidence estimation*. Journal of clinical microbiology. 2012;50(12):3968-76.
13. Hanson DL, Song R, Masciotra S, Hernandez A, Dobbs TL, Parekh BS, et al. *Mean Recency Period for Estimation of HIV-1 Incidence with the BED-Capture EIA and Bio-Rad Avidity in Persons Diagnosed in the United States with Subtype B*. PloS one. 2016;11(4):e0152327.
14. Suligoi B, Galli C, Massi M, et al. *Precision and accuracy of a procedure for detecting recent human immunodeficiency virus infections by calculating the antibody avidity index by an automated immunoassay-based method*. Journal of clinical microbiology. 2002;40(11):4015-20.
15. Hauser A, Santos-Hoevener C, Meixenberger K, Zimmermann R, Somogyi S, Fiedler S, et al. *Improved testing of recent HIV-1 infections with the BioRad avidity assay compared to the limiting antigen avidity assay and BED Capture enzyme immunoassay: evaluation using reference sample panels from the German Seroconverter Cohort*. PloS one. 2014;9(6):e98038.
16. Hofmann A, Hauser A, Zimmermann R, Santos-Hovener C, Batzing-Feigenbaum J, Wildner S, et al. *Surveillance of recent HIV infections among newly diagnosed HIV cases in Germany between 2008 and 2014*. BMC Infect Dis. 2017;17(1):484.
17. Scheer S, Chin CS, Buckman A, McFarland W. *Estimation of HIV incidence in San Francisco*. Aids. 2009;23(4):533-4.
18. Kim S, Lee JH, Choi JY, Kim JM, Kim HS. *False-positive rate of a "fourth-generation" HIV antigen/antibody combination assay in an area of low HIV prevalence. Clinical and vaccine immunology* : CVI. 2010;17(10):1642-4.
19. Hauser A, Heiden MA, Meixenberger K, Han O, Fiedler S, Hanke K, et al. *Evaluation of a BioRad Avidity assay for identification of recent HIV-1 infections using dried serum or plasma spots*. Journal of virological methods. 2019;266:114-20.
20. Verhofstede C, Fransen K, Van Den Heuvel A, Van Laethem K, Ruelle J, Vancutsem E, et al. *Decision tree for accurate infection timing in individuals newly diagnosed with HIV-1 infection*. BMC Infect Dis. 2017;17(1):738.
21. Carlisle LA, Turk T, Kusejko K, Metzner KJ, Leemann C, Schenkel C, et al. *Viral diversity from next-generation sequencing of HIV-1 samples provides precise estimates of infection recency and time since infection*. The Journal of infectious diseases. 2019.
22. Puller V, et al. *Estimating time of HIV-1 infection from next-generation sequence diversity*. PLoS Comput Biol. 2017;13(10):e1005775.
23. Park SY, Love TMT, Reynell L, Yu C, Kang TM, Anastos K, et al. *The HIV Genomic Incidence Assay Meets False Recency Rate and Mean Duration of Recency Infection Performance Standards*. Sci Rep. 2017;7(1):7480.
24. Keele BF, Giorgi EE, Salazar-Gonzalez JF, et al. *Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection*. Proceedings of the National Academy of Sciences of the United States of America. 2008;105(21):7552-7.
25. Abrahams MR, Anderson JA, Giorgi EE, Seoighe C, Mlisana K, Ping LH, et al. *Quantitating the multiplicity of infection with human immunodeficiency virus type 1 subtype C reveals a non-poisson distribution of transmitted variants*. Journal of virology. 2009;83(8):3556-67.
26. Fischer W, Ganusov VV, Giorgi EE, Hraber PT, Keele BF, Leitner T, et al. *Transmission of single HIV-1 genomes and dynamics of early immune escape revealed by ultra-deep sequencing*. PloS one. 2010;5(8):e12303.
27. Simon V, Ho DD. *HIV-1 dynamics in vivo: implications for therapy*. Nat Rev Microbiol. 2003;1(3):181-90.
28. Mansky LM, Temin HM. *Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase*. Journal of virology. 1995;69(8):5087-94.
