

Entwicklung einer Software zur Identifizierung
neuartiger und bekannter Infektionserreger in klinischen Proben

Dissertation zur Erlangung des Doktorgrades
an der Fakultät für Mathematik, Informatik und Naturwissenschaften
Fachbereich Biologie
der Universität Hamburg
vorgelegt von Malik Alawi
Hamburg, 2020

Vorsitzender der Prüfungskommission

Dr. PD Andreas Pommerening-Röser

Gutachter

Professor Dr. Adam Grundhoff

Professor Dr. Stefan Kurtz

Datum der Disputation

30. April 2021

Abstract

Sequencing of diagnostic samples is widely considered a key technology that may fundamentally improve infectious disease diagnostics. The approach can not only identify pathogens already known to cause a specific disease, but may also detect pathogens that have not been previously attributed to this disease, as well as completely new, previously unknown pathogens. Therefore, it may significantly increase the level of preparedness for future outbreaks of emerging pathogens.

This study describes the development and application of methods for the identification of pathogenic agents in diagnostic samples. The methods have been successfully applied multiple times under clinical conditions. The corresponding results have been published within the scope of this thesis. Finally, the methods were made available to the scientific community as an open source bioinformatics tool.

The novel software was validated by conventional diagnostic methods and it was compared to established analysis pipelines using authentic clinical samples. It is able to identify pathogens from different diagnostic entities and often classifies viral agents down to strain level. Furthermore, the method is capable of assembling complete viral genomes, even from samples containing multiple closely related viral strains of the same viral family.

In addition to an improved method for taxonomic classification, the software offers functionality which is not present in established analysis pipelines. It is, for example, able to annotate protein domains and it performs the classification of sequences based on these annotations. The conserved, functional domains provide an additional level of evidence for the presence of putatively pathogenic agents and they may aid especially the detection of novel pathogens.

Asides from the analysis of individual samples, the software can perform cohort-based analyses. In this mode cross-sample comparisons are carried out to identify sequence signatures which are overrepresented in a group of samples in comparison to a control group. This approach neither requires previous taxonomic classification nor sequence homology searches in external databases and thus enables the detection of truly novel pathogenic agents.

Zusammenfassung

Die Sequenzierung von diagnostischen Proben gilt als eine Schlüsseltechnologie, welche die Diagnostik von Infektionskrankheiten grundlegend verbessern kann. Mit diesem Ansatz können nicht nur Infektionserreger identifiziert werden, von denen bereits bekannt ist, dass sie mit einer bestimmten Krankheit assoziiert sind, sondern auch solche, welche bisher nicht mit dieser Krankheit in Verbindung gebracht wurden. Zudem ist der Ansatz geeignet vollständig unbekannte Infektionserreger zu identifizieren. Er kann daher dazu beitragen, zukünftigen Ausbrüchen neuartiger Infektionserreger besser vorbereitet zu begegnen.

Diese Studie beschreibt die Entwicklung und Anwendung von Methoden für die Identifizierung von Infektionserregern in diagnostischen Proben. Die Methoden wurden mehrfach erfolgreich unter klinischen Bedingungen eingesetzt und die entsprechenden Ergebnisse wurden im Rahmen dieser Doktorarbeit publiziert. Die entwickelten Methoden stehen der wissenschaftlichen Gemeinschaft in Form einer quelloffenen Software zur Verfügung. Unter Verwendung klinischer Proben wurde die neue Software mit Methoden konventioneller Diagnostik validiert und mit etablierten bioinformatischen Analyse-Pipelines verglichen. Sie detektiert Infektionserreger aus verschiedenen diagnostischen Entitäten zuverlässig und klassifiziert virale Erreger oft bis zur Ebene des Stammes. Darüber hinaus ist die Software in der Lage virale Genome vollständig zu rekonstruieren. Dies gelingt sogar in Proben, welche mit mehreren nah verwandten Stämmen infiziert sind.

Zusätzlich zu einer verbesserten Methode der taxonomische Klassifikation, bietet die neue Software auch Funktionen, welche in etablierten Analyse-Pipelines nicht vorhanden sind. Sie ist beispielsweise in der Lage, Proteindomänen zu annotieren und sie kann Sequenzen anhand dieser Annotationen klassifizieren. Als konservierte, funktionale Einheiten, können Proteindomänen zusätzliche Evidenz für das Vorhandensein möglicher Infektionserreger bieten. Sie können insbesondere dabei helfen, neuartige Infektionserreger zu identifizieren.

Neben der Auswertung einzelner Proben kann die Software auch kohortenbasierte Analysen durchführen. In diesem Modus werden probenübergreifende Vergleiche durchgeführt, um Sequenzsignaturen zu identifizieren welche in einer Gruppe von Proben im Vergleich zu einer Kontrollgruppe überrepräsentiert sind. Dieser Ansatz erfordert weder eine vorangegangene taxonomische Zuordnung, noch Homologie zu bereits beschriebenen Sequenzen. Er ermöglicht somit den Nachweis gänzlich neuartiger Infektionserreger.

Inhaltsverzeichnis

1	Einleitung	1
1.1	Molekularbiologische Methoden zum Nachweis von Infektionserregern in klinischen Proben	1
1.2	Amplikon-basierte Sequenzanalysen	2
1.2.1	Nachweis und Quantifizierung prokaryotischer und eukaryotischer Infektionserreger mittels ribosomaler RNA	2
	Detektion viraler Infektionserreger mittels Multiplex-PCR	5
1.3	Metagenom und Metatranskriptom Analysen	6
1.3.1	Taxonomischen Klassifikation kurzer Sequenzen	6
	Alignment-basierte Analysen	6
	Alignment-freie Analysen	8
	Taxonomische Klassifikation von Contigs und Scaffolds	10
	Zielsetzung	12
2	Im Rahmen der Promotion entstandene Publikationen	15
3	Diskussion	21
3.1	Auswertung klinischer Proben mit DAMIAN	22
3.1.1	Module der Auswertung	23
	Verifikation der Eingabedaten	23
	Qualitätsprüfung und -kontrolle	23
	Digitale Subtraktion und approximative Quantifizierung	25
	Assemblierung und Erfassung der Eigenschaften von Contigs	26
	Funktionelle Annotation und Ranking von Contigs	26
	Taxonomische Zuordnung	26
	Erstellung von Ergebnisberichten	27
	Kohortenanalyse	28
3.1.2	Integrierte Funktionalität zum Verteilten Rechnen	30

Inhaltsverzeichnis

3.2	Ausblick	31
3.2.1	Integrierte Software	31
3.2.2	Gesonderte Prozessierung von längeren Reads	32
3.2.3	Benutzungsschnittstelle und Ergebnisberichte	33
3.2.4	Klassifikation anhand von 16S-, ITS- oder anderen Marker-Genen	33
3.3	Auswertung eines COVID-19 Datensatzes	34
	Literatur	39
4	Appendix	45

1 Einleitung

Ziel der vorliegenden Arbeit ist die Entwicklung von Methoden zur Detektion von Infektionserregern in klinischen Proben. Die entwickelten Methoden wurden implementiert und der Öffentlichkeit als quelloffene Software zur freien Nutzung und Weiterentwicklung zur Verfügung gestellt [1].

Im Folgenden werden zunächst etablierte Methoden zur Detektion von Infektionserregern betrachtet. Der Fokus wird dabei auf Methoden gelegt, welche auf der Sequenzierung von Nukleinsäuren beruhen, denn zu ihnen zählen auch die in dieser Arbeit entwickelten Methoden. Am Ende der Einleitung wird die Zielsetzung dieser Arbeit im Kontext der bestehenden Ansätze erläutert und es wird beschrieben, welche zusätzlichen Anforderungen eine Software erfüllen muss.

1.1 Molekularbiologische Methoden zum Nachweis von Infektionserregern in klinischen Proben

In der klinischen Diagnostik von Infektionskrankheiten wurden kulturbasierte Nachweisverfahren weitgehend von molekularbiologischen Methoden abgelöst [2].

Insbesondere auf Nukleinsäuren basierte molekularbiologische Methoden haben in der mikrobiologischen Diagnostik einen hohen Stellenwert eingenommen, da sie sensitiv, schnell und kostengünstig sind. Meist basieren sie auf der gezielte Amplifikation bestimmter Moleküle und sie setzen keine Anzucht der Infektionserregers voraus. Dies ist ein erheblicher Vorteil gegenüber anderen, ebenfalls sensitiven, diagnostischen Verfahren, zumal sich nicht alle Erreger anzüchten lassen.

Wie auch mit klassische diagnostische Methoden, lässt sich mit auf Nukleinsäuren basierten molekularbiologischen Methoden sensitiv und spezifisch prüfen, ob ein bestimmter Infektionserreger in einer Probe vorhanden ist oder nicht. Es ist also vorab zu entscheiden, auf welche Infektionserreger eine Probe zu untersuchen ist. Daher bleiben diese

1 Einleitung

Methoden auf solche Fälle beschränkt, bei denen aus der Klinik eines Patienten eindeutige Rückschlüsse auf die mögliche Anwesenheit bestimmter Infektionserreger abgeleitet werden können. Hinzu kommt, dass bei häufig vorkommende Syndromen, wie Pneumonie, Sepsis und Enzephalitis, verschiedene Pathogene kaum unterscheidbare Symptome verursachen [3]. Diesem Umstand kann dadurch Rechnung getragen werden, dass größere diagnostische Arrays für häufig auftretende Pathogene verwendet werden. Der inhärente Bias der molekularbiologischen Methode bleibt dabei aber grundsätzlich weiter bestehen. Sollte im ersten Untersuchungsschritt kein Infektionserreger detektiert werden, so können zusätzlich aufwändige Nachuntersuchungen notwendig werden.

1.2 Amplikon-basierte Sequenzanalysen

Zusammen mit den molekularbiologischen Methoden werden zunehmend auch bioinformatische Methoden eingesetzt. Hierzu werden oftmals kurze hypervariable Regionen in konservierten Genen zunächst amplifiziert und dann sequenziert, um mikrobielle Organismen in einer Probe zu detektieren und deren relative Abundanz zu ermitteln

Gut etabliert sind Methoden welche die ribosomale RNA (rRNA) untersuchen. Für die Analyse von Prokaryoten wird insbesondere die 16S rRNA verwendet. Eukaryoten können anhand von 18S und Internal Transcribed Spacer (ITS) rRNA untersucht werden. Ein Vorteil dieser Methoden ist, dass sie prinzipiell nicht auf bestimmte Bakterien, Archebakterien oder Pilze beschränkt sind.

Anders als bei den Pro- und Eukaryoten verhält es sich bei Viren. Ein Analogon zur rRNA gibt es bei ihnen nicht. Konservierte genomische Regionen, welche für einen vergleichbaren Ansatz geeignet wären, existieren bei ihnen höchstens innerhalb viraler Familien. Die Anzahl an Viren, die gleichzeitig mit einer Amplikon-basierter Methode detektiert werden kann, ist deshalb vergleichsweise gering.

1.2.1 Nachweis und Quantifizierung prokaryotischer und eukaryotischer Infektionserreger mittels ribosomaler RNA

Das prokaryotische 16S rRNA Gen umfasst ungefähr 1 500 bp. Die darin enthaltenen neun variablen Regionen sind jeweils flankiert von konservierten Regionen. Mit Primern, welche in den konservierten Regionen binden, können entsprechend die variablen Regionen amplifiziert werden [4]. Die dadurch entstehenden Amplikons werden nachfolgend

sequenziert. 16S rRNA wird verwendet, um bakterielle und archaeobakterielle Metagenome zu untersuchen. 18S und ITS rRNA eignet sich hingegen zur Identifizierung pilzlicher Organismen in metagenomischen Proben.

Methoden zur Amplifikation und Sequenzierung kurzer (< 500 bp) ribosomaler Regionen sind gut etabliert, kostengünstig, sowie schnell und einfach in der Anwendung. Die Auswertung der dabei entstehenden Daten ist weitgehend standardisiert und wenig rechenintensiv, so dass sie vollautomatisch auf den Rechnern ausgeführt werden kann, welche in Sequenzierplattformen integriert sind.

Nicht immer ist es mit diesen Methoden möglich, Organismen niedrigen taxonomischen Ebenen eindeutig zuzuordnen. Die Sequenz eines Amplikons kann beispielsweise ein Substring der 16S rRNA Sequenz von mehreren Spezies sein. In solchen Fällen ist eine Zuordnung nur auf derjenigen taxonomischen Ebene möglich, welche allen diesen Spezies gemein ist. Diese Ebene wird als 'Lowest Common Ancestor' (LCA) bezeichnet. Die Häufigkeit mit der es nicht gelingt ein Amplikon auf Ebene der Spezies zuzuordnen ist dabei abhängig von den amplifizierten variablen Regionen. Wird beispielsweise allein die variable Region V4 betrachtet, so lassen sich mehr als die Hälfte der Sequenzen nicht auf Speziesebene zuordnen [5]. Abbildung 1A zeigt diesen Anteil für Sequenzen, welche verschiedenen Kombinationen von variablen Regionen entsprechen. Ein Primerpaar für die V3–V4 Region erzeugt ein Amplikon von ungefähr 460 bp Länge. Es eignet sich damit gut für eine Sequenzierung auf weit verbreiteten Sequenzierplattformen wie dem Illumina MiSeq und es wird entsprechend häufig verwendet. Bei diesem Protokoll lassen sich jedoch 20–56% der Sequenzen nicht auf Ebene der Spezies zuordnen. Diese Spanne ergibt sich aus den von Johnson et al. ermittelten Werten für die V3–V5 und die V4 Region.

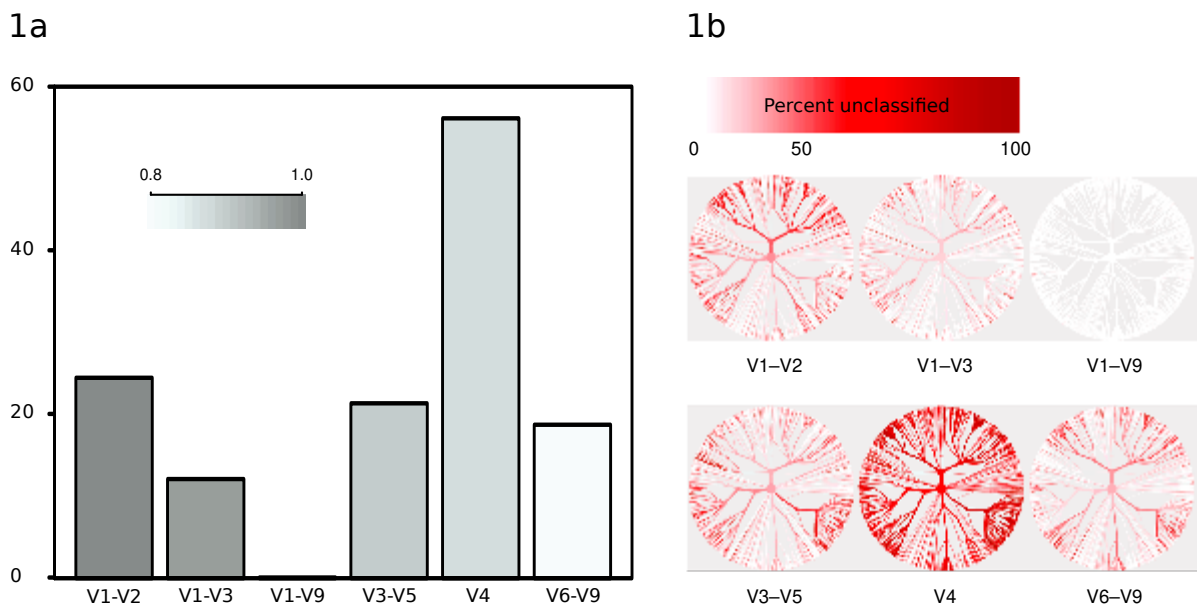


Abbildung 1: Vergleich variabler Regionen der 16S rRNA. (A) Anteil derjenigen Sequenzen, welche bei Auswahl bestimmter variabler Regionen nicht auf Speziesebene zugeordnet werden können (Schwellenwert für die Konfidenz 80%). (B) Die dargestellten Bäume wurden basierend auf einem Subset der Sequenzen der Greengenes Datenbank berechnet [6]. Für jede variable Region wird der selbe Baum gezeigt. Je stärker die Rotfärbung, umso größer ist der Anteil an Sequenzen, der nicht auf Speziesebene zugeordnet werden kann. Quelle: [5]

Die Verteilung der Sequenzen, die sich nicht zuordnen lassen, ist nicht zufällig. Je nach den verwendeten Regionen gibt es einen Bias zu bestimmten Taxa (Abb. 1B). So werden beispielsweise bei Verwendung der V1–V2 Region Proteobakterien schlecht klassifiziert und bei der Verwendung von V3–V5 Actinobakterien.

Darüber hinaus liegen rRNA Gene oftmals in mehreren Kopien im Genom vor. Wäre diese Kopienzahl stets bekannt, so könnte man sie bei der Quantifizierung berücksichtigen [7]. Für die meisten Organismen ist sie jedoch nicht beschrieben und eine exakte Quantifizierung ist mit den bestehenden Methoden deshalb kaum möglich.

Die Sequenzierung sämtlicher variabler Regionen (V1–V9, ca. 1 500 bp) in einem Read ist mit aktuellen Short-Read Plattformen (z.B. von Illumina oder dem BGI) nicht möglich. Die Sequenzierplattformen von Pacific Biosciences und Oxford Nanopore sind hingegen in der Lage die entsprechenden Moleküle in voller Länge zu sequenzieren. Aus Abbildung 1 wird ersichtlich, dass durch den Einsatz der längeren Amplikons fast alle Sequenzen auf Ebene der Spezies zugeordnet werden können.

In der zuvor genannten Publikation zeigen Johnson et al. zudem exemplarisch, dass sich

Kopien des rRNA Genes eines Organismus anhand von Einzelnukleotid-Polymorphismen unterscheiden lassen, wenn eine ausreichende Sequenziertiefe und -qualität erreicht wird. Dadurch gewonnene Informationen könnten künftig eine genauere Quantifizierung ermöglichen.

Zusammenfassend lässt sich feststellen, dass die Sequenzierung ribosomaler rRNA durchaus das Potential besitzt in Zukunft eine wichtigere Rolle bei der Identifizierung von Infektionserregern in klinischen Proben einzunehmen. Insbesondere in Kombination mit einer kompakten und mobilen Sequenzierplattform, wie dem Oxford Nanopore MinION, können sich interessante, mögliche Szenarien für einen solchen Einsatz ergeben. Da im Idealfall alle beschriebenen Bakterien, Archebakterien oder Pilze detektiert werden, bieten sich erhebliche Vorteile gegenüber den rein molekularbiologischen Methoden. Ein Nachteil bleibt jedoch selbst dann bestehen, wenn komplette rRNA Gene sequenziert werden: Taxa, deren rRNA-Sequenz nicht bekannt oder schlichtweg nicht vorhanden ist, können grundsätzlich nicht mit dieser Methode detektiert werden - dies trifft auch auf Viren zu.

Detektion viraler Infektionserreger mittels Multiplex-PCR

Die zuvor beschriebenen auf rRNA basierten Ansätze können für die Detektion von Viren nicht genutzt werden. Geeignete konservierte genomischen Regionen finden sich höchstens innerhalb einzelner Virusfamilien, nicht jedoch über die Grenzen viraler Familien hinweg. Multiplex-PCR (mPCR) erlaubt es mehrere genomische Regionen auf einmal zu amplifizieren. So gibt es mPCR Systeme welche beispielsweise die gleichzeitige Untersuchung auf das Humanen Respiratorischen Synzytial-Virus und das Humanen Metapneumovirus erlauben [8]. Genomische Segmente von Influenza A und Influenza B Viren werden hierzu in einem Ansatz amplifiziert [9].

Die Anzahl an Viren, die sich so gleichzeitig untersuchen lässt ist begrenzt. Mit der Anzahl der verwendeten Primer erhöht sich auch die Wahrscheinlichkeit unerwünschter Wechselwirkungen. Weit mehr als die zuvor beschriebenen rRNA-basierten Methoden, erfordern diese Ansätze deshalb *a-priori* Wissen zu den möglichen Infektionserregern im konkreten klinischen Fall und zu deren Nukleotidsequenzen im Allgemeinen [10].

1.3 Metagenom und Metatranskriptom Analysen

Zusätzlich zu Amplikon-basierten Methoden, kommen zunehmend solche zum Einsatz, welche auf eine selektive Anreicherung gänzlich verzichten und stattdessen das gesamte Metagenom beziehungsweise Metatranskriptom betrachten. Virale, prokaryotische und eukaryotische Infektionserreger können prinzipiell gleichzeitig detektiert werden. Das für diese Ansätze notwendige *a-priori* Wissen zu möglichen Krankheitserregern, und damit der methodische Bias, ist soweit reduziert, dass man diese Ansätze als 'unbiased' bezeichnet [11].

Methoden, die diesen Ansatz nutzen, lassen sich in zwei Kategorien gliedern. In der ersten Kategorie finden sich Methoden, die darauf abzielen, einzelne Sequenz-Reads taxonomisch zuzuordnen. Vertreter der zweiten Kategorie assemblieren die Reads zunächst. Zugeordnet werden erst die bei der Assemblierung entstehenden Contigs oder Scaffolds.

1.3.1 Taxonomischen Klassifikation kurzer Sequenzen

Im Folgenden werden zunächst die Ansätze der ersten Kategorie betrachtet. Sie lassen sich weiter unterteilen in solche, die im Wesentlichen auf Alignments beruhen und solche, welche ohne Alignments auskommen.

Alignment-basierte Analysen

Ein typischer Vertreter der ersten Kategorie ist die Software SURPI [12]. Eine Übersicht zu den einzelnen Schritten einer Auswertung mit dieser Software ist in Abbildung 2 gezeigt.

Die Sequenz-Reads werden zunächst vorprozessiert. Artificielle Sequenzen (Sequenzier-Adapter), sowie Sequenzen mit niedriger Qualität und niedriger Komplexität werden dabei entfernt. Dieser erste Schritt ist noch nicht spezifisch für diese Art von Methoden, sondern wird auch bei anderen, im Folgenden beschriebenen, Methoden vorangestellt.

Im nächsten Schritt werden semiglobale Alignments der Reads mit humanen Nukleotidsequenzen berechnet. Neben Sequenzen genomischer DNA (hg19), werden hierbei auch rRNA, mRNA und mitochondriale RNA Sequenzen, aus der NCBI reference sequence Datenbank (RefSeq) [13] verwendet. Es wird dabei angenommen, dass Reads, welche sich wenigstens einmal mit kleiner (< 12) Levenshtein-Distanz mit den Referenzsequenzen ali-

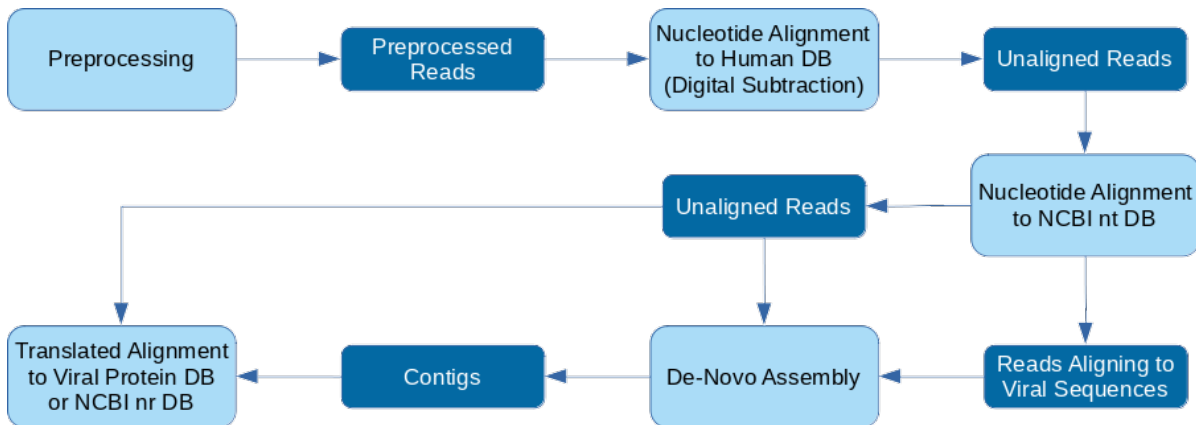


Abbildung 2: Schematische Darstellung der Auswertung mit SURPI als Vertreter Alignment-basierter Auswertungsstrategien. Gezeigt ist der erweiterte ('comprehensive') Modus. Zunächst werden die Reads vorprozessiert. In diesem Schritt werden artifizielle Sequenzen sowie Sequenzen niedriger Qualität und niedriger Komplexität entfernt. Danach erfolgt die Digitale Subtraktion. Die verbleibenden Reads werden mit den Sequenzen der NCBI Nukleotiddatenbank (nt) [14] aligniert. Reads welche auch in diesem Schritt nicht aligniert werden können oder aber taxonomisch den Viren zugeordnet wurden, werden anschließend assembliert. Zusammen mit den bisher nicht alignierten Reads werden Contigs danach auf Peptidebene entweder mit Sequenzen aus einer viralen Referenzdatenbank oder der umfassenderen NCBI Proteindatenbank (nr) aligniert. Vereinfacht nach [12].

gnieren lassen, dem Wirtsgenom zugehören und somit nicht pathogenen Ursprungs sind. Dieses Vorgehen wird auch als 'Digitale Subtraktion' bezeichnet. Wenn Sequenzhomologien zwischen einem Pathogen und dem Wirt bestehen, so kann die Digitale Subtraktion dazu führen, dass Reads pathogenen Ursprungs frühzeitig dem Wirt zugeordnet werden und dann bei der weiteren Analyse nicht mehr berücksichtigt werden. Insbesondere Proteindomänen sind oftmals stark konserviert und Sequenzhomologien bestehen mitunter selbst über die Grenzen von taxonomischen Domänen hinweg.

Nach der Digitalen Subtraktion folgt in der SURPI Pipeline ein Alignment der Reads mit Sequenzen der NCBI Nukleotiddatenbank (nt) [14]. Reads die sich mit Sequenzen von Viren, Bakterien, Pilzen oder Parasiten alignieren lassen, werden als potentielle Pathogene betrachtet. Reads, welche auch in diesem Schritt nicht aligniert werden können, werden hingegen assembliert. Für die Assemblierung werden gleichzeitig auch solche Reads verwendet, welche auf Nukleotidebene mit viralen Sequenzen aligniert werden konnten.

Da Assemblierungs-basierte Methoden im Folgenden noch separat betrachtet werden

1 Einleitung

und sie nicht typischerweise Bestandteil der Alignment-basierten Methoden sind, soll an dieser Stelle nur kurz darauf eingegangen werden. Die von SURPI ausgeführte Assemblierung ist auch deshalb nicht mit der von Assemblierungs-basierten Ansätzen vergleichbar, weil nur ein kleiner Anteil von Reads zur Assemblierung genutzt wird. Nicht nur diejenigen Sequenzen, welche homolog zum Wirtsgenom sind, sondern auch solche, welche Sequenzähnlichkeit zu nicht-viralen Infektionserregern in der NCBI Nukleotiddatenbank (nt) aufweisen, werden nicht assembliert. Der Nutzen der Assemblierung ist damit eingeschränkt, denn Regionen mit Homologie zum Wirtsgenom oder dem Genom eines nicht-viralen Infektionserregers können so grundsätzlich nicht assembliert werden.

Die als Ergebnis der Assemblierung entstehenden Contigs werden, wenn ihre Länge wenigstens dem 1,75-fachen der Readlänge entspricht, zusammen mit den nicht alignierten Reads des vorangegangenen Schrittes in allen sechs Leserahmen translatiert und dann entweder mit der gesamten NCBI Proteindatenbank (nr) oder aber nur mit viralen Referenzsequenzen aligniert. Ein Read gilt dabei als 'aligniert', wenn der resultierende E-value kleiner als 0.1 ist.

Alignment-freie Analysen

Exemplarisch für die Alignment-freien Methoden wird im Folgenden das Programm Taxonomer [3] vorgestellt. Ähnliche Methoden werden beispielsweise auch von Kraken beziehungsweise Kraken 2 [15, 16] verwendet. Das primäre Ziel der letztgenannten Programme ist jedoch nicht die Detektion von Infektionserregern.

Der Informationsfluss in Taxonomer ist in Abbildung 3 dargestellt. Bei dieser Methode wird die Abundanz von 21-meren in einem Set von Referenzsequenzen ermittelt. Es werden sowohl die Sequenzen des Wirtsgenoms als auch diejenigen von möglichen Pathogenen berücksichtigt. Bakterielle und pilzliche Genome werden dabei nur über deren rRNA repräsentiert. Jedes k -mer wird mit einer eindeutigen 'bit flag' versehen. Sie signalisiert die Zugehörigkeit des k -mers zu einer oder mehreren breiten taxonomischen Kategorien. Diese Kategorien repräsentieren beispielsweise sämtliche Bakterien, Viren und Phagen oder Pilze. Im ersten Schritt der Analyse eines Datensatzes wird jeder Read der Kategorie zugeordnet, mit der er am meisten k -mere teilt. Es erfolgt also keine Subtraktion von Sequenzen wie bei SURPI. Stattdessen werden Wirtsgenom und die Genome möglicher Pathogene gleichzeitig betrachtet. Dies ist ein erheblicher Vorteil, da Sequenzhomologien besser berücksichtigt sind. Im nächsten Schritt erfolgt innerhalb der Kategorien die taxonomische Klassifikation der darin enthaltenen Reads. Diese basiert

ebenfalls auf k -meren, allerdings mit einer Gewichtung. Das Gewicht eines jeden k -mers ist dabei ein für die Referenzdatenbank spezifisches Maß für die Wahrscheinlichkeit, dass ein bestimmtes k -mer zu einer bestimmten Referenz gehört. Dieses Maß ergibt sich aus der Häufigkeit des k -mers in der Referenzsequenz, seiner Häufigkeit in allen Referenzsequenzen und der Anzahl aller k -mere.

Die taxonomische Zuweisung von Reads erfolgt, indem für jede Referenzsequenz die Summe über alle gewichteten k -mere gebildet wird, die sowohl in der Referenzsequenz als auch in dem Read vorkommen. Falls mehrere gleichwertige Möglichkeiten der Zuordnung existieren, so erfolgt, wie bereits weiter oben beschrieben, eine Zuordnung zum 'Lowest Common Ancestor' (LCA).

Die bisher beschriebene Klassifikation erfolgt unmittelbar mit den k -meren der Reads und damit auf Nukleotidebene. Da diese Klassifikation auf dem gemeinsamen Vorkommen von identischen DNA-Sequenzen der Länge 21 in der Referenzsequenz und dem Read beruht, können Polymorphismen innerhalb einer Spezies oder eines Stammes dazu führen, dass ein Read nicht korrekt zugeordnet werden kann. Insbesondere die genetische Variabilität von Viren, ihre hohen Mutationsraten und ihre oft unvollständigen Referenzsequenzen können die Klassifikation auf Nukleotidebene erheblich erschweren. Diesem Umstand wird zumindest teilweise dadurch begegnet, dass Kategorien mit viralen und unbekanntem Taxa (und gegebenenfalls zusätzlich solche mit benutzerdefinierten Eigenschaften) auf Peptidebene betrachtet werden. Hierzu wird ein Subset von viralen Sequenzen aus UniRef90 und ein Subset bakterieller Sequenzen aus Uniref50 verwendet [17]. Aus diesen Peptidsequenzen wird zunächst eine Menge artifizierlicher Nukleotidsequenzen erzeugt. Jede Aminosäure wird durch genau ein Codon repräsentiert. Zur Klassifikation wird jeder Read in alle sechs Leserahmen übersetzt. Aus den resultierenden Peptidsequenzen werden ebenfalls artifizierliche Nukleotidsequenzen erzeugt. Die weitere Prozessierung erfolgt dann wie zuvor für die Zuordnung auf Nukleotidebene beschrieben, allerdings werden 30-mere (zehn Aminosäuren) verwendet.

Um auch homologe virale Proteine mit geringer Sequenzähnlichkeit zu detektieren, nutzt Taxonomer darüber hinaus eine Methode welche an DIAMOND [18] angelehnt ist und ein reduziertes Aminosäurealphabet der Länge 11 verwendet.

Der Ansatz von Taxonomer hat zahlreiche Vorteile gegenüber dem von SURPI. Ein Geschwindigkeitsvorteil ergibt sich daraus, dass nur auf das Vorhandensein identischer k -mere in Read und Referenzdatenbank geprüft wird und kein Alignment zu berechnen ist. Zudem nutzt Taxonomer nicht die vollständige NCBI Nukleotid- (nt) beziehungsweise

1 Einleitung

NCBI Proteindatenbank (nr), sondern innerhalb der meisten Kategorien nur Marker-Gene und das Transkriptom des Wirtes. Ein weiterer Vorteil gegenüber SURPI ist, dass Sequenzen potentiell pathogenen und insbesondere viralen Ursprungs nicht frühzeitig entfernt werden. Sequenzen welche homolog zum Wirtsgenom und zu einem Pathogen sind, werden von SURPI mit letzterem grundsätzlich nie aligniert. Gleichsam werden auch Sequenzen, die homolog zu einem bekannten Infektionserreger sind, aber stattdessen einem nicht beschriebenen Pathogen entstammen, niemals auf Peptidebene aligniert. Beide Probleme müssen jedoch nicht zwangsläufig in jedem auf Alignments basierten Ansatz zur Erkennung von Infektionserregern auftreten.

Sowohl Taxonomer als auch SURPI zielen darauf ab, Reads zu klassifizieren. Dies erlaubt eine schnelle Prozessierung, welche bei Taxonomer meist nur im Bereich von wenigen Minuten liegt. Die Reads sind jedoch kurz und der Informationsgehalt entsprechend gering. Insbesondere auf Peptidebene kommt es dadurch zu vielen falschpositiven Ergebnissen. Dies wird auch von den Autoren von Taxonomer eingeräumt und es wird bei der Beschreibung der Peptid-basierten Klassifikation darauf hingewiesen, dass an Stelle von Reads auch längere Contigs verwendet werden könnten, um diesem Problem zu begegnen.

Taxonomer und SURPI gehen über die Ebene der Klassifikation von einzelnen Sequenzen nicht hinaus. Jeder Read wird zudem als eigenständige Entität und losgelöst von der Gesamtheit aller in einer Probe vorhandenen Sequenzen betrachtet. Durch die damit verbundene isolierte Betrachtung lokaler Ähnlichkeit besteht grundsätzlich die Gefahr, dass die Anzahl der Taxa in einer Probe überschätzt und fälschlich die Präsenz zusätzlicher Taxa angenommen wird.

Taxonomische Klassifikation von Contigs und Scaffolds

Alternativ zur taxonomischen Klassifikation einzelner Reads, können diese zunächst zu Contigs oder Scaffolds assembliert werden. Dadurch kann im Folgenden mit längeren Sequenzen gearbeitet werden, welche einen höheren Informationsgehalt aufweisen und sich dadurch zuverlässiger taxonomisch zuordnen lassen. Insbesondere neue Pathogene, welche nur geringe Sequenzähnlichkeit zu bekannten Infektionserregern aufweisen, können besser detektiert werden. Unabhängig von einer taxonomischen Zuordnung erlauben es die längeren Sequenzen zudem oftmals einzelne Proteindomänen und auch aus mehrere Domänen bestehende Domänenarchitekturen zu identifizieren.

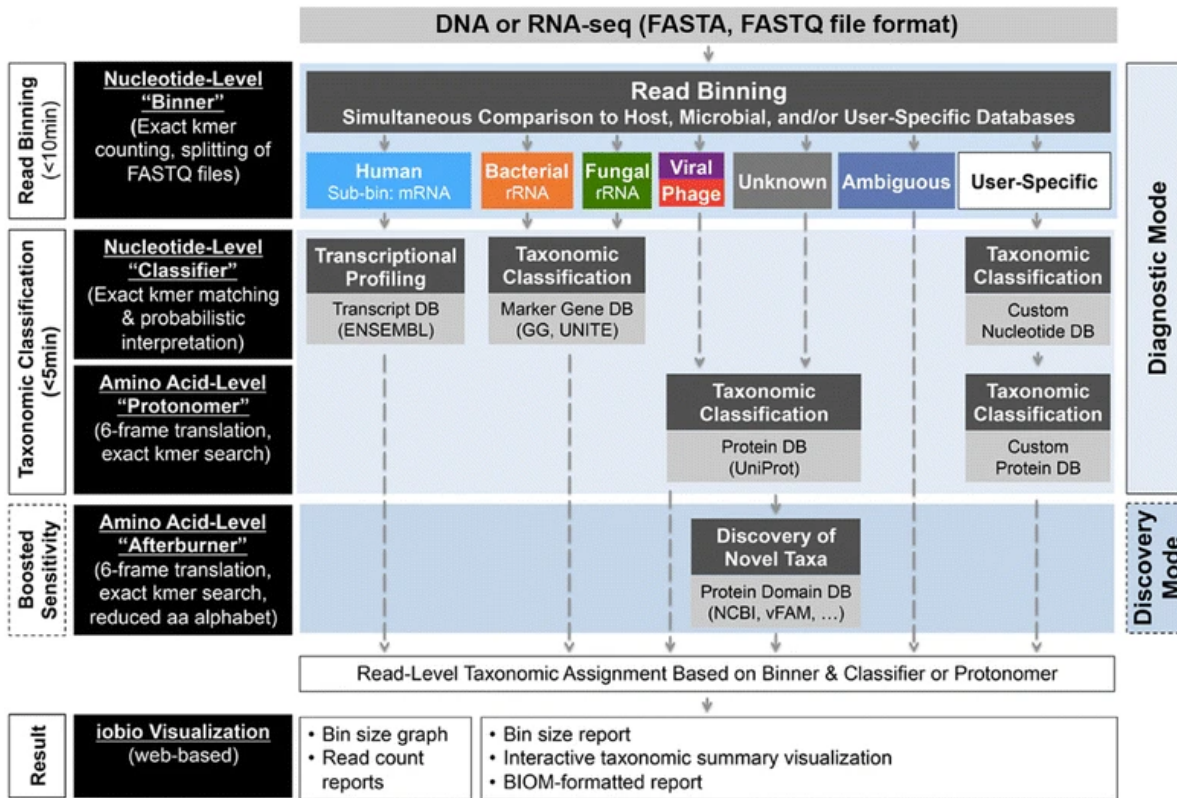


Abbildung 3: Schematische Darstellung der Auswertung mit Taxonomer als Beispiel Alignment-freier Auswertestrategien. Zunächst werden die Reads Kategorien zugeordnet ('Binner'). Die weitere Prozessierung erfolgt dann innerhalb dieser Kategorien. Reads welche als bakteriell oder pilzlich kategorisiert wurden, werden beispielsweise anhand von Marker-Genen taxonomisch auf Nukleotidebene klassifiziert. Virale und unbekannte Taxa werden hingegen auf Peptidebene klassifiziert. Der exakten Klassifikation auf dieser Ebene ('Protonomer') kann sich eine Klassifikation anhand eines reduzierten Aminosäurealphabetes anschließen ('Afterburner'). Gegenstand der Klassifikation sind stets einzelne Reads. Quelle: [3].

1 Einleitung

Diese funktionalen Einheiten können wichtige Hinweise auf die Pathogenität selbst gänzlich unbekannter Taxa geben.

Mit Hilfe der Assemblierung ist es zudem oftmals möglich, vollständige oder nahezu vollständige Genomsequenzen von Infektionserregern zu erhalten [19]. Ein möglicher Nachteil von Methoden, welche eine Assemblierung durchführen, ist deren höherer Zeitaufwand. Dieser erklärt vielleicht auch, warum in SURPI nur ein kleiner Anteil von Reads assembliert wird und warum die Autoren von Taxonomer zwar die Vorteile von längeren Contigs erwähnen, entsprechende Funktionalität für die Assemblierung aber nicht integrieren. Vor dem Hintergrund, dass für vorangehende Analyseschritte, wie die Präparation der Sequenzierungs-Library und die Sequenzierung selbst, wenigstens ein Tag einzuplanen ist, bleibt jedoch abzuwägen, ob eine zuverlässigere Analyse, welche statt weniger Minuten vielleicht wenige Stunden dauert, nicht zielführender ist.

Ein anderer möglicher Nachteil ergibt sich bei der Analyse von Proben in denen Infektionserreger nur mit sehr wenigen Reads vertreten sind. In solchen Fällen könnten vielleicht noch einzelne Reads zugeordnet werden, es ist aber nicht zu erwarten, dass sich deutlich längere Contigs assemblieren lassen. Es wäre jedoch immer noch möglich mit den ursprünglichen, kürzeren Reads zu arbeiten, so dass sich nicht zwangsläufig ein Nachteil gegenüber anderen Sequenz-basierten Ansätzen ergibt.

Die im Rahmen dieser Arbeit entwickelte Software verfolgt ebenfalls einen Ansatz, welcher auf Assemblierung beruht. Das Thema wird deshalb in der Diskussion wieder aufgegriffen und vertieft.

Zielsetzung

Ziel der vorliegenden Arbeit ist die Entwicklung einer Software zur Detektion neuer und bekannter Infektionserreger. Ihre Fähigkeiten sollen deutlich über diejenigen bereits bestehender Ansätze hinausgehen.

Es gibt mehrere Ansatzpunkte um dieses Ziel zu erreichen. Insbesondere das Arbeiten mit längeren Sequenzen kann künftig dabei helfen, Limitierungen bestehender Ansätze zu überwinden. Nicht nur die taxonomische Zuordnung auf Nukleotidebene wird dadurch zuverlässiger. Es können auch längere Peptidsequenzen übersetzt und analysiert werden. Die Integration von Informationen, beispielsweise zu Proteindomänen, kann zusätzliche Anhaltspunkte für das mögliche Vorhandensein von Pathogenen liefern. Die Möglichkeiten zur Detektion neuer Pathogene können dadurch erheblich verbessert werden.

Wichtig ist auch, verbesserte Methoden der taxonomischen Zuordnung zu entwickeln. Die bereits beschriebene LCA Methode ist einfach und bewährt. Daher erscheint es sinnvoll, sie nicht gänzlich zu ersetzen, sondern um neue Ansätze zu ergänzen. Ein solcher neuer Ansatz besteht darin, die Sequenzen nicht isoliert bestimmten Taxa zuzuordnen. Stattdessen sollte das Ziel sein, alle beobachteten Sequenzen mit einem möglichst einfachen Modell der taxonomischen Zusammensetzung der Probe zu erklären.

Etablierte Methoden bieten zudem keine oder nur sehr begrenzte Möglichkeiten für das Arbeiten mit Kohorten und für den Vergleich mehrerer Proben. Probenübergreifende Analysen können insbesondere dann vorteilhaft sein, wenn eine taxonomische Zuordnung nicht möglich ist oder wenn sie keine Rückschlüsse auf das Vorhandensein eines Pathogens erlaubt. Unabhängig von einer taxonomischen Zuordnung kann eine probenübergreifende Auswertung in so einem Fall Sequenzen identifizieren, welche nur in Proben vorkommen, die mit dem Ausbruch einer Krankheit assoziiert sind, nicht jedoch in Kontrollproben. Solche Sequenzen sind ein wichtiger Ausgangspunkt für weiterführende Analysen.

Eine neue Software, welche die zu entwickelnden Methoden integriert, ist an den spezifischen Einsatz in der klinischen Diagnostik anzupassen. Sie sollte weder informatische Fachkenntnisse beim Benutzer voraussetzen noch hohe Anforderungen an die vorhandene IT-Infrastruktur mit sich bringen. Idealerweise umfasst sie zudem alle notwendigen Analyseschritte, so dass eine Vor- oder Nachprozessierung von Daten nicht notwendig ist und die gesamte Auswertung automatisch und standardisiert ausgeführt wird.

2 Im Rahmen der Promotion entstandene Publikationen

Im formellen Rahmen der Promotion sind die im Folgenden aufgeführten sechs Veröffentlichungen entstanden. In der ersten Publikation [1] werden die im Rahmen der Promotion entwickelten bioinformatischen Methoden beschrieben und evaluiert. In den fünf anderen genannten Publikationen werden sie im Hinblick auf verschiedene Fragestellungen angewandt. Im Kontext dieser Fragestellungen wurden die Methoden weiterentwickelt und sie konnten sich in der Praxis bewähren. Damit spiegelt jede der Publikationen eine unterschiedliche Entwicklungsstufen der Methoden wider.

Der Inhalt der Publikationen wird jeweils kurz zusammengefasst. Das Hauptaugenmerk wird dabei auf Aspekte gelegt, die in Bezug zum Promotionsthema stehen. Zudem wird für jede Publikation der eigene Beitrag beschrieben. Die Publikationen selbst finden sich im Appendix.

Alawi, M., Burkhardt, L., Indenbirken, D., Reumann, K., Christopheit, M., Kröger, N., Lütgehetmann, M., Aepfelbacher, M., Fischer, N. & Grundhoff, A. DAMIAN: an open source bioinformatics tool for fast, systematic and cohort based analysis of microorganisms in diagnostic samples. Scientific reports 9, 1–17 (2019)

In dieser Publikation werden die Software DAMIAN, die darin verwendeten Methoden und ihre grundlegende Anwendungsszenarien vorgestellt. Die Software ist in der Lage neue und bekannte Infektionserreger in klinischen Proben schnell und zuverlässig zu detektieren. Die Leistungsfähigkeit wird anhand von zehn diagnostischen Proben demonstriert. Drei dieser Proben wurden zuvor von den Autoren von Taxonomer zur Evaluation ihrer Software verwendet. Bei den anderen Proben handelt es sich um neue diagnostische Proben, die mit DAMIAN erstmals ausgewertet wurden. Die Ergebnisse von DAMIAN

2 Im Rahmen der Promotion entstandene Publikationen

werden anhand von konventioneller molekularbiologischer Methoden (kulturbasierte Methoden und PCR) validiert und mit den Ergebnissen von etablierten bioinformatischen Werkzeugen verglichen. Es wird gezeigt, dass DAMIAN virale und bakterielle Infektionserreger akkurat detektiert und zuordnet. Insbesondere virale Infektionserreger können zuverlässig auf Speziesebene und in den meisten Fällen sogar auf Ebene des Stammes identifiziert werden.

Die Entwicklung der in dieser Publikation beschriebenen bioinformatischen Methoden geht im Wesentlichen auf den Autor der vorliegenden Dissertation zurück. Er hat sämtliche Teile der veröffentlichten Software selbstständig entwickelt und die bioinformatischen Analysen durchgeführt. Im Manuskript hat er in erster Linie die technischen Aspekte beschrieben.

Zapatka, M., Borozan, I., Brewer, D. S., Iskar, M., Grundhoff, A., Alawi, M., Desai, N., Sültmann, H., Moch, H., Cooper, C. S., et al. The landscape of viral associations in human cancers. *Nature genetics* 52, 320–330 (2020)

In dieser Studie werden Whole-Genome-Sequencing (WGS) und Transkriptom Daten von 2374 Spendern analysiert. Ziel ist es, Zusammenhänge zwischen Virusinfektionen und Krebs zu finden und zu charakterisieren. In 382 WGS- und 68 Transkriptom-Datensätzen werden Viren detektiert.

Zum Einsatz kommen drei unterschiedliche Methoden zur Detektion von viralen Infektionserregern. Die in der Publikation beschriebene 'Pathogen Discovery Pipeline' (P-DiP) basiert auf zwei verschiedenen Versionen der Software, welche inzwischen als DAMIAN publiziert wurde. Für die zuerst durchgeführten, genomischen Analysen wird noch eine Version verwendet, die das Programm Trinity [21] für Assemblierungen nutzt. Für die Prozessierung der Transkriptome wird eine neuere Version verwendet, welche bereits IDBA-UD [22] integriert.

DAMIAN bereitet umfangreiche Statistiken zu einzelnen Proben auf, um die klinische Diagnostik zu unterstützen. In dieser Studie wurden jedoch mehrere hundert Proben prozessiert und die einzelnen Statistiken konnten deshalb nicht manuell evaluiert werden. Aus diesem Grund kommen bei P-DiP R-Skripte zum Einsatz mit denen Übersichtsstatistiken aufbereitet und mögliche Kontaminationen anhand der NCBI Taxonomy ID gefiltert werden. Einträge unterhalb von 'Plasmid' und 'synthetic virus' werden beispielsweise als mögliche Kontaminationen betrachtet. DAMIAN verfügt zwar ebenfalls über

die Möglichkeit Kontaminationen zu identifizieren, es nutzt dafür jedoch eine kleinere, händisch kurierte Datenbank.

Die erwähnten R-Skripte wurden vom Erstautor der Publikation entwickelt. Sie zeigen, wie DAMIAN Schnittstellen für externe Werkzeuge bereitstellt. Dieser Aspekt wird im letzten Kapitel dieser Dissertation noch einmal aufgegriffen und als Ansatzpunkt für künftige Verbesserungen, insbesondere für neue Benutzungsschnittstellen, diskutiert.

Die Publikation entstand im Rahmen des ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium [23]

Der Autor der vorliegenden Dissertation hat zu dieser Publikation durch Entwicklung der dort als P-DiP beschriebenen Methode beigetragen und diese auch im Manuskript beschrieben. Die verwendete Implementation dieser Methoden stammt von ihm. Lediglich die in der Publikation zusätzlich verwendeten R-Skripte wurden nicht von ihm entwickelt.

Günther, T., Haas, L., Alawi, M., Wohlsein, P., Marks, J., Grundhoff, A., Becher, P. & Fischer, N. Recovery of the first full-length genome sequence of a parapoxvirus directly from a clinical sample. Scientific reports 7, 1–8 (2017)

In dieser Publikation wird beschrieben, wie erstmalig die vollständige Genomsequenz eines Parapoxvirus aus einer klinischen Probe isoliert und sequenziert wurde. Die Probe stammt aus der Hautläsion einer Kegelrobbe (*Halichoerus grypus*). Eine Analyse der vom Virus kodierten Proteine identifizierte Gene, welche spezifisch für die Anpassung an die Haut und die Pathogenese von Parapoxviren sind. Aufgrund dieses Befundes wurde das detektierte Virus als neue Spezies innerhalb der Gattung *Parapoxvirus* klassifiziert.

Es erfolgte eine Sequenzierung auf einer Illumina MiSeq Plattform und eine Auswertung mit einer frühen, unveröffentlichten Version der im Rahmen dieser Promotion entwickelten Software. Da für die Kegelrobbe keine Referenzassemblierung vorhanden war, wurde stattdessen eine Assemblierung des Genoms der Wendellrobbe (*Leptonychotes weddellii*) als Referenz verwendet. Bei der Auswertung wurde ein Parapoxvirus detektiert, sowie die fast vollständige Sequenz des Virusgenoms. Offene Leserahmen wurden automatisch annotiert. Um auch die Hairpin-Termini des Genoms zu bestimmen, erfolgte eine zusätzliche Sequenzierung mit der Oxford Nanopore MinION-Technologie und anschließend eine Hybrid-Assemblierung mit SPAdes [24].

Der Autor der vorliegenden Dissertation hat zu dieser Publikation durch Entwicklung der bioinformatischen Methoden, welche zur Detektion und Annotation des Parapoxvirus

2 Im Rahmen der Promotion entstandene Publikationen

fürten, beigetragen. Er hat die Analyse der Daten mit der von ihm geschriebenen Software sowie die Hybrid-Assemblierung durchgeführt und die entsprechenden Methoden beschrieben.

Coronado, L., Liniger, M., Muñoz-González, S., Postel, A., Pérez, L. J., Pérez-Simó, M., Perera, C. L., Frias-Lepoureau, M. T., Rosell, R., Grundhoff, A., Indenbirken, D., Alawi, M., Fischer, N., Becher, P., Ruggli, N. & Ganges, L. Novel poly-uridine insertion in the 3 UTR and E2 amino acid substitutions in a low virulent classical swine fever virus. *Veterinary microbiology* 201, 103–112 (2017)

Gegenstand dieser Publikation ist die Untersuchung der Virulenz zweier Stämme des Klassischen Schweinepest-Virus. Es kommen Methoden zum Einsatz, die später Bestandteil von DAMIAN wurden. So erfolgte eine Digitale Subtraktion von Reads, die sich dem Hausschwein (*Sus scrofa*) zuordnen lassen.

Es wurden ausschließlich Nukleotidsequenzen aligniert. Eine Assemblierung wurde mit Trinity durchgeführt und die taxonomische Zuweisung erfolgte allein anhand des Lowest Common Ancestor-Prinzips.

Der Autor der vorliegenden Dissertation hat zu dieser Publikation durch Entwicklung von bioinformatischen Methoden zur Detektion und Bestimmung der Genomsequenz des untersuchten Virus beigetragen. Er hat die Analyse der Daten mit der von ihm entwickelten Software durchgeführt.

Postel, A., Hansmann, F., Baechlein, C., Fischer, N., Alawi, M., Grundhoff, A., Derking, S., Tenhündfeld, J., Pfankuche, V. M., Herder, V., Baumgärtner, W., Wendt, M. & Becher, P. Presence of atypical porcine pestivirus (APPV) genomes in newborn piglets correlates with congenital tremor. *Scientific reports* 6, 1–9 (2016)

Hier erfolgt ein Screening der Sera von 369 dem Anschein nach gesunden Schweinen auf das Vorhandensein des Atypischen Porcinen Pestivirus (APPV). Erstmals wurde die komplette Polyprotein-kodierende Sequenz eines europäischen APPV bestimmt. Auf Nukleotidebene ist sie zu 88,2% mit einer zuvor beschriebenen APPV Sequenz aus den USA identisch. In der Studie konnte das Virus in neugeborenen Ferkeln mit kongenitalem Tremor nachgewiesen wurde, in gesunden neugeborenen Ferkeln wurde es hingegen nicht

detektiert.

Der Autor hat zu dieser Publikation durch Entwicklung von bioinformatischen Methoden zur Detektion und Bestimmung der Genomsequenz des Atypischen Porcinen Pestivirus beigetragen. Er hat die Analyse der Daten mit der von ihm geschriebenen Software durchgeführt.

Baechlein, C., Grundhoff, A., Fischer, N., Alawi, M., Hoeltig, D., Waldmann, K.-H. & Becher, P. Pegivirus infection in domestic pigs, Germany. *Emerging infectious diseases* 22, 1312 (2016)

In dieser Publikation wird die Detektion eines Pegivirus im Serum von Schweinen beschrieben. Mit im Rahmen dieser Promotion entwickelten Methoden wurden in einem Pool von Sera verschiedener Tiere zunächst Sequenzen detektiert, welche auf das Vorhandensein von viralen Pathogenen mit entfernter Verwandtschaft zu einem bei Fledermäusen vorkommenden Pegivirus hindeuten. Auf Basis der ermittelten Sequenzen konnten Primer für den Nachweis der Viren mittels RT-PCR entwickelt werden. So konnte diejenige Einzelprobe identifiziert werden, welche das Pegivirus beinhaltet. Diese Probe wurde erneut sequenziert und wie zuvor ausgewertet, so dass schließlich ein Contig mit 9 145 bp pegiviralen Ursprungs assembliert werden konnte.

Der Autor der vorliegenden Dissertation hat zu dieser Publikation durch Entwicklung von bioinformatischen Methoden zur Detektion und Bestimmung der Genomsequenz des Pegivirus beigetragen. Er hat die Analyse der Daten mit der von ihm geschriebenen Software durchgeführt.

3 Diskussion

Das wichtigste Ergebnis der vorliegenden Arbeit ist die Software DAMIAN. Sie integriert im Rahmen der Promotion entwickelte Methoden und stellt sie der Öffentlichkeit unter der GNU General Public License zur Verwendung und Weiterentwicklung zur freien Verfügung. In der zugehörigen Publikation wird die Software unter Verwendung authentischer klinischer Proben mit Methoden konventioneller Diagnostik validiert und mit etablierten bioinformatischen Analyse-Pipelines verglichen. Es wird gezeigt, dass die Ergebnisse bestehender Methoden übertroffen werden und virale sowie bakterielle Infektionserreger zuverlässig detektiert werden. Virale Erreger können auf Speziesebene und meist auch auf Stammebene korrekt taxonomisch klassifiziert werden. Unter den zehn in der Publikation analysierten Proben finden sich drei, welche zuvor von den Autoren von Taxonomer für die Evaluation ihres Programmes verwendet wurden. DAMIAN konnte auch in diesen Proben alle von Taxonomer detektierten Infektionserreger identifizieren. In den beiden Proben, die mit viralen Infektionen assoziiert sind, gelang es zudem die Genome der Erreger in voller Länge zu rekonstruieren. Auch in weiteren Publikation konnte mehrfach gezeigt werden, wie sich die Software im klinischen Einsatz bewährt. Gleichzeitig war die Arbeit an diesen Publikationen und insbesondere der damit verbundene Dialog mit klinische arbeitenden Wissenschaftlern, wichtig für die konsequente Weiterentwicklung der Software und für ihre Ausrichtung an klinischen Anforderungen. Die Software kam in unterschiedlichen Szenarien zum Einsatz. Sie wurde sowohl für DNA als auch RNA-Proben verwendet und neben verschiedenen Proben menschlicher Herkunft wurden auch zahlreiche Proben anderer eukaryotischer Wirtsspezies erfolgreich vom Autor mit der Software ausgewertet [1, 19, 20, 25–27].

Die flexible und gut dokumentierte Benutzungsschnittstelle von DAMIAN und der hohe Grad an Portabilität erlauben es anderen Wissenschaftlern die Software selbstständig und in ihrem eigenen wissenschaftlichen Kontext zu verwenden. So wurden Adenoviren im Lungengewebe von Meerschweinchen (*Cavia porcellus*) identifiziert [28]. Eine unveröffentlichte Version von DAMIAN wurde zur Detektion von Infektionserregern in

Fischen genutzt [29]. Und bei einem an Enzephalitis erkrankten Patienten konnte ein pilzlicher Krankheitserreger identifiziert werden [30].

Im Folgenden werden zunächst, analog zu den Beschreibungen in der Einleitung, die einzelnen Schritte einer Auswertung mit DAMIAN diskutiert. Es wird dabei auch auf Aspekte eingegangen, welche bisher nicht in Veröffentlichungen beschrieben sind. Danach folgt ein Ausblick, welcher konkret und ausführlich aufzeigt, wie DAMIAN künftig weiter verbessert werden könnte. Abschließend wird die Analyse eines COVID-19 Datensatzes beschrieben und die praktischen Aspekte der Aufbereitung der Ergebnisse werden an diesem Beispiel diskutiert.

3.1 Auswertung klinischer Proben mit DAMIAN

DAMIAN integriert sämtliche Schritte der Datenauswertung, eine zusätzlich Vor- oder Nachprozessierung ist nicht notwendig. Die Auswertung beginnt mit Sequenz-Reads und sie endet mit der Generierung umfassender Ergebnisberichte. DAMIAN wurde von Beginn an für den Einsatz in einem klinischen Umfeld entwickelt. So können DNA- und RNA-Sequenzen verarbeitet werden die von verschiedener Formen von Probenmaterial stammen. Zudem liefert DAMIAN unabhängig vom Taxon des Wirtsorganismus zuverlässige Ergebnisse. Die effiziente Nutzung von DAMIAN erfordert weder eine spezielle IT-Infrastruktur noch Expertise in Bioinformatik.

Wie eingangs beschrieben, basieren viele andere Werkzeuge, darunter SURPI, Taxonomer und KRAKEN 2, auf der taxonomischen Klassifikation von einzelnen Reads. Ein solcher Ansatz erlaubt eine schnelle Klassifikation der Reads, die bei Verfügbarkeit geeigneter Referenzsequenzen zudem zuverlässig ist. DAMIAN basiert auf einem grundlegend anderen Ansatz. Die Reads werden zunächst zu längeren Contigs assembliert und erst diese werden taxonomisch zugeordnet und annotiert. Die längeren Sequenzen verfügen über einen höheren Informationsgehalt, so dass eine höhere Sensitivität und Spezifität bei der taxonomischen Zuordnung erreicht werden kann. Zudem erlauben sie weiterführende Analysen, welche sich mit kürzeren Reads nicht oder nur sehr eingeschränkt durchführen ließen. Hierzu zählt einerseits die funktionale Annotation mit Proteindomänen, welche zusätzliche Informationen zu einer Sequenz liefert. Andererseits können anhand der Contigs aber auch verschiedene Proben verglichen werden. Beide weiterführenden Analysen liefern selbst dann wertvolle Anhaltspunkte für die Präsenz von Infektionserregern, wenn eine taxonomische Zuordnung anhand von Sequenzhomologien nicht möglich ist. Sie ver-

bessern damit insbesondere die Fähigkeiten zur Detektion neuer, bisher unbeschriebener Infektionserreger.

Abbildung 4 zeigt eine schematische Übersicht der im Folgenden beschriebenen Analyse-schritte. Konkrete Beispiele der von DAMIAN generierten Ausgaben einzelner Schritte finden sich in der Beschreibung der Auswertung des COVID-19 Datensatzes am Ende der Diskussion.

3.1.1 Module der Auswertung

Ausgangspunkt einer Auswertung sind Reads im FASTQ-Format, dem Standardausgabeformat aller Short-Read Sequenzierplattformen. Die Dateien mit den Reads können komprimiert oder unkomprimiert verwendet werden. DAMIAN arbeitet mit einer beliebigen Anzahl von Dateien und einer beliebigen Kombination aus paired-end und single-end Reads. Lange Reads, beispielsweise aus Sequenzierungen auf Oxford Nanopore und Pacific Biosciences Plattformen, können zwar ebenfalls genutzt werden, DAMIAN wurde für ihre Verwendung jedoch bisher nicht optimiert. Dieser Aspekt wird ausführlicher in Paragraph 3.2.2 auf Seite 32 diskutiert.

Verifikation der Eingabedaten

Zu Beginn der Auswertung wird zunächst überprüft, ob die notwendigen Voraussetzungen bezüglich der Daten und Software-Komponenten erfüllt sind. Hierzu werden einerseits die Eingaben des Nutzers überprüft. Es wird evaluiert, ob die gesetzten Parameter kompatibel zueinander sind, ob Pfade zu Dateien korrekt sind und, ob notwendige Lese- und Schreibrechte im Dateisystem vorhanden sind. Andererseits wird geprüft, ob eine Verbindung zur Datenbank aufgebaut werden kann und, ob Softwareabhängigkeiten erfüllt sind. In diesem Schritt werden zudem Informationen zu verwendeten Softwareversionen, den Nutzereingaben und zum Ausführungszeitpunkt gesammelt und in der Datenbank abgelegt. Dadurch wird die Analyse jederzeit reproduzierbar.

Qualitätsprüfung und -kontrolle

Die Reads werden zunächst einer Qualitätsprüfung unterzogen. Neben allgemeinen Werten, wie der Verteilung der Readlängen, wird dabei auch die Qualität der Basen, sowie das Vorhandensein artifizieller Sequenzen und möglicher PCR-Artefakte festgestellt.

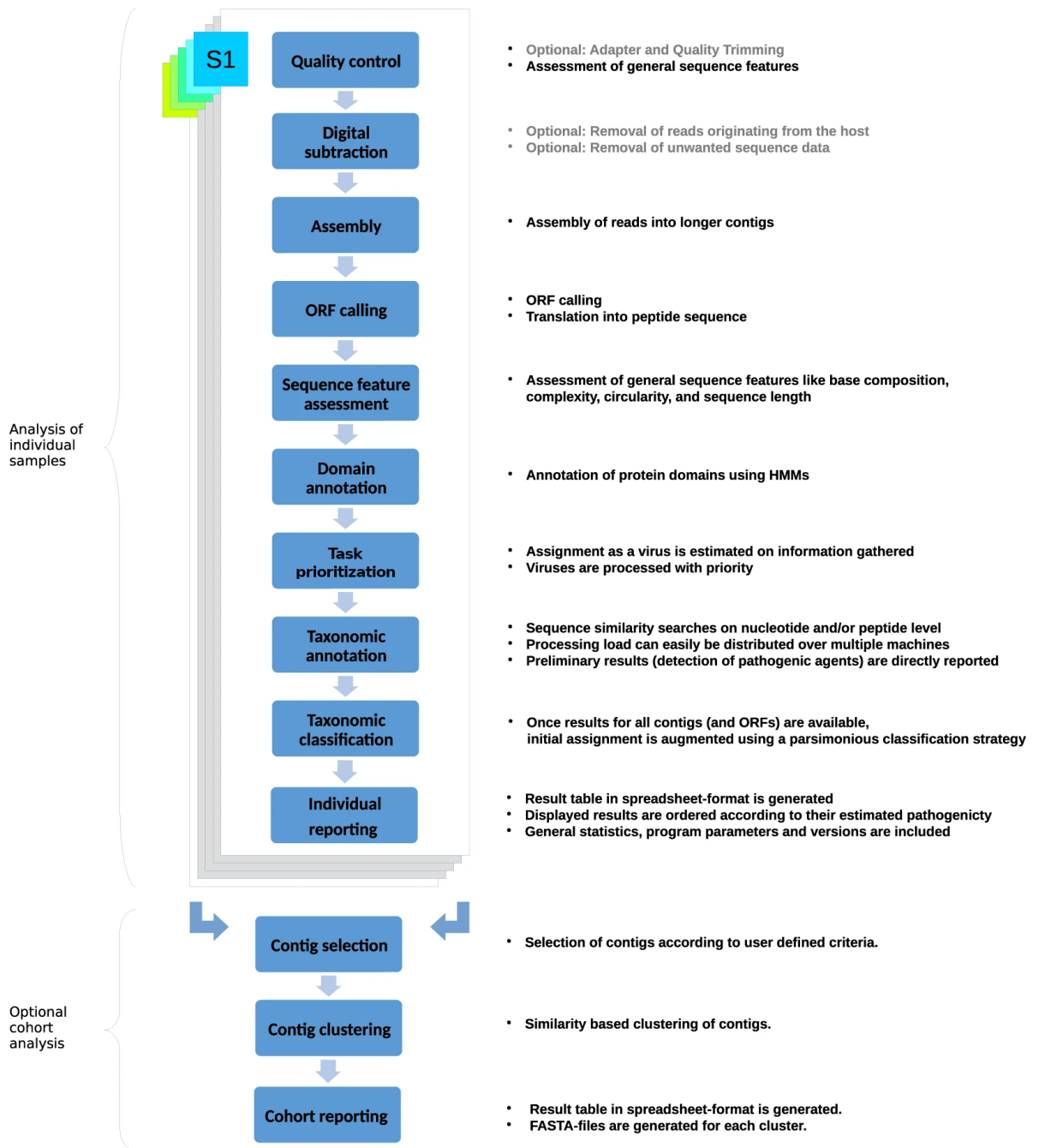


Abbildung 4: Schematische Darstellung der Auswertung mit DAMIAN. Jede Probe wird zunächst einzeln ausgewertet. Eine optionale Kohortenanalyse kann nachträglich für eine beliebige Anzahl zuvor prozessierter Proben durchgeführt werden. Quelle: [1].

Danach werden Sequenzen von Sequenzier-Adaptoren, sehr kurze Sequenzen (≤ 40 bp) und Sequenzen niedriger Qualität (Phred-Qualitäts-Score ≤ 10) automatisch für die weiteren Analyseschritte ausgeschlossen. Die Qualitätsprüfung wird wiederholt und alle erfassten Daten werden in der Datenbank abgelegt. Sie werden später Teil des Ergebnisberichtes. Auch in den im Folgenden beschriebenen Schritten der Auswertung werden jeweils umfangreich Daten und Metadaten zu diesem Zweck gesammelt, auch dann wenn es hier im Einzelnen nicht mehr erwähnt wird. Die Kriterien für die Qualitätskontrolle und Parameter für die weiteren Schritte, können weitestgehend vom Benutzer definiert werden. Dies kann entweder über Optionen beim Programmaufruf oder mit Hilfe einer Konfigurationsdatei geschehen. Es besteht keine grundsätzliche Notwendigkeit diese Werte anzupassen, da die Voreinstellungen den meisten Anforderungen gerecht werden.

Digitale Subtraktion und approximative Quantifizierung

Optional kann DAMIAN den Anteil an Reads schätzen, die große Übereinstimmungen mit Referenzsequenzen haben. Hierzu kann der Benutzer beliebige Referenzsequenzen (z.B. das Genom des Wirtsorganismus) definieren. DAMIAN unterscheidet zwischen RNA- und DNA-Daten und geeignete Referenzsequenzen werden entsprechend ausgewählt. Neben den genannten Schätzwerten wird, falls paired-end Reads zur Verfügung stehen, auch die durchschnittliche Länge und Standardabweichung der sequenzierten DNA-Fragmente bestimmt. Standardmässig wird für die Berechnung der Schätzwerte nur ein Teil ($N = 10^5$) jeder Sequenzier-Library betrachtet.

Zudem kann, wie in der Einleitung beschrieben, eine Digitale Subtraktion von Reads, deren Alignment mit dem Wirtsgenom oder -transkriptom bestimmten Kriterien genügt, vorgenommen werden. Die Subtraktion kann insbesondere bei Proben mit einem großen Anteil von Reads aus Sequenzen des Wirtsgenoms zu einer reduzierten Laufzeit führen. Es ist jedoch nicht ausgeschlossen, dass in diesem Schritt Reads pathogenen Ursprungs fälschlicherweise entfernt werden, wenn es entsprechende Sequenzhomologien zwischen Wirt und Pathogen gibt. Daher sollten die Vor- und Nachteile der optionalen Digitale Subtraktion abgewogen werden. Bei allen in dieser Arbeit beschriebenen Anwendungen von DAMIAN wurde eine Digitale Subtraktion durchgeführt.

Assemblierung und Erfassung der Eigenschaften von Contigs

Reads, die nicht in den vorangegangenen Schritten verworfen wurden, werden assembliert. Zu den hierbei entstehenden Contigs werden Daten zur Länge, dem GC-Gehalt, der Zirkularität und der Sequenz-Komplexität ermittelt. Zusätzlich werden die Contigs in alle sechs Leserahmen übersetzt um Offene Leserahmen einer gegebenen Mindestlänge zu identifizieren. Die Peptidsequenzen dieser Offenen Leserahmen werden für weiterführende Analysen in der Datenbank gespeichert. Die Verwendung Offener Leserahmen setzt, im Gegensatz zu einer Genvorhersage, kein *a-priori* Wissen voraus und funktioniert gleichwohl bei prokaryotischen, viralen und eukaryotischen Organismen.

Funktionelle Annotation und Ranking von Contigs

DAMIAN arbeitet mit einer Datenbank von Proteindomänen, die auf PFAM [31] basiert. Die in PFAM enthaltenen Domänen werden von DAMIAN zusätzlich anhand ihrer taxonomischen Zugehörigkeit zu einer oder mehreren taxonomischen Kategorien annotiert. So gibt es beispielsweise Proteindomänen, die bisher ausschließlich in Viren oder sogar nur in bestimmten Viren gefunden werden. Andere Domänen sind spezifisch für Bakterien oder für Pilze. Da die Proteindomänen funktionelle Einheiten repräsentieren, können sie auch dann eindeutige Ergebnisse liefern, wenn andere Methoden taxonomischer Zuordnung fehlschlagen. Zusätzlich werden Informationen zu Domänen genutzt, um die Reihenfolge festzulegen, in welcher die Contigs im weiteren Verlauf prozessiert werden. Ziel ist es, mögliche Infektionserreger und insbesondere Viren möglichst frühzeitig zu identifizieren. Neben der taxonomischen Zuordnung der Proteindomänen wird dabei auch deren Länge berücksichtigt.

Taxonomische Zuordnung

Die taxonomische Zuordnung von DAMIAN unterscheidet sich von derjenigen anderer Methoden nicht nur dadurch, dass an Stelle von kürzeren Reads mit längere Contigs gearbeitet wird. DAMIAN nutzt zudem auch die gesamten NCBI Protein- (nr) und NCBI Nukleotiddatenbanken (nt) [14].

Programme der BLAST+ Suite [32] werden verwendet um Alignments auf Nukleotidebene und Peptidebene unabhängig voneinander oder iterativ durchzuführen. Dabei wird jedes Contig gegebenenfalls mehreren Sequenzen der genannten Datenbanken zugeordnet.

Jede dieser Sequenzen ist einem Knoten im NCBI-Taxonomie-Baum zugeordnet. Diese Knoten sind dann Gegenstand der zuvor beschriebenen LCA-Methode.

Vorläufige Ergebnisse werden direkt ausgegeben, sobald ein mögliches virales Pathogen detektiert wird. Eine Detektion nicht-viraler Infektionserreger ist zu diesem Zeitpunkt noch nicht zuverlässig möglich, so dass in diesem Fall keine Ausgabe erfolgt.

Zusätzlich zur LCA-Methode wurde für DAMIAN eine weitere Methode der taxonomischen Klassifikation entwickelt und implementiert. Sie geht über die individuelle Zuordnung einzelner Contigs hinaus.

Zunächst wird anhand aller Contigs bestimmt, welche Taxa in einer Probe möglicherweise vorhanden sind. Hierzu werden neben dem Alignment mit dem höchsten Bitscore auch weitere, auf suboptimalen Alignments beruhende, taxonomische Zuweisungen eines jeden Contigs betrachtet. Für jede Zuweisung wird ein Score berechnet, welcher sich aus dem Produkt der Abundanz des Contigs und dem Bitscore des Alignments ergibt. Dieser Score wird dem zugeordneten Taxon zugerechnet. So erhält man, nach Betrachtung aller Contigs, Scores für die möglicherweise vorhandenen Taxa.

Zur taxonomischen Zuordnung der Contigs, werden diese nun ein zweites Mal betrachtet. Es werden erneut alle taxonomischen Zuordnungen, auch solche suboptimaler Alignments, berücksichtigt. Diejenige dieser Zuordnungen, welche zu dem Taxa mit dem höchsten zuvor ermittelten Score erfolgte, wird als finale Zuordnung des Contigs übernommen. Falls sich mit dieser Methode mehrere gleichwertige Zuordnungen ergeben, so wird der LCA dieser Zuordnungen verwendet.

Dadurch, dass Contigs nicht isoliert betrachtet werden, ist diese Methode exakter und sparsamer bei den Zuordnungen. Fast immer sind Zuordnungen auf den niedrigsten taxonomischen Ebenen möglich. Die Zuordnungen entsprechend der LCA Methode und die Zuordnungen entsprechend der neu entwickelten Methode sind Bestandteil der Ergebnisse von DAMIAN.

Erstellung von Ergebnisberichten

Für jede analysierte Probe wird eine Datei im Excel-Format erzeugt. Das erste Tabellenblatt enthält eine Übersicht zu den detektierten Taxa. Die Zeilen der Tabelle sind so sortiert, dass auf potentielle Pathogene verweisende Einträge zuerst gezeigt werden. Farbkodierungen erlauben es zudem schnell einen Überblick über die Ergebnisse zu erhalten.

3 Diskussion

DAMIAN arbeitet mit sechs Kategorien denen jeweils eine Farbe zugeordnet ist. Die erste Kategorie (hell lila) beinhaltet Einträge, die sowohl anhand von Sequenzhomologien zu Einträgen in der NCBI Nukleotid- (nt) beziehungsweise NCBI Proteindatenbank (nr) als auch anhand von Proteindomänen als Virus klassifiziert wurden. In der zweiten (dunkel lila) und dritten (hellblau) Kategorie erfolgt die Zuordnung nur anhand eines der beiden Kriterien. Phagen werden generell in einer separaten Kategorie (dunkelblau) dargestellt. Erkannte Artefakte und Kontaminationen werden ebenfalls markiert (grau). Die sechste Kategorie (schwarz) umfasst alle weiteren Einträge, sie schließt auch solche ein, die nicht taxonomisch zugeordnet werden konnten.

Die Tabelle ist interaktiv und verlinkt jeden Eintrag mit weiteren Tabellenblättern. Auf diesen sind detaillierte Informationen zu den entsprechenden Contigs, Offenen Leserahmen und Proteindomänen zusammengestellt. Es sind sowohl naive taxonomische Zuordnungen (basierend auf dem Alignment oder den Alignments mit dem kleinsten E-Value), als auch Ergebnisse der LCA-Methode und der weiter oben beschriebenen neuen Methode enthalten. Nukleotidsequenzen sämtlicher Contigs einer benutzerdefinierten Mindestlänge und alle aus ihnen abgeleiteten Peptidsequenzen sind im FASTA Format gespeichert. Positionen und weitere Informationen zu den Proteindomänen sind im BED Format verfügbar. Sämtliche dieser Dateien sind mit der Excel-Datei verknüpft und können so direkt geöffnet werden.

Der Report enthält darüber hinaus allgemeine Statistiken, wie beispielsweise zur Anzahl der Reads, zu dem Anteil an Reads, welcher dem Wirtsgenom zugeordnet wurde und auch zur Länge der sequenzierten Fragmente. Die Versionen der genutzten Programme und die verwendeten Parameter sind ebenfalls aufgeführt. Ergebnisberichte können jederzeit erstellt werden. Auch dann, wenn die Auswertung noch nicht abgeschlossen ist. Zu jedem Zeitpunkt fließen alle bis dahin verfügbaren Informationen ein. Der Status der Analyse, welcher beispielsweise angibt, ob sie abgeschlossen wurde, noch ausgeführt wird oder abgebrochen wurde, ist aus dem Report ersichtlich.

Kohortenanalyse

Für die probenübergreifende Analyse von Kohorten wird, wie oben beschrieben, zunächst jede einzelne Probe ausgewertet. Der Nutzer teilt die prozessierten Proben in Gruppen ein. Eine Gruppe umfasst alle Proben, die eindeutig mit einer Infektion assoziiert sind. Eine zweite Gruppe umfasst alle Proben, die eindeutig nicht mit einer Infektion assoziiert sind. Eine letzte Gruppe umfasst alle anderen Proben. Die Analyse kann mit allen Contigs der

gewählten Proben oder auf einem Subset von Contigs durchgeführt werden. Die Kriterien für die Auswahl von Contigs können vom Nutzer definiert werden. Dabei können die Länge, die Komplexität der Sequenz, die Anzahl an Proteindomänen, die Anzahl Offener Leserahmen und die taxonomische Zuordnung berücksichtigt werden. Jedes so gewählte Contig wird mit jedem anderen gewählten Contig aligniert. Hierzu werden Programme aus der BLAST+ Suite [32] verwendet. Voreingestellt ist die Nutzung von MEGABLAST, es können stattdessen aber auch DCMEGABLAST, BLASTN oder TBLASTX verwendet werden.

Der Bitscore jedes Alignments wird in einer $n \times n$ -Matrix gespeichert, wobei n die Anzahl der Contigs ist. Die Contigs werden nun anhand dieser paarweisen Bitscores mit einem Single-Linkage Clustering Verfahren zu Clustern zusammengefasst. Sobald der Bitscore bei der nächsten Fusionierung von Clustern unter einen zuvor bestimmten Wert fällt, ist die Zuordnung der Cluster final und das Clustering-Verfahren bricht ab.

Für jeden Cluster wird ein Score berechnet. Dieser Score ist ein Mass dafür, wie stark der Cluster mit der Infektion assoziiert ist und er berechnet sich wie folgt:

$$Score_{cluster} = \frac{C_{infection}}{N_{infection}} - \frac{C_{control}}{N_{control}}$$

Hierbei ist $C_{infection}$ die Anzahl von Proben im Cluster, welche mit der Infektion assoziiert sind. $N_{infection}$ ist die Gesamtzahl aller mit der Infektion assoziierter Proben. $C_{control}$ und $N_{control}$ sind die entsprechenden Werte für die Kontrollgruppe.

Es ergibt sich also ein Score zwischen +1 bis -1. Ein Cluster mit einem Score von +1 enthält wenigstens eine Sequenz aus jeder mit der Infektion assoziierten Probe und keine einzige Sequenz aus einer Probe der Kontrollgruppe.

Die Kohortenanalyse erlaubt es mit Infektionen assoziierte Sequenzen auch dann zu identifizieren, wenn deren taxonomische Zuordnung nicht möglich oder wenn die taxonomische Zuordnung allein keinen Aufschluss über die Pathogenität gibt.

In einer Kohortenanalyse von fünf Proben von Patienten mit Enzephalitis und 22 Kontrollproben von anderen Personen konnte gezeigt werden, wie ein Infektionserreger auch dann identifiziert wird, wenn er nicht zuvor taxonomisch zugeordnet wird [1]. In diesem Fall wurde ein einziger Cluster mit einem Score von +1 berechnet. Die Contigs darin konnten nachträglich eindeutig dem Erreger, Enterovirus B, taxonomisch zugeordnet werden.

3.1.2 Integrierte Funktionalität zum Verteilten Rechnen

In DAMIAN ist es für die taxonomische Zuordnung von Contigs notwendig, diese und die gegebenenfalls daraus abgeleitet Proteinsequenzen mit den Sequenzen der NCBI Nukleotid- (nt) beziehungsweise NCBI Proteindatenbank (nr) zu alignieren. Dieser Schritt ist wegen der grossen Menge an Contigs und der grossen Menge an Referenzsequenzen zeitaufwändig. Bei der in Paragraph 3.3 auf Seite 34 beschriebenen Auswertung dauerte es bei Verwendung von acht Threads 40 Minuten bis er abgeschlossen war. Werden zusätzlich Alignments auf Peptidebene durchgeführt, sind Laufzeiten von mehreren Stunden zu erwarten.

Eine Vielzahl von mitunter sehr langen Nukleotid- und gegebenenfalls auch Peptidsequenzen wird mit den Sequenzen der NCBI Nukleotid- (nt) beziehungsweise NCBI Proteindatenbank (nr) aligniert. Jede zu alignierende Sequenz wird dabei unabhängig von den anderen betrachtet und aligniert. Es ist demnach möglich die Sequenzen parallel zu prozessieren. Zwar lassen sich die verwendeten Programme aus der BLAST+ Suite [32] parallelisieren und es stehen Werkzeuge wie CrocoBLAST [33] zur Verfügung, welche die Effizienz der Parallelisierung erhöhen, es lässt sich jedoch nicht ohne Weiteres mehr als ein Rechner für diese Aufgabe verwenden. Da DAMIAN ohnehin eine PostgreSQL voraussetzt und der entsprechende Datenbankserver nicht auf dem selben Rechner wie DAMIAN laufen muss, wurde die Datenbank genutzt um eine vollständig integrierte Lösung zum Verteilten Rechnen zu implementieren. Zur Verwendung ist deshalb keine spezifische Hard-oder Software und kein informatisches Fachwissen erforderlich. Diese Lösung wird im Folgenden erstmalig beschrieben.

DAMIAN implementiert eine Klasse *TaskManager*, welcher unter Anderem die Aufgabe zufällt, mehrere zu alignierende Sequenzen sogenannten 'Jobs' zuzuordnen. Ein Job referenziert sämtliche für die Prozessierung dieser Sequenzen notwendigen Informationen. Hierzu zählen beispielsweise die Sequenzen selbst, aber auch, mit welcher Methode und welcher Referenzdatenbank die Auswertung erfolgen soll. Zudem hat jeder Job einen Status und eine Priorität. Der Status gibt an, ob ein Job gerade ausgeführt wird, neu erstellt, erfolgreich beendet, abgebrochen oder erneut eingestellt wurde. Die Priorität wird vom Nutzer oder automatisch gesetzt und gegebenenfalls vom *TaskManager* erhöht. Die Reihenfolge, in denen die Jobs bearbeitet werden, ergibt sich aus der Priorität und der seit der Erstellung eines Jobs verstrichenen Zeit. Der *TaskManager* erstellt nicht nur neue Jobs, er verwaltet auch abgebrochene Jobs und solche, deren Prozessierungsdauer ein gesetztes Zeitlimit übersteigt. Beide Arten von Jobs werden mit erhöhter Priorität erneut

eingestellt. Erfolgreich beendete Jobs werden abhängig von den gesetzten Einstellungen bearbeitet. Je nachdem, wie DAMIAN konfiguriert wird, kann dies bedeuten, dass die Contigs hintereinander in mehreren verschiedenen Jobs mit MEGABLAST, BLASTN oder BLASTP aligniert werden. Gibt es keine Jobs in den zuvor beschriebenen Kategorien, so führt der *TaskManager* selbst einen Job aus oder er beendet diesen Schritt der Auswertung, wenn alle zu berechnenden Ergebnisse vorliegen.

Der *TaskManager* sucht in der Datenbank generell nur nach solchen Jobs, die er selbst erstellt hat. Das Skript *damian_node.rb* wurde geschrieben, um unabhängig vom *TaskManager* Jobs zu bearbeiten. Es sucht ebenfalls nach zu bearbeitenden Jobs in der Datenbank. Es kann sich dabei jedoch an verschiedenen gleichzeitig laufenden Auswertungen mit DAMIAN beteiligen. Das Skript kann so parametrisiert werden, dass es nur neue Jobs annimmt, wenn die Auslastung des Systems, auf dem es gestartet wurde, ein bestimmtes Niveau unterschreitet.

Zur Nutzung der beschriebenen Funktionalität ist keine zusätzliche Software notwendig. Jedes System, welches die Anforderungen zur Ausführung von DAMIAN erfüllt, kann auf diese Weise Ressourcen zur Verfügung stellen. Es müssen lediglich der Zugang zum Datenbankserver und zu den NCBI Datenbanken möglich sein.

So können freie Rechenkapazitäten flexibel und ohne informatische Fachkenntnisse zur Auswertung genutzt werden und - auch über das Internet - dazu beitragen, die Laufzeiten erheblich zu verkürzen.

3.2 Ausblick

Es konnte gezeigt werden, dass mit DAMIAN eine nützliche Software zur Verfügung steht, welche einen erheblichen Mehrwert gegenüber etablierten Methoden bietet. Gerade deshalb sollte die Software weiterentwickelt werden. Im Folgenden werden mögliche Ansätze für eine Weiterentwicklung diskutiert.

3.2.1 Integrierte Software

DAMIAN integriert mehrere etablierte Programme. Hierzu zählen beispielsweise der Assembler IDBA-UD [22], der Short-Read Aligner Bowtie2 [34] und die BLAST+ Suite [32]. Diese Programme sind über Schnittstellen integriert, welche es vielfach erlauben von den konkret integrierten Programmen und ihrem spezifischen Syntax zu abstrahieren.

Beispielsweise sind Schnittstellen für die drei Assembler IDBA-UD, Trinity [21] und SPAdes [24] vorhanden. Diese Schnittstellen erlauben es einerseits, selbst zur Laufzeit verschiedene Programme für die Assemblierung auszuwählen. Andererseits erleichtern sie auch das Einbinden und Evaluieren von weiteren Programmen erheblich. Der gleiche modulare Aufbau wurde auch für andere integrierte Programme übernommen. Dadurch ist es generell sehr einfach, integrierte Programme auszutauschen oder im Vergleich zu neueren Programmen zu evaluieren.

Die in DAMIAN integrierten Programme der BLAST+ Suite [32] und HMMer [35] zum Alignieren von Sequenzen und Sequenzprofilen haben lange Laufzeiten und sie bilden daher den Flaschenhals bei der Zuordnung von Contigs. Es ist deshalb stets auf die Frage zu hinterfragen, ob sich die Laufzeit durch Verwendung anderer Programme zum Alignment verringern lässt, ohne dabei die Qualität der Ergebnisse zu verschlechtern. Die Verwendung von Programmen wie DIAMOND [18] oder MMSEQS2 [36] könnte die Berechnung von Alignments erheblich beschleunigen und ihre Integration sollte deshalb unbedingt in Erwägung gezogen werden.

Für die Qualitätskontrolle wird derzeit das Programm Trimmomatic [37] verwendet. Das neuere Programm FASTP [38] erlaubt eine schnellere Prozessierung, es erkennt artifizielle Sequenzen automatisch und generiert Berichte, die auf einfache Weise weiterverarbeitet werden können. Daher erscheint es sinnvoll, in Zukunft Trimmomatic durch FASTP zu ersetzen.

3.2.2 Gesonderte Prozessierung von längeren Reads

Aktuelle Versionen von DAMIAN unterstützen die Verwendung von beliebigen Nukleotidsequenzen im FASTQ-Format und damit grundsätzlich die längeren Reads, welche aus Sequenzierungen mit beispielsweise Pacific Biosciences und Oxford Nanopore Plattformen stammen. Die spezifischen Eigenschaften der letztgenannten, längeren Sequenzen bleiben bei der Prozessierung jedoch noch unberücksichtigt. Eine wichtige Grundlage für die Vorteile, welche DAMIAN gegenüber anderen Ansätzen aufweist, ist die Evaluation längerer Sequenzen. Wie bereits beschrieben, werden die Reads nicht unmittelbar taxonomisch klassifiziert und annotiert, sondern erst zu längeren Contigs assembliert. Entsprechend kann DAMIAN insbesondere von einer guten Assemblierung (vor Allem gekennzeichnet durch einen hohen N50-Wert und wenige Fehler bei der Assemblierung) profitieren. Zukünftige Versionen könnten insbesondere dadurch aufgewertet werden,

dass eine Hybrid-Assemblierung durchgeführt wird. Diese kombiniert kurze und lange Reads jeweils unter Berücksichtigung ihrer spezifischen Eigenschaften [39, 40]. Darüber hinaus wäre es empfehlenswert, die für längere Reads spezifischen Methoden der Qualitätsprüfung, der Qualitätskontrolle und der Quantifizierung anzuwenden.

In diesem Zusammenhang kann es sinnvoll sein weitere Programme zur Qualitätskontrolle, zur Assemblierung oder auch zum Alignieren der Reads zu integrieren. Auch diese Erweiterungen wären über die zuvor genannten Schnittstellen einfach zu implementieren.

3.2.3 Benutzungsschnittstelle und Ergebnisberichte

Die Auswertung mit DAMIAN lässt sich in zwei Hauptteile gliedern. Die Analyse von Proben ist getrennt von der Datenaufbereitung in Form von Ergebnisberichten. Die Schnittstelle zwischen beiden Teilen bildet die Datenbank.

DAMIAN orientiert sich an den Bedürfnissen klinisch arbeitender Wissenschaftler und seine Verwendung setzt kein bioinformatisches Fachwissen voraus. Die Ergebnisberichte werden, wie weiter oben beschrieben, im weit verbreiteten Excel-Format bereitgestellt und sie sind zudem innerhalb der Tabellen und darüber hinaus mit (FASTA und BED) Dateien verlinkt, so dass sie interaktiv verwendet werden können.

Neben der Kommandozeilenschnittstelle könnte jedoch künftig auch eine graphische Benutzungsschnittstelle implementiert werden. Dadurch liesse sich voraussichtlich eine höhere Akzeptanz der Benutzer erreichen. Es bietet sich insbesondere eine web-basierte Benutzungsoberfläche an. Sie könnte weitestgehend unabhängig vom Betriebssystem der Nutzer verwendet werden.

3.2.4 Klassifikation anhand von 16S-, ITS- oder anderen Marker-Genen

Es wurde gezeigt, dass es oftmals möglich ist, vollständige oder nahezu vollständige virale Genome mit DAMIAN zu rekonstruieren. Prokaryotische und eukaryotische Genome lassen sich hingegen normalerweise nicht aus metagenomischen Proben rekonstruieren. Neben anderen Eigenschaften, wie der Repetitivität dieser Genome, reicht die Tiefe der Sequenzierung meist aus, um diese Genome auch nur einfach abzudecken. Bei der Sequenzierung von RNA kommt hinzu, dass bei diesen Organismen ohnehin nur Transkripte sequenziert werden. Ob ein Transkript assembliert wird, hängt wesentlich von der Höhe

seiner Expression ab.

Ribosomale RNA macht stets einen erheblichen Anteil einer Gesamt-RNA Probe aus. Wie in der Einleitung im Kontext von Amplikonsequenzierungen beschrieben, kann sie zur Identifizierung von prokaryotischen und pilzlichen Organismen genutzt werden. Die Verwendung von Marker-Genen in metagenomischen und metatranskriptomischen Proben ist nicht neu. Auch in Verbindung mit einer frühen Version von DAMIAN wurde bereits unterstützend mit 16S und 23S rRNA gearbeitet. Die Sequenzen wurden mit Hidden Markov Modellen in den Contigs detektiert und dann weiter analysiert [11]. Auch Taxonomer nutzt, wie eingangs beschrieben, 16S- und ITS-Sequenzen zur taxonomischen Zuordnung.

Die Integration dieser Funktionalität in künftige Versionen von DAMIAN könnte die Detektion prokaryotischer und eukaryotischer Infektionserreger weiter verbessern.

3.3 Auswertung eines COVID-19 Datensatzes

Aus aktuellem Anlass wird im Folgenden die Auswertung eines COVID-19 Datensatzes mit DAMIAN beschrieben. Der Datensatz stammt aus einer Anfang 2020 publizierten, zahlreich zitierten Studie, deren Ziel es war, mögliche Erreger derjenigen Krankheit zu finden, welche heute als COVID-19 bekannt ist [41].

Die analysierte RNA wurde aus Probenmaterial extrahiert, welches einem Patienten aus Wuhan, China mittels bronchoalveolärer Lavage (BAL) entnommen wurde. Es erfolgte eine rRNA-Abreicherung und eine Sequenzierung auf einer Illumina MiniSeq Plattform im paired-end (2×150 bp) Modus. Es wurden 28,3 Millionen Readpaare sequenziert. Sie sind unter der Accession PRJNA603194 beim European Nucleotide Archive (ENA) [42] archiviert und wurden für die hier beschriebene Auswertung von dort bezogen.

Für die Auswertung wurde die Version GRCh38 als Wirtsgenom verwendet. Taxonomische Zuordnungen erfolgten anhand einer NCBI Nukleotid- beziehungsweise Taxonomiedatenbank vom 28. September 2020. Für die Berechnung wurden acht Threads genutzt und darüber hinaus Standardparameter verwendet. Dies bedeutet, dass keine Peptidsequenzen mit der NCBI Proteindatenbank aligniert wurden. Offene Leserahmen und Proteindomänen wurden hingegen annotiert.

SARS-CoV-2 wurde nach 148 Minuten detektiert. Zu diesem Zeitpunkt der Auswertung waren noch nicht alle Contigs aligniert. Die Zuordnung ist deshalb vorläufig und beruht

3.3 Auswertung eines COVID-19 Datensatzes

auf der LCA Methode. Die Identifizierung von zahlreichen Proteindomänen, welche in Offenen Leserahmen detektiert wurden, unterstützt bereits das auf Nukleotidebene erzielte Ergebnis. Assembliert wurde ein Contig mit 29 872 bp. Die Referenzsequenz in der Datenbank ist 29 882 bp lang. Nach 241 Minuten standen alle Ergebnisse zur Verfügung und es konnte ein abschliessender Bericht im Excel-Format generiert werden. Tabelle 1 zeigt einen Ausschnitt der Übersichtsseite des Berichtes. Das assemblierte Contig stimmt zu 100% mit der Sequenz des Virus aus der Datenbank überein. 38 für Viren spezifische Domänen wurden detektiert. Das Virus erscheint deshalb in der Tabelle an erster Stelle und der Eintrag ist entsprechend farbkodiert. Mit jedem Eintrag ist ein detaillierter Bericht verknüpft, welcher über den mit *view* bezeichneten Verweis aufgerufen werden kann. Einen kleinen Ausschnitt des Berichtes für SARS-CoV-2 zeigt Tabelle 2. Dargestellt sind Informationen zu einem Teil der im Contig identifizierten Proteindomänen. Die Mehrzahl dieser Domänen ist spezifisch für *Coronaviridae*. Sie liefern zusätzliche Evidenz für das Vorhandensein von SARS-CoV-2 in der Probe.

Species	Report	Absolute Abundance	Relative Abundance	Assembly Length	Contigs	Orfs	Perc. Maximum Identity (nucl)	Distinct Viral Domains	Viral Domains	Bacterial Domains	Eukaryotic Domains	Other Domains	Tax ID
Severe acute respiratory syndrome-related coronavirus	view	113786	8.745%	29872	1	47	100	38	46	11	14	0	694009
CRESS virus sp.	view	1234	0.095%	5475	5	25	99.779	0	1	1	2	0	2202563
Microviridae sp.	view	174	0.013%	532	1	3	85.098	0	1	1	1	0	2202644
uncultured virus	view	146	0.011%	2324	3	5	94.663	0	2	2	2	0	340016
Arequatovirus	view	98	0.008%	1295	1	7	97.222	0	0	0	0	0	1982881
Pacmanvirus A23	view	72	0.006%	1018	1	1	97.143	0	0	0	0	0	1932881
Panoviridae sp.	view	46	0.004%	671	1	5	86.944	0	1	1	1	0	1940570
EBPR siphovirus 1	view	20	0.002%	550	1	2	87.597	0	0	0	0	0	1048520
Marine virus AFVG_250M1101	view	16	0.001%	541	1	2	97.872	0	0	0	0	0	2693129
Crucivirus-like circular genetic element-85	view	12	0.001%	572	1	3	93.958	0	0	0	0	0	2761503
Myoviridae	view	12	0.001%	603	1	2	91.304	0	1	1	0	0	10662
[Unassigned]	view	287806	22.118%	2560104	3359	9729	-1	73	2197	3713	3682	0	-1
Prevotella jejuni	view	208056	15.989%	843402	890	2593	100	23	1112	2129	2084	0	1177574

Tabelle 1: Abschließender Ergebnisbericht. Gezeigt ist ein Ausschnitt der Übersichtseite. SARS-CoV-2 wird automatisch am erster Stelle und entsprechend farbkodiert dargestellt.

3.3 Auswertung eines COVID-19 Datensatzes

Contig_ID	Orf_ID	Accession	Name	Description	Viral Root	Bacterial Root	Eukaryotic Root	Other Root	Start_Pos	End_Pos	Evalue
27134	77943	PF09401	CoV_NSP10	Coronavirus RNA synthesis protein NSP10	Nidovirales				4266	4388	0
27134	77943	PF19213	CoV_NSP6	Coronavirus replicase NSP6	Coronaviridae				3602	3863	0
27134	77943	PF11501	bCoV_NSP1	Betacoronavirus replicase NSP1	Coronaviridae				13	147	4.2E-34
27134	77943	PF08710	CoV_NSP9	Coronavirus replicase NSP9	Coronaviridae				4145	4257	6.9E-52
27134	77943	PF08716	CoV_NSP7	Coronavirus replicase NSP7	Coronaviridae				3864	3946	5.1E-40
27134	77943	PF08717	CoV_NSP8	Coronavirus replicase NSP8	Coronaviridae				3947	4143	0
27134	77943	PF12379	bCoV_NSP3_N	Betacoronavirus replicase NSP3, N-terminal	Coronaviridae				884	1054	2.6E-55
27134	77943	PF16251	bCoV_NAR	Betacoronavirus nucleic acid-binding (NAR)	Betacoronavirus				1926	2023	2.7E-41
27134	77943	PF16348	CoV_NSP4_C	Coronavirus replicase NSP4, C-terminal	Coronaviridae				3170	3265	6.6E-44
27134	77943	PF19218	CoV_NSP3_C	Coronavirus replicase NSP3, C-terminal	Coronaviridae				2265	2752	0
27134	77943	PF19212	CoV_NSP2_C	Coronavirus replicase NSP2, C-terminal	Coronaviridae				656	822	3.1E-30
27134	77943	PF19217	CoV_NSP4_N	Coronavirus replicase NSP4, N-terminal	Coronaviridae				2792	3145	0
27134	77943	PF19211	CoV_NSP2_N	Coronavirus replicase NSP2, N-terminal	Coronaviridae				187	427	2E-52
27134	77943	PF12124	bCoV_SUD_C	Betacoronavirus SUD-C domain	Coronaviridae				1502	1565	1.4E-34
27134	77943	PF11633	bCoV_SUD_M	Betacoronavirus single-stranded poly(A) binding	Coronaviridae				1355	1497	0
27134	77943	PF01661	Macro	Macro domain	Viruses	Bacteria	Eukaryota	Archaea	1062	1168	9.5E-21
27134	77943	PF05409	Peptidase_C30	Coronavirus endopeptidase C30	Coronaviridae				3296	3586	0
27134	77943	PF08715	CoV_peptidase	Coronavirus papain-like peptidase	Coronaviridae				1568	1886	0
27134	77974	PF13086	AAA_11	AAA domain	Aureococcus ant	Bacteria	Eukaryota	Archaea	1201	1269	1.2E-08
27134	77974	PF13087	AAA_12	AAA domain	Viruses	Bacteria	Eukaryota	Archaea	1427	1504	1E-09
27134	77974	PF13604	AAA_30	AAA domain	Viruses	Bacteria	Eukaryota	Archaea	1205	1336	5.4E-09
27134	77974	PF06471	CoV_Methyltr_1	Coronavirus guanine-N7 methyltransferase	Coronaviridae				1534	2055	0
27134	77974	PF06460	CoV_Methyltr_2	Coronavirus 2'-O-methyltransferase	Nidovirales				2405	2700	0
27134	77974	PF19215	CoV_NSP15_C	Coronavirus replicase NSP15, uridylylate-specific	Nidovirales				2249	2401	0

Tabelle 2: Abschließender Ergebnisbericht. Gezeigt ist ein Ausschnitt der Details zum SARS-CoV-2 Eintrag. Dieser Teil des Berichtes kann über den entsprechenden Verweis *view* auf der Übersichtsseite aufgerufen werden. Die Proteinomänen unterstützen die taxonomische Zuordnung auf Nukleotidebene. Domänen welche nur in Viren vorkommen, wurden automatisch lila eingefärbt. Viele der detektierten Domänen sind zudem spezifisch für *Coronaviridae*.

Literatur

1. Alawi, M., Burkhardt, L., Indenbirken, D., Reumann, K., Christopeit, M., Kröger, N., Lütgehetmann, M., Aepfelbacher, M., Fischer, N. & Grundhoff, A. DAMIAN: an open source bioinformatics tool for fast, systematic and cohort based analysis of microorganisms in diagnostic samples. *Scientific reports* **9**, 1–17 (2019).
2. Westblade, L. F., van Belkum, A., Grundhoff, A., Weinstock, G. M., Pamer, E. G., Pallen, M. J. & Dunne, W. M. Role of clinicogenomics in infectious disease diagnostics and public health microbiology. *Journal of clinical microbiology* **54**, 1686–1693 (2016).
3. Flygare, S., Simmon, K., Miller, C., Qiao, Y., Kennedy, B., Di Sera, T., Graf, E. H., Tardif, K. D., Kapusta, A., Ryneerson, S., *et al.* Taxonomer: an interactive metagenomics analysis portal for universal pathogen detection and host mRNA expression profiling. *Genome biology* **17**, 1–18 (2016).
4. Weisburg, W. G., Barns, S. M., Pelletier, D. A. & Lane, D. J. 16S ribosomal DNA amplification for phylogenetic study. *Journal of bacteriology* **173**, 697–703 (1991).
5. Johnson, J. S., Spakowicz, D. J., Hong, B.-Y., Petersen, L. M., Demkowicz, P., Chen, L., Leopold, S. R., Hanson, B. M., Agresta, H. O., Gerstein, M., *et al.* Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nature communications* **10**, 1–11 (2019).
6. DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P. & Andersen, G. L. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and environmental microbiology* **72**, 5069–5072 (2006).
7. Kembel, S. W., Wu, M., Eisen, J. A. & Green, J. L. Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. *PLoS Comput Biol* **8**, e1002743 (2012).

8. You, H.-L., Chang, S.-J., Yu, H.-R., Li, C.-C., Chen, C.-H. & Liao, W.-T. Simultaneous detection of respiratory syncytial virus and human metapneumovirus by one-step multiplex real-time RT-PCR in patients with respiratory symptoms. *BMC pediatrics* **17**, 89 (2017).
9. Zhou, B., Deng, Y.-M., Barnes, J. R., Sessions, O. M., Chou, T.-W., Wilson, M., Stark, T. J., Volk, M., Spirason, N., Halpin, R. A., *et al.* Multiplex reverse transcription-PCR for simultaneous surveillance of influenza A and B viruses. *Journal of Clinical Microbiology* **55**, 3492–3501 (2017).
10. Wang, D., Coscoy, L., Zylberberg, M., Avila, P. C., Boushey, H. A., Ganem, D. & DeRisi, J. L. Microarray-based detection and genotyping of viral pathogens. *Proceedings of the National Academy of Sciences* **99**, 15687–15692 (2002).
11. Fischer, N., Indenbirken, D., Meyer, T., Lütgehetmann, M., Lellek, H., Spohn, M., Aepfelbacher, M., Alawi, M. & Grundhoff, A. Evaluation of unbiased next-generation sequencing of RNA (RNA-seq) as a diagnostic method in influenza virus-positive respiratory samples. *Journal of clinical microbiology* **53**, 2238–2250 (2015).
12. Naccache, S. N., Federman, S., Veeraraghavan, N., Zaharia, M., Lee, D., Samayoa, E., Bouquet, J., Greninger, A. L., Luk, K.-C., Enge, B., *et al.* A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. *Genome research* **24**, 1180–1192 (2014).
13. O’Leary, N. A., Wright, M. W., Brister, J. R., Ciufu, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic acids research* **44**, D733–D745 (2016).
14. Sayers, E. W., Beck, J., Brister, J. R., Bolton, E. E., Canese, K., Comeau, D. C., Funk, K., Ketter, A., Kim, S., Kimchi, A., *et al.* Database resources of the national center for biotechnology information. *Nucleic acids research* **48**, D9 (2020).
15. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology* **15**, 1–12 (2014).
16. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome biology* **20**, 257 (2019).

17. Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., Wu, C. H. & Consortium, U. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932 (2015).
18. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nature methods* **12**, 59–60 (2015).
19. Günther, T., Haas, L., Alawi, M., Wohlsein, P., Marks, J., Grundhoff, A., Becher, P. & Fischer, N. Recovery of the first full-length genome sequence of a parapoxvirus directly from a clinical sample. *Scientific reports* **7**, 1–8 (2017).
20. Zapatka, M., Borozan, I., Brewer, D. S., Iskar, M., Grundhoff, A., Alawi, M., Desai, N., Sülthmann, H., Moch, H., Cooper, C. S., *et al.* The landscape of viral associations in human cancers. *Nature genetics* **52**, 320–330 (2020).
21. Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., *et al.* Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature biotechnology* **29**, 644 (2011).
22. Peng, Y., Leung, H. C., Yiu, S.-M., Lv, M.-J., Zhu, X.-G. & Chin, F. Y. IDBA-tran: a more robust de novo de Bruijn graph assembler for transcriptomes with uneven expression levels. *Bioinformatics* **29**, i326–i334 (2013).
23. ICGC/TCGA Pan-Cancer Analysis Whole Genomes of Consortium. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
24. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology* **19**, 455–477 (2012).
25. Coronado, L., Liniger, M., Muñoz-González, S., Postel, A., Pérez, L. J., Pérez-Simó, M., Perera, C. L., Frias-Lepoureau, M. T., Rosell, R., Grundhoff, A., Indenbirken, D., Alawi, M., Fischer, N., Becher, P., Ruggli, N. & Ganges, L. Novel poly-uridine insertion in the 3' UTR and E2 amino acid substitutions in a low virulent classical swine fever virus. *Veterinary microbiology* **201**, 103–112 (2017).
26. Postel, A., Hansmann, F., Baechlein, C., Fischer, N., Alawi, M., Grundhoff, A., Derking, S., Tenhüdnfeld, J., Pfankuche, V. M., Herder, V., Baumgärtner, W., Wendt, M. & Becher, P. Presence of atypical porcine pestivirus (APPV) genomes in newborn piglets correlates with congenital tremor. *Scientific reports* **6**, 1–9 (2016).



27. Baechlein, C., Grundhoff, A., Fischer, N., Alawi, M., Hoeltig, D., Waldmann, K.-H. & Becher, P. Pegivirus infection in domestic pigs, Germany. *Emerging infectious diseases* **22**, 1312 (2016).
28. Hofmann-Sieber, H., Gonzalez, G., Spohn, M., Dobner, T. & Kajon, A. E. GenOMIC AND Phylogenetic analysis of TWO Guinea Pig Adenovirus STRAINS RECOVERED from archival LUNG tissue. *Virus Research*, 197965 (2020).
29. Fux, R., Arndt, D., Langenmayer, M. C., Schwaiger, J., Ferling, H., Fischer, N., Indenbirken, D., Grundhoff, A., Dölken, L., Adamek, M., *et al.* Piscine orthoreovirus 3 is not the causative pathogen of proliferative darkening syndrome (pds) of brown trout (*salmo trutta fario*). *Viruses* **11**, 112 (2019).
30. Christopeit, M., Grundhoff, A., Rohde, H., Belmar-Campos, C., Grzyska, U., Fiehler, J., Wolschke, C., Ayuk, F., Kröger, N. & Fischer, N. Suspected encephalitis with *Candida tropicalis* and *Fusarium* detected by unbiased RNA sequencing. *Annals of hematology* **95**, 1919–1921 (2016).
31. El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., *et al.* The Pfam protein families database in 2019. *Nucleic acids research* **47**, D427–D432 (2019).
32. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, T. L. BLAST+: architecture and applications. *BMC bioinformatics* **10**, 421 (2009).
33. Tristão Ramos, R. J., de Azevedo Martins, A. C., da Silva Delgado, G., Ionescu, C.-M., Ürményi, T. P., Silva, R. & Koča, J. CrocoBLAST: Running BLAST efficiently in the age of next-generation sequencing. *Bioinformatics* **33**, 3648–3651 (2017).
34. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* **10**, R25 (2009).
35. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput Biol* **7**, e1002195 (2011).
36. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology* **35**, 1026–1028 (2017).
37. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

38. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
39. Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., Ganapathy, G., Wang, Z., Rasko, D. A., McCombie, W. R., Jarvis, E. D., *et al.* Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature biotechnology* **30**, 693–700 (2012).
40. Giani, A. M., Gallo, G. R., Gianfranceschi, L. & Formenti, G. Long walk to genomics: History and current approaches to genome sequencing and assembly. *Computational and Structural Biotechnology Journal* **18**, 9–19 (2020).
41. Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., Hu, Y., Tao, Z.-W., Tian, J.-H., Pei, Y.-Y., *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020).
42. Amid, C., Alako, B. T., Balavenkataraman Kadhivelu, V., Burdett, T., Burgin, J., Fan, J., Harrison, P. W., Holt, S., Hussein, A., Ivanov, E., *et al.* The European Nucleotide Archive in 2019. *Nucleic acids research* **48**, D70–D76 (2020).

4 Appendix

OPEN

DAMIAN: an open source bioinformatics tool for fast, systematic and cohort based analysis of microorganisms in diagnostic samples

Malik Alawi ^{1,2}, Lia Burkhardt¹, Daniela Indenbirken¹, Kerstin Reumann¹, Maximilian Christopeit³, Nicolaus Kröger³, Marc Lütgehetmann⁴, Martin Aepfelbacher⁴, Nicole Fischer^{4,5*} & Adam Grundhoff ^{1,5*}

We describe DAMIAN, an open source bioinformatics tool designed for the identification of pathogenic microorganisms in diagnostic samples. By using authentic clinical samples and comparing our results to those from established analysis pipelines as well as conventional diagnostics, we demonstrate that DAMIAN rapidly identifies pathogens in different diagnostic entities, and accurately classifies viral agents down to the strain level. We furthermore show that DAMIAN is able to assemble full-length viral genomes even in samples co-infected with multiple virus strains, an ability which is of considerable advantage for the investigation of outbreak scenarios. While DAMIAN, similar to other pipelines, analyzes single samples to perform classification of sequences according to their likely taxonomic origin, it also includes a tool for cohort-based analysis. This tool uses cross-sample comparisons to identify sequence signatures that are frequently present in a sample group of interest (e.g., a disease-associated cohort), but occur less frequently in control cohorts. As this approach does not require homology searches in databases, it principally allows the identification of not only known, but also completely novel pathogens. Using samples from a meningitis outbreak, we demonstrate the feasibility of this approach in identifying enterovirus as the causative agent.

Nucleic acid based detection of pathogens has widely replaced culture based laboratory methods for the identification of putative pathogens in samples from patients with infectious diseases^{1,2}. These procedures are commonly amplification-based and biased because they require a correct hypothesis with regard to the specific infectious agents involved in an infectious disease. Less biased approaches interrogate highly conserved regions (e.g. 16S rRNA bacteria and ITS sequences for fungi) or employ amplification protocols with pan-primer mixes for individual viral families³⁻⁵. Alternatively, multiplex PCR approaches with multiple primer sets and detection probes in a single tube may be used for specific infectious syndromes (e.g. encephalitis, acute gastroenteritis, pneumonia or severe respiratory distress syndrome). Still, a priori knowledge of specific pathogen is necessary and very often these methods, although highly sensitive, remain negative.

Unbiased next-generation sequencing (NGS) of diagnostic samples is now widely considered a key technology that will fundamentally improve infectious disease diagnostics^{2,6-9}. Due to the principal potential to identify not only known but also novel pathogens, such methods are also expected to strengthen the level of preparedness for future outbreaks of emerging pathogens¹⁰. Decreasing reagent cost and availability of affordable bench top sequencing instruments with relatively low infrastructure demands have promoted the establishment of

¹Heinrich-Pette-Institute (HPI), Leibniz Institute for Experimental Virology, Research Group Virus Genomics, Hamburg, Germany. ²Bioinformatics Core, University Medical Center Hamburg-Eppendorf, Hamburg, Germany. ³Department of Stem Cell Transplantation, University Medical Center Hamburg-Eppendorf (UKE), Hamburg, Germany. ⁴Institute of Medical Microbiology, Virology and Hygiene, University Medical Center Hamburg-Eppendorf (UKE), Hamburg, Germany. ⁵German Center for Infection Research, DZIF, partner site Hamburg-Borstel-Lübeck-Riems, Germany. *email: nfischer@uke.de; adam.grundhoff@leibniz-hpi.de

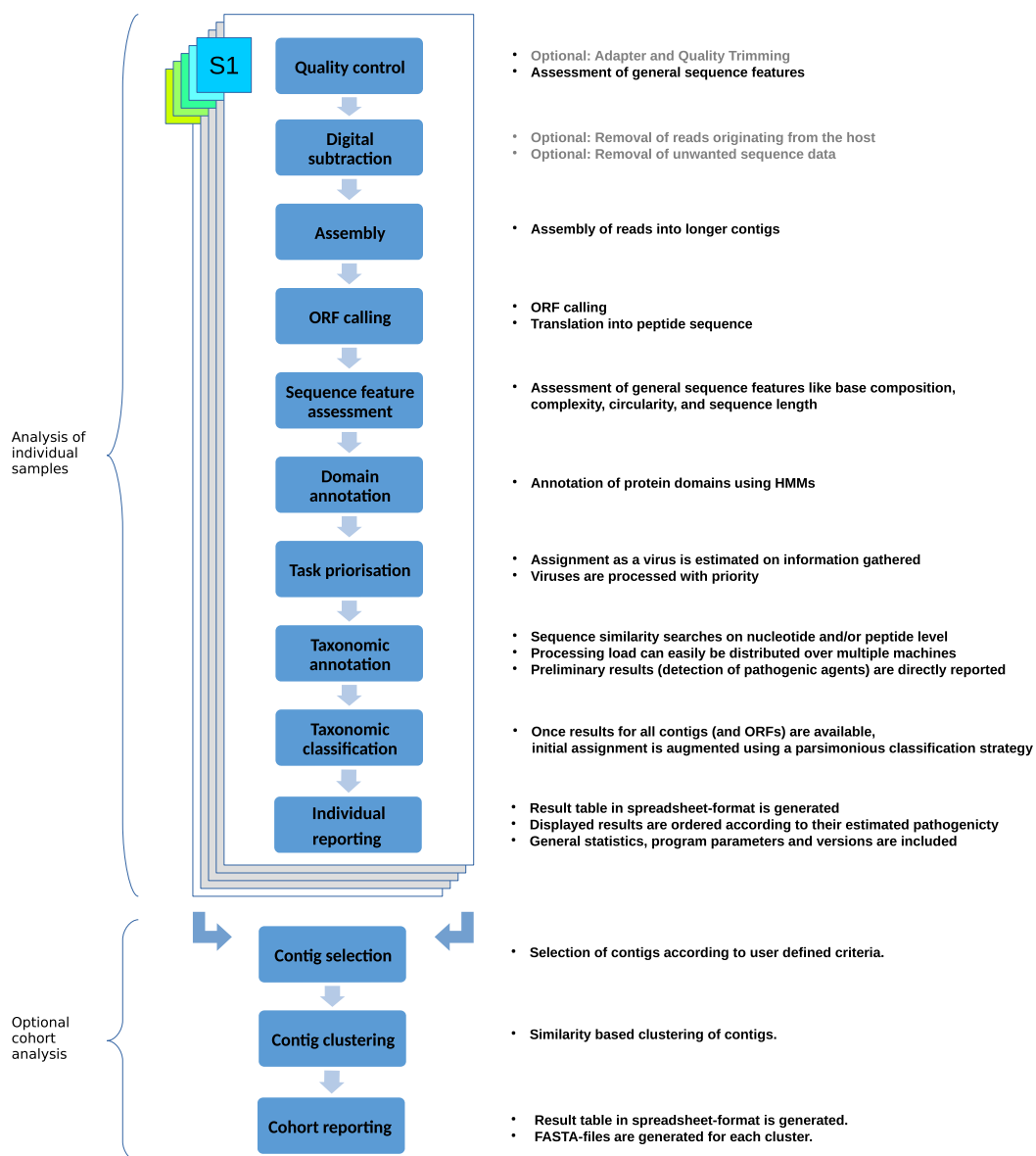


Figure 1. Schematic representation of the data processing steps performed by DAMIAN. Depicted are the individual modules in the DAMIAN workflow starting with FASTQ as input files for the analysis. Individual samples are generally processed independently first. An optional cohort analysis can later be performed on any number of previously processed samples.

next-generation sequencing platforms in many hospitals or microbiology laboratories and make this technique highly attractive to improve pathogen detection in diagnostics^{11–13}. However, there is still a lack in open source bioinformatic tools that are specifically designed for clinical settings.

Here we describe a user-friendly open source software, which enables clinical personnel without a background in bioinformatics to accurately, and rapidly identify potentially pathogenic agents in clinical specimen. Notably, DAMIAN (Detection & Analysis of viral and Microbial Infectious Agents by NGS) goes beyond taxonomic classification of sequence reads. Its capabilities include functional sequence analysis, which allows for reliable results even in the case of truly novel emerging pathogens not represented in sequence databases. Furthermore, the ability to process cohorts make it a valuable tool for the analysis of outbreak samples. To the best of our knowledge, this is the first software for the detection of pathogens to provide such features. Here, we demonstrate that DAMIAN achieves excellent detection capabilities and an unprecedented level of guidance in the interpretation of analysis results.

Results

Description of DAMIAN features and data processing steps. DAMIAN provides capabilities to rapidly identify known and novel infectious agents in samples of various sources. It integrates all required processing steps, ranging from the quality control of raw reads to the generation of comprehensive reports, into a single user-friendly software system. Being intended for the employment in clinical diagnostics, DAMIAN does require

neither specialized computational infrastructure nor expertise in bioinformatics to accomplish its tasks. It works for both DNA and RNA samples and, if desired, takes into account almost any host organism to subtract background reads.

Many taxonomic classification tools (e.g. Taxonomer, SURPI or Kraken^{11,14,15}) aim at taxonomically classifying single reads. Such an approach is able to deliver results quickly and, at least in cases where adequate reference sequences are available also allows for a solid classification. By contrast, DAMIAN pursues a different strategy and assembles reads into longer contigs prior to classification and annotation. The longer sequences increase the sensitivity and specificity of sequence similarity searches and therewith the quality of taxonomic assignments. Moreover, they allow for a functional annotation, which provides valuable information even when sequence similarity searches do not yield significant matches, and permit cross-comparison of sequence contig signatures across multiple sample cohorts.

The minimum requirement for starting an analysis with DAMIAN are reads in (gzip-compressed or uncompressed) FASTQ-format. Any number and combination of paired-end reads and single-end reads is supported. In the following, we briefly describe the features and processing steps in their actual order of execution during analysis with DAMIAN (Fig. 1).

Upon starting, DAMIAN performs checks to ensure that all requirements for successfully conducting an analysis are met. User input, database connectivity, file permissions and software dependencies are validated and at the same time information, like software versions and parameters given on the command line, are aggregated and stored.

Quality control and self-documentation. DAMIAN automatically removes low quality bases and sequencing adapter sequences. Prior processing with external tools is not required. DAMIAN automatically documents every single analysis step and provides gathered information in its analysis results. This information comprehensively describes an analysis and allows for exact reproduction. During the quality control step, for example, information is collected on how quality and adapter trimming effects read properties. Although not explicitly mentioned in the following paragraphs, a similar behavior was implemented for all analysis steps.

Digital subtraction and abundance estimation of unwanted sequence reads. In general, sequence reads originating from the host organism are removed and counted. However, this is optional and DAMIAN can be used with any number of different reference sequences or none at all. DAMIAN is able to discriminate between RNA and DNA data and suitable reference sequences can be selected accordingly.

Assembly and assessment of basic contig features. Reads remaining after the preceding steps are assembled into larger contigs. Features like length, circularity, GC-content and sequence complexity are determined for each contig and sequences of ORFs are translated into amino acid sequences.

Functional annotation and contig ranking. The amino acid sequences are screened for known protein domains. The domains are classified according to the taxonomic entities they are associated with. Some domains, for example, are only found in viruses while others are specific to bacteria or fungi. As the protein domains are functional regions, they can provide unmistakable results even when BLAST searches yield no significant matches with the sequences of known pathogens. Additionally, information on functional domains allows to rank the contigs for subsequent processing such that contigs potentially originating from pathogenic agents are processed first.

Taxonomic assignment. DAMIAN employs the complete NCBI nt and nr databases to perform classifications. Searches with nucleotide and derived amino-acid sequences can be performed independently, iteratively or redundantly. Preliminary results are reported whenever a contig yields significant matches to a known microbial or viral agent. In addition to lowest common ancestor (LCA) based taxonomic assignments, DAMIAN also incorporates a two-pass method for taxonomic assignment. It aims at determining which species are present in a sample based on aggregated information from all contigs instead of assigning each contig individually.

Reporting. DAMIAN provides a comprehensive report in spreadsheet format for each sample (see *Diagnostic Application* section and Datasets S2–S11 for examples). The main page provides an overview of detected taxonomic entities. Entries are sorted and color-coded to allow for a quick identification of potentially pathogenic agents. The color-code depicts six different categories, red, pink, light blue, dark blue, grey and black. The first category (red) contains entries, which were classified as viruses based on sequence similarity and protein domains. For the second (pink) and third category (light blue) there is only evidence for viral sequences from either sequence similarity or protein domains. Phages are generally listed within a separate category (dark blue). The fifth category is for known artifacts or contaminants (grey), which can be defined by the user and the sixth and final category for everything else, bacteria, fungi and parasites (black). Additionally, DAMIAN enables its users to further investigate sequences, which did not yield significant alignments (see cohort analysis below).

The report is interactive and links entries of the main page to detailed views. These views display detailed data regarding the corresponding contigs, ORFs and protein domains. Additionally, nucleotide and amino acid sequences can be accessed. Other pages of the report contain information on general statistics like number of reads, amount of reads originating from the host and the size sequenced fragments. Program versions and parameters used are also part of the report. It is not necessary to wait for the preceding steps to complete before generating a report. Preliminary reports, integrating all information available so far, can be generated at any time.

sample ID ^a	diagnostic entity	detected pathogen(s)	time ^b
104	bronchoalveolar lavage	Influenza A	73
3157	bronchoalveolar lavage	Influenza A	40
4505	bronchoalveolar lavage	Chlamydomphila psittaci	n/a
9790	stool	human Parechovirus	44
9792	stool	Sapporovirus	38
		human Parechovirus	38
1	stool	Norwalk Virus	162
7653	cerebrospinal fluid	Enterovirus B	17
SRR1553464	serum	Zaire Ebolavirus	13
SRR533978	serum	Bas Congovirus	18
SRR1564804	plasma	Chlamydomphila psittaci	n/a

Table 1. Time frame in which clinically relevant results were obtained by DAMIAN. ^aDiagnostic sample or public available dataset (SRR1553464, SRR533978, SRR1564804) analyzed by DAMIAN. ^bTime (in minutes) until the first report of a putative pathogen was received. n/a: not applicable; time frames are only calculated for viral contigs since DAMIAN prioritizes viral sequences. The analysis was performed using 12 threads of a server with two Intel Xeon E5-2687W v3 CPUs.

Cohort analysis. The optional cohort-based analysis allows the identification of sequences which may originate from pathogenic agents shared among groups of samples from individuals showing a given disease phenotype. This analysis does not depend on reference databases, taxonomic assignments or similar prior knowledge. Rather, the user assigns any number of samples to a group of known positives, known negatives or of unclassified samples. For example, all samples that belong to a suspected outbreak can be assigned to the group of positives while samples, which are known to be unrelated to the outbreak, would be assigned to the group of negatives. Finally, samples for which it is unsure whether they are part of the outbreak could be assigned to the group of unclassified samples. The pipeline then performs pairwise BLAST alignment amongst all assembled contigs and sorts them into bins according to their sequence similarity. Within each sequence cluster, a score reflecting the degree to which the cluster is preferentially associated with the positive phenotype is calculated. By sorting the clusters according to their score, the user can easily identify those contigs, which are most likely linked to the phenotype in question, and thus select the most promising candidates that may represent causally related pathogenic agents. Results are reported in spreadsheet format and additionally FASTA files are generated from contig sequences for each cluster. While the results table (see Supplementary Dataset S1 for an example) contains taxonomic assignments for those clusters in which individual contig members could be classified, the clustering itself is completely independent of the success or failure of contig classification. Hence, this approach allows for the identification of completely novel pathogens, provided that they are overrepresented in the positive phenotype group.

Diagnostic application and comparison with existing software (Taxonomer, PathoScope and metaMix). To verify the ability of our tool to detect pathogens in diagnostic and putative outbreak settings, we applied DAMIAN to a number of specimens derived from patients suspected to suffer from common community- or hospital acquired infections or in the context of public health emergencies (Table 1). Results were compared to those results obtained via Taxonomer BETA and PathoScope pipelines^{11,16,17} (Tables 2–5). While Taxonomer, PathoScope and DAMIAN each incorporate all analysis steps, metaMix requires the results of sequence similarity searches as an input. The way the pre-processing is performed may immediately impact the results of metaMix. Here we used an IDBA-UD assembly and MEGABLAST results to perform the analysis. IDBA-UD was employed since it is also integrated in DAMIAN, and MEGABLAST was used to allow the analysis to complete within a similar time frame as the other tools. metaMix performance may improve if it is run with different, yet computationally more demanding, pre-processing steps. We included it in the comparison, because like DAMIAN and unlike the other two aforementioned tools, it is able to perform an analysis, which is based on contigs. All specimens were pre-analyzed by state of the art diagnostic tests as part of routine analysis procedures. The routine specimens included two respiratory (bronchoalveolar lavages (BALs) 104 and 3157) and one cerebrospinal fluid samples (CSF 7653), while the public health emergency-related specimens comprised one respiratory (BAL 4505) and three stool samples (1, 9792 and 9790). For all samples, we constructed strand specific RNA-Seq libraries from total nucleic acids extracted in a routine diagnostic environment. Libraries were multiplex sequenced on MiSeq or HiSeq2500 instruments with 2.4 to 3.3 million or ~25 million reads per sample, respectively. In general, DAMIAN reported first results after 10–20 minutes. Pathogenic agents were reported within less than an hour in most cases (Table 1).

Respiratory (BAL) samples. DAMIAN readily detected Influenza A in the two routine diagnostic samples investigated in this study (BALs 104 and 3157). The presence of Influenza A was first called after 73 and 40 minutes in samples 104 and 3157, respectively, and inspection of the analysis report identified H1N1 and H3N2 strains as the most likely source of infection (Table 2, Supplementary Datasets S2 and S3). As expected for BAL material, all samples exhibited high abundance of human sequences, with significant variation between the individual samples ranging from approximately 52 to 99% of sequence reads (Figs 2, 3 and Table 2). Routine diagnostic PCRs for a standard panel of respiratory viruses was performed in parallel and yielded positive Ct values of 26 and 30 for Influenza A in BALs 104 and 3157, respectively. All other respiratory viruses included in the PCR panel (hPIV

	104	3157	4505
	23,828,285 reads 99,23% human sequences	3,265,314 reads 51,74% human sequences	2,370,210 reads 98,13% human sequences
DAMIAN	<ul style="list-style-type: none"> ✓ Influenza A (39,810 reads; 42%) • H1N1 all 8 segments (8 contigs) 97–100% id. ○ Influenza A ○ A/Singapore/TT198/2011 (H1N1) ○ A/Swine/France/71-130116/2013 (H1N1) ○ A/Swine/France/71-130116/2013 (H1N1) ○ A/Singapore/TT198/2011 (H1N1) ○ A/Santa Clara/YGA_03065/2013(H1N1), ○ A/Arizona/M2/2012(H1N1) ○ A/Swine/France/71-130116/2013 (H1N1) ✓ Candida albicans (14,249 reads, 15.15%) 	<ul style="list-style-type: none"> ✓ Influenza A (1,886 reads, 1.57%) • H3N2 all 8 segments (8 contigs) 99% id. ○ PB2, A/Connecticut/Flu140/2013(H3N2) ○ PB1, A/Connecticut/Flu140/2013(H3N2) ○ PA, A/Connecticut/Flu140/2013(H3N2) ○ HA, A/Connecticut/Flu140/2013(H3N2) ○ NP, A/Connecticut/Flu140/2013(H3N2) ○ NA, A/Connecticut/Flu140/2013(H3N2) ○ M2, M1, A/Connecticut/Flu140/2013(H3N2) ○ NEP, NS1, A/Connecticut/Flu140/2013(H3N2) ✓ human parainfluenza 3 virus (1 read) 99% id. ✓ human herpes simplex virus 1 (4 reads) 99% id. ✓ Candida albicans (10,207 reads, 7.65%) 	<ul style="list-style-type: none"> ✓ Chlamydomophila psittaci (237 reads; 4.13%) 100%id. • Chlamydomophila psittaci 6BC, 4 contigs, 16S and 23S rRNA • Chlamydomophila VS225, 1 contig, 16S rRNA • Chlamydomophila Mat116, 1 contig, 16S rRNA
Taxonomer BETA	9,900,000 reads sampled [‡] ; 5% classified Bacteria: 62,128 reads; Viruses: 10,723 reads; Fungi: 86 reads <ul style="list-style-type: none"> ✓ Influenza A (1,227 reads) • H1N1 (1,227 reads) ✓ α-retrovirus (9,677 reads) ✓ dsDNA virus (505 reads) 	3,200,000 reads samples; 13% classified Bacteria: 300,309 reads; Viruses: 7,686 reads; Fungi: 17,878 reads <ul style="list-style-type: none"> ✓ Influenza A (1,433 reads) • H3N2 (354 reads) ✓ Human parainfluenza 3 virus (13 reads) ✓ α-retrovirus (907 reads) ✓ Caudovirales (1,299 reads) ✓ Herpesviridae (184 reads) ✓ Candida albicans (15,537 reads) 	2,300,000 reads samples; 4% classified Bacteria: 24,960 reads; Viruses: 950 reads; <ul style="list-style-type: none"> ✓ Chlamydia (75 reads) • Chlamydia psittaci (33 reads) • Chlamydia trachomatis (30 reads) ✓ Proteobacteria (510 reads) ✓ Firmicutes (33 reads) ✓ α-retrovirus (169 reads) ✓ Herpesviridae (52 reads)
PathoScope	45,576 aligned reads; 2,234 hits ✓ Influenza A (12,234 reads) • Subtypes H3N2; H5N1; H1N1; H9N2; H2N2 ✓ Hepatitis C (224 reads) • Genotype 2; 1; 6 ✓ Encephalomyocarditis Virus (115 reads)	184,719 aligned reads; 2,434 hits ✓ Influenza A (1,337 reads) ✓ Subtypes H3N2 ✓ Avian leukosis virus (1,256 reads) ✓ human herpes simplex virus 1 (102 reads) ✓ Veillonella parvula (130,363 reads) ✓ Enterococcus faecium (21,178 reads)	4,408 aligned reads; 1,752 hits ✓ Chlamydomophila psittaci (23.83 reads)#
metaMix	26 hits; 7,328,046 human reads ✓ Influenza A (1 contig; 46,698 reads) • H1N1, 1 contig; A/Canela/LACENRS-418/2013 ✓ Candida albicans SC5314 2 contigs (601,527 reads) Bacteria 3 contigs (372 reads)	44 hits; 201,764 human reads ✓ Influenza A (3 contigs; 2,244 reads) • H3N2, 2 contigs; A/Bage/LACENRS-205/2013; A/Porto Alegre/LACENRS-275/2013 ✓ Candida albicans, 1 contig (97,816reads) Bacteria; 15 contigs (109,048 reads)	16 hits; 142,965 human reads ✓ Chlamydomophila psittaci (1 contig; 252 reads)

Table 2. Comparison of BAL sample analysis results obtained by DAMIAN, Taxonomer BETA, PathoScope and metaMix. [‡]Files >5GB are not supported by taxonomer BETA version; 10,000,000 reads were randomly sampled to meet 5 GB maximum size for upload. *Files >5GB are not supported by taxonomer BETA version; 10,000,000 reads were randomly sampled to meet 5 GB maximum size for upload. # fractional read abundance given by PathoScope.

1–3, hRV, Enteroviruses, Adenovirus, hRSV) were negative (Suppl. Table S1). The significantly lower Ct value observed for Influenza A in BAL 104 is in agreement with the fact that the relative fraction of Influenza A reads was much higher in this sample compared to BAL 3157 (approximately 42 and 1.6%, respectively). The assembled contigs allowed recovery and strain assignments for all influenza genomic segments (Fig. 2A,B, Table 2), thus permitting immediate identification of putative reassortment events between the individual segments. We performed lineage assignment with the FluGenome tool¹⁸, which reported genotype H1N1 (C (PB2), D (PB1), E (PA), 1A (HA), A (NP), 1F NA), F (MP), 1A (NS)) for sample 104 and H3N2 (A, D, B, 3A, A, 2A, B, 1A) for sample 3157. In addition to Influenza A virus, DAMIAN detected a putative coinfection with *Candida albicans* (15.13% and 7.65% of all non-host reads, respectively; see Fig. 2 and Table 2) in both BAL samples. BAL 3157 also displayed one shorter contig (505nt) unambiguously assigned to the human parainfluenza virus 3 genome (sequence identity 98.75%), and a shorter contig (458nt) with 99.36% identity to human herpesvirus 1 (HSV-1; Fig. 2B, Table 2). The co-infections with both *Candida albicans* and parainfluenzavirus 3 were confirmed by conventional diagnostic methods (fungal culture and PCR). We also included a third BAL sample (BAL 4505) which was one out of three samples of a suspected infectious disease outbreak published earlier^{2,6} in our analysis. In accord with our previous results, DAMIAN correctly identified *Chlamydomophila psittaci* and assigned 4.13% of all non-host reads to rRNA moieties originating from the intracellular bacterium (Fig. 3, Table 2).

The comparative analysis results obtained with Taxonomer, PathoScope and metaMix for the three BAL sample datasets are shown in Table 2. While all tools identified *Chlamydomophila psittaci* in sample BAL 4505, they differed substantially in the number of assigned reads (237 reads for DAMIAN, 75 reads for Taxonomer, 23.83 reads for PathoScope and 252 for metaMix). The same was true for Influenza A in sample 104 (1,227, 12,234 and 46,698 reads, respectively). Only DAMIAN was able to assign the correct genotype and strain for each individual

	9790	9792	1
	1,667,291 reads 0,64% human sequences	1,347,375 reads 0,16% human sequences	23,292,070 reads 1,36% human sequences
DAMIAN	<ul style="list-style-type: none"> ✓ human parechovirus 6 (3 contigs, 132 reads, 0.1%) 96–97% id. ✓ Bacteroides ✓ Bifidobacterium 	<ul style="list-style-type: none"> ✓ Sapporovirus (10,028 reads, 14.62%) • Sapovirus Hu/G1/BE-HPI01/DE/2012 ✓ human parechovirus 1 (3 contigs; 370 reads, 0.54%) 90–97% id ✓ Bifidobacterium 	<ul style="list-style-type: none"> ✓ Norwalk Virus (1,163,565 reads, 8.22%) • Primate Norovirus strain simianNoV-nj, complete genome 98% id (1 contig, 757,078 reads)# • Chiba Virus genomic RNA, complete genome 93% id (1 contig, 402,355 reads) • Norovirus Hu/GII.P16/GII.13/New/Taipei/13-BA-1/2013/TW complete genome 99% id (1 contig, 4,132 reads) ✓ Bacteria
	1,600,000 reads samples; 77% classified Bacteria: 1,207,356 reads; Viruses: 2,493 reads; Fungi: 419 reads	1,300,000 reads samples; 86% classified Bacteria: 925,978 reads; Viruses: 13,005 reads;	9,900,000 reads sampled ^d ; 85% classified Bacteria: 8,200,128 reads; Viruses: 100,803 reads; Fungi: 86 reads
Taxonomer BETA	<ul style="list-style-type: none"> ✓ human parechovirus (89 reads) • human parechovirus 6 (10 reads) • human parechovirus 1 (7 reads) ✓ α-retrovirus (1166 reads) ✓ ds DNA viruses (1,334 reads) ✓ ss DNA viruses (353 reads) ✓ Bacteroidetes (492,974 reads) ✓ Actinobacteria (99,321 reads) ✓ Proteobacteria (89,830 reads) ✓ Firmicutes (388,188 reads) 	<ul style="list-style-type: none"> ✓ Calciviridae (7,913 reads) • Sapporovirus (562 reads) GI (80 reads); GI.2 (53 reads) ✓ human parechovirus 1 (46 reads) ✓ Pandoravirus (247 reads) ✓ Actinobacteria (538,844 reads) Bifidobacteriales (463,853 reads) ✓ Firmicutes (101,073 reads) 	<ul style="list-style-type: none"> ✓ Calicivirus (2,570 reads) • GI/10360/2010/NM 950 reads • GI/DH1751/2009/IND 30 reads • GI.3/13440/2007/RJ/BRA 80 reads • GI.3/C9/GF/1978 10 read • GI.4/1643/2008/US 70 reads • GI.4/15waterBS/T11/ITA 10 read • GII.4 Beijing 40 reads ✓ α-retrovirus (150 read) ✓ Parvovirus NIH-CQV (10 read) Bacteria
PathoScope	1,226,383 aligned reads; 2,210 hits <ul style="list-style-type: none"> ✓ α-retrovirus (113 reads) ✓ HHV8 (7 reads) ✓ Pepper mild mottle virus (2 reads) ✓ Actinobacteria, Bifidobacteriaceae (166,773 reads) ✓ Bacteroidetes (784,119 reads) ✓ Firmicutes, Clostridiales (138,760 reads) 	1,116,813 aligned reads, 2,205 hits <ul style="list-style-type: none"> ✓ Sapovirus_Hu/Dresden/pJG-Sap01/DE (831 reads) ✓ human parechovirus (19 reads) ✓ α-retrovirus (16 reads) ✓ human herpesvirus 6A (2 reads) >900,000 reads Bifidobacterium 	16,713,832 aligned reads; 2,558 hits <ul style="list-style-type: none"> ✓ Norovirus GI (20,880 reads) ✓ human papillomavirus (8 reads) ✓ polyomavirus (4 read) ✓ Hepatitis C Virus (35 reads) ✓ Human Herpesvirus (89 reads) ✓ α-retrovirus (2,423 reads) >5,000,000 reads Bacteroides
metaMix	88 hits; 1,771 human reads Bacteria; 658,685 reads Bacteroides Bifidobacterium	45 hits; 0 human reads <ul style="list-style-type: none"> ✓ Human parechovirus (1 contig; 398 reads) Bifidobacterium 	138 hits; 0 human reads <ul style="list-style-type: none"> ✓ Norovirus GI (1 contig; 1,169,366 reads) ✓ Norovirus Hu/GII.P16/GII.13/New/Taipei/13-BA-1/2013/TW (1 contig; 3,912 reads) Circoviridae (1 contig; 5,433 reads)

Table 3. Comparison of stool sample analysis results obtained by DAMIAN, Taxonomer, PathoScope and metaMix. *Files >5GB are not supported by taxonomer BETA version; 10,000,000 reads were randomly sampled to meet 5 GB maximum size for upload. # Genbank entry KX396056 is identical to NC_031324 describing a human norovirus in diarrhetic chimps; next closest assignment NC_039897.1, human Norovirus GI, 92% sequence identity.

segment. PathoScope and Taxonomer were both unable to differentiate between H1N1 and H3N2 in samples 104 or 3157, respectively. MetaMix correctly assigned H1N1 to one contig. Furthermore, the observed co-infections of *Candida albicans* and parainfluenzavirus 3 were only identified by DAMIAN or Taxonomer for sample 3157 (Table 2), whereas co-infections in sample 104 were detected by DAMIAN and metaMix.

Stool samples. We included three stool samples collected during a large outbreak of acute gastroenteritis (AGE) occurring in fall of 2012 in Germany^{19,20}, in our comparative analysis (Figs 4, 5 and Table 3). RNA from two samples (9790 and 9792) was sequenced with approximately 1.5 million reads per sample on a MiSeq instrument, while RNA material from the third (sample 1) was sequenced at a depth of 23.3 million on a HiSeq instrument. As expected for most stool samples²¹, only few host sequences were present (generally between 0.2 and 1.4%). Contigs aligning to caliciviral sequences were assembled in two of the three libraries: Sample 1 contained Norovirus (hNoV) sequences, whereas Sapovirus sequences were detected in sample 9792. In both cases, contigs representing complete or near-complete caliciviral genome sequences were recovered. In sample 1, inspection of the contigs furthermore readily revealed co-infection with three Norovirus strains. Sequences were assigned to two different genotype I strains (98.35% and 92.91% sequence identity to primate norovirus strain Simian NoV-nj (gb|KX396056) and the next closest relative, Chiba virus (gb|AB042808), respectively), and a third contig representing recombinant norovirus of genotype GII.16/GII.13 with 98.79% sequence identity to the Taipei/13-BA-1 isolate (gb|KM036380) (Fig. 5).

Interestingly, samples 9790 and 9792 also contained reads from picornaviruses with significant nucleotide homologies to human parechovirus type 6 (hPeV6) or human parechovirus type 1 (hPeV1) (Table 3). Sample 9792 yielded three contigs of 1,648; 2,130 and 3,463 nt covering approximately 95% of the most closely related hPEV1 strain (97.44%, 90.59% and 97.22% sequence identity to isolate 550163, accession GQ183021.1, respectively). In

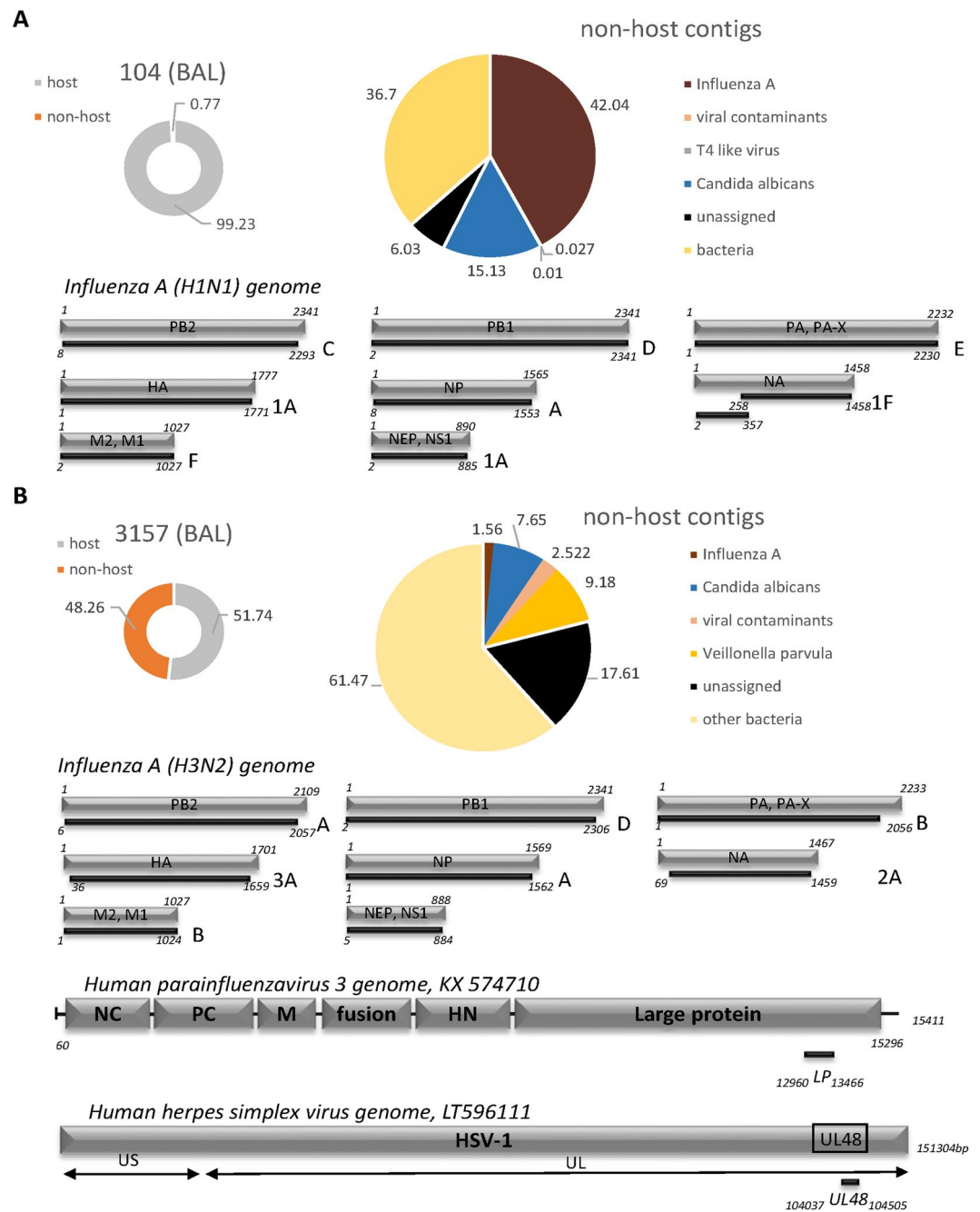


Figure 2. Application of DAMIAN to RNA-Seq libraries from diagnostic BAL samples from patients with viral respiratory infections. Donut shaped charts represent the distribution of host (grey) versus non-host (orange) reads. The pie chart illustrated the taxonomic classification of non-host reads; represented are the relative abundance of contigs assigned to these species. Reads not aligning to sequences in the NCBI database are indicated in black, bacterial sequences are represented in yellow, viral contaminants are shown in pink. The pathogen most likely contributing to the clinical symptoms is indicated in read. In each sample, the contigs of the putative pathogen identified in the sample are aligned to the closest relative: (A) Influenza A, H1N1 (full-length segments); (B) Influenza A, H3N2 (full-length segments); PIV3 and HSV-1.

sample 9790, contigs of 1,010; 1,063 and 3,650 nt aligned to approx. 80% of human parechovirus type 6 (isolate 2005-823, accession EU077518.1) with 96.43%, 95.83% and 96.63% sequence identity.

Similar to the respiratory samples, the stool sample datasets were also analyzed by Taxonomer, PathoScope and metaMix. Results are summarized in Table 3. DAMIAN, Taxonomer and PathoScope tools identified Sapovirus GI together with human parechovirus in sample 9792, but only DAMIAN and Taxonomer specified the human parechovirus as a type 1 strain. However, metaMix did not identify Sapovirus under the conditions used. The tools identified different Sapovirus strains (see Table 3) with DAMIAN identifying Sapovirus Hu/G1/BE-HPI01/DE/2012, the sequence which was originally identified with DAMIAN from this sample and submitted

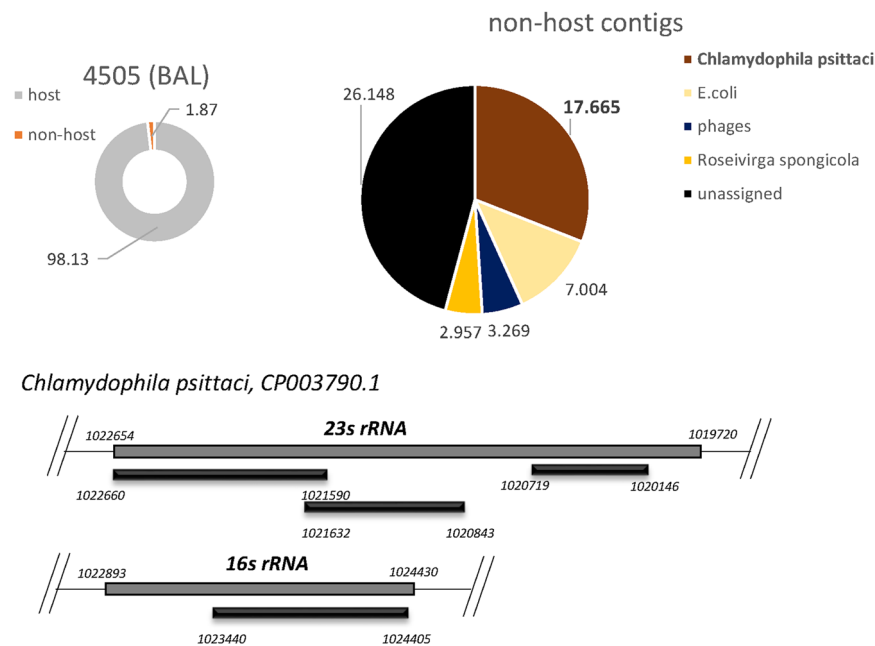


Figure 3. Application of DAMIAN to RNA-Seq libraries from diagnostic BAL samples from patients with bacterial respiratory infections. Similar to Fig. 2, the pie charts represent the distribution of host and non-host reads (left) and the taxonomic classification of non-host reads (right). The contigs of the putative pathogen identified are aligned to the closest relative, *Chlamydomphila psittaci*.

to Genbank (accession number JX993277.1). Taxonomer reported Sapovirus Hu/GI.2/BR-DF-01/BRA/2009 and PathoScope listed Sapovirus Hu/Dresden/pJG-Sap01/DE (GenBank accession number NC_006269.1) instead, with the latter showing 73% sequence identity and 84% coverage to the original Hu/G1/BE-HPI01/DE/2012 sequence present in sample 9792.

DAMIAN, taxonomer both identified human parechovirus sequences in sample 9790. However, the three contigs assembled by DAMIAN unequivocally aligned to human parechovirus type 6, whereas Taxonomer assigned 46 sequence reads to human parechovirus type 1. PathoScope and metaMix did not detect any parechovirus sequences in sample 9790 (Fig. 4, Table 3) at all.

The fact that DAMIAN assembled full-length contigs for 3 different norovirus genotypes in sample 1 suggests that this patient acquired an infection in the course of the 2012 norovirus outbreak, the largest recorded food-borne outbreak in Germany with more than 4,000 cases registered by the public health agencies^{19,20}. Most of the samples analyzed during this outbreak showed co-infection with multiple Norovirus genotypes, indicative of massive fecal contamination of food sources representing the origin of the outbreak^{19,20}. In accordance with the public health data, DAMIAN recovered two discrete full-length Noroviruses of genotype I as well as a recombinant GII.16/GII.13 genome from the sample. Together, over one million reads were mapped to the three genomes. MetaMix successfully classified two contigs as Calicivirus sequences of genotypes GI and recombinant GII.16/GII.13, with the GI sequence being much more abundant compared to the recombinant genotype II. In contrast, Taxonomer assigned 2,570 reads to seven different Norovirus strains of genotypes I and II, whereas PathoScope classified 20,880 reads as originating exclusively from norovirus genotype I (Table 3).

CSF samples. We included one routine diagnostic CSF sample in the comparison. The sample was submitted by the clinic with the request to detect viruses known to induce encephalitis in immune competent patients. Parallel to quantitative PCR for HSV, Enteroviruses, Mumps, Measles and Rubella, the sample was analyzed by DAMIAN, Taxonomer, PathoScope and metaMix. DAMIAN and metaMix both reported Echovirus 30, a call that is concordant with results obtained by diagnostic PCR (Supplementary Table S1) and subsequent Sanger sequencing of the 250 bp fragment. Two contigs covering nearly the complete genome, were recovered (Fig. 6). Taxonomer identified 123 reads as Enterovirus B, with 10 reads assigned to Coxsackievirus B2 and 49 reads to Enterovirus 30. PathoScope identified Enterovirus sequences (39 reads in total), however none of the reads was assigned to Echovirus 30 (Table 4).

SRR samples. In addition to the the diagnostic samples collected in this study, we applied DAMIAN to three datasets (SRR533978, SRR1553464 and SRR1564804) which have been which have been used by Flygare and colleagues to evaluate the ability of Taxonomer to detect viruses in public health emergency samples¹¹. Similar to our analysis of CSF, stool and BAL samples we compared the DAMIAN results of these datasets to those obtained with Taxonomer, PathoScope and metaMix (Table 5, Supplementary Fig. S1A–C and Suppl. Datasets S9–S11). SRR533978 represent RNA-Seq data from a serum of a patient with hemorrhagic fever caused by Bas Congo Virus (Suppl. Fig. S1A). SRR1553464 is a plasma sample from a patient with Ebola virus infection (Suppl. Fig. S1B), and

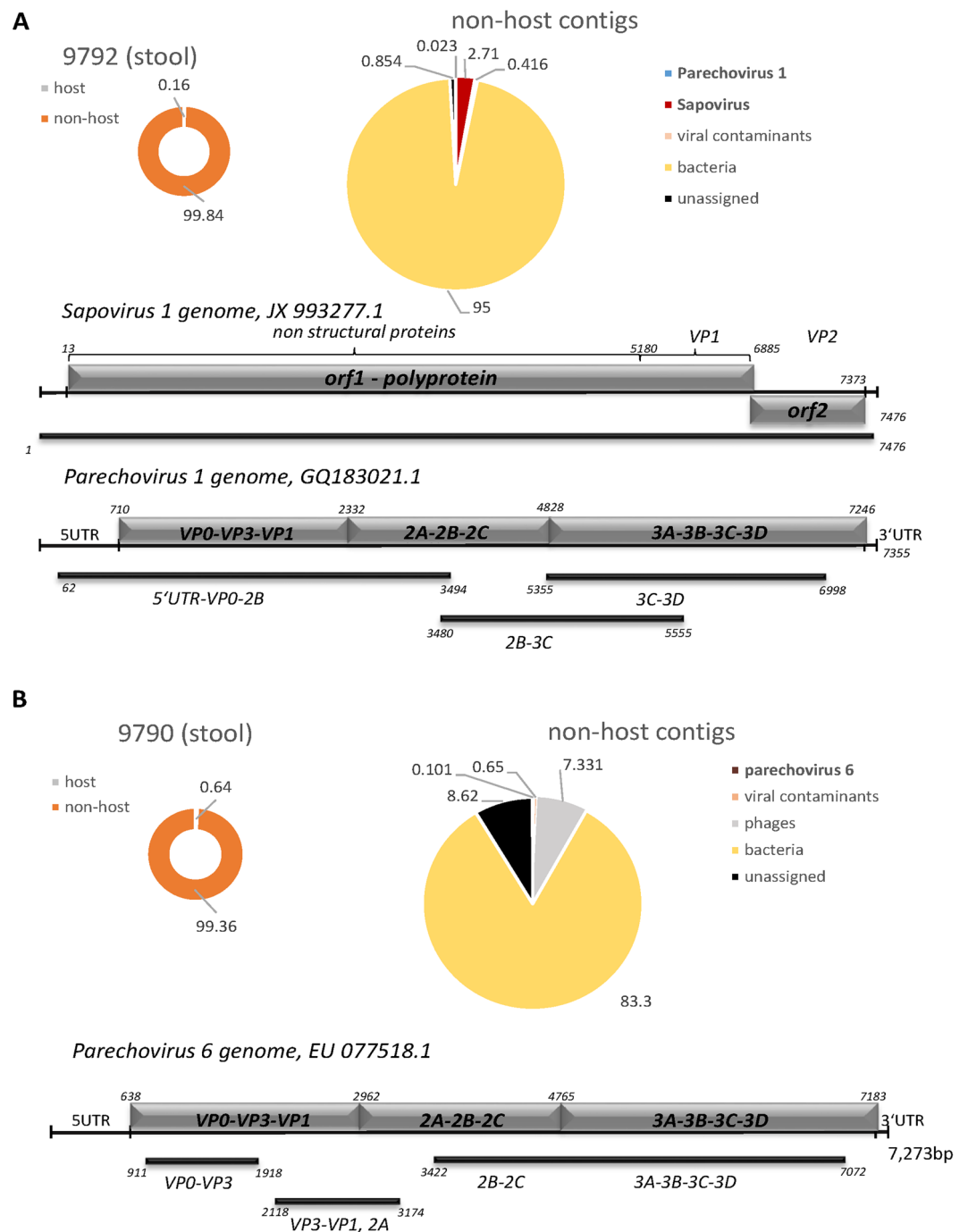


Figure 4. Application of DAMIAN to RNA-Seq libraries from diagnostic stool samples from patients with acute gastroenteric disease. Sequences are depicted as described in Fig. 2. The putative pathogens identified are (A) Sapovirus 1, Parechovirus 1 and (B) Parechovirus 6.

SRR1564804 represent a plasma sample from a patient with *Chlamydomphila psittaci* infection (Suppl. Fig. S1C). All tools identified Bas Congo Virus in sample SRR533978, Ebolavirus Zaire in SRR1553464 and *Chlamydomphila psittaci* as well as GB-Virus C in SRR1564804. In the case of the viral infections, DAMIAN recovered whole viral genomes for Bas Congo Virus (7 contigs, 467 bp–4,977 bp) and Ebolavirus (1 contig, 18,839 bp). GB-Virus C in sample SRR1564804 was only represented by two small contigs of ~600 bp, indicating it may have been present in relatively low copy numbers. Detection of *Chlamydomphila psittaci* in the sample was based on contigs aligning to 16S and 23S rRNA. Differences between the individual tools were observed with regard to the number of reads assigned to the individual taxons. In addition, only DAMIAN, PathoScope and metaMix identified equine infectious anemia virus in SRR1564804.

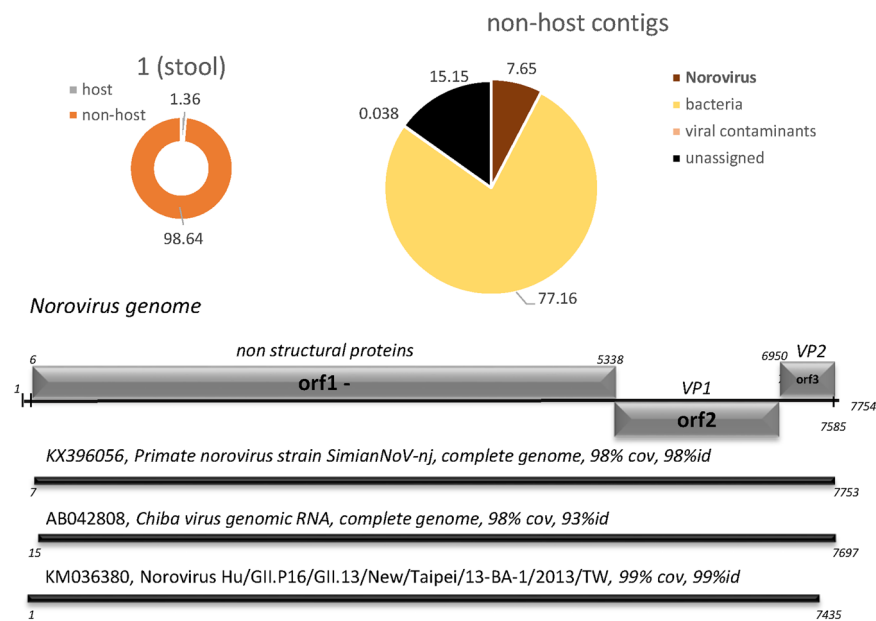


Figure 5. Identification of full-length genomes of three different Norovirus strains from a stool sample from a patient with acute gastroenteric disease. Primate Norovirus, Genbank entry KX396056 is identical to NC_031324 describing a human norovirus in diarrhetic chimps; next closest assignment NC_039897.1, human Norovirus GI, 92% sequence identity.

Cohort based analysis. *Identification of pathogen signatures shared among outbreak samples.* To demonstrate the ability of the cohort-based analysis tool to identify pathogens that may be responsible of infectious disease outbreaks, we analyzed five CSF samples derived from an enterovirus meningitis outbreak occurring in the Hamburg region during summer 2015 (Supplementary Table S1). CSF samples were negative by diagnostic PCR for HSV, VZV, EBV and *Borrelia burgdorferi* while samples showed Ct values between 31 and 33 for Enterovirus B PCR. As a negative control group in our cluster analysis, we used 22 unrelated routine diagnostic CSF samples that had tested negative in diagnostic taqman-PCR for a panel of viruses commonly involved in encephalitis. Table S3 summarizes the sequencing data of all samples included. Figure 5A depicts a schematic outline of our analysis. In total, more than 16,500 contigs were assembled across the 27 samples. The single linked cluster analysis tool integrated in the DAMIAN pipeline (see Material & Methods for details) produced 13,457 sequence clusters from these contigs. For each individual cluster, the fraction of positive samples in the outbreak and control cohorts was determined, and a cluster score was calculating by summation of the positive outbreak fraction value and negative value of the control fractions. Accordingly, the resulting score can take a maximum value of +1 if all samples in the outbreak cohort are positive while all controls are negative, or minimally reach a value of -1 if all control but no outbreak samples are positive. Overall, we observed 267 discrete patterns of positive and negative samples among the 13,457 sequence clusters, with scores that ranged from +1.00 to -0.45. A map of all signature patterns (sorted by descending score) along with their observed frequencies is shown in Fig. 7B. The full distribution of clusters and assignment of sequences within the cluster can be found in Supplementary Dataset S1.

Overall, a total of 30 sequence clusters were shared among all five outbreak samples; of these, 15 were not present in any of the control samples and consequently were awarded the highest score of +1.00 (see annotated top-scoring pattern in Fig. 7B and Supplementary Dataset S1). Only one of these fifteen clusters was assigned to a pathogenic species, namely Enterovirus B. Interestingly this cluster contained 14 contig sequences, with the longest contig encompassing 7,337 nt (and thus extending over the entire length of the Enterovirus B genome). The contig contained one single ORF with proteins clearly identified as Enterovirus protein domains (see Supplementary Dataset 1). The other eleven clusters were either of environmental or commensal bacterial origin ($n = 6$), unknown origin (no match in NCBI database, $n = 4$) or unclear origin (*Calidris pugnax*, $n = 1$). Thus, while DAMIAN readily classified the assembled Enterovirus B contigs taxonomically due to their nucleotide homology to existing NCBI database entries, even if the taxonomic classification had failed the approach presented here would have reduced the number of candidates that may be responsible for the outbreak to just a handful.

Reoccurring viral contaminants. In addition to its value for identifying putatively novel pathogens, the cohort based analysis tool is also useful to identify and flag common contaminants that are frequently present in NGS data. Such contaminants, for example, may reflect environmental bacteria that are introduced by excessive handling of the diagnostic specimen. In addition, contaminants may be introduced via laboratory materials and reagents, for example, retroviral sequences that originate from reverse transcriptase enzyme preparations in library kits, or parvoviral sequences that likely stem from silica gel columns used for nucleic acid extraction^{14,22,23}. By virtue of the fact that they register in all (or nearly all sequences), such sequences can be easily identified by

	7653
DAMIAN	1,618,480 reads 86,86% human sequences ✓ Enterovirus B (660 reads, 1.28%) ● Echovirus 30 (2 contigs; 660 reads) 98% id.
Taxonomer BETA	1,600,000 reads sampled; 5% classified ✓ Enterovirus B (123 reads) ● Echovirus 30 (49 reads) ● Coxsackievirus B2 (10 reads) ✓ α-retrovirus (742 reads) ✓ Caudovirales (745 reads) ✓ Herpesvirales (57 reads) ● HHV6A (21 reads)
PathoScope	6,021 aligned reads; 2,234 hits ✓ Enterovirus (39 reads) ● Enterovirus 107 (30 reads) ● Enterovirus 100 (4 reads) ● Enterovirus B (5 reads) ✓ Encephalomyocarditis Virus (85 reads) ✓ Adenovirus (1 read) ● Adenovirus F (1 read) ✓ Hepatitis C (1 read) ● Genotype 1 (1 read) ✓ α-retrovirus (398 reads)
metaMix	its; 182,276 human reads ✓ Echovirus 30 (1 contig; 794 reads)

Table 4. Comparison of CSF sample analysis results obtained by DAMIAN, Taxonomer BETA, PathoScope and metaMix.

DAMIAN, and subsequently can be excluded from downstream analyses. By default, DAMIAN filters for a number of viral sequences (mostly representing unclassified circular DNA viruses; see complete list in Supplementary Table S2) that we have frequently detected in our metagenomic DNA or RNA shotgun sequencing experiments. These sequences are identified by DAMIAN and flagged as putative contaminants in the DAMIAN output files (for examples, see entries in light grey color code in Supplementary Datasets S2–S11). To our knowledge, no other tools aimed at diagnostic NGS applications recognize such contaminants. For example, both PathoScope and Taxonomer report alpharetroviral sequences in BAL sample 104, whereas DAMIAN clearly flags the corresponding contigs as putative contaminants (Supplementary Dataset S2).

Discussion

DAMIAN is a publicly available, comprehensive software tool for the fast and reliable detection of pathogens specifically in diagnostic samples. To our knowledge, it is the first software to include a tool for cohort based analyses, a feature which can be highly valuable in infectious disease outbreak scenarios where multiple samples have to be compared for presence of shared pathogen sequences. DAMIAN is easy to use and easy to install. Its output provides an interpretation of its findings (including flagging of commensals and technical artifacts) and allows for fast decision making in clinical context. Assembled sequences, which often represent complete or near-complete viral genomes, are a part of the output. DAMIAN automatically documents its analyses. Software and database versions, parameters and similar information is stored and allows to quickly describe or reproduce an analysis.

Using primary/authentic diagnostic samples that have been well characterized by conventional diagnostic (culture and PCR) methods (Figs 2–6), as well as publicly available benchmarking data sets originally used to validate the Taxonomer pipeline (Table 5; Supplementary Fig. S1 and Supplementary Datasets S9–S11), we have verified that DAMIAN accurately identifies viral and bacterial pathogens. Furthermore, DAMIAN allows reliable classification of viral sequences at the species level and, in most cases, even at the strain level. Compared to DAMIAN, the other tools tested here provided strain level assignments which were substantially more error prone or incomplete. This is especially true for those tools, which are based on classification of single reads (PathoScope, Taxonomer). For example, only DAMIAN was able to assign Sapovirus, Chiba Virus and Norovirus strains in human stool samples. DAMIAN is furthermore superior in detecting and differentiating between individual strains of multiple viral species present in a single sample, as demonstrated by the analysis of a stool sample (sample 1) originating from a large AGE outbreak in Germany that had been caused by sewage-contaminated food sources. Indeed, DAMIAN was not only able to identify the individual strains, but also assemble complete (or near-complete) genomes of the GI and GII Norovirus genotype viruses, a feature which is highly valuable when investigating infectious disease outbreak situations such as the 2012 AGE outbreak^{19,20}.

Of note, while the data presented in Tables 2–5 demonstrate complete or near complete recovery of RNA virus genomes (or genome segments) with a size of 20 kb or less, DAMIAN is also able to assemble considerably larger viral genomes. For example, we recently used a previous version of the pipeline to help recover the full sequence of a novel seal parapoxvirus from DNA-seq reads derived from a skin lesion²⁴. In Supplementary Fig. S1D and Dataset S12 we furthermore demonstrate that RNA-seq reads can be used to recover near-complete DNA-virus genomes. In this case, unbiased RNA sequencing of a human stool sample from an immunosuppressed patient allowed recovery of 12 contigs (1,929 bp to 10,132 bp) which covered the full genome of human adenovirus type 31. Of course, successful assembly of complete DNA viruses from RNA-seq reads will require abundant

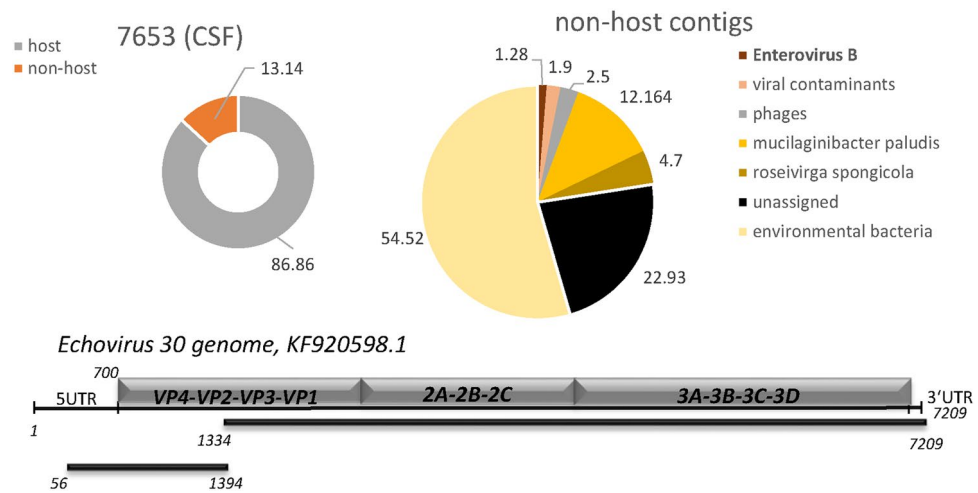


Figure 6. Application of DAMIAN to RNA-Seq libraries from diagnostic stool samples from patients with encephalitis. Sequences are depicted as described in Fig. 2. Two contigs representing significant sequence homology to Echovirus 30 were identified.

	SRR533978	SRR1553464	SRR1564804
DAMIAN	2,538,346 reads 7.7% human sequences ✓ Bas Congo Virus (7 contigs, 8,366 reads, 0.6%) 99.94% id. ✓ Parabrucella tropica (335,121 reads, 23.9%) ✓ Parabrucella fungorum (186,569 reads, 13.3%)	1,752,608 reads 1.34% human sequences ✓ Zaire Ebolavirus (1 contig, 1,740,198 reads, 98.1%) 98.45% id. ✓ Ralstonia (8,482 reads, 489 contigs)	627,013 reads 0.83% human sequences ✓ Equine infectious anemia virus (1 contig 89 reads, 0.02%) ✓ GB virus C (2 contigs, 48 reads, 0.01%) ✓ Chlamydomonas psittaci (59 contigs; 328,399 reads, 71.3%)
Taxonomer BETA	196 K reads sampled, 23% classified threshold 50 reads ✓ Bas Congo Virus (477 reads) ✓ Human Rotavirus A (94 reads) ✓ Neisseria meningitidis 6,334 reads ✓ Mycobacteria (2,814 reads) ✓ Microbacterium laevaniformans (13,314 reads) ✓ Candida albicans (199 reads)	179 K reads sampled, 67% classified threshold 50 reads ✓ Zaire Ebolavirus (86,872 reads) ✓ Bradyrhizobium (1,595 reads) ✓ Actinobacteria (1,776 reads)	184 K reads sampled, 50% classified threshold 50 reads ✓ Chlamydomonas psittaci (4,999 reads) ✓ GB Virus C (121 reads) ✓ Actinobacteria (7,830 reads) ✓ Bacilli (4,025 reads) ✓ Alphaproteobacteria (10,169 reads)
PathoScope	537 hits ✓ Burkholderia gladioli BSR 3 (88,777 reads) ✓ Staphylococcus epidermidis (15,834 reads) ✓ Acidovorax sp. JS42 (15,306 reads) ✓ Hepatitis C Virus (58 reads)	843 hits ✓ Ebolavirus Zaire 1976 strain (1,309,786 reads) ✓ Ralstonia pickettii (12,823 reads) ✓ HHV-4 (4 reads) ✓ Human Adenovirus C (1 read)	802 hits ✓ Chlamydomonas psittaci (392,837 reads) ✓ Ralstonia pickettii (30,029 reads) ✓ Staphylococcus aureus (10,436 reads) ✓ Equine infectious anemia virus (115 reads) ✓ GB virus C (73 reads)
metaMix	109 hits; 11,130 human reads ✓ Bas-Congo Tibrovirus (1 contig; 8,849 reads) ✓ Parabrucella 11 contigs; 89,485 reads ✓ Burkholderia 24 contigs; 79,270 reads	141 hits; 406 human reads ✓ Zaire Ebolavirus (1 contig; 2,230,594 reads) ✓ Ralstonia (9 contigs; 6,090 reads) ✓ Bradyrhizobium (11 contigs; 2,334 reads)	31 hits; 123 human reads ✓ Chlamydomonas psittaci (1 contig; 587,653 reads) ✓ Equine infectious anemia virus (1 contig; 146 reads) ✓ GB virus C (1 contig; 45 reads)

Table 5. Comparison of analysis results for stool samples obtained by DAMIAN, Taxonomer, PathoScope and metaMix.

transcription across the majority of the viral genome. Hence, RNA-seq of samples in which viral transcription is restricted (e.g., latently infected cells) are very unlikely to yield complete viral sequences.

The possibility to perform cohort-based analysis of multiple samples represents a unique advantage of the DAMIAN pipeline. Independent of taxonomic classification, this tool allows the identification of sequence signatures that are uniquely (or preferentially) associated with a given sample (e.g., disease-associated) cohort when compared to a collection of control samples. While information from external database can be integrated, the main advantage of this approach is that such information is not at all required to detect pathogenic agents.

We have previously used a similar approach to help resolve a suspected outbreak involving three patients suffering from severe pneumonia. As initial routine diagnostics failed to detect an infectious agent, it was speculated

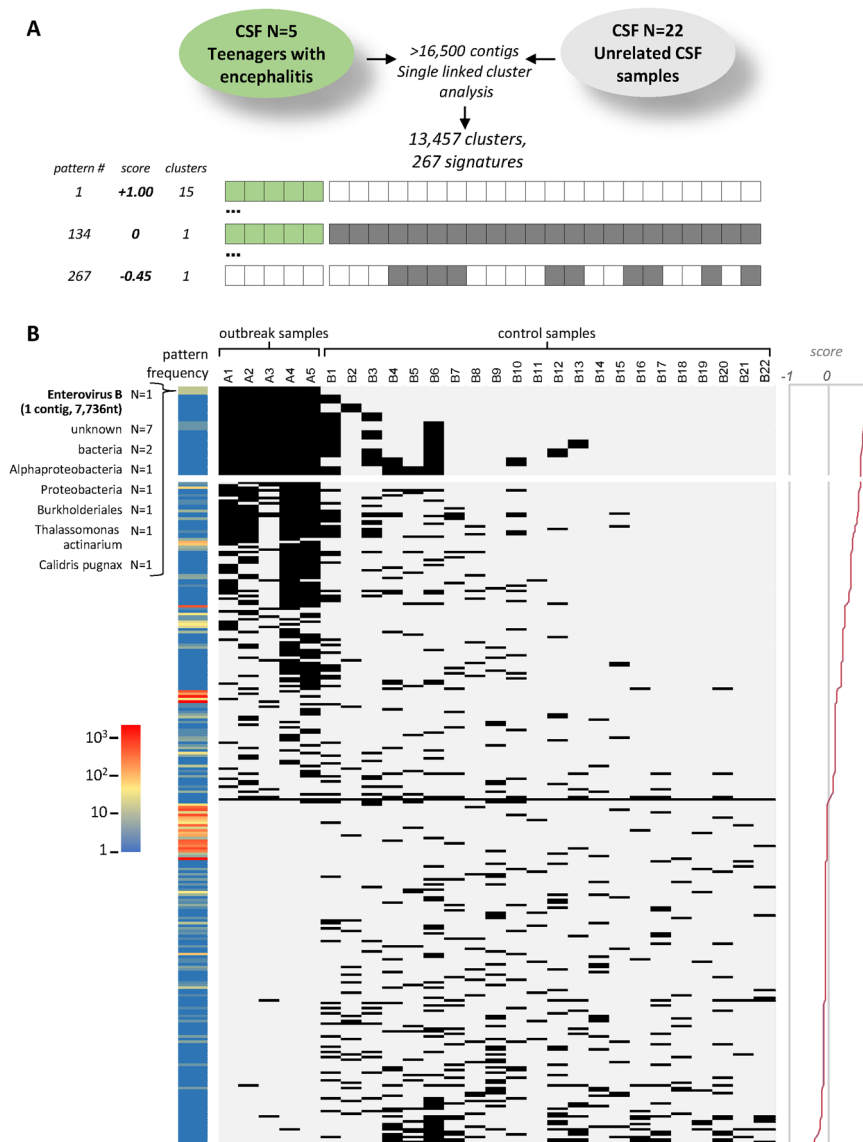


Figure 7. Cluster analysis of CSF samples from encephalitis and control cases. **(A)** Schematic depiction of the cohort analysis performed on five samples derived from an enterovirus outbreak and 22 unrelated control samples. Single linked cluster analysis produced 13,457 clusters from ~16,500 individual samples. Depending on the distribution of samples that do or do not contribute contigs to a given cluster, these can be assigned to one of a total 267 observed ‘signature’ patterns. The lower panel schematically depicts the highest (score = +1), lowest (−0.45) and neutral (0) scoring signature patterns, with filled (dark green or grey for encephalitis or cohort samples, respectively) or empty squares symbolizing samples that do or do not contribute contigs, respectively. The total numbers of clusters assigned to each of the three signatures is shown to the left. **(B)** Distribution map and frequencies of observed signature patterns. Each row depicts one of the 267 observed signature patterns as described above under (A). Signatures (black and light gray rectangles for positive and negative samples, respectively) are ordered by their score (plotted to the right). The ten signatures in which all encephalitis samples contribute contigs are shown enlarged at the top. The colored heat map bar to the left indicates the number of clusters that share a given signature pattern. The taxonomic annotation (lowest common ancestor of individual contig assignments, or ‘unknown’ if contigs do not have significant hits) of the 15 clusters with the highest scoring pattern are indicated at the top.

that the cases may represent an outbreak of a novel pathogen. Upon NGS-based analysis of BAL material, however, our pipeline readily called the presence of *Chlamydophila psittaci* in one of the samples, an infection which was subsequently confirmed by routine diagnostic procedures as the cause of the observed clinical symptoms. Importantly, neither on the level of taxonomic assignments nor after performing pairwise BLAST alignments did we find any evidence of a potential shared pathogen sequence signature among the three samples, strongly arguing against the hypothesis that the cases represented an outbreak of a novel pathogen⁶.

While the above example highlights the usefulness of combining taxonomic assignment with cross-sample sequence alignments to rule out an infectious disease outbreak, we here also demonstrate the ability of the

DAMIAN cohort analysis tool to identify a causative pathogen in authentic outbreak samples. Remarkably, the assembled Enterovirus B genomes represented one of only a handful of clusters that were shared among all five outbreak samples, but were not present in the control cohort. Notably, while Enterovirus B was also identified taxonomically, the clustering result *per se* is completely independent of taxonomic classification. Even if Enterovirus sequences were not present in the database (or if no reference databases were available at all), it would be fairly straightforward to hunt for the causative agent among the top-scoring fifteen candidates that were ranked solely due to their pairwise sequence homology across the sample cohorts.

Naturally, depending on the given type of disease or diagnostic specimen it will not be always feasible to presume that a causative pathogen must be present in 100% of the outbreak samples at the time of diagnosis, while being completely absent from the controls. Even in such scenarios, however, ranking of the contigs according to the scores awarded by the cohort analysis tool will allow identification of those sequences which are preferentially associated with a given disease cohort. Hence, especially in cases where the presence of a potentially novel pathogen is suspected, we expect that researchers as well as clinicians will find DAMIAN a valuable tool to help eliminate contigs originating from common microorganisms or contaminants, and thus aid in focusing on those sequences that represent the most promising candidates for a causative pathogen.

Materials and Methods

Quality control. Trimmomatic²⁵ was integrated for the optional removal of low quality bases and sequencing adapter sequences. DAMIAN executes the program with predefined parameters, which can be modified. Information on read properties prior and after this step is collected by DAMIAN and stored in its database.

Digital subtraction and abundance estimations. Digital subtraction and abundance estimation of unwanted sequence reads is optional and DAMIAN can be used with any number of different host reference genomes or no host genome at all. Bowtie2^{26,27} was integrated for read alignment tasks. Host abundance estimation is performed on a subset of sequence reads (default 1 M reads) using Bowtie2's 'sensitive-local' parameter preset. Reads aligning without insertions and deletions and with a minimal mapping quality of 10 are used to estimate the size of sequenced fragments and its standard deviation. Digital subtraction is performed on all reads. Here the 'fast' preset is applied, which enforces end-to-end alignments. Bowtie2, like all other tools, was integrated and the user is not required to be familiar with its functionality. Sequence indices, for example, are built automatically.

Assembly and assessment of basic contig features. Sequence reads are assembled using IDBA-ud²⁸. Following its author's instructions, the source code of the program was slightly modified to support reads up to a length of 250 bp. DAMIAN processes the assembled contigs individually. It extracts open reading frames by translating the contig sequences in the six possible reading frames and subsequently identifying putative amino acid sequences of a given minimal length (75 bp per default) which are not interrupted by stop codons. Sequence complexity is assessed using dustmasker from the NCBI Blast + suite. Contig abundance is calculated based on the alignment of sequence reads to the contigs. This task is performed with Bowtie2. Coverage tracks for every contig are stored in the database.

Functional annotation and contig ranking. Derived amino acid sequences are screened for known protein domains using HMMER²⁹ and the PFAM³⁰ database. DAMIAN classifies PFAM domains according to their taxonomic occurrences. Besides from being an additional level of evidence for the detection and classification of pathogens, the domain annotation is also used to determine the order in which contigs are processed in the subsequent analysis steps. Contigs are ranked according to the number of bases located in annotated domains and according to whether or not these domains are known to be exclusively present in viruses. Reordering the contigs does not affect the results, but since DAMIAN reports important findings immediately and since preliminary results in Excel-format can be generated before an analysis is finished, the ranking leads to putative pathogens being reported earlier.

Taxonomic assignment. Ranked contigs are processed with BLAST³¹ to identify similar sequences in NCBI's nt and nr database. The default strategy is to perform only MEGABLAST searches. Optionally BLASTN and BLASTP can each be used on every contig or only on contigs, which did not yield a match with the preceding less sensitive search. For every contig, all matches with a bitscore at least as high as 90 percent of the highest observed bitscore for a given contig, are stored in the database. Additionally, NCBI's taxnames and taxnodes are used to traverse taxonomic lineages and determine the lowest common ancestor (LCA) for the matches with the single highest bitscore and, separately, for those matches in the stratum defined above. If the LCA is a viral taxon and neither a phage nor a deliberately excluded taxonomic entity, then it is being reported immediately. Once all contigs have been processed with BLAST, the initial analysis is complete. At this point, final reports can be generated for individual samples or a cohort of previously processed samples can be analyzed jointly.

Refined taxonomic assignment and reporting. For the report of individual samples, contigs are taxonomically assigned a second time using a strategy which incorporates information from all contigs. First, all species observed in any BLAST match of any contig are ranked according to the bitscore of the matches and abundance of the contigs yielding these matches. Then every contig is assigned to the highest scoring species it yielded a match for. This strategy follows the assumption that if a contig C1 can be unambiguously assigned to a species S1 and another contig C2 could equally well be assigned to species S1 or S2 then the most parsimonious explanation for this observation is, that both C1 and C2 originate from species S1. Only in rare cases, where this algorithm does not yield an unambiguous match on species level, an LCA is computed from all optimal matches.

The output contains, the taxonomic assignment based on the procedure described above, the preceding LCA assignment and the underlying initial BLAST assignments.

Cohort analysis. The user assigns any number of samples to a group of known positives, known negatives or of unclassified samples. For example all samples which belong to one outbreak can be assigned to the group of positives while samples which are known to be unrelated the outbreak would be assigned to the group of negatives. Finally, samples for which it is unsure whether they are part of the outbreak could be put in the group of unclassified samples. Contigs not meeting user-defined criteria can be excluded from the analysis. These criteria include the information content, the length of the contig, the number of detected protein domains, the number of ORFs and the taxonomic assignment. Remaining contigs are used in an all-versus-all BLAST. From the pairwise results, a (bit-)score matrix is calculated. Single-linkage clustering is performed until the score of the two most similar pair of clusters is lower than a defined threshold. After clustering, a score is calculated for each cluster. The score is based on the number of contigs belonging to each of the three groups and group specific predefined weight, such that clusters with a higher score contain more sequences from the group of positive sample and less sequences from the group of the control group.

Results are reported in spreadsheet format and additionally FASTA files are generated from contig sequences for each cluster (see exemplary results in Supplementary Dataset S1).

Implementation and availability. The software was written in Ruby and meant to be deployed in Linux environments. A PostgreSQL database is used to store analysis results and associated metadata. Results presented in this publication were achieved using DAMIAN's standard settings. DAMIAN's source code is available at <https://sourceforge.net/projects/damian-pd>.

Taxonomer. Taxonomer analyses were performed using the web based metagenomics analysis tool provided on <http://taxonomer.iobio.io/>.

PathoScope. PathoScope analyses were performed using version 2.06 with default parameters. The optional PathoDB, and PathoReport modules were included and the optional PathoQC module was omitted.

metaMix. Sequence reads were assembled with IDBA-UD. Contigs longer than 399 bp were aligned to the NCBI nt database with MEGABLAST. Read length and taxon identifiers were incorporated in the BLAST output as described in the metaMix user guide. Finally, metaMix v0.3 was run with standard settings¹⁷.

Diagnostic sample. Samples were collected during routine diagnostic analysis performed at the UKE. Respiratory BAL samples derived from patients with respiratory illness and suspected influenza infection. All samples were screened by standard diagnostic quantitative RT-PCR for known respiratory pathogens.

Stool samples, collected during the gastroenteritis outbreak in Germany, were received from the Robert Koch Institute (RKI) and from the DRK hospital (Berlin).

CSF samples were received from the Department of hematopoietic stem cell transplantation. The CSF samples were collected due to neurological complications in these patients. All samples were screened for pathogens involved in CNS infections applying conventional diagnostics. Only samples tested negative were included here. The five CSF samples from teenager with encephalitis were received from the UKE. The study was approved in compliance with relevant laws and institutional guidelines by the local ethics committee, Freie Hansestadt Hamburg, WF-012/15; WF-026/13; WF025/12. The study was conducted retrospectively on anonymously stored clinical samples. Information which would allow the identification of the patient (human sequences, name, address, birth date, hospitalization number) was removed. Under these conditions the ethics committee approved the study on diagnostic samples without an informed consent.

Datasets SRR533978, SRR1553464 and SRR1564804 are derived from SRA, sequence read archive: SRX173233, SRX674125, SRX691917.

Diagnostic PCRs. The PCR primers and specific probes for influenza virus quantitative PCR used have been described previously^{6,32–38}. The following primers and probes were used: Infl.A_F: GACAAGACCAATCCTGTC ACYTCTG, Infl.A_R: AAGCGTCTACGCTGCAGTCC, HEX-TTCACG-CTCACCGTGCCAGTGAGC-BHQ2 and Infl. B_F: TCGCTGTTT-GCAGACACAAT, Infl. B_R TTCTTTCCACCGAACCA, Cyan500-AGAAGAT-GG AGAAGGCAAAGCAGAACT-DB. Norovirus PCR was performed individually for NoV GI and GII sequences. The following primer and probe sequences were used for GI PCR: NV192 (s) 5'-GCYATGTTCCGCTGGATGC, NV193 (as) 5'-CGTCCTTAGACGCCATCATCA, TM9-MGB probe 5'-VIC-TGGACAGGAGATCGC-MGB-NFQ. For GII PCR the primer sequences NV107c (s) 5'-AICCIATGTTYAGITGGATG and NV119 (as) 5'-TCGACGCCATCTT CATTAC were used together with the MGB probe TM3AP 5'-6'FAM-TGGGAGGGCGATCGCAATCTGGC-MGB-NQF. Sapovirus PCR was performed using SaV124F 5'-GAY CAS GCT CTC GCY ACC TAC, SaV1245R 5'-CCCTCCATYTCAAACACTA; SaV124TP FAM-CCR CCT ATR AAC CA-MGB-NQF. PCR reactions were performed using the Quantifast pathogen RT-PCR Kit + IC (Qiagen). 5 µl eluate was amplified (Roche Lightcycler 480 instrument) using the following conditions: 20 min @50°C, 5 min @95°C, 45 × 15 sec @95°C, 30 sec @ 60°C.

RNA extraction. Stool samples were homogenized in 1.4 ml DNA/RNA buffer (ZR Viral DNA/RNA Kit, Zymo Research) using MP matrix C tubes (Millipore) applying 2 × 30 s at 6,000 rpm in a Precellys 24 tissue homogenizer. Cleared supernatant was transferred to the column and nucleic acid was extracted following manufacturer's instructions. Nucleic acid was eluted in 30 µl DNAase/RNase free water.

BAL samples and CSF samples (200 µl) were automatically extracted using QIASymphony (Qiagen Hilden)^{6,31}. Nucleic acid was eluted in 100 µl final volume.

Library preparation and high-throughput sequencing. RNA Illumina NGS libraries were prepared from each sample. Illumina library from RNA was generated using a modified protocol of the SCRIPT SEQTM v2 RNA Seq Kit (Epicentre Biotechnologies) which was described recently^{6,33}. All libraries were multiplexed sequenced on an Illumina HiSeq. 2500 instrument (300 cycles, 2 × 150 bp on a paired-end protocol) or MiSeq (300 cycles) according to the manufacturer's protocol.

Received: 10 May 2019; Accepted: 24 October 2019;

Published online: 14 November 2019

References

1. Basein, T. *et al.* Microbial Identification Using DNA Target Amplification and Sequencing: Clinical Utility and Impact on Patient Management. *Open forum infectious diseases* 5, ofy257, <https://doi.org/10.1093/ofid/ofy257> (2018).
2. Westblade, L. F. *et al.* Role of Clinicogenomics in Infectious Disease Diagnostics and Public Health Microbiology. *Journal of clinical microbiology* 54, 1686–1693, <https://doi.org/10.1128/JCM.02664-15> (2016).
3. Rampini, S. K. *et al.* Broad-range 16S rRNA gene polymerase chain reaction for diagnosis of culture-negative bacterial infections. *Clinical infectious diseases: an official publication of the Infectious Diseases Society of America* 53, 1245–1251, <https://doi.org/10.1093/cid/cir692> (2011).
4. Salipante, S. J. *et al.* Rapid 16S rRNA next-generation sequencing of polymicrobial clinical samples for diagnosis of complex bacterial infections. *PloS one* 8, e65226, <https://doi.org/10.1371/journal.pone.0065226> (2013).
5. Wagner, K., Springer, B., Pires, V. P. & Keller, P. M. Molecular detection of fungal pathogens in clinical specimens by 18S rDNA high-throughput screening in comparison to ITS PCR and culture. *Scientific reports* 8, 6964, <https://doi.org/10.1038/s41598-018-25129-w> (2018).
6. Fischer, N. *et al.* Rapid metagenomic diagnostics for suspected outbreak of severe pneumonia. *Emerging infectious diseases* 20, 1072–1075, <https://doi.org/10.3201/eid2006.131526> (2014).
7. Loman, N. J. *et al.* A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxinigenic *Escherichia coli* O104:H4. *Jama* 309, 1502–1510, <https://doi.org/10.1001/jama.2013.3231> (2013).
8. Naccache, S. N. *et al.* Diagnosis of neuroinvasive astrovirus infection in an immunocompromised adult with encephalitis by unbiased next-generation sequencing. *Clinical infectious diseases: an official publication of the Infectious Diseases Society of America* 60, 919–923, <https://doi.org/10.1093/cid/ciu912> (2015).
9. Wilson, M. R. *et al.* Actionable diagnosis of neuroleptospirosis by next-generation sequencing. *The New England journal of medicine* 370, 2408–2417, <https://doi.org/10.1056/NEJMoal401268> (2014).
10. Chiu, C. Y. & Miller, S. A. Clinical metagenomics. *Nat Rev Genet* 20, 341–355, <https://doi.org/10.1038/s41576-019-0113-7> (2019).
11. Flygare, S. *et al.* Taxonomer: an interactive metagenomics analysis portal for universal pathogen detection and host mRNA expression profiling. *Genome biology* 17, 111, <https://doi.org/10.1186/s13059-016-0969-1> (2016).
12. Miller, S. *et al.* Laboratory validation of a clinical metagenomic sequencing assay for pathogen detection in cerebrospinal fluid. *Genome research* 29, 831–842, <https://doi.org/10.1101/gr.238170.118> (2019).
13. Schlager, R. *et al.* Validation of Metagenomic Next-Generation Sequencing Tests for Universal Pathogen Detection. *Archives of pathology & laboratory medicine* 141, 776–786, <https://doi.org/10.5858/arpa.2016-0539-RA> (2017).
14. Naccache, S. N., Hackett, J. Jr., Delwart, E. L. & Chiu, C. Y. Concerns over the origin of NIH-CQV, a novel virus discovered in Chinese patients with seronegative hepatitis. *Proceedings of the National Academy of Sciences of the United States of America* 111, E976, <https://doi.org/10.1073/pnas.1317064111> (2014).
15. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology* 15, R46, <https://doi.org/10.1186/gb-2014-15-3-r46> (2014).
16. Francis, O. E. *et al.* Pathoscope: species identification and strain attribution with unassembled sequencing data. *Genome research* 23, 1721–1729, <https://doi.org/10.1101/gr.150151.112> (2013).
17. Morfopoulou, S. & Plagnol, V. Bayesian mixture analysis for metagenomic community profiling. *Bioinformatics* 31, 2930–2938, <https://doi.org/10.1093/bioinformatics/btv317> (2015).
18. Lu, G., Rowley, T., Garten, R. & Donis, R. O. FluGenome: a web tool for genotyping influenza A virus. *Nucleic acids research* 35, W275–279, <https://doi.org/10.1093/nar/gkm365> (2007).
19. Hohne, M., Niendorf, S., Mas Marques, A. & Bock, C. T. Use of sequence analysis of the P2 domain for characterization of norovirus strains causing a large multistate outbreak of norovirus gastroenteritis in Germany 2012. *Int J Med Microbiol* 305, 612–618, <https://doi.org/10.1016/j.ijmm.2015.08.010> (2015).
20. Made, D., Trubner, K., Neubert, E., Hohne, M. & John, R. Detection and Typing of Norovirus from Frozen Strawberries Involved in a Large-Scale Gastroenteritis Outbreak in Germany. *Food and environmental virology*. <https://doi.org/10.1007/s12560-013-9118-0> (2013).
21. Vincent, C., Mehrotra, S., Loo, V. G., Dewar, K. & Manges, A. R. Excretion of Host DNA in Feces Is Associated with Risk of *Clostridium difficile* Infection. *J Immunol Res* 2015, 246203, <https://doi.org/10.1155/2015/246203> (2015).
22. Friis-Nielsen, J. *et al.* Identification of Known and Novel Recurrent Viral Sequences in Data from Multiple Patients and Multiple Cancers. *Viruses* 8, <https://doi.org/10.3390/v8020053> (2016).
23. Smuts, H., Kew, M., Khan, A. & Korsman, S. Novel hybrid parvovirus-like virus, NIH-CQV/PHV, contaminants in silica column-based nucleic acid extraction kits. *Journal of virology* 88, 1398, <https://doi.org/10.1128/JVI.03206-13> (2014).
24. Gunther, T. *et al.* Recovery of the first full-length genome sequence of a parapoxvirus directly from a clinical sample. *Scientific reports* 7, 3734, <https://doi.org/10.1038/s41598-017-03997-y> (2017).
25. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120, <https://doi.org/10.1093/bioinformatics/btu170> (2014).
26. Langdon, W. B. Performance of genetic programming optimised Bowtie2 on genome comparison and analytic testing (GCAT) benchmarks. *BioData Min* 8, 1, <https://doi.org/10.1186/s13040-014-0034-0> (2015).
27. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* 10, R25, <https://doi.org/10.1186/gb-2009-10-3-r25> (2009).
28. Peng, Y., Leung, H. C., Yiu, S. M. & Chin, F. Y. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28, 1420–1428, <https://doi.org/10.1093/bioinformatics/bts174> (2012).
29. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS computational biology* 7, e1002195, <https://doi.org/10.1371/journal.pcbi.1002195> (2011).
30. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic acids research* 47, D427–D432, <https://doi.org/10.1093/nar/gky995> (2019).
31. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC bioinformatics* 10, 421, <https://doi.org/10.1186/1471-2105-10-421> (2009).

32. Dierssen, U., Rehren, F., Henke-Gendo, C., Harste, G. & Heim, A. Rapid routine detection of enterovirus RNA in cerebrospinal fluid by a one-step real-time RT-PCR assay. *Journal of clinical virology: the official publication of the Pan American Society for Clinical Virology* **42**, 58–64, <https://doi.org/10.1016/j.jcv.2007.11.016> (2008).
33. Fischer, N. *et al.* Evaluation of Unbiased Next-Generation Sequencing of RNA (RNA-seq) as a Diagnostic Method in Influenza Virus-Positive Respiratory Samples. *Journal of clinical microbiology* **53**, 2238–2250, <https://doi.org/10.1128/JCM.02495-14> (2015).
34. Jansen, R. R. *et al.* Development and evaluation of a four-tube real time multiplex PCR assay covering fourteen respiratory viruses, and comparison to its corresponding single target counterparts. *Journal of clinical virology: the official publication of the Pan American Society for Clinical Virology* **51**, 179–185, <https://doi.org/10.1016/j.jcv.2011.04.010> (2011).
35. Li, L. *et al.* Multiple diverse circoviruses infect farm animals and are commonly found in human and chimpanzee feces. *Journal of virology* **84**, 1674–1682, <https://doi.org/10.1128/JVI.02109-09> (2010).
36. Panning, M. *et al.* Detection of influenza A(H1N1)v virus by real-time RT-PCR. *Euro surveillance: bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin* **14** (2009).
37. Schibler, M. *et al.* Critical analysis of rhinovirus RNA load quantification by real-time reverse transcription-PCR. *Journal of clinical microbiology* **50**, 2868–2872, <https://doi.org/10.1128/JCM.06752-11> (2012).
38. Ward, C. L. *et al.* Design and performance testing of quantitative real time PCR assays for influenza A and B viral load measurement. *Journal of clinical virology: the official publication of the Pan American Society for Clinical Virology* **29**, 179–188, [https://doi.org/10.1016/S1386-6532\(03\)00122-7](https://doi.org/10.1016/S1386-6532(03)00122-7) (2004).

Acknowledgements

We are grateful to Susanne Pfefferle, Martin Mielke (RKI), Martina Höhne (RKI) and the DRK, Berlin for providing reagents. We thank Anja Koppe and Svenja Reucher for excellent technical support. We thank Alexis Robitaille (IARC; Lyon) for reading the manuscript and for valuable discussion. This work was supported in part by the German Center for Infection Research, project grant given to N.F. and A.G. and Leibniz Competition grant given to A.G. (T57/2015).

Author contributions

M.A., N.F. and A.G. developed and benchmarked the tool. N.F. and M.L. extracted the diagnostic samples and performed PCR validation. L.B., D.I., K.R. prepared the NGS libraries and performed the NGS sequencing. N.K. and M.C. provided CFS samples; M.Ae. provided stool and respiratory samples. N.F., M.A. and A.G. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41598-019-52881-4>.

Correspondence and requests for materials should be addressed to N.F. or A.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019

The landscape of viral associations in human cancers

Marc Zapatka^{1,20,21}, Ivan Borozan^{1,2,21}, Daniel S. Brewer^{1,3,4,21}, Murat Iskar^{1,21}, Adam Grundhoff⁵, Malik Alawi^{5,6}, Nikita Desai^{7,8}, Holger Sültmann^{9,10}, Holger Moch¹¹, PCAWG Pathogens¹², Colin S. Cooper^{4,13}, Roland Eils^{14,15,16}, Vincent Ferretti^{17,18}, Peter Lichter^{1,10,20*} and PCAWG Consortium¹⁹

Here, as part of the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium, for which whole-genome and—for a subset—whole-transcriptome sequencing data from 2,658 cancers across 38 tumor types was aggregated, we systematically investigated potential viral pathogens using a consensus approach that integrated three independent pipelines. Viruses were detected in 382 genome and 68 transcriptome datasets. We found a high prevalence of known tumor-associated viruses such as Epstein-Barr virus (EBV), hepatitis B virus (HBV) and human papilloma virus (HPV; for example, HPV16 or HPV18). The study revealed significant exclusivity of HPV and driver mutations in head-and-neck cancer and the association of HPV with APOBEC mutational signatures, which suggests that impaired antiviral defense is a driving force in cervical, bladder and head-and-neck carcinoma. For HBV, HPV16, HPV18 and adeno-associated virus-2 (AAV2), viral integration was associated with local variations in genomic copy numbers. Integrations at the *TERT* promoter were associated with high telomerase expression evidently activating this tumor-driving process. High levels of endogenous retrovirus (ERV1) expression were linked to a worse survival outcome in patients with kidney cancer.

The World Health Organization estimates that 15.4% of all cancers are attributable to infections and 9.9% are linked to viruses^{1,2}. Cancers that are attributable to infections have a greater incidence than any individual type of cancer worldwide. Eleven pathogens have been classified as carcinogenic agents in humans by the International Agency for Research on Cancer (IARC)³. After *Helicobacter pylori* (associated with 770,000 cases worldwide), the four most prominent infection-related causes of cancer are estimated to be viral⁴: HPV⁴ (associated with 640,000 cases), HBV⁵ (420,000 cases), hepatitis C virus (HCV)⁶ (170,000 cases) and EBV⁷ (120,000 cases). It has been shown that viruses can contribute to the biology of multistep oncogenesis and are implicated in many of the hallmarks of cancer⁸. Notably, the discovery of links between infection and cancer types has provided actionable opportunities, such as the use of HPV vaccines as a preventive measure, to reduce the global impact of cancer. The following characteristics have been proposed to define human viruses that cause cancer through direct or indirect carcinogenesis⁹: (1) presence and persistence of viral DNA in tumor biopsies; (2) growth-promoting activity of viral genes in model systems; (3) dependence of a malignant phenotype on continuous viral oncogene expression or modification of

host genes; and (4) epidemiological evidence that a virus infection represents a major risk for the development of cancer.

The worldwide efforts of comprehensive genome and whole-transcriptome analyses of tissue samples from patients with cancer have generated appropriate facilities for capturing information not only from human cells but also from other—potentially pathogenic—organisms or viruses that are present in the tissue. A comprehensive collection of whole-genome and whole-transcriptome data from cancer tissues has been generated within the International Cancer Genome Consortium (ICGC) project PCAWG¹⁰, providing a unique opportunity for a systematic search for tumor-associated viruses.

The PCAWG Consortium aggregated whole-genome sequencing (WGS) data from 2,658 cancers across 38 tumor types that have been generated by the ICGC and The Cancer Genome Atlas (TCGA) projects. These sequencing data were reanalyzed with standardized, high-accuracy pipelines to align to the human genome (build hs37d5) and identify germline variants and somatically acquired mutations¹⁰. The PCAWG working group 'Pathogens' analyzed the WGS and whole-transcriptome sequencing (RNA-sequencing (RNA-seq)) data of the PCAWG consensus cohort (2,656 donors). Focusing on viral pathogens, we applied

¹Division of Molecular Genetics, German Cancer Research Center (DKFZ), Heidelberg, Germany. ²Informatics and Bio-computing Program, Ontario Institute for Cancer Research, Toronto, Ontario, Canada. ³Norwich Medical School, University of East Anglia, Norwich, UK. ⁴Earlham Institute, Norwich, UK. ⁵Heinrich-Pette-Institute, Leibniz Institute for Experimental Virology, Hamburg, Germany. ⁶Bioinformatics Core, University Medical Center Hamburg-Eppendorf, Hamburg, Germany. ⁷Bioinformatics Group, Department of Computer Science, University College London, London, UK. ⁸Biomedical Data Science Laboratory, Francis Crick Institute, London, UK. ⁹Division of Cancer Genome Research, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Heidelberg, Germany. ¹⁰German Cancer Consortium (DKTK), Heidelberg, Germany. ¹¹Department of Pathology and Molecular Pathology, University and University Hospital Zürich, Zurich, Switzerland. ¹²A list of members and affiliations appears in the Supplementary Note. ¹³The Institute of Cancer Research, London, UK. ¹⁴Division of Theoretical Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, Germany. ¹⁵Department of Bioinformatics and Functional Genomics, Institute of Pharmacy and Molecular Biotechnology, Heidelberg University and BioQuant Center, Heidelberg, Germany. ¹⁶Center for Digital Health, Berlin Institute of Health and Charité Universitätsmedizin Berlin, Berlin, Germany. ¹⁷Ontario Institute for Cancer Research, MaRS Centre, Toronto, Ontario, Canada. ¹⁸Department of Biochemistry and Molecular Medicine, University of Montreal, Montreal, Québec, Canada. ¹⁹A list of members and affiliations appears at the end of the paper. ²⁰These authors jointly supervised this work: Marc Zapatka, Peter Lichter. ²¹These authors contributed equally: Marc Zapatka, Ivan Borozan, Daniel S. Brewer, Murat Iskar. *e-mail: peter.lichter@dkfz-heidelberg.de

three independently developed pathogen-detection pipelines 'Computational Pathogen Sequence Identification' (CaPSID)¹¹, 'Pathogen Discovery Pipeline' (P-DiP) and 'Searching for Pathogens' (SEPATH) to generate a large compendium of viral associations across 38 cancer types. We extensively characterized the known and novel viral associations by integrating driver mutations, mutational signatures, gene expression profiles and patient survival data of the same set of tumors analyzed by the PCAWG Consortium.

Results

Identification of tumor-associated viruses. To identify the presence of viral sequences, we explored the WGS data of 5,354 tumor-normal samples across 38 cancer types, and 1,057 tumor RNA-seq data across 25 cancer types (Supplementary Tables 1, 2, 20). In total, 195.8 billion reads were considered for analysis, as they were not sufficiently aligned to the human reference genome in the PCAWG-generated alignment. The remaining reads ranged from 28,036 to 800 million reads per WGS and up to 120 million reads per RNA-seq tumor sample (Fig. 1a, Extended Data Fig. 1a–c). Viral sequences were detected and quantified independently by the three recently developed pathogen-discovery pipelines CaPSID, P-DiP and SEPATH. The estimated relative abundance of a virus was calculated as viral reads per million extracted reads (PMER) at the genus level to improve consistency between pipelines. To minimize the rate of false-positive hits in virus detection, we applied a strict threshold of PMER > 1 supported by at least three viral reads as suggested in previous studies^{11,12}. Virus detection in a sample by at least two pipelines was considered to be a consensus hit. In total, 532 genera were considered for the extensive virus search in at least two of the pipelines (Extended Data Fig. 1d, Supplementary Table 18). Filtering of suspected viral laboratory contaminants was achieved through P-DiP, by examining each assembled contig of viral sequence segments for artificial, non-viral vector sequences and inspecting virus genome coverage across all positive samples (Extended Data Fig. 2a). The most frequent hits prone to suspected contamination were lambdavirus, alphabaculovirus, microvirus, simplexvirus, hepacivirus, cytomegalovirus (CMV), orthopoxvirus and punalikevirus; these were observed across many tumor types (Fig. 1b). For example, mastadenovirus showed an uneven genome coverage that could result from contaminating vector sequences. Therefore, we analyzed the virus detections across sequencing dates (Extended Data Fig. 2b) to assess any batch effect indicative of a contaminant; in mastadenovirus, we identified an association with sequencing date in early-onset prostate cancer regardless of tumor-normal state. We conclude that our mastadenovirus detections are due to a contamination that occurred across projects worldwide for which similar patterns could be identified.

We generally observed a strong overlap of the genera identified across pipelines (Extended Data Fig. 1e, Supplementary Tables 6, 7, 11). From the WGS dataset, we identified 321, 598 and 206 virus-tumor pairs using P-DiP, CaPSID and SEPATH, respectively (Fig. 2a; overlap after random permutation of detections, Extended Data Fig. 3a, Supplementary Tables 3–5). The number of hits derived from the RNA-seq dataset differed between the pipelines (virus-tumor pairs: 101 for P-DiP, 83 for CaPSID, 41 for SEPATH; Fig. 2b, Supplementary Tables 8–10). SEPATH, which used a *k*-mer approach, detected the lowest number of virus hits and was the least sensitive. Despite this, the identified viruses matched well with the consensus (DNA 90%, RNA 95%). P-DiP, which was based on an assembly and BLAST approach, detected more hits with 59% of the DNA and 54% of the RNA hits in the consensus set, whereas CaPSID, which was the most sensitive, implemented a two-step alignment process complemented with an assembly step and identified 60% (DNA) and 80% (RNA) hits within the consensus set. Although the majority of the virus hits from RNA-seq ($n=61$ out of 68 consensus hits based on RNA-seq) overlapped with the WGS

data, a lower fraction of detections from the WGS data were present in the RNA-seq data ($n=61$ out of 168 of 382 consensus hits based on WGS with available RNA-seq data), emphasizing the importance of DNA sequencing for generating an unbiased catalog of tumor-associated viruses. This difference can also be attributed to the viral life cycle, as viral gene expression can be minimal during incubation or latent phases¹³. Contrasting virus-positive and virus-negative samples within each organ type shows that the organ system, as expected, has a significant influence, but virus positivity does not ($P < 2 \times 10^{-16}$, analysis of variance modeling of candidate reads that are dependent on organ system and virus positivity; Extended Data Fig. 1c). This indicates that virus-positive tumors were not detected owing to a higher number of candidate reads; this is consistent with the fact that the viral reads in most cases do not substantially contribute to the reads analyzed. In total, 86% of the sequence hits detected in WGS and RNA-seq data were found to be from double-stranded DNA viruses and double-stranded DNA viruses with reverse transcriptase (Fig. 1c, Supplementary Table 19). This could be attributed to (1) a higher frequency of tumor-associated viruses from these genome types³, (2) a larger sequencing dataset for WGS compared with RNA-seq, (3) a potential limitation of our analysis due to DNA and RNA extraction protocols that are less likely to include single-stranded DNA or RNA viruses or (4) the selection bias of tumor entities included in the PCAWG study (Fig. 1c).

The virome landscape across 38 distinct tumor types. We used a consensus approach that resulted in a reliable set of 389 distinct virus-tumor pairs from WGS and RNA-seq data (Fig. 2a–d). Overall, 23 virus genera were detected across 356 patients with cancer (13%). The top-five most-prevalent viruses (lymphocryptovirus, orthohepadnavirus, roseolovirus, alphapapillomavirus and CMV) account for 85% of the consensus virus hits in tumors ($n=329$ out of 389). Among these five prevalent virus genera, three have been well described in the literature as drivers of tumor initiation and progression³: (1) lymphocryptovirus ($n=145$ samples (5.5%); for example, EBV) is the most common viral infection across a variety of tumor entities that mainly occur in the gastrointestinal tract and shows a much lower prevalence in the matched non-malignant control samples ($n=82$ (3%); Fig. 2c); (2) orthohepadnavirus ($n=67$ (2.5%); for example, HBV) is—as expected—the most frequent among liver cancer with HBV present in 62 of 330 donors (18.9%); and (3) alphapapillomavirus (discussed below). Lymphocryptovirus ($n=11$), orthohepadnavirus ($n=18$) and alphapapillomavirus ($n=32$) were detected in both RNA-seq and DNA-sequencing data (Fig. 2c, left), of which alphapapillomavirus was the most frequent (32 out of 39 consensus hits). This is consistent with the constitutive expression of viral oncogenes in cancers associated with these viruses, a parameter that supports a direct role in carcinogenesis⁹. An in-depth analysis of the virus genome equivalents per human tumor genome equivalent, which considers genome sizes, coverage and tumor purity, showed overall low viral genome equivalents even for established tumor viruses (Extended Data Fig. 3c, Supplementary Table 12). Evidence of a mouse mammary tumor virus (MMTV, PMER = 3.4) was detected in one renal carcinoma sample and in none of the 214 analyzed breast cancer samples. Previous work has suggested that MMTV may have a role in breast cancer but our comprehensive search of viral sequences could not identify any MMTV-positive case in breast cancer that would support this claim.

Roseolovirus and alphatorquevirus show a higher number of hits in non-malignant control samples, which were mainly derived from blood cells (Fig. 2c). For example, we identified 59 patients as roseolovirus-positive (human herpesvirus (HHV)-6A, HHV-6B and HHV-7) in their tumors (pancreas, 6%; stomach, 8%; colon/rectum, 8.3%) and 90 patients positive in the non-malignant control samples. Considering the known cell tropism of roseolovirus for B and

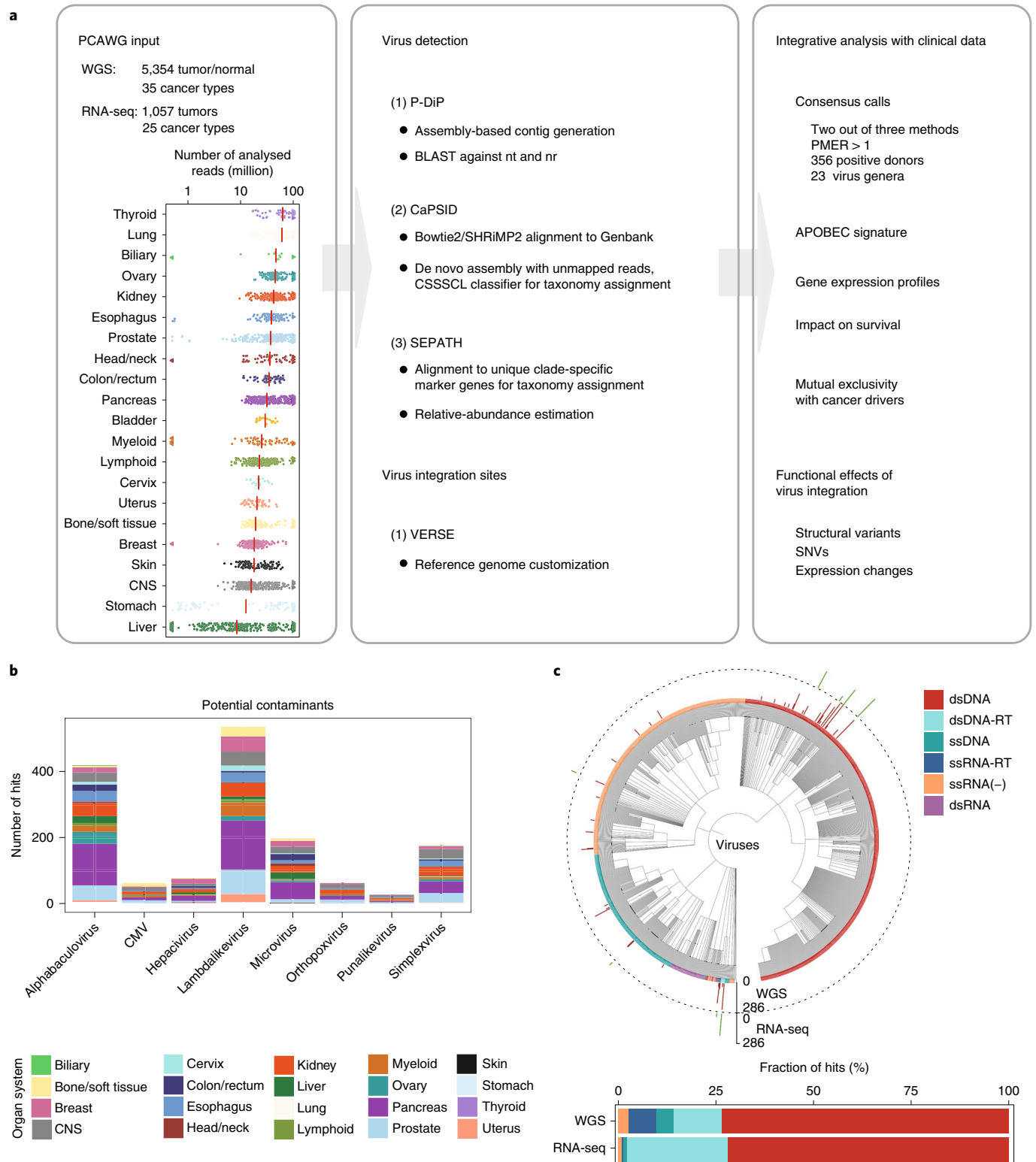


Fig. 1 | Overview, design and summary statistics. **a**, Workflow to identify and characterize viral sequences from the WGS and RNA sequencing of tumor and non-malignant samples. Viral hits were characterized in detail by using several clinical annotations and resources generated by PCAWG. The red line represents the median. CNS, central nervous system. **b**, Identified viral hits in contigs that showed higher viral reads PMER for artificial sequences such as vectors than for the virus. All viruses that occurred in at least 20 primary tumor samples in the same contig together with an artificial sequence are shown. **c**, Summary of the viral search space used in the analysis grouped by virus genome type. The number of virus-positive tumor samples is indicated in the outer rings (PMER log scale for WGS and RNA-seq data) as detected by any of the pipelines. Taxonomic relations between the viruses are indicated by the phylogenetic tree. dsDNA, double-stranded DNA; dsDNA-RT, double-stranded DNA with reverse transcriptase; dsRNA, double-stranded RNA; ssDNA, single-stranded DNA; ssRNA-RT: single-stranded RNA with reverse transcriptase; ssRNA, single-stranded RNA; dsRNA, double-stranded RNA. The fractions of hits in WGS and RNA-seq data are depicted as stacked bar graphs.

T cells¹⁵, we asked whether immune infiltration would be higher in roseolovirus-positive tumors. However, we could not identify a stronger contribution of immune cells in virus-positive tumor cases as estimated using CIBERSORT¹⁴ (false-discovery rate (FDR)-corrected $P > 0.05$ for pancreas; Extended Data Fig. 4a). Therefore, consistently with current knowledge (reviewed in ref. ¹⁶), we cannot confirm a link between roseolovirus and immune-cell content or tumor development. Furthermore, we could not identify actively transcribed viral genes for roseolovirus and alphatorquevirus at the transcriptome level. This is in agreement with the latent state of these viruses in blood mononuclear cells¹⁵, and their transmission through blood transfusions¹⁷. CMV was found, as expected¹⁸, after identification and removal of contaminations in both stomach tumors ($n = 13$) and the adjacent non-malignant tissue ($n = 11$). In line with a recent publication¹⁹, we could not detect CMV in the 294 tumors of the central nervous system (146 medulloblastomas, 89 pilocytic astrocytoma, 41 glioblastomas and 18 oligodendrogliomas) that were analyzed. Therefore, a previously debated role of this virus is not supported. Notably, we did not identify a significant enrichment of co-infection of multiple viruses in any tumor type (Extended Data Fig. 3d).

Incidence of HBV. HBV was most frequently detected in liver cancers ($n = 62$). Compared with the histopathological gold-standard HBV PCR test^{20,21} ($n = 228$), the WGS-based consensus detections had the same high specificity (96.1%) and a high sensitivity (84.0%), indicating that HBV detection using WGS is reliable (Fig. 3a, Extended Data Fig. 4b, Supplementary Table 13). Furthermore, five out of the seven cases that were positive using WGS but negative for HBV PCR showed positivity for HBsAg, indicating that the WGS analysis has a high sensitivity. In summary, the precision (85.7%) and recall (84%) for the detection of HBV based on around 30-fold-coverage WGS data were comparable to those of targeted PCR. We confirmed a significant exclusivity between HBV infection and mutations in *CTNNT1*, *TP53* and *ARID1A* that was found in a larger liver cancer cohort analyzed by high-throughput sequencing (FDR-corrected $P = 5.35 \times 10^{-6}$, 0.0023 and 0.0023, respectively; DISCOVER²²).

Detection of EBV. EBV was detected in many different tumor entities and normal samples (Fig. 2c). When comparing the PMER of EBV in tumor and matched normal samples, we see a stronger contribution in matched normal samples from matched solid tissue or tissue adjacent to the tumor (Extended Data Fig. 4c). For samples that contained reads for EBV in WGS and with available RNA-seq data, the absolute score for immune cells based on CIBERSORT¹⁴ was not significantly different between virus-positive and virus-negative samples (FDR-corrected $P > 0.05$ for colon/rectum, head-and-neck, lymphoid and stomach; Extended Data Fig. 4a). In summary, there is no evidence that the detection of EBV is due to infiltrating immune cells. This indicates the presence of EBV in the respective organs. On the basis of the expression data available for the tumor samples, we identified viral transcripts of the latent as well as lytic phase of the viral life cycle (Fig. 3b, Extended Data Fig. 4d, Supplementary Table 13). Eight of the nine tumors that expressed lytic EBV transcripts were from stomach cancers, confirming the active contribution of EBV to gastric cancer²⁴.

Identification of alphapapillomaviruses. Alphapapillomaviruses were mainly detected in head-and-neck cancers ($n = 18$ out of 57), cervical cancers ($n = 19$ out of 20) and in two bladder cancer cases out of 23, in agreement with previous studies^{4,25,26}. There is also supporting evidence for 32 out of 39 alphapapillomavirus hits in the whole-transcriptome data (Fig. 2c). We observed only one HPV subtype per tumor according to the P-DiP results and HPV16 was the dominant type in cervical ($n = 11$) and head-and-neck ($n = 15$) tumors, followed by HPV18, which was present in only cervical cancer ($n = 6$). As reported previously²⁷, HPV33 was identified in head-and-neck ($n = 3$) and cervical ($n = 1$) tumors. Different HPV variants, type 6 and 45, were detected in bladder cancer.

In head-and-neck cancer, HPV-positive tumors exhibited an almost complete mutual exclusivity with mutations in known drivers such as *TP53*, *CDKN2A* and *TERT* (FDR-corrected $P = 1.73 \times 10^{-5}$, 1.73×10^{-5} and 0.012, respectively; multiple testing corrected for presented mutations in EBV and HPV, DISCOVER²²) (Fig. 3c, Supplementary Table 13), as reported previously²⁵, which could be explained by the mutation-independent inactivation of TP53 due to the human papillomaviruses^{28,29,30}. Furthermore, we found that mutational signature 2 was enriched in alphapapillomavirus-positive cases of head-and-neck cancer³¹ (FDR-corrected $P = 0.02$; Fig. 3d, Supplementary Tables 12, 22). In addition, the expression of APOBEC3B is significantly higher in virus-positive head-and-neck cancers compared with virus-negative cancers³² ($P = 1.6 \times 10^{-4}$; Fig. 3f). However, we did not observe enrichment of APOBEC signatures and changes in expression in EBV-positive samples found in the cervix or in other tissues.

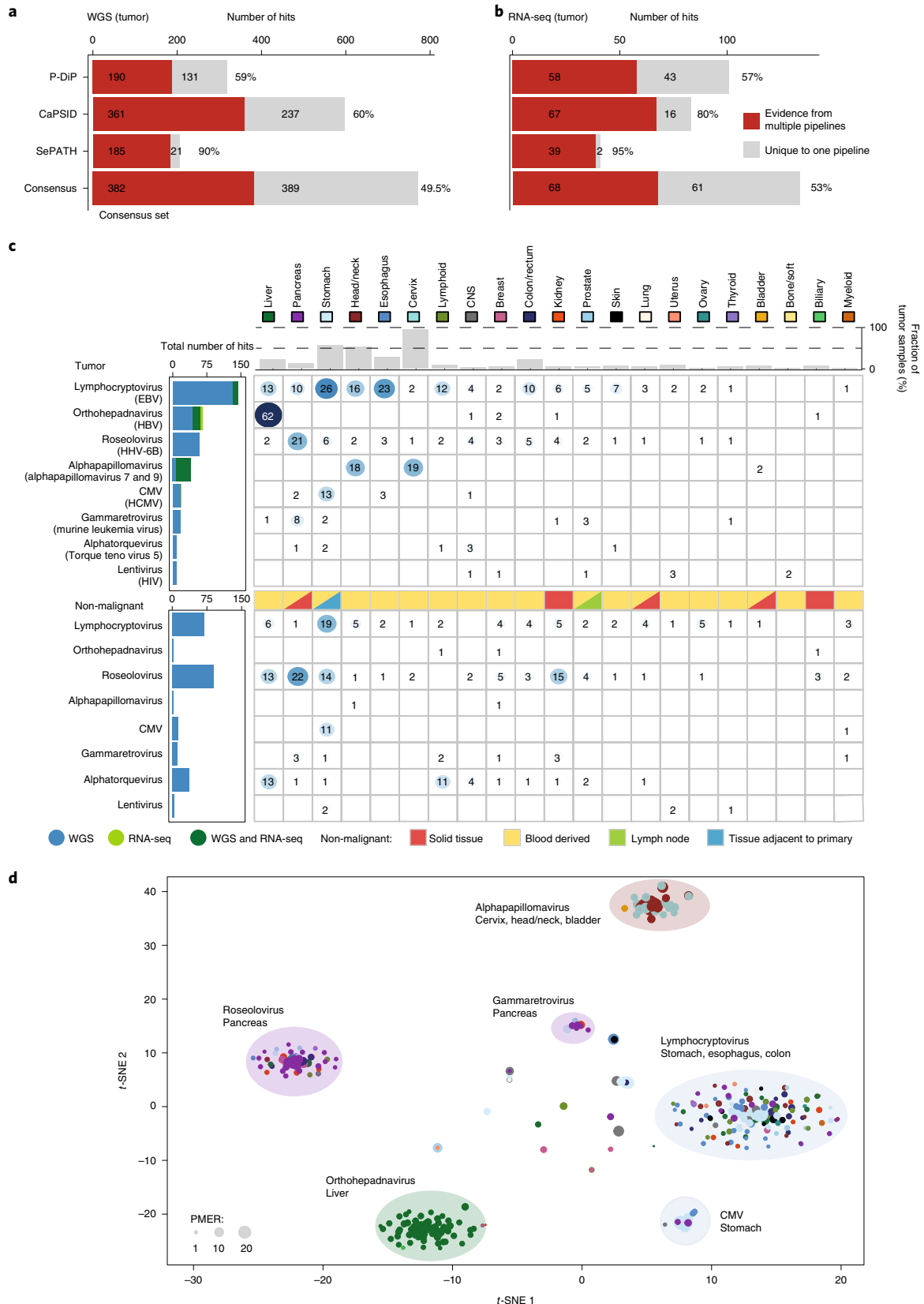
Distinct expression profiles between virus-positive and virus-negative tumors in head-and-neck cancer were observed³³ (Fig. 3e, Supplementary Table 23). Analyzing the immune cells estimated by CIBERSORT, we identified a significant increase in macrophages and T-cell signals in alphapapillomavirus-positive head-and-neck cancers ($P = 0.004$, 0.012 and 0.012 for follicular helper, CD8 and regulatory T cells, respectively, and $P = 0.018$ for M1 macrophages; FDR corrected for all viruses and cell types tested; Fig. 3g, Supplementary Table 24). Our integrative analysis of HPV reconfirms many of the findings related to HPV infection, illustrating the potential of our systematic approach in identifying and characterizing tumor-associated viruses.

Activation of endogenous retroviruses linked to outcome. Human endogenous retroviruses (HERV) are integrations in the human DNA that originate from infection of germline cells by retroviruses over millions of years³⁴ and contribute over 500,000 individual sites, or 2.7% of the overall sequence the human genome^{35,36}. ERVs were identified by all three pathogen-detection pipelines but were filtered by CaPSID and SEPATH. In addition, an alignment-based approach was used to detect HERV sequences that were embedded in the human reference genome that could be missed by the pipelines by focusing only on non-human reads. In this study, we quantified the expression of HERV-like long terminal repeat retrotransposons that were categorized into several clades by Repbase³⁷ as ERVL, ERVL-MaLR, ERV1, ERVK and ERV (Supplementary Table 14). In comparison to the other HERV families, ERV1 shows

Fig. 2 | Consensus for detected viruses in WGS and RNA-seq data. Number of genus hits among tumor samples for the three independent pipelines and the consensus set defined by evidence from multiple pipelines. **a**, Analysis based on WGS. **b**, Analysis based on whole-transcriptome sequencing. **c**, Heat map showing the total number of viruses detected across various cancer entities. The sequencing data used for detection are indicated among the total number of hits (WGS, blue; RNA sequencing, green). The fraction of virus-positive samples is shown at the top and the type of non-malignant tissue used in the analysis is indicated if more than 15% of the analyzed samples are from a respective tissue type (solid tissue, lymph node, blood or adjacent to primary tumor). **d**, t-SNE clustering of the tumor samples based on PMER of their consensus virome profiles, using Pearson correlation as the distance metric. Major clusters are highlighted by indicating the strongest viral genus and the dominant tissue types that are positive in that cluster. Dot size represents the viral reads PMER.

the strongest expression on average (Fig. 4a) and ERVK the highest fraction of active loci (Fig. 4b). By analyzing the expression of HERVs, we could identify strong expression of ERV1 in chronic lymphocytic leukemia compared with all other tumor tissues and

adjacent normal tissues (Fig. 4c). However, we could not identify a link between transcriptionally active stemness markers (OCT3/4, SOX2 and KLF4) and increased HERV expression, in contrast to a previous report³⁸ (Spearman rank correlation <0.35; Extended Data Fig. 5).



New data suggest that expression of HERVs is associated with prognosis in clear cell renal cell carcinoma³⁹. Analyzing HERV expression in relation to patient survival, we found that high ERV1 expression in kidney cancer was linked to worse survival outcome ($P=0.0081$; log-rank test; Fig. 4d, Extended Data Fig. 6, Supplementary Table 15).

Genomic integration of viral sequences. Viral integration into the host genome has been shown to be a causal mechanism that can lead to the development of cancer⁴⁰. This process is well-established for HPVs in cervical, head-and-neck and several other carcinomas, and for HBV in liver cancer^{41,42}.

Low-confidence integration events were detected for HHV4 (gastric cancer and malignant lymphoma) and HPV6b (head-and-neck and bladder carcinoma), whereas integration events with high confidence were demonstrated for HBV (liver cancer), AAV2 (liver), HPV16 and HPV18 (in both cervical and head-and-neck carcinoma). Most of these integration events were found to be distributed across chromosomes and a significant number of viral integrations occurred in the intronic (40%) regions whereas only 3.4% of integrations was detected in gene coding regions ($n = 84$ intronic versus $n = 31$ other regions excluding intergenic regions, two-sample test for equality of proportions, $P = 7.0 \times 10^{-12}$; Extended Data Fig. 7a–d).

HBV was found to be integrated in 36 liver cancer specimens out of 61 patients who were identified to be HBV positive. Notably, genomic clusters of viral integrations were identified in *TERT* (number of integration sites within a genomic cluster (NGC) of 6), *KMT2B* (NGC=4)—which was recently identified to be a likely cancer driver gene^{43,44}—and *RGS12* (NGC=3) (Extended Data Fig. 7e). Furthermore, two or more integration events in individual samples were observed in the gene (or gene promoter) regions of *CCNE1*, *CDK15*, *FSIP2*, *HEATR6*, *LINC01158* (also known as *PANTR1*), *MARS2* and *SLC1A7* (Fig. 5a). Additional events with two integration sites were also detected within a distance of 50 kb from *CLMP*, *CNTNAP2* and *LINC00359* genes. Integration events at *TERT* were found to recur in five different liver cancer samples. One sample had a genomic cluster of three viral integration events within *TERT* and four samples contained a single integration event in the *TERT* promoter, or 3' or 5' untranslated regions (UTR) (Supplementary Table 17). When comparing gene expression in samples with virus integration to those without, we found that only *TERT* was overexpressed (fold change ≥ 2.0) in two liver cancer samples (Fig. 5e). Additional genes with increased expression that were influenced by integration events include *TEKT3*, *CCNA2*, *CDK15* and *THRB* (Fig. 5a).

There was a significant association between HBV viral integrations and somatic copy-number alterations (SCNAs, Fig. 5c). For

samples with HBV integration events, the number of SCNAs was higher on average in the vicinity of viral integration sites (within 1 Mb) compared with samples without HBV integration (mean 4.2 versus 2.3, $P = 7.4 \times 10^{-3}$; two-sided paired t -test). No evidence of an SCNA association was seen for other integrated viruses like HPV16 and HPV18 (Extended Data Fig. 8a,b).

HPV18 integration events were detected in seven tumors in total (Fig. 5b), with the most notable clusters of integration events that affected *TALDO1* (NGC=4) in cervical cancer samples (Extended Data Fig. 7g).

In 20 samples, HPV16 integration events were detected. Genomic clusters of viral integration sites were identified in cervical and head-and-neck cancer samples (Extended Data Fig. 7f). None of these multiple integration events were observed to recur across patients (Fig. 5b). Integration events were also observed in two different long noncoding RNAs (lncRNAs), *LINC00111* and the plasmacytoma variant translocation 1 gene (*PVT1*), an oncogenic lncRNA^{45,46}. Expression of both genes is strongly increased in the cases with HPV16 integration (Extended Data Fig. 8f, Supplementary Table 17).

Using the PCAWG SNV calls¹⁰, we found a significant increase in the number of mutations that occurred within $\pm 10,000$ bp of high-confidence viral integration sites (average number of mutations per sample, 0.41 (HPV16⁺) versus 0.14 (HPV16⁻), $P = 0.02$; one-sided paired t -test, alternative greater, Extended Data Fig. 8c,d). Notably, the integration sites are—compared with a random genome background—enriched in proximity (<1,000 bp) to common fragile sites ($P = 0.0018$, Kolmogorov–Smirnov test). These results suggest that HPV16 integration reflects either characteristics of chromatin features that favor viral integration, such as fragile sites or regions with limited access to DNA repair complexes, or the influence of integrated HPV16 on the host genome. Such a correlation was not seen for the integration sites of other viruses (Extended Data Fig. 8e). Finally, a single AAV2 integration event located in the intronic region of the cancer driver gene *KMT2B*⁴⁷ was detected in one liver cancer sample.

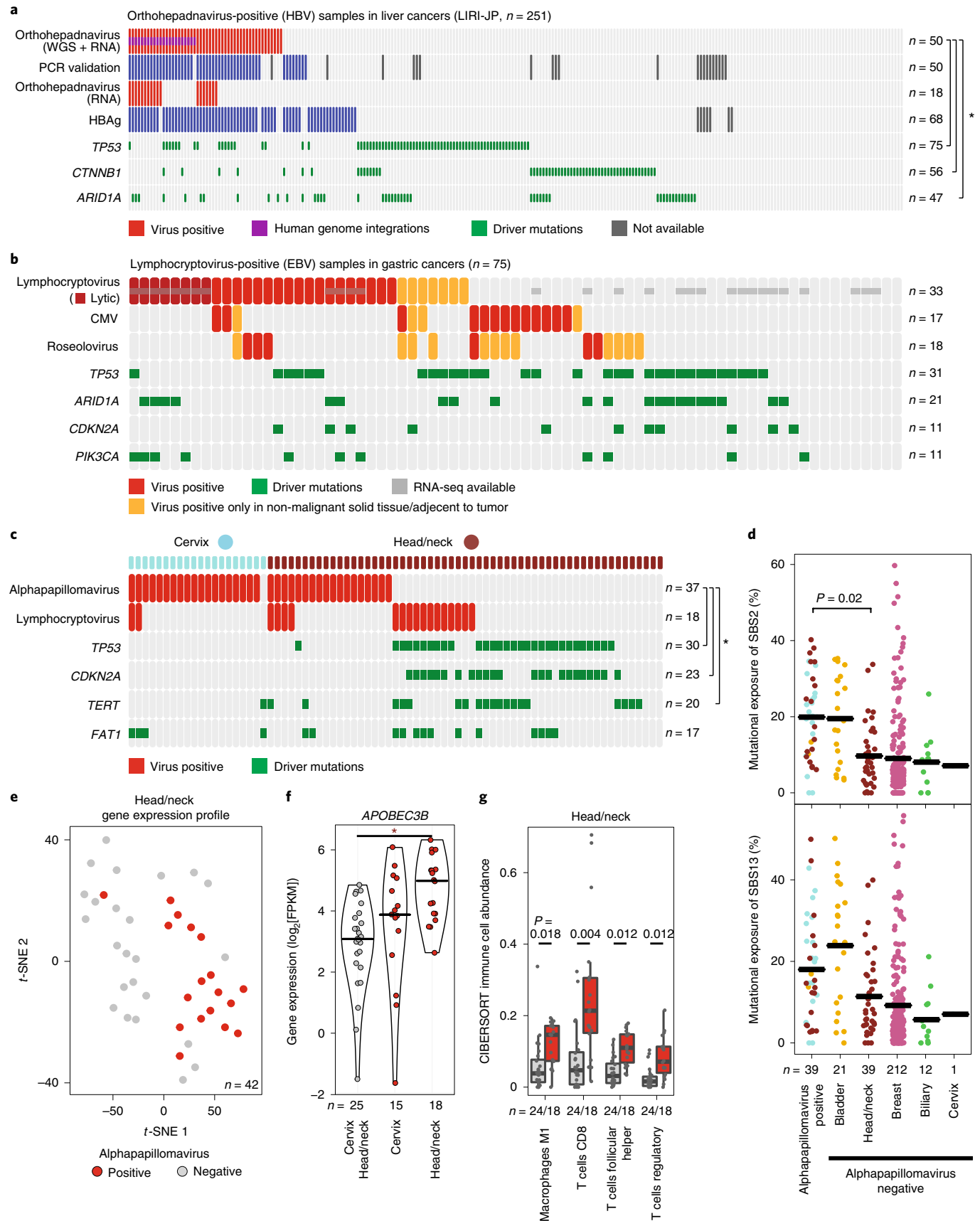
Identification of novel viral species or strains. De novo analysis using the CaPSID pipeline has generated 56 different contigs that have been classified into taxonomic groups at the genus level by CSSSCL⁴⁸. After filtering de novo contigs for their homology to known reference sequences, we identified 29 contigs in 28 different tumor samples that showed low sequence similarity (on average 63%) to any nucleotide sequence contained in the BLAST database. In this respect, our analysis has shown that WGS and RNA-seq can be used to identify isolates from potentially new viral species. However, the total numbers of novel isolates were low in comparison to viral hits to well-defined genera (Fig. 2c). These de novo

Fig. 3 | Virus-specific findings. **a**, HBV detections, validations and driver mutations in liver cancer. The asterisk indicates mutual exclusivity between HBV detection and somatic driver gene mutations. Red boxes represent virus-positive tumor samples, purple boxes show viral genomic integrations, green boxes indicate driver mutations and gray boxes represent missing data. **b**, Virus detections in gastric cancer samples, indication of virus phase (lytic/latent, dark red) and driver mutations (green). A yellow color indicates donors with virus-positive non-malignant samples. The gray box refers to samples with available RNA-seq data. **c**, Virus detections (red) and driver mutations (green) in cervix (blue) and head-and-neck cancer (brown). The asterisk indicates mutual exclusivity between alphapapillomavirus detections and somatic driver gene mutations. **d**, Alphapapillomavirus detection and exposures of mutational APOBEC signatures SBS2 and SBS13. Sample sizes are shown at the bottom. A two-sided Wilcoxon rank-sum test showed a significant difference ($P = 0.02$) of mutational signature exposure between virus-positive and virus-negative head-and-neck tumor samples. The black line indicates the median for each group. **e**, Gene expression analysis based a t -SNE map of head-and-neck cancer samples shows a distinct gene expression profile for virus-positive samples. Virus-positive and virus-negative samples are shown as red and gray dots, respectively. **f**, The violin plot of *APOBEC3B* gene expression for alphapapillomavirus-positive and alphapapillomavirus-negative samples in cervix and head-and-neck cancer (FDR-corrected two-sided Wilcoxon rank-sum test, $P = 1.6 \times 10^{-4}$). FPKM, fragments per kilobase of transcript per million mapped reads. The center line represents the median, and the upper and lower boundaries of the violin plot refer to the maximum and minimum values, respectively. **g**, Tumor-infiltrating immune cells as quantified by CIBERSORT using RNA-seq samples from patient with head-and-neck cancer. All four cell types showed significant enrichment of immune cells in virus-positive samples (FDR-corrected two-sided Wilcoxon rank-sum test, $n = 24$ virus negative versus 18 virus positive). Tukey box plots show the median (the middle line) and the 25–75th percentiles (the box); the whiskers show 1.5x the interquartile range from the lower and upper quartile.

contigs were not enriched for a specific tumor entity but rather were distributed across cancer types including bladder, head-and-neck and cervical cancers (Extended Data Fig. 9).

Discussion

Searching large pan-cancer genome and whole-transcriptome datasets enabled the identification of a high percentage of virus-associated



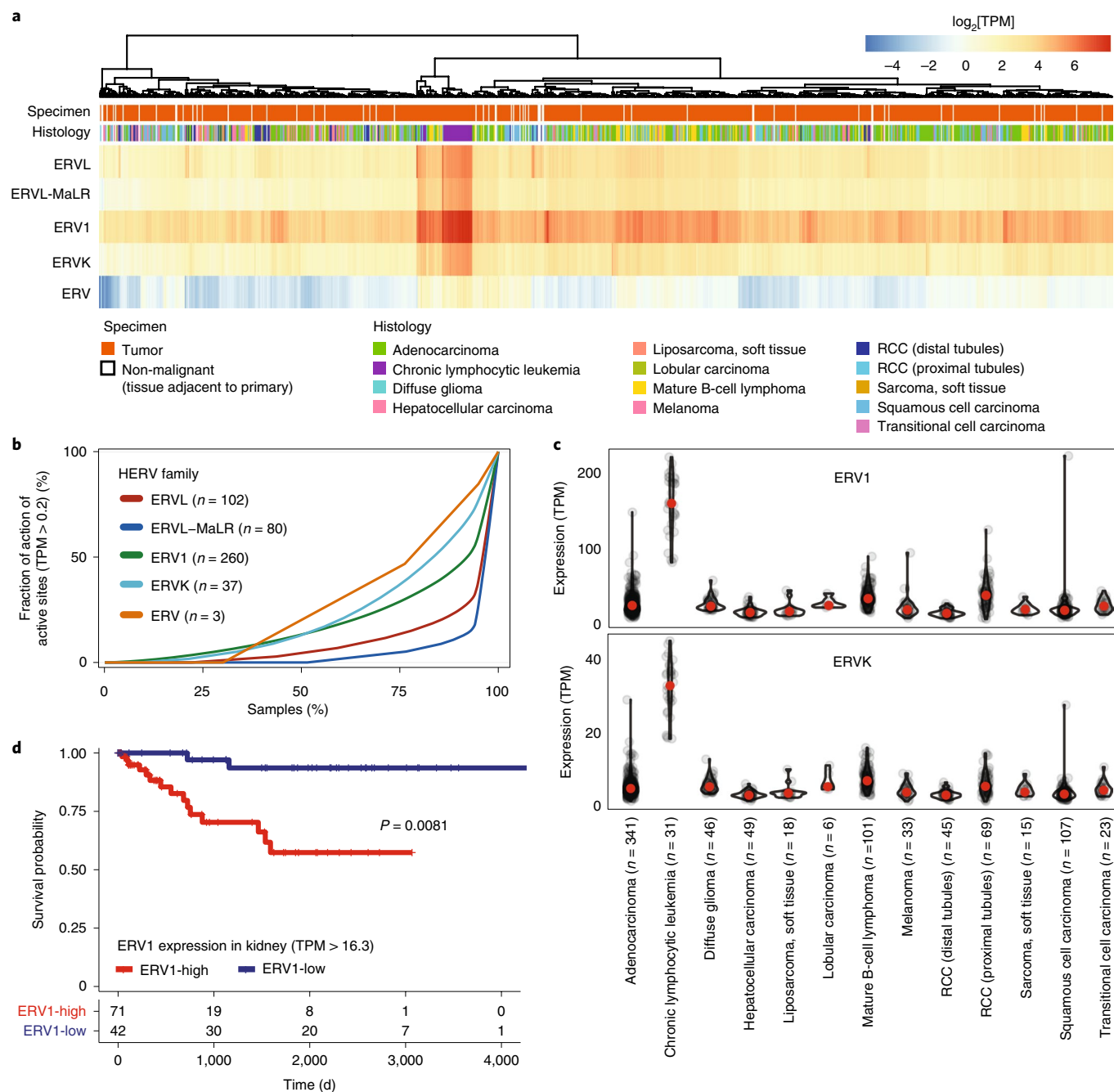


Fig. 4 | Expression of ERVs. **a**, Heat map showing the expression of HERV across all tumor samples. HERV transcripts per million (TPM) were grouped by family and summed up. Hierarchical clustering was performed by family according to Manhattan distance with complete linkage after \log_2 transformation of HERV TPM expression values. (RCC, renal cell carcinoma). **b**, Fraction of active loci in the genome with a TPM > 0.2 plotted against the fraction of samples. **c**, TPM-based expression of the highly expressed HERVs ERV1 and ERVK across tumor types. n , number of analyzed tumor samples. Violin plots are shown; red dots indicate the median. The upper and lower boundaries of the violin plot extend to the maximum and minimum values. **d**, Survival difference between patients with kidney cancer expressing high (red) and low levels (blue) of ERV1. Kaplan-Meier curve shows the overall survival of patients ($n = 113$) with high and low levels of ERV1 with a cut-off of 16.3 TPM (log-rank test $P = 0.0081$). The number of patients at risk is shown at the bottom.

cases (16%). In particular, analysis of tumor genomes, which were sequenced on average to a depth of at least 30-fold coverage, identified considerably more virus-positive cases than investigations of whole-transcriptome data alone, which is the search space analyzed in most previous virome studies. This is probably mainly due to viruses with no or only weak transcriptional activity in the given tumor tissue. Co-infections, generally believed to indicate a weak immune system, were very rare (Extended Data Fig. 3d).

This could, however, also be the result of selection processes during tumorigenesis.

Although universal criteria for a causality of viral pathogens are prone to errors, it is worthwhile to look at individual features that might support a potentially pathomechanistic contribution of a given pathogen. These include aspects that affect the expression of host factors (for example, after viral integration) or the mutual exclusivity of the presence of viral genomes and other host factors,

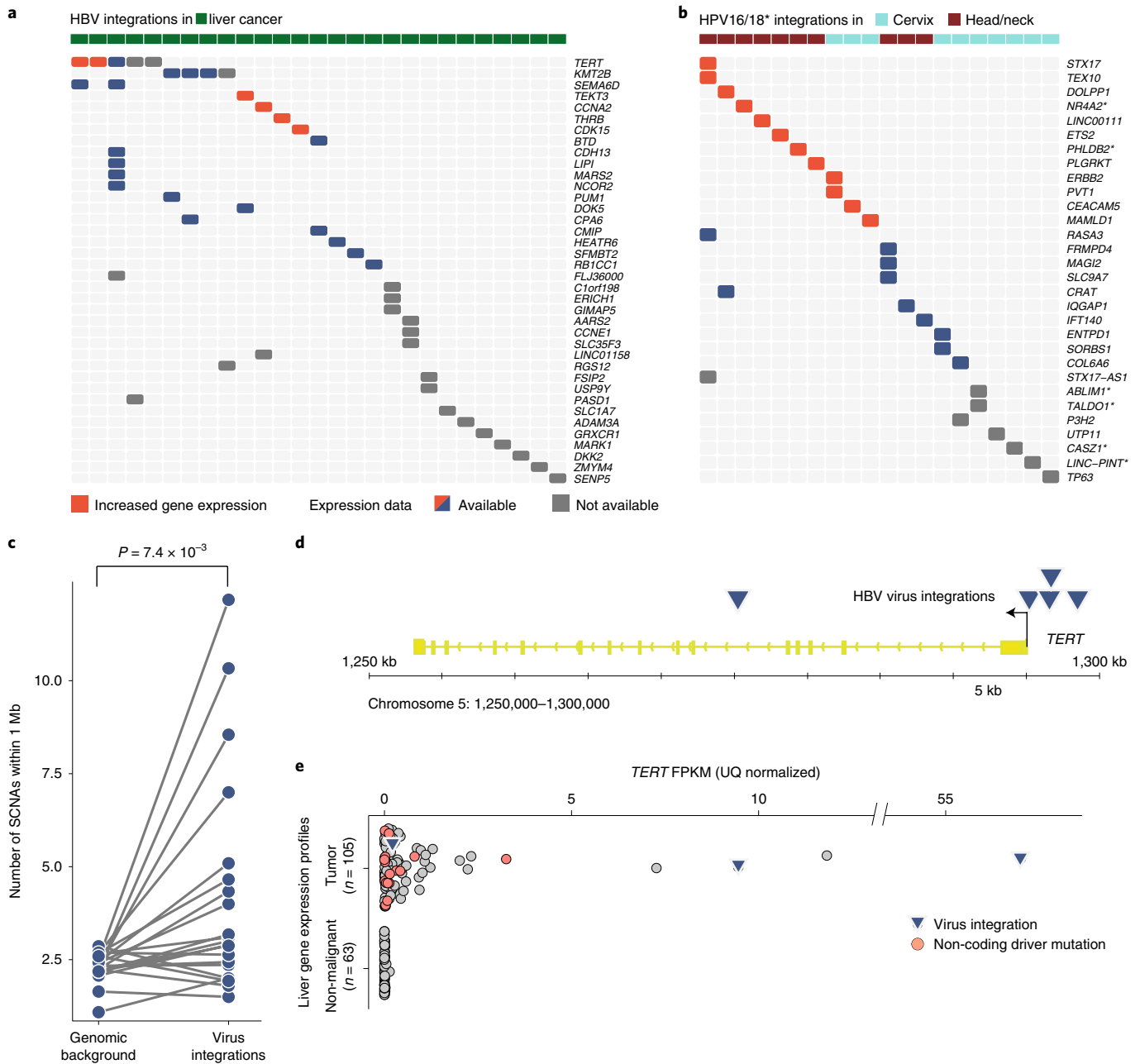


Fig. 5 | The effect of virus integration. **a**, Integration sites detected in gene regions (including promoter, exon, intron and 5' UTR regions) are labeled in red for increased gene expression and blue for expression measured. Rows of each heat map designate the nearest genes to the integration sites, and columns represent individual ICGC donor and project IDs. Intragenic HBV integration sites detected in liver cancers (ICGC project codes: LIRI, LIHC and LINC). For *TERT* and *SEMA6D*, intergenic integrations are also shown. **b**, Integration sites detected for HPV16 and HPV18 in head-and-neck (magenta) and cervical (blue) cancers (ICGC project codes: HNSC and CESC). Gene labels with an asterisk indicate HPV18 as opposed to HPV16 viral integrations. **c**, A local increase in the number of SCNAs was shown in the vicinity of HBV integrations ($n = 21$ viral integrations in individual patients, $P = 7.4 \times 10^{-3}$; two-sided paired *t*-test). **d**, Genomic visualization of the HBV integration sites relative to the *TERT* gene in five patients with liver tumors. **e**, The increased gene expression (in FPKM, upper-quartile normalization, UQ) of *TERT* in two liver tumors with HBV integrations in comparison to the expression of *TERT* in tumor and non-malignant adjacent tissues. Tumor samples with a non-coding driver mutation are labeled in orange.

which are already known to have a role in the etiology of a given tumor type. Such aspects need to be carefully considered when discussing what strengthens the potentially pathogenic role of a virus.

Not surprisingly, known tumor-associated viruses, such as EBV, HBV, HPV16 and HPV18, were among the most frequently detected targets. Notably, viral detection based on WGS showed similar performance with respect to precision and recall as a targeted PCR for HBV, indicating that this approach is sensitive to detect viruses.

This is particularly true for the common integration verified for HBV, HPV16 and HPV18 in our study. In addition, the common theme of potential pathomechanistic effects by the genomic integration of viruses, which were also supported by the observations of multiple nearby integration sites in a given tumor genome that we report in the present study, has gained further momentum. By analyzing the effect of viral integrations on gene expression, we identified several links to genes nearby the integration site. In this

regard, the frequently observed integration of HBV at the *TERT* promoter accompanied with the transcriptional upregulation of *TERT* constitutes an intriguing mechanistic example, as the increased activity of *TERT* is a well-understood driver of carcinogenesis⁴⁹. Furthermore, we also linked viral integrations to increased mutations (SNVs and SCNAs) nearby the integration site.

The known causal role of HPV16 and HPV18 in several tumor entities, which triggered one of the largest measures in cancer prevention, has been the motivation for extensive elucidation of the pathogenetic processes involved. Nevertheless, comprehensive analyses of WGS and RNA-seq datasets revealed additional novel findings. While we confirmed the exclusivity of HPV infection and *TP53*, *CDKN2A* and *TERT* mutations in head-and-neck tumors, we could also link virus presence to an increase in mutations attributed to the mutational signature 2 (ref.⁵⁰). These are explained by the activity of APOBEC, which—among other effects—changes viral genome sequences as a mechanism of cellular defense against viruses^{51,52}. This activation could have an important function in introducing further host genome alterations and, thus, constitute an important mechanism that drives tumorigenesis^{32,52}. In liver cancer, mutations in *CTNBN1*, *TP53* and *ARID1A*, major primary oncogenes in this cancer type and HBV infections were confirmed to occur significantly mutually exclusive²³. Furthermore, the virus-positive head-and-neck cancer samples had a significantly higher abundance of T-cell and M1 macrophage expression signals, which is in agreement with recently described subtypes of head and neck squamous cell carcinoma that differ—among other features—in virus infection and inflammation features.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41588-019-0558-9>.

Received: 30 November 2018; Accepted: 22 November 2019;

Published online: 5 February 2020

References

- Parkin, D. M. The global health burden of infection-associated cancers in the year 2002. *Int. J. Cancer* **118**, 3030–3044 (2006).
- Plummer, M. et al. Global burden of cancers attributable to infections in 2012: a synthetic analysis. *Lancet Glob. Health* **4**, e609–e616 (2016).
- Bouvard, V. et al. A review of human carcinogens—part B: biological agents. *Lancet Oncol.* **10**, 321–322 (2009).
- Muñoz, N., Castellsagué, X., de González, A. B. & Gissmann, L. Chapter 1: HPV in the etiology of human cancer. *Vaccine* **24**, S1–S10 (2006).
- Bialecki, E. S. & Di Bisceglie, A. M. Clinical presentation and natural course of hepatocellular carcinoma. *Eur. J. Gastroenterol. Hepatol.* **17**, 485–489 (2005).
- Hermine, O. et al. Regression of splenic lymphoma with villous lymphocytes after treatment of hepatitis C virus infection. *N. Engl. J. Med.* **347**, 89–94 (2002).
- Thompson, M. P. & Kurzrock, R. Epstein–Barr virus and cancer. *Clin. Cancer Res.* **10**, 803–821 (2004).
- Mesri, E. A., Feitelson, M. A. & Munger, K. Human viral oncogenesis: a cancer hallmarks analysis. *Cell Host Microbe* **15**, 266–282 (2014).
- zur Hausen, H. Oncogenic DNA viruses. *Oncogene* **20**, 7820–7823 (2001).
- The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* <https://doi.org/10.1038/s41586-020-1969-6> (2020).
- Borozan, I. et al. CaPSID: a bioinformatics platform for computational pathogen sequence identification in human genomes and transcriptomes. *BMC Bioinformatics* **13**, 206 (2012).
- Borozan, I., Watt, S. N. & Ferretti, V. Evaluation of alignment algorithms for discovery and identification of pathogens using RNA-seq. *PLoS ONE* **8**, e76935 (2013).
- Nicoll, M. P. et al. The HSV-1 latency-associated transcript functions to repress latent phase lytic gene expression and suppress virus reactivation from latently infected neurons. *PLoS Pathog.* **12**, e1005539 (2016).
- Newman, A. M. et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).
- Krug, L. T. & Pellett, P. E. Roseolovirus molecular biology: recent advances. *Curr. Opin. Virol.* **9**, 170–177 (2014).
- Eliassen, E. et al. Human herpesvirus 6 and malignancy: a review. *Front. Oncol.* **8**, 512 (2018).
- Spandole, S., Cimponeriu, D., Berca, L. M. & Mihăescu, G. Human anelloviruses: an update of molecular, epidemiological and clinical aspects. *Arch. Virol.* **160**, 893–908 (2015).
- van de Berg, P. J. et al. Human cytomegalovirus induces systemic immune activation characterized by a type 1 cytokine signature. *J. Infect. Dis.* **202**, 690–699 (2010).
- García-Martínez, A. et al. Lack of cytomegalovirus detection in human glioma. *Virol. J.* **14**, 216 (2017).
- Fujimoto, A. et al. Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing. *Nat. Genet.* **42**, 931–936 (2010).
- Furuta, M. et al. Characterization of HBV integration patterns and timing in liver cancer and HBV-infected livers. *Oncotarget* **9**, 25075–25088 (2018).
- Canisius, S., Martens, J. W. M. & Wessels, L. F. A. A novel independence test for somatic alterations in cancer shows that biology drives mutual exclusivity but chance explains most co-occurrence. *Genome Biol.* **17**, 261 (2016).
- Kawai-Kitahata, F. et al. Comprehensive analyses of mutations and hepatitis B virus integration in hepatocellular carcinoma with clinicopathological features. *J. Gastroenterol.* **51**, 473–486 (2016).
- Borozan, I., Zapatka, M., Frappier, L. & Ferretti, V. Analysis of Epstein–Barr virus genomes and expression profiles in gastric adenocarcinoma. *J. Virol.* **92**, e01239-17 (2018).
- Mork, J. et al. Human papillomavirus infection as a risk factor for squamous-cell carcinoma of the head and neck. *N. Engl. J. Med.* **344**, 1125–1131 (2001).
- Li, N. et al. Human papillomavirus infection and bladder cancer risk: a meta-analysis. *J. Infect. Dis.* **204**, 217–223 (2011).
- Cao, S. et al. Divergent viral presentation among human tumors and adjacent normal tissues. *Sci. Rep.* **6**, 28294 (2016).
- Travé, G. & Zanier, K. HPV-mediated inactivation of tumor suppressor p53. *Cell Cycle* **15**, 2231–2232 (2016).
- Werness, B. A., Levine, A. J. & Howley, P. M. Association of human papillomavirus types 16 and 18 E6 proteins with p53. *Science* **248**, 76–79 (1990).
- Scheffner, M., Werness, B. A., Huibregtse, J. M., Levine, A. J. & Howley, P. M. The E6 oncoprotein encoded by human papillomavirus types 16 and 18 promotes the degradation of p53. *Cell* **63**, 1129–1136 (1990).
- Henderson, S., Chakravarthy, A., Su, X., Boshoff, C. & Fenton, T. R. APOBEC-mediated cytosine deamination links PIK3CA helical domain mutations to human papillomavirus-driven tumor development. *Cell Rep.* **7**, 1833–1841 (2014).
- Burns, M. B., Temiz, N. A. & Harris, R. S. Evidence for APOBEC3B mutagenesis in multiple human cancers. *Nat. Genet.* **45**, 977–983 (2013).
- Schlecht, N. et al. Gene expression profiles in HPV-infected head and neck cancer. *J. Pathol.* **213**, 283–293 (2007).
- Nelson, P. N. et al. Demystified. Human endogenous retroviruses. *Mol. Pathol.* **56**, 11–18 (2003).
- Paces, J. et al. HERVd: the human endogenous retroviruses database: update. *Nucleic Acids Res.* **32**, D50 (2004).
- Pavlicek, A., Paces, J., Elleder, D. & Hejnar, J. Processed pseudogenes of human endogenous retroviruses generated by LINEs: their integration, stability, and distribution. *Genome Res.* **12**, 391–399 (2002).
- Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11 (2015).
- Ohnuki, M. et al. Dynamic regulation of human endogenous retroviruses mediates factor-induced reprogramming and differentiation potential. *Proc. Natl Acad. Sci. USA* **111**, 12426–12431 (2014).
- Smith, C. C. et al. Endogenous retroviral signatures predict immunotherapy response in clear cell renal cell carcinoma. *J. Clin. Invest.* **128**, 4804–4820 (2018).
- Tang, K.-W. & Larsson, E. Tumour virology in the era of high-throughput genomics. *Phil. Trans. R. Soc. Lond. B* **372**, 20160265 (2017).
- Jiang, Z. et al. The effects of hepatitis B virus integration into the genomes of hepatocellular carcinoma patients. *Genome Res.* **22**, 593–601 (2012).
- Hu, Z. et al. Genome-wide profiling of HPV integration in cervical cancer identifies clustered genomic hot spots and a potential microhomology-mediated integration mechanism. *Nat. Genet.* **47**, 158–163 (2015).
- Zhao, L.-H. et al. Genomic and oncogenic preference of HBV integration in hepatocellular carcinoma. *Nat. Commun.* **7**, 12992 (2016).
- Li, X. et al. The function of targeted host genes determines the oncogenicity of HBV integration in hepatocellular carcinoma. *J. Hepatol.* **60**, 975–984 (2014).
- Shen, C.-J., Cheng, Y.-M. & Wang, C.-L. lncRNA PVT1 epigenetically silences miR-195 and modulates EMT and chemoresistance in cervical cancer cells. *J. Drug Target.* **25**, 637–644 (2017).
- Tang, K.-W., Alaei-Mahabadi, B., Samuelsson, T., Lindh, M. & Larsson, E. The landscape of viral expression and host gene fusion and adaptation in human cancer. *Nat. Commun.* **4**, 2513 (2013).

47. Nault, J.-C. et al. Recurrent AAV2-related insertional mutagenesis in human hepatocellular carcinomas. *Nat. Genet.* **47**, 1187–1193 (2015).
48. Borozan, I. & Ferretti, V. CSSSCL: a Python package that uses combined sequence similarity scores for accurate taxonomic classification of long and short sequence reads. *Bioinformatics* **32**, 453–455 (2015).
49. Sung, W. K. et al. Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma. *Nat. Genet.* **44**, 765–769 (2012).
50. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.* **3**, 246–259 (2013).
51. Wallace, N. A. & Münger, K. The curious case of APOBEC3 activation by cancer-associated human papillomaviruses. *PLoS Pathog.* **14**, e1006717 (2018).
52. Roberts, S. A. et al. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat. Genet.* **45**, 970–976 (2013).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020

PCAWG Consortium

Ivan Borozan², Daniel S. Brewer^{3,4}, Colin S. Cooper^{4,13}, Nikita Desai^{7,8}, Roland Eils^{14,15,16}, Vincent Ferretti^{17,18}, Adam Grundhoff⁵, Murat Iskar¹, Kortine Kleinheinz^{14,15}, Peter Lichter^{1,10}, Hidewaki Nakagawa²², Akinyemi I. Ojesina^{23,24,25}, Chandra Sekhar Pedamallu^{26,27,28}, Matthias Schlesner^{14,29}, Xiaoping Su³⁰ and Marc Zapatka¹

²²RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. ²³Department of Epidemiology, University of Alabama at Birmingham, Birmingham, AL, USA. ²⁴HudsonAlpha Institute for Biotechnology, Huntsville, AL, USA. ²⁵O'Neal Comprehensive Cancer Center, University of Alabama at Birmingham, Birmingham, AL, USA. ²⁶Broad Institute of MIT and Harvard, Cambridge, MA, USA. ²⁷Harvard Medical School, Boston, MA, USA. ²⁸Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA. ²⁹Bioinformatics and Omics Data Analytics, German Cancer Research Center (DKFZ), Heidelberg, Germany. ³⁰University of Texas MD Anderson Cancer Center, Houston, TX, USA.

Methods

Identifying potential pathogenic reads. To reduce the number of reads to be considered for the pathogen search, we identified potential pathogenic reads by using P-DiP (<https://github.com/mzapatka/p-dip>). On the basis of reads aligned to hg19 by BWA⁵³ or STAR⁵⁴ using the standard PCAWG approach, we identified read pairs for which at least one read did not map well to the human genome (longest stretch of mapped bases from 20 to 30 bases) and read pairs that were unmapped or mapped to NC_007605 (human herpesvirus 4, which is contained in the 1000 Genomes version of the hg19 human reference genome), and extracted these for further processing. To speed up the extraction, we used bamcollate2 from Biobambam2 (v.2.08)⁵⁵ as an input stream to the Python script.

Identification of ERVs. The expression of ERVs was analyzed using RNA-seq data and aligned STAR sequences based on the settings developed within PCAWG (hg19 and Gencode 19). In contrast to the standard pipeline, the reference transcripts from Gencode 19 were enriched by adding HERV locations extracted from RepeatMasker (<http://www.repeatmasker.org>, rmsk from UCSC, version 17/08/03) and Featurecounts (subread-1.5.3)⁵⁶ applied to identify reads mapping to the modified reference transcripts. Resulting reads counts were converted into TPM according to Wagner et al.⁵⁷

The SEPATH pipeline. Our starting point is to take reads that are not mapped to the human genome, using the extracted potentially pathogenic reads. Low quality bases ($q < 30$) were trimmed from the read ends and the TruSeq indexed adapter and TruSeq universal adapter were removed using Cutadapt (v.1.8.1)⁵⁸. Reads less than 32 bp were discarded. Additional filtering was performed to remove reads that contained more than 5% of Ns or those with low complexity (dust method with maximum score of 10) by using Prinseq (v.0.20.3)⁵⁹. Metagenomic Phylogenetic Analysis (MetaPhlan)^{60,61} was then applied to identify and quantify the presence of bacterial and viral populations. MetaPhlan comes with a curated marker database of around 1 million unique clade-specific marker genes identified from reference genomes (version 2.0 of the database was used). Reads were aligned against the unique marker gene database by using Bowtie2 (v.2.2.1)⁶² with presets set to sensitive. Reads were then counted and normalized giving an estimation of the relative abundance for each level of the phylogenetic tree.

Detection and analysis of microbial infectious agents by NGS P-DiP. The assembly-based pipeline (P-DiP) was further developed based on a version implemented by M.A. and A.G.⁶³. In summary, the pipeline runs preprocessing, assembly and BLAST searches and stores processing details and final results in a PostgreSQL database. For the WGS and RNA-seq analyses, we started with the potentially pathogenic reads extracted from the BWA-aligned WGS BAM files. As a first step, reads were trimmed based on quality using trimmomatic. Thereafter, host reads were subtracted by aligning to the human reference genome (WGS: hg19 excluding NC_007605 and hs37d5 and adding phiX; RNA sequencing: Homo_sapiens.GRCh37.dna.primary_assembly) using Bowtie2 (v.2.2.8)⁶². Trinity (v.2.0.6)⁶⁴ was used for the read assembly of WGS reads that were not aligned by Bowtie with sufficient quality (not aligned with --very-fast (-D 5 -R 1 -N 0 -L 22 -i S,0,2.50) to Homo_sapiens.GRCh37.ncrna, Homo_sapiens.GRCh37.cdna.all or PhiX); for the RNA-seq data we applied idba assembler (v.1.1.3)⁶⁵. Assembled contigs were filtered by size (minimal length of 300 bp). Abundance was estimated by remapping all of the reads that did not align to the human reference to the assembled contigs by using Bowtie2. Putative PCR duplicates identified by mapping location were removed from the abundance count. The taxonomic classification of the size-filtered contigs was performed using the BLAST+ package (v.2.2.30)⁶⁶ and nucleotide databases nt (<ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nt.gz>, accessed 15 May 2015) and nr (<ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz>, accessed 20 April 2015). For the extraction of pathogen hits R-scripts were used to filter the BLAST results (<https://github.com/mzapatka/p-dip>). In summary, for each of the contigs, the best BLAST hits for each segment of the contig were considered and the reads aligning to these segments identified. Potential contaminants were defined based on the taxonomy annotation in NCBI taxonomy. Any taxonomy ID below plasmids (36,549), transposons (2,387), midvariant sequences (31,896), insertion sequences (2,673), artificial sequences (81,077) and synthetic viruses (512,285) was annotated as potential contamination. Segments with higher read counts of these sequences compared to pathogen hits were flagged as contaminants and not further considered.

CaPSID description of the analysis workflow. The metagenomic analysis pipeline of CaPSID¹¹ starts by first processing a BAM file that contains the reads sequenced from a tumor (or normal) sample aligned to the human reference sequence (GRCh37/hg19). Reads that did not map to the human reference were extracted and filtered for low complexity and quality using the SGA⁶⁷ preprocessing module and then aligned in single-end mode using the Bowtie2 aligner⁶² to 5,652 NCBI⁶⁸ viral reference sequences (RefSeq) and a filter sequence reference database composed of 5,242 bacterial and 1,138 fungal reference sequences that were also downloaded from the NCBI. To improve the sensitivity and specificity with which viral sequences were detected, reads that did not map to any reference with Bowtie2 were realigned to the same viral RefSeq database, using the more-sensitive

aligner SHRiMP2 in local alignment mode⁶⁹. At the completion of this two-step alignment process, reads that aligned to viral reference sequences were annotated using the information stored in the genome database of CaPSID, which contains full NCBI GenBank and taxonomic information. Using information from each aligned read, CaPSID then calculates the following four metrics: (1) the total number of reads (or hits) that aligned across any given viral genome, (2) the total number of reads that aligned only across gene regions within any given viral genome, (3) the total coverage across each viral genome and (4) the maximum coverage across any of the genes in a given viral genome.

Filtering of viral candidates with low significance. For the analysis of the tumor WGS or RNA-seq samples, CaPSID reports candidate sequences from dozens of different viral genomes, some of which are not related to the cancer phenotype. Some of these reported viral hits are also due to a series of experimental and computational artifacts. To reduce the number of potential false-positive hits, the CaPSID pipeline flags viral genomes that could be the result of artifacts present in the sequencing data or those with no obvious relation to cancer phenotype and that could be filtered in subsequent steps. The following criteria were used to flag and filter for potential viral candidates: (1) flag viral candidates with low coverage, (2) flag bacteriophage viral genome sequences, (3) report only viral candidates with a read composition different from the one expected when generated from the host's reference GRCh37/hg19 sequence, (4) flag viral candidates that are typically not known to infect humans and those with low read abundance and/or low overall alignment read accuracy.

In the first step, CaPSID flagged viral genomes with low read count and/or coverage using three metrics, including total number of uniquely aligned reads <3, total genome coverage <10% and maximum gene coverage <50%. Viral genomes with low read count can arise as a result of (1) low read/transcript abundance in the human sequenced sample, (2) unspecific alignment between sequenced short reads (for example, low complexity reads) and viral reference sequences and (3) for RNA-seq library preparation in which highly expressed transcripts generally dominate over low abundance targets. To limit the reporting of viral genomes with very low coverage, we chose to flag all genomes for which the maximum gene coverage was <50%. As this lower bound on the maximum gene coverage applied to individual genes and not to the complete viral genome, it appears to be unlikely that viruses with such low coverage are biologically important. The second step in our filtering approach was to flag bacteriophage viral genomes that are most likely not related to any cancer phenotype. Bacteriophages are detected as a result of the presence of bacteria (or bacterial contamination) in human sequenced samples. The third step was used to determine whether the genome coverage observed for each viral candidate was different from the one expected to arise from reads that originated exclusively from the human reference DNA GRCh37/hg19 sequence. To build the CaPSID background model, we used the ART NGS read simulator. The entire GRCh37/hg19 reference file is first fed to the ART⁷⁰ simulator (parameters: art_illumina [Illumina platform] -l [read length=100 bp] -f [the fold of read coverage to be simulated=100] with default values for indels and substitution rates), which then generates single-end (or paired-end) reads and base quality values.

Reads simulated by ART were then aligned to the viral reference sequence database using the same alignment approach for reads that originated from tumor samples (see above). CaPSID then calculated the four metrics for the GRCh37/hg19 background model using the alignment information from simulated reads that aligned to viral reference sequences. The fourth step consisted of flagging viral candidates that were typically not known to infect humans using a dictionary of around 130 terms that we compiled from a database of all viruses known to infect humans. In addition to the above filtering criteria, CaPSID also considered the read abundance associated with each viral candidate sequence (abundance is expressed in terms of aligned reads in parts-per million of total number of unmapped reads) and the average read percentage identity with which reads aligned to a given viral candidate reference sequence.

De novo assembly and taxonomic classification of contigs. The purpose of this analysis step is to attempt to characterize potential novel viral sequences at the species or subspecies level. Unaligned reads that could not be aligned to any of the filter/host or viral reference sequences were assembled into contigs using the IDBA algorithm⁶⁵. Assembled contigs were then masked for repeat regions by using RepeatMasker and then filtered for their size and read coverage (contig length ≥ 500 bp and coverage $>5\times$). Resulting contigs were then assigned to taxonomic groups at the genus level by using the CSSSCL algorithm⁴⁸. Contigs lacking sequence homology to reference sequences contained in the CaPSID or BLAST nucleotide databases with percentage identity <90% were then selected as suggestive of the presence of new viral strains/isolates or species.

Defining consensus hits. Identification of the consensus hits was achieved by optimizing two features of the individual genus hits: PMER 1 as cut-off (Supplementary Note) and percentage identity >90%. The 90% percentage identity threshold was determined based on our benchmarking study¹² that indicated that an alignment-based approach can still accurately characterize viral sequences with up to 10% mutation rate (compared with sequences stored in a reference database). Lowering the threshold, with which short reads align to any given

reference sequence below 90% identity on average, results in a drop of sequence coverage due to a high attrition rate of aligned reads, lowering the detection rate and thus providing more uncertain characterizations of viral candidates. Notably, there was no difference in the PMER distribution of common hits across the three pipelines, indicating that a common detection cut-off is reasonable (Extended Data Fig. 3b).

The consensus set was restricted to genera that were covered in at least two detection pipelines (Extended Data Fig. 1b). Notably, we could not detect any more hits with high PMER using the unique search space of P-DiP, indicating that almost all of the viral hits from individual pipelines were also screened by another pipeline.

Virus integration detection analysis. A subset of viral candidates identified to be present in tumor samples by the CaPSID analysis pipeline (parameters used: PMER ≥ 1.1 and genome coverage $>$ simulated background model) was selected for the detection of viral integration events using the VERSE⁷¹ algorithm. This subset of viruses included: herpesviruses (HHV1, HHV2, HHV4, HHV5, HHV6A/B), simian virus 40 (SV40) and 12 (SV12), human immunodeficiency virus (HIV1), human and simian T-cell lymphotropic virus type 1 (HTLV1 and STLV1), BK polyomavirus (BKP), human parvovirus B19, mouse mammary tumor virus, murine type C retrovirus, Mason–Pfizer monkey virus, HBV, HPV (HPV16, HPV18 and HPV6a) and AAV2. Below we describe the steps used for the viral integration detection analysis.

Viral integration events in the host can be detected by using paired-end NGS technologies that facilitate the detection of genomic rearrangements, as well as gene fusions and novel transcripts. VERSE is capable of determining virus integration sites within a single base resolution by requiring the presence of both chimeric and soft clipped reads. In addition, VERSE improves the detection through customizing reference genomes and was shown to substantially enhance the sensitivity of the detection of virus integration sites⁷¹. VERSE categorizes its predictions into one of two classes: (1) a high confidence hit with a single base resolution—if there was a sufficient number of soft-clipped reads to support an integration locus so that CREST was able to detect it; or (2) a low confidence hit with a 10-bp resolution for which CREST failed to detect an integration event because of the lack of high-quality soft-clipped reads.

To further limit the false-positive rate associated with viral integration sites, we compared results obtained with VERSE to those from a previous study⁷². Out of 64 WGS liver cancer samples with HBV integration events that were reported previously⁷², 50 were part of the PCAWG dataset analyzed in this study. Of those, 45 out of 50 tested positive for HBV when analyzed by CaPSID (filtering criteria used: PMER ≥ 1.0 , genome coverage $>$ host background model and read percentage identity $\geq 89\%$). In addition, 50 of these WGS samples had 23 matching whole-transcriptome samples and 22 of these were identified to be positive for HBV by CaPSID (filtering criteria used: maximum gene coverage $\geq 50\%$, read percentage identity $\geq 89\%$ and PMER ≥ 1.0). By combining WGS and RNA-seq tumor samples, 47 out of 50 samples tested positive for HBV when analyzed by CaPSID.

Using VERSE, virus integration sites were detected in 28 out of 47 (60%) of these. This result indicates that for a subset of viral integration events, VERSE might be a more stringent approach compared to previously used methods⁷². This can be explained by the fact that VERSE requires both the presence of paired-end chimeric and soft clipped reads whereas the previously described method⁷² relied only on paired-end reads. To explore these results further, we compared integration sites obtained with VERSE and those described previously⁷² with an overlapping window of 10 bp. Our analysis indicates that among 23 integration sites identified by VERSE in RNA-seq data and that overlap with the previously published results⁷², 91% were classified with high confidence hits and only 9% with low (N total overlap = 23, high = 21 (91%) and low = 2 (9%)). However, a similar result was not observed for integration events found using WGS data (N total overlap = 14, high = 6 (43%), low = 8 (57%)), for which the proportion of integration events classified as high and low was similar.

Thus, our analysis indicates that one important factor for improving the agreement between these two datasets is the confidence level assigned by VERSE to each candidate integration site—but only in the case when integration sites are detected using RNA-seq data. To reduce the potential number of false-positive hits, we decided to use all integration sites predicted by VERSE when these were obtained using WGS data and only high-confidence calls when using RNA-seq data.

Contaminations. On the basis of the presence of vector sequences in the contig assembled by P-DiP and the background model from CaPSID, we could identify which virus hits originated from common laboratory contaminants or were due to sequence similarities to the human genome. In addition, we filtered known contaminants (see below). For P-DiP, we filtered all hits that did not have more target reads than any artificial sequence (excluding artificial viruses) on an individual contig region. Hits caused by vector and other artificial sequences were identified by analyzing the assembled contigs for combined hits to viral pathogens and artificial sequences. Checking viral hits that occurred at least 40 times in such a contig, we could clearly separate contaminants from viral pathogens.

The gammaretrovirus hits (NCBI taxonomy ID 153135; species, murine leukemia virus) were also marked as artifacts, on the basis of the additional BLAST hits of the corresponding contigs to the *Mus musculus* genome by P-DiP as well as the background model of the CaPSID pipeline, which was designed to limit the number of spurious hits. Most of the frequent virus hits prone to contamination by artificial sequences were lambda-likevirus, alphabaculovirus, microvirus, simplexvirus, hepacivirus, CMV, orthopoxvirus and punalikevirus. However, restricting to at least 1 PMER for the potential virus hit contaminants reduced these to one CMV case.

Filtering contaminants. We filtered all Microviridae (taxonomy ID 10841) because of the phix174 spike-in used during sequencing. Caudovirales (taxonomy ID 28883), tailed bacteriophages, were removed as they typically infect bacterial hosts. Baculoviridae were filtered because these infect insect cells and are commonly used in the laboratory. The virus coverage was analyzed by aligning the potentially pathogenic reads with BWA-mem to the human hg19 reference genome after adding the respective virus reference sequence that was most frequently detected within the genus. Coverage was thereafter calculated base specific using BEDTools coverage. As we identified EBV in all 14 normal blood controls from ovarian cancer that were EBV immortalized, these were removed from the virus hits.

Integration of external PCAWG datasets. We tested for mutual exclusivity, for example, between virus detections and driver gene mutations by applying DISCOVER²². On the basis of the gene expression data, immune-cell proportions were analyzed by CIBERSORT¹⁴. For survival analysis, Cox proportional hazards analysis was performed using R libraries 'survival' and 'survminer' for the figures. The optimal cut points were identified by maxstat using a previously described method⁷³ (library maxstat).

Virus load. The viral load in relation to the human genome equivalents was calculated based on the human bases sequenced (read length \times number of reads mapped to the human genomes), tumor sample purity (if available or 100% otherwise) assuming a ploidy of two and using a human genome size of 2,897,310,462 bases (the mappable part of the human genome). This number of human genome equivalents was then related to the viral genome equivalents that were calculated based on the number of identified viral reads, read length and virus genome size.

$$\text{tumor genome equivalents} = \frac{\text{read length} \times \text{number of reads mapped to the human genome}}{\text{mappable human genome size} \times \text{tumor ploidy}} \times \text{tumor purity}$$

$$\text{virus genome equivalents} = \frac{\text{read length} \times \text{number of viral sequences}}{\text{virus genome size}}$$

$$\text{virus load} = \frac{\text{virus genome equivalents}}{\text{tumor genome equivalents}}$$

Human research participants. The ethics oversight for the PCAWG protocol was undertaken by the TCGA Program Office and the Ethics and Governance Committee of the ICGC. Each individual ICGC and TCGA project that contributed data to PCAWG had its own local arrangements for ethics oversight and regulatory alignment.

Statistics. If not specified otherwise, we used two-sided Wilcoxon rank-sum tests for groups with $n > 3$. Further details can be found in the Nature Research Reporting Summary.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Somatic and germline variant calls, mutational signatures, subclonal reconstructions, transcript abundance, splice calls and other core data generated by the ICGC/TCGA PCAWG Consortium are described in an associated paper¹⁰ and are available for download at <https://dcc.icgc.org/releases/PCAWG>. Additional information on accessing the data, including raw read files, can be found at <https://docs.icgc.org/pcawg/data/>. In accordance with the data-access policies of the ICGC and TCGA projects, most molecular, clinical and specimen data are in an open tier that does not require access approval. To access potentially identifying information, such as germline alleles and underlying sequencing data, researchers will need to apply to the TCGA Data Access Committee (DAC) via dbGaP (<https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=login>) for access to the TCGA portion of the dataset, and to the ICGC Data Access Compliance Office (DACO; <http://icgc.org/daco>) for the ICGC portion. In addition, to access somatic SNVs derived from TCGA donors, researchers will need to obtain dbGaP authorization. Datasets described specifically in this manuscript can be found in the Supplementary Tables.

Code availability

The core computational pipelines used by the PCAWG Consortium for alignment, quality control and variant calling are available to the public at <https://dockstore.org/search?search=pcawg> under the GNU General Public License v.3.0, which enables the reuse and distribution of the pipelines. The pathogen-discovery pipeline P-DiP is available on GitHub (<https://github.com/mzapatka/p-dip>). CaPSID is available from GitHub (pipeline, <https://github.com/capsid/capsid-pipeline>; webapp, <https://github.com/capsid/capsid-webapp>). The taxonomic classifier CSSSCL is available from GitHub (<https://github.com/oicr-ibc/cssscl>).

References

53. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at <https://arxiv.org/abs/1303.3997v2> (2013).
54. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
55. Tischler, G. & Leonard, S. Biobambam: tools for read pair collation based algorithms on BAM files. *Source Code Biol. Med.* **9**, 13 (2014).
56. Liao, Y., Smyth, G. K. & Shi, W. FeatureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
57. Wagner, G. P., Kin, K. & Lynch, V. J. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.* **131**, 281–285 (2012).
58. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10 (2011).
59. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**, 863–864 (2011).
60. Truong, D. T. et al. MetaPhlan2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12**, 902–903 (2015).
61. Segata, N. et al. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* **9**, 811–814 (2012).
62. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
63. Fischer, N. et al. Rapid metagenomic diagnostics for suspected outbreak of severe Pneumonia. *Emerg. Infect. Dis.* **20**, 1072–1075 (2014).
64. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
65. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420–1428 (2012).
66. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
67. Simpson, J. T. & Durbin, R. Efficient de novo assembly of large genomes using compressed data structures. *Genome Res.* **22**, 549–556 (2012).
68. Pruitt, K. D., Tatusova, T., Klimke, W. & Maglott, D. R. NCBI reference sequences: current status, policy and new initiatives. *Nucleic Acids Res.* **37**, D32–D36 (2009).
69. David, M., Dzamba, M., Lister, D., Ilie, L. & Brudno, M. SHRiMP2: Sensitive yet practical short read mapping. *Bioinformatics* **27**, 1011–1012 (2011).

70. Huang, W., Li, L., Myers, J. R. & Marth, G. T. ART: a next-generation sequencing read simulator. *Bioinformatics* **28**, 593–594 (2012).
71. Wang, Q., Jia, P. & Zhao, Z. VERSE: a novel approach to detect virus integration in host genomes through reference genome customization. *Genome Med.* **7**, 2 (2015).
72. Fujimoto, A. et al. Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. *Nat. Genet.* **48**, 500–509 (2016).
73. Lausen, B. & Schumacher, M. Maximally selected rank statistics. *Biometrics* **48**, 73–85 (1992).

Acknowledgements

We thank the IT Core Facility at the DKFZ for technical assistance, M. Hain and R. Kabbe for computational support, S. Gerhardt for technical assistance with the validation experiments. We acknowledge the contributions of the many clinical networks across the ICGC and TCGA who provided samples and data to the PCAWG Consortium, and the contributions of the Technical Working Group and the Germline Working Group of the PCAWG Consortium for collation, realignment and harmonized variant calling of the cancer genomes used in this study. We thank the patients and their families for their participation in the individual ICGC and TCGA projects. V.F. and I.B. received support for their work from the Ontario Institute for Cancer Research (OICR) through funding provided by the government of Ontario. A.G. received support for his work from the Leibniz Association (grant number SAW-2015-IPB-2) and the German Center for Infection Research (grant number TTU 01.801). P.L. and A.G. received support for this work from the German Federal Ministry of Education and Research (BMBF BioTop grant number 01EK1502C, ICGC-DE-Mining grant number 01KU1505A-G). D.S.B. and C.S.C. received support from Cancer Research UK C5047/A14835/A22530/A17528, the Dallaglio Foundation, Bob Champion Cancer Trust, The Masonic Charitable Foundation successor to The Grand Charity, The King Family and the Stephen Hargrave Trust. H.M. was supported by a Swiss National Science Foundation grant (number S-87701-03-01).

Author contributions

M.Z. and P.L. jointly supervised research. V.F., R.E., C.S.C., M.I., I.B., M.Z. and P.L. conceived and designed the experiments. H.S. performed experiments. M.I., D.S.B., I.B. and M.Z. performed statistical analysis. N.D., M.I., A.G., D.S.B., I.B. and M.Z. analyzed the data. V.F., R.E., C.S.C., H.M., M.A., A.G., D.S.B., I.B. and M.Z. contributed reagents, materials and/or analysis tools. M.I., D.S.B., I.B., M.Z. and P.L. wrote the paper. V.F., A.G., C.S.C., D.S.B., M.I., I.B., M.Z. and P.L. critiqued manuscript for intellectual content.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41588-019-0558-9>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41588-019-0558-9>.

Correspondence and requests for materials should be addressed to P.L.

Reprints and permissions information is available at www.nature.com/reprints.

SCIENTIFIC REPORTS

OPEN

Recovery of the first full-length genome sequence of a parapoxvirus directly from a clinical sample

Thomas Günther¹, Ludwig Haas², Malik Alawi^{1,3}, Peter Wohlsein⁴, Jerzy Marks⁵, Adam Grundhoff^{1,6}, Paul Becher^{1,2,7} & Nicole Fischer^{6,8}

We recovered the first full-length poxvirus genome, including the terminal hairpin region, directly from complex clinical material using a combination of second generation short read and third generation nanopore sequencing technologies. The complete viral genome sequence was directly recovered from a skin lesion of a grey seal thereby preventing sequence changes due to *in vitro* passaging of the virus. Subsequent analysis of the proteins encoded by this virus identified genes specific for skin adaptation and pathogenesis of parapoxviruses. These data warrant the classification of seal parapoxvirus, tentatively designated SePPV, as a new species within the genus *Parapoxvirus*.

Parapoxviruses (PPVs) form a genus of the family *Poxviridae*. Poxviruses are large double stranded DNA viruses with genomes of approximately 135 to 360 kbp, which contain up to 328 open reading frames¹. According to the International Committee on Taxonomy of Viruses (ICTV; <http://ictvonline.org/virusTaxonomy.asp>), the genus *Parapoxvirus* comprises the following species members: the Orf virus (ORFV), considered the prototype parapoxvirus causing contagious pustular dermatitis in small ruminants, the Bovine papular stomatitis virus (BPSV), the Pseudocowpox virus (PCPV) and the Parapoxvirus of red deer in New Zealand (PVNZ). Beside these viruses with known full-length nucleotide sequences, a number of tentative species have been proposed based on the amplification of smaller genome fragments using pan-PCR primers encompassing 250–550 bp of the DNA polymerase, DNA topoisomerase I and major envelope protein encoding regions. These species comprise reindeer musk ox parapoxviruses^{2,3}, cattle parapoxviruses⁴, pinniped parapoxviruses⁵ and a very recently described, putative novel parapoxvirus in horses⁶. PPVs are considered as zoonotic and can cause circumscribed skin lesions in humans, historically best known as milker's nodules. Typical lesions of this type have also been described in humans who have come in contact with PPV infected seals^{7,8}.

Parapoxvirus infections have been reported in different seal species and sea lions of the Atlantic and Pacific oceans including habitats in the sub-arctic, arctic, and antarctic waters^{5,9–16}. They typically cause pustular skin lesions and ulcerations around the mouth and on the flippers of the animals as well as mucosal lesions with ulcerations in the oral cavity. Parapoxvirus infections are diagnosed by clinical evaluation of skin and mucosal lesions together with electron microscopy or immunohistochemistry analyses, isolation of viral particles and/or detection of viral sequences by polymerase chain reaction (PCR). Infections are generally self-limiting after 1–6 weeks, but in the case of bacterial superinfections they may result in severe ulcerative and necrotizing lesions. Since the full-length genome sequence of seal parapoxvirus had previously not been determined, classification of the virus was derived from short sequence fragments amplified from infected tissues^{5,7–10,12–19}. Based on phylogenetic analysis of these sequences, it was suggested that the parapoxviruses of seals may belong to a separate species within the genus *Parapoxvirus*⁵.

Here, we report a clinical case of a PPV infection in a grey seal (*Halichoerus grypus*) found in the Baltic Sea in Poland in April 2015. Primary laboratory diagnosis was based on histologic evaluation and *in situ* hybridization, as well as electron microscopy and PCR analysis. Next generation sequencing (NGS) was performed on DNA

¹Heinrich-Pette Institute, Leibniz Institute for Experimental Virology, Hamburg, Germany. ²Institute of Virology, University of Veterinary Medicine, Hannover, Germany. ³Bioinformatics Core, University Medical Center Hamburg-Eppendorf, Hamburg, Germany. ⁴Department of Pathology, University of Veterinary Medicine, Hannover, Germany. ⁵Profesor Krzysztof Skóra Hel Marine Station, Institute of Oceanography, University of Gdańsk, Gdańsk, Poland. ⁶German Center for Infection Research, Hamburg – Borstel – Lübeck – Riems, Germany. ⁷German Center for Infection Research, Hannover – Braunschweig, Germany. ⁸Institute for Medical Microbiology, Virology and Hygiene, University Medical Center Hamburg-Eppendorf, Hamburg, Germany. Correspondence and requests for materials should be addressed to P.B. (email: paul.becher@tiho-hannover.de) or N.F. (email: nfischer@uke.de)

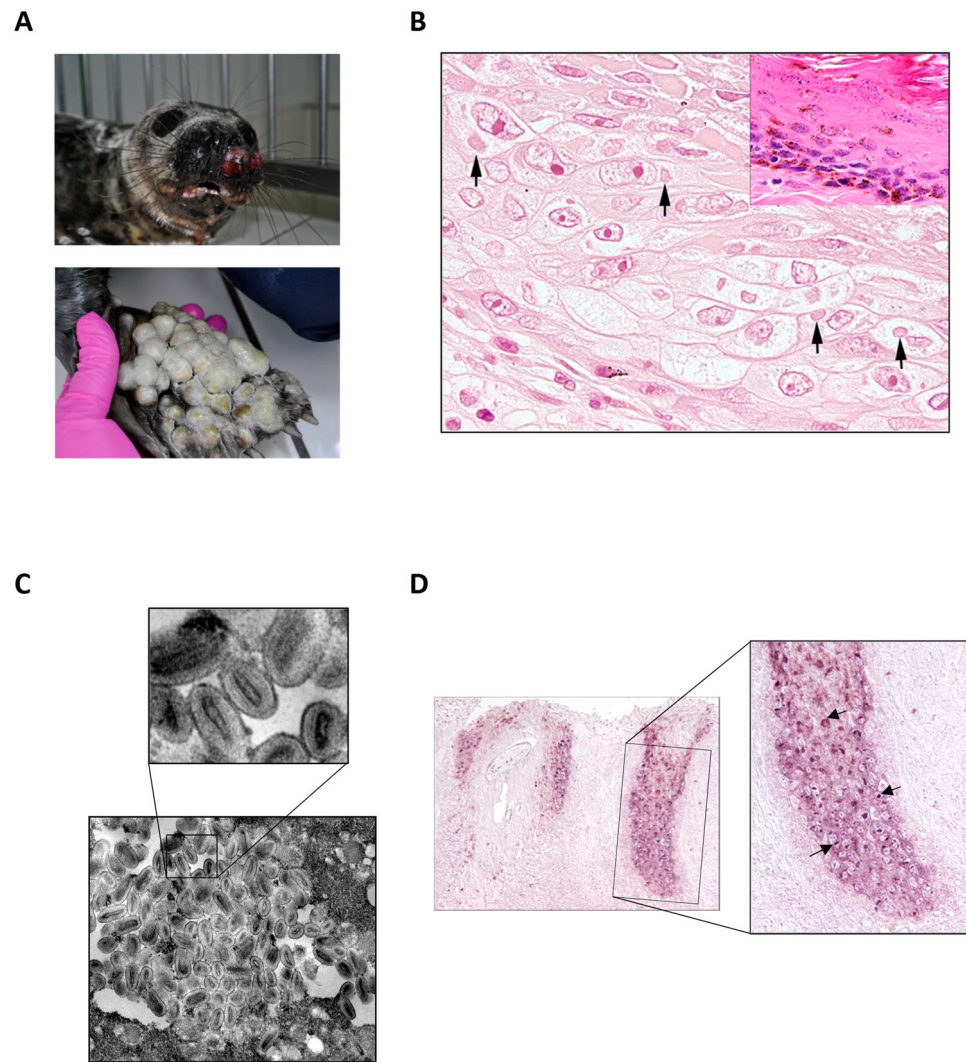


Figure 1. Macroscopic (A), electronmicroscopical (B) and histological analysis (C) of a harbor seal infected with seal parapoxvirus. (A) Ulcerative nodular skin lesion were identified on both front fins and the muzzle of the animal. (B) Histological changes of hair follicle epithelium characterized by ballooning degeneration and cytoplasmic eosinophilic inclusion bodies (arrows), magnification 600x. Normal hair follicle epithelium of an unaffected grey seal is shown in the upper right corner, HE, magnification 600x. (C) Electron microscopy pictures of the cytoplasm of a hair follicle epithelial cell isolated from an infected seal skin lesion. Magnification 37,500x. Mature and immature virus like particles were densely packed. The ovoid to lanceolate shape of the core virions are clearly visible. (D) *In situ* hybridization using a digoxigenin-labeled parapoxvirus-specific DNA probe: parapox virus specific signal is detected in hair follicle epithelial cells of the suprabasal layers (arrow), while no specific signal was detected in the basal layers. Left panel represents an overview with a 100 x magnification, right panel shows a blown up of the highlighted area, magnification: 200x.

extracted directly from skin lesion material. We employed a combination of second (Illumina MiSeq) and third (Oxford Nanopore MinION²⁰) generation sequencing to recover the full-length genome of this seal parapoxvirus, including telomere sequences and hairpin structure. To our knowledge, this is the first report of a full-length seal parapoxvirus genome sequence, as well as the first report of recovery of a full-length parapoxvirus sequence directly from clinical material. Thus, the full-length genome sequence reported here can be assumed to faithfully reflect the naturally acquired pathogenic parapoxvirus strain, without any adaptations resulting from *in vitro* culturing^{18,21}.

Results and Discussion

Identification of a parapoxvirus from a grey seal (*Halichoerus grypus*). In April 2015, a young grey seal taken to a seal rehabilitation center showed large spherical dermal lesions with severe ulcerations at both flippers and in the mucosa of the oral cavity (Figure 1A). Histological examination of the affected skin areas showed extensive ulcerations and necrosis with epidermal loss and infiltration of neutrophils and few macrophages. Ballooning degeneration with severe swelling of hair follicular epithelial cells was observed. Occasionally,

Name	Genbank accession number	Size (bp)	GC content (%)	number of annotated ORFs
Orf Virus (ORFV)	AY386264.1	139,981	65.0	134
Seal Parapoxvirus (SePPV)	KY382358	127,941	55.9	119
Red Deer Parapoxvirus (PVNZ)	NC_025963.1	139,962	63.4	132
Bovine papular stomatitis virus (BPSV)	NC_005337.1	134,431	64.5	134
Pseudocowpoxvirus (PCPV)	NC_013804.1	145,289	65.0	134

Table 1. Summary of size, relative GC content and number of open reading frames in all full-length genomes of the genus parapoxvirus.

cytoplasmic eosinophilic inclusion bodies were present (Figure 1B). Transmission electron microscopical analysis of follicular epithelial cells clearly identified densely packed, multifocal clusters of viral particles in the cytoplasm (Figure 1C) indicative of parapoxvirus particles based on the elongated, ovoid shaped core surrounded by a membrane and a superficial membrane. Enveloped particles had a length of approximately 200 to 250 nm and a width of 100 to 150 nm.

Parapoxvirus infection was confirmed by performing a pan-parapoxvirus PCR targeting a 552 bp sequence of the genomic region encoding the major envelope protein⁵. Sequence analysis established 97.3% homology to the partial seal parapoxvirus (SePPV) sequences described earlier⁵. *In situ*-hybridization (ISH) with a probe specific for the amplified region revealed abundant parapoxvirus DNA-positive hair follicle epithelial cells corresponding to the cells with ballooning degeneration (Figure 1D). In contrast, cells of the basal cell layer lacked a specific hybridization signal (Figure 1D).

Unfortunately, virus isolation using a seal tissue-derived cell line was not successful. Limiting amount of tissue and suboptimal condition of the tissue most likely contributed to the toxic effects observed during cultivation processes.

Recovery of the full-length seal parapoxvirus genome applying high throughput short read sequencing together with Nanopore MinION sequencing.

DNA from skin lesion was subjected to high throughput multiplex sequencing on an Illumina MiSeq Instrument, as well as nanopore sequencing on a third generation Oxford Nanopore MinION device²⁰. Approximately 5% out of a total of 2,272,653 short reads generated on the MiSeq instrument did not map to host sequences and thus were considered to be of potential exogenous origin. De novo assembly and iterative mapping of the non-host reads yielded 19 contigs (sizes between 590 bp and 23,767 bp) that showed distant sequence similarity to the virus family *Poxviridae*, genus *Parapoxvirus*. These contigs accounted for 68.33% (71,680 reads) of all non-host reads; there were no other contigs or reads indicative of the presence of other pathogenic viruses in this sample. Iterative mapping of all sequences to full-length genome of ORFV allowed the assembly of a single contig of 127,941 bp (minimal coverage: 260 over 99% of the contig) (Table 1). The sequence was classified as a seal parapoxvirus due to its close homology to a short fragment from the major envelope protein-coding sequence described in 2002⁵. We used nanopore sequencing to verify the assembly of short sequencing reads along the coding region and the termini of the virus. Nanopore sequencing produces relatively high error rates and thus is of limited use for de novo sequencing. However, the long read lengths which can be produced by this technique make it ideally suited to confirm the overall structure of sequence contigs. As shown in Figure S1 and Table S1, nanopore sequencing of the primary clinical material produced a total of 48 reads which mapped across at least 30 kbp of the viral genome, with the longest read covering a continuous stretch of 56.2 kbp. Together, although there were a few gaps at the right end of the genome, the nanopore reads covered more than 92% of the assembled genome, thus confirming the accuracy of the short read assembly.

Nucleotide sequence alignments among all fully sequenced parapoxvirus genomes revealed that the seal virus is only distantly related to the other genus members, showing the closest homology (77.3% sequence identity) to the bovine papular stomatitis virus (BPSV) sequence (Figure 2A, Table 2). The seal parapoxvirus, tentatively named SePPV, has the smallest genome (128 kbp) among the fully sequenced members within this genus, followed by bovine papular stomatitis virus with 138 kbp (Figure 2B). Similar to other parapoxviruses, the SePPV genome sequence shows a relatively high GC content. However, with 55.9% the GC content is the lowest known so far within this genus (Figure 2B; Table 1). In addition to the core sequence, we were able to resolve the hairpin termini of the parapoxvirus genome, including the telomere resolution sequence important for effective replication of the virus (Figure 2C). The inverted terminal repeats (ITR) of SePPV encompass 2,087 bp, coordinates 1–2,087 sense orientation and 125,855–127,941 antisense orientation. Interestingly the ITR contains a complete duplication of the ORF encoding for a dUTPase (similar to ORF007 dUTPase, Supplementary Dataset S1). In addition, the ITR contains a partial duplication of the ankyrin repeat (similar to ORF008 ankyrin repeat, Supplementary Dataset S1). While a partial duplication of an ankyrin repeat has been described for Camelpox virus (AF438165), the observed duplication of the ORF007 has not been described for parapoxviruses or poxviruses in general.

The contig generated by de novo assembly contained 230 single open reading frames (ORFs) of which 120 were identified as putative SePPV genes. Of these, 116 ORFs are coding for proteins with significant homology to annotated ORFs in other parapoxviruses (Table 1; Supplementary Datasets S1, S2). The relative order of the genes is similar to other parapoxviruses thereby supporting the classification as a novel species within the genus *Parapoxvirus*. Similar to the other members of the genus *Parapoxvirus*, the SePPV contains ORFs encoding for proteins involved in pathogenesis (Supplementary Datasets S1, S2). SePPV encodes a viral homologue of IL10

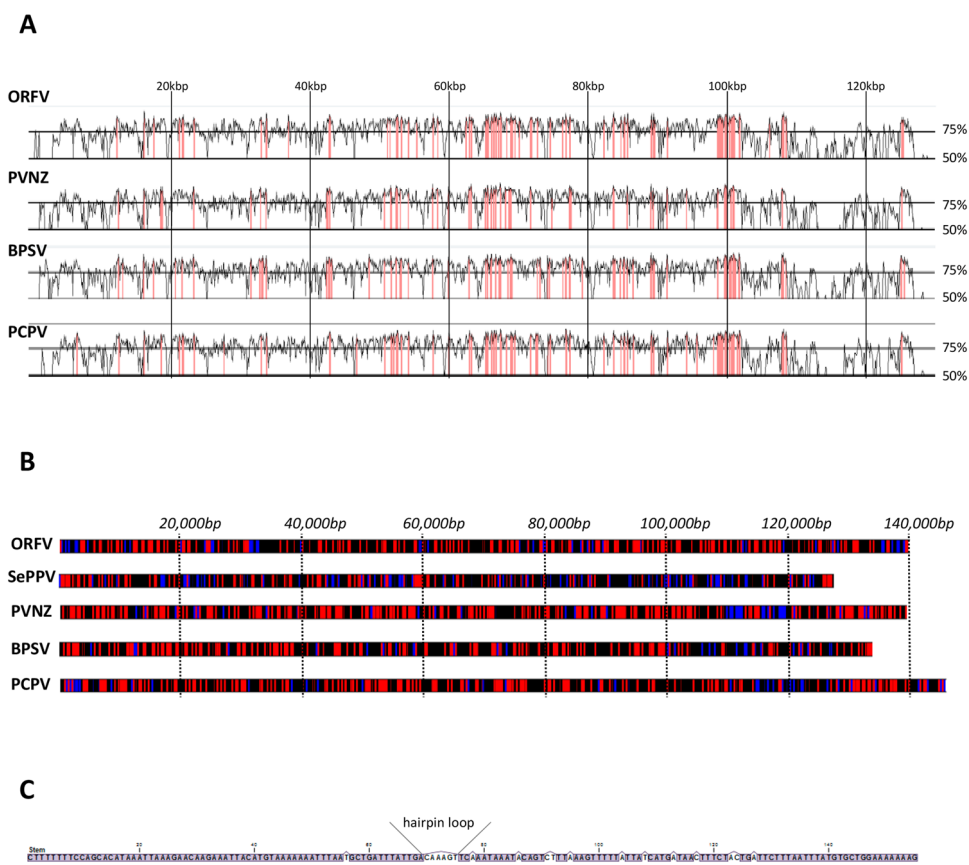


Figure 2. Genome Characterization of the full-length genome seal parapoxvirus sequence. (A) Sequence alignment of seal parapoxvirus (KY382358) to all four reference genomes (ORFV: AY386264.1, PVNZ: NC_025963.1, BPSV: NC_005337.1, PCPV: NC_013804.1) in the genus parapoxvirus. The alignment was performed using the global alignment program AVID implemented in VISTA (tool for comparative genomics). Alignments were visualized with VISTA point (Calc Window, bp: 100; Min Cons Width, bp: 100; Cons Identity, %: 90 Minimum Y, %:50). The graph represents percent conservation between the aligned sequences at a given coordinate on the base sequence. Highly conserved regions, with a conservation higher than 90%, are shown in pink. (B) G + C genome profile of all reference genomes (AY386264.1, NC_025963.1, NC_005337.1, NC_013804.1) listed in Genbank for parapoxviruses together with the newly identified seal parapoxvirus. Each trace represents the % G + C content of the indicated viral genome. GC content is indicated by the color scheme with blue representing a GC content range from 0–33.3%, black from 33.3–66.6% and red from 66.6–100%. (C) Terminal hairpin sequences of the seal parapoxvirus genome. SePPV hairpin terminus consists of an incomplementary base-paired and AT rich sequence. Telomere resolution sequence is underlined.

	SePPV	ORFV	BPSV	PCPV	PVNZ
SePPV	100	76.9	77.3	77.1	76.0
ORFV	76.9	100	81.2	90.9	81.0
BPSV	77.3	81.2	100	82.0	81.0
PCPV	77.1	90.9	82.0	100	81.5
PVNZ	76.0	81.0	81.0	81.5	100

Table 2. Percentage nucleotide identity between the full-length genome sequences within the genus parapoxvirus. SePPV (KY382358); ORFV (AY386264.1); PVNZ (NC_025963.1); BPSV (NC_005337.1); PCPV (NC_013804.1).

(SePPVgORF114), which has been shown in ORFV to be a potent anti-inflammatory virokinin since deletion of the ORF results in an attenuated virus^{22,23}. In addition, SePPVgORF101 expresses an anti-inflammatory chemokine binding protein CBP, which in ORFV plays a role in disrupting chemotactic recruitment of leukocytes²⁴. SePPV contains an ORF, SePPVgORF013, of which the gene product has significant homology to an inhibitor of interferon response which blocks activation of the dsRNA dependent protein kinase²⁵. SePPV also encodes proteins, encoded by SePPVgORF017 and SePPVgORF109 with significant homology to factors described for

	SePPV	ORFV	BPSV	PCPV	PVNZ
SePPV	100	84.08	84.77	83.78	84.57
ORFV	0.17	100	86.72	94.05	85.43
BPSV	0.16	0.14	100	88.11	87.12
PCPV	0.17	0.06	0.13	100	85.53
PVNZ	0.17	0.16	0.14	0.16	100

Table 3. Percentage amino acid identity between the DNA polymerase within the genus parapoxvirus. Upper comparison gradient indicates percentage identity between two sequences; lower comparison gradient indicates the distance between two sequences as calculated by the distance measure Jukes-Cantor. Percentage identity higher than 90% is shown in bold numbers.

	SePPV	ORFV	BPSV	PCPV	PVNZ
SePPV	100	83.96	86.79	83.65	83.44
ORFV	0.18	100	86.79	95.91	83.75
BPSV	0.14	0.14	100	88.36	84.69
PCPV	0.18	0.04	0.12	100	84.06
PVNZ	0.18	0.17	0.16	0.17	100

Table 4. Percentage amino acid identity between the DNA topoisomerase I within the genus parapoxvirus. Upper comparison gradient indicates percentage identity between two sequences; lower comparison gradient indicates the distance between two sequences as calculated by the distance measure Jukes-Cantor. Percentage identity higher than 90% is shown in bold numbers.

other parapoxviruses involved in NF κ B signaling, ORF24 and ORF121. The function of the ORF24 gene product in ORFV is described to decrease TNF α induced phosphorylation whereas ORF121 of ORFV most likely is involved in the inhibition of NF κ B-p65 phosphorylation and nuclear translocation^{26,27}. As described for all established parapoxvirus species, we also identified an ORF encoding a vascular endothelial growth factor (VEGF) homologue^{28–31}. The SePPVgORF118 encoded VEGF shows closest homology to BPSV, however different to BPSV with regard to the location of the ORF coding for VEGF, SePPV VEGF is located at the right end of the genome^{29,32}. Viral VEGF most likely enhances viral growth by promoting cellular regeneration of the epidermis. Viruses devoid of VEGF do not induce extensive blood vessel formation and dermal swelling which is discussed to provide protection for immune cells²⁴. Interestingly, BPSV encoded VEGF different to all other parapoxvirus encoded VEGFs shows higher homology to mammalian VEGF-A^{24,29}.

SePPV encodes for 4 unique open reading frames not identified in ORFV and other parapoxviruses. The hypothetical proteins encoded by these ORFs do not show any significant homology with known proteins from the family *Poxviridae* (Supplementary Dataset S1). In comparison to ORFV, 16 ORFs including 9 hypothetical proteins, 2 ankyrin repeat proteins, 2 putative IMV membrane proteins and 2 proteins involved in virion morphogenesis are not present in SePPV (Supplementary Table S2). In addition, different to ORFV and other parapoxviruses we did not identify inhibitors of granulocyte-macrophage colony stimulating factor and interleukin-2, which are known as GIF (Supplementary Table S2) and play a role in the regulation of the adaptive immune response of the host^{33,34}.

As shown in Tables 3 and 4, on the protein level SePPV demonstrates relatively uniform distances to other members of the genus (84.3 and 84.5% mean amino acid identity for DNA polymerase and DNA topoisomerase, respectively), with the closest relative again being BPSV.

For phylogenetic analysis all proteins in SePPV sequence were aligned with proteins identified in 14 representative genomes of the subfamily Chordopoxvirinae. Proteins which showed an alignment of >90% of the length of the individual protein within SePPV were used to construct a concatenated polyprotein. This polyprotein consisting of 47 proteins was used for the phylogenetic tree analysis (Figure 3). In addition, phylogenetic analysis was also applied with complete coding sequences of the DNA polymerase and DNA Topoisomerase I confirming the results obtained with the concatenated polyprotein (supplementary material Figure S2A,B, Tables 3 and 4). Thus, the results of our phylogenetic tree analyses together with the description of the gene order and annotated ORFs clearly warrant the classification of SePPV as a distinct species within the genus *Parapoxvirus*.

Conclusion

We report the first reconstruction of a full-length genome sequence of a member of the family *Poxviridae* directly from a clinical sample using Illumina short read sequencing combined with Oxford Nanopore sequencing. Poxviruses are challenging to grow in culture and culture adapted genomes might not faithfully represent those of virulent field strains. The power of combining high throughput short read sequencing with nanopore sequencing to recover a large and complex viral genome directly from complex clinical material as demonstrated here should be useful also for other virus families in particular those with large viral genomes.

Methods

Clinical case. In April 2015, a young grey seal was taken to the Hel Marine Station, a field station belonging to the Institute of Oceanography at the University of Gdansk, Poland. A few days later, skin lesions appeared on

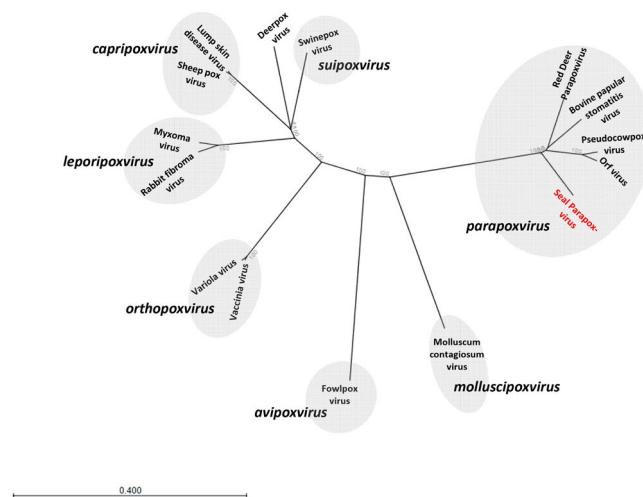


Figure 3. Phylogenetic tree analysis based on 47 proteins. Protein sequences were considered being conserved, if the corresponding sequence of SePPV yielded BLASTP alignments over at least 90% of the SePPV protein sequence length with sequences of all 14 representative genomes. The following sequences were used: Red Deer Parapoxvirus (PVNZ) HL953 (NC_025963.1); ORFV (AY386264.1); PCPV (NC_013804.1); BPSV (NC_005337.1); Vaccinia Virus (NC_006998.1); Variola Virus (NC_001611.1); Myxoma Virus (NC_001132.2); swinepox virus (NC_003389.1); deerpox virus (NC_006966.1); sheep pox virus (NC_004002.1); lumpy skin disease virus (NC_003027.1); fowlpox virus (NC_002188.1); rabbit fibroma virus (NC_001266.1); molluscum contagiosum virus (NC_001731.1).

the front flippers followed by lesions around mouth and nose (Figure 1A). Skin samples from lesions of the left front flipper were sent to the University of Veterinary Medicine, Hannover, Germany, for further diagnosis in the framework of veterinary microbiological diagnostics in accordance to the German legislation. Therefore no ethical approval was required for the use of these samples. Upon symptomatic treatment, the lesions declined and the seal could eventually be discharged in June 2015.

Polymerase chain reaction (PCR) and Sanger Sequencing. Total DNA was extracted from skin lesions using a High Pure PCR Template Preparation Kit (Roche Diagnostics).

Conventional PCR was done as described before⁵, amplifying a part of the putative major envelope gene. To confirm the specificity of PCR, amplicons were sequenced (LGC Genomics GmbH, Berlin, Germany) and aligned using the Clustal W multiple alignment tool implemented in BioEdit³⁵.

Histology and *in-situ* hybridization. Formalin-fixed tissue biopsies of the altered skin were processed routinely into paraffin wax. Tissue section of 3–4 µm thickness were cut and stained with hematoxylin and eosin (HE) for light microscopical examination.

In situ-hybridization was performed as previously described using an universal parapox primer pair, P1 (5'-GTCTGCCACGATGAGCAGCT-3') and P2 (5'-TACGTGGGAAGCGCCTCGCT-3'; GeneBank accession number U06671), and a digoxigenin-labeled DNA probe¹⁹.

For transmission electron microscopy a direct pop-off technique from the HE-stained slide was used as previously described³⁶.

Illumina library preparation and MiSeq short read sequencing. Total DNA from 75 mg tissue was mechanically homogenized in PBS using disposable tissue grinder pestles. After low speed pelleting, supernatant was filtered (0.2 µm) and DNA was subsequently isolated using DNeasy blood & tissue Kit (Qiagen).

DNASeq library compatible with short read Illumina sequencing was generated using the NEB Ultra DNA library Kit (NEB) starting with 500 ng DNA, as measured by Qubit (Invitrogen) and following the manufacturer's instructions. Briefly, DNA was fragmented, end repaired and subsequently the adapter were ligated. Agencourt AMPure XP beads were used to size select the DNA fragments containing the adapters. Finally, the library was amplified by 15 PCR cycles. The fragment size distribution of the library was analyzed on a BioAnalyzer High Sensitivity LabChip showing a size range between 400 and 446 bp with the main peak of the library at 401 bp. The library was diluted to 2 nM and multiplex-sequenced together with five samples on the Illumina MiSeq (2 × 250 bp paired end run, estimated 4.3 million reads/sample).

Oxford Nanopore library preparation and MinION sequencing. Nanopore sequencing library preparations using Nanopore Sequencing Kits SQK-MAP005 and SQK-MAP006 were essentially performed as described in the protocols and guidelines provided by Oxford Nanopore Technologies (ONT). Briefly, 1 µg of the genomic DNA isolated from skin lesions was fragmented to an average size of 8–15 kb using g-TUBEs (Covaris). DNA fragments were end-repaired and adenylated using NEBNext Ultra II End-Repair/dA-tailing Module (NEB) followed by cleanup with Ampure XP beads (Beckmann Coulter). Sequencing and hairpin adapters (ONT) were ligated using NEB Blunt/TA Ligase Master Mix (NEB) followed by incubation with the hairpin tether (ONT).

Cleanup of libraries was done either with His-Tag beads or MyOne C1 streptavidin beads (Invitrogen) depending on the respective Sequencing Kit and flowcell version. Prepared libraries were eluted in 25 µl of the ONT-supplied elution buffer.

Prior to sequencing, 6 µl of the eluate (pre-sequencing mix), 75 µl running buffer (ONT), 60 µl nuclease free H₂O and 4 µl fuel mix (ONT) were combined gently and were immediately loaded onto the prepared MinION flowcells. Sequencing was performed using 48 hr sequencing run scripts with addition of freshly prepared input material to the MinION flowcell every 12 hrs until no further active pores were available anymore.

Sequence assembly. Illumina reads were aligned to the *Leptonychotes weddellii* (Wedell seal) reference assembly (GCF_000349705.1) using Bowtie2 (v2.2.3)²⁴. Reads yielding significant alignments with the reference assembly were excluded from further analysis. The remaining short reads were initially assembled into contigs using Trinity (r20140717).

Nanopore events were converted into FastQ containing Fast5 data using Metrichor basecalling with the respective 2D workflows. Fasta files were extracted from Fast5 data using poretools²⁵. Long reads (>3000 bp) were subsequently aligned with LAST²⁶ to the SePPV assembly generated from Illumina sequencing data using the following parameters: -s2 -T0 -Q0 -a1 -f1. Alignments were further filtered by a minimum length of 3,000 bp to reduce false positive results due to low complexity region alignment. The joint assembly of MinION and MiSeq was performed using SPAdes (v3.6.0) using the 'careful' option and otherwise standard parameters³⁷.

Phylogenetic analysis. Nucleotide sequences of viruses classified to the genus *Parapoxvirus* or to the family of *Poxviridae* were downloaded from GenBank: Red Deer Parapoxvirus (PVNZ) HL953 (NC_025963.1); ORFV (AY386264.1); PCPV (NC_013804.1); BPSV (NC_005337.1); Vaccinia Virus (NC_006998.1); Variola Virus (NC_001611.1); Myxoma Virus