29. Suspene R, Petit V, Puyraimond-Zemmour D, et al. *Double-stranded RNA adenosine deaminase ADAR-1-induced hypermutated genomes among inactivated seasonal influenza and live attenuated measles virus vaccines*. Journal of virology. 2011;85(5):2458-62.
30. Lehmann KA, Bass BL. *Double-stranded RNA adenosine deaminases ADAR1 and ADAR2 have overlapping specificities*. Biochemistry. 2000;39(42):12875-84.
31. Traulsen A, Iwasa Y, Nowak MA. *The fastest evolutionary trajectory*. J Theor Biol. 2007;249(3):617-23.
32. Gokhale CS, Iwasa Y, Nowak MA, Traulsen A. *The pace of evolution across fitness valleys*. J Theor Biol. 2009;259(3):613-20.
33. Smyth RP, Smith MR, Jousset AC, Despons L, Laumond G, Decoville T, et al. *In cell mutational interference mapping experiment (in cell MIME) identifies the 5' polyadenylation signal as a dual regulator of HIV-1 genomic RNA production and packaging*. Nucleic acids research. 2018;46(9):e57.
34. Smyth RP, Despons L, Huili G, Bernacchi S, Hijnen M, Mak J, et al. *Mutational interference mapping experiment (MIME) for studying RNA structure and function*. Nat Methods. 2015;12(9):866-72.
35. Mortimer SA, Kidwell MA, Doudna JA. *Insights into RNA structure and function from genome-wide studies*. Nat Rev Genet. 2014;15(7):469-79.
36. Park SY, et al.. *Developing high-throughput HIV incidence assay with pyrosequencing platform*. Journal of virology. 2014;88(5):2977-90.
37. Nei M, Li WH. *Mathematical model for studying genetic variation in terms of restriction endonucleases*. Proceedings of the National Academy of Sciences of the United States of America. 1979;76(10):5269-73.
38. Meixenberger K, Hauser A, Jansen K, Yousef KP, Fiedler S, von Kleist M, et al. *Assessment of ambiguous base calls in HIV-1 pol population sequences as a biomarker for identification of recent infections in HIV-1 incidence studies*. Journal of clinical microbiology. 2014;52(8):2977-83.
39. Hanke K, Faria NR, Kuhnert D, Pouran Yousef K, Hauser A, Meixenberger K, et al. *Reconstruction of the genetic history and the current spread of HIV-1 subtype A in Germany*. Journal of virology. 2019.
40. Meixenberger K, Yousef KP, Smith MR, Somogyi S, Fiedler S, Bartmeyer B, et al. *Molecular evolution of HIV-1 integrase during the 20 years prior to the first approval of integrase inhibitors*. Virol J. 2017;14(1):223.
41. Zu Knyphausen F, Scheufele R, Kucherer C, et al. *First line treatment response in patients with transmitted HIV drug resistance and well defined time point of HIV infection: updated results from the German HIV-1 seroconverter study*. PLoS One. 2014;9(5):e95956.
42. Machnowska P, Meixenberger K, Schmidt D, Jessen H, Hillenbrand H, Gunsenheimer-Bartmeyer B, et al. *Prevalence and persistence of transmitted drug resistance mutations in the German HIV-1 Seroconverter Study Cohort*. PloS one. 2019;14(1):e0209605.
43. Smith M. *Methods to detect Evolutionary Constraints*: Application to HIV, PhD Thesis, FU-Berlin. 2019
44. Wensing AM, Calvez V, Gunthard HF, et al. 2017 *Update of the Drug Resistance Mutations in HIV-1*. Top Antivir Med. 2017;24(4):132-3.
45. Fu YX, Li WH. *Statistical tests of neutrality of mutations*. Genetics. 1993;133(3):693-709.
46. Zanini F, Brodin J, Thebo L, Lanz C, Bratt G, Albert J, et al. Population genomics of intrapatient HIV-1 evolution. Elife. 2015;4.
47. Kafando A, Fournier E, Serhir B, Martineau C, Doualla-Bell F, Sangare MN, et al. *HIV-1 envelope sequence-based diversity measures for identifying recent infections*. PloS one. 2017;12(12):e0189999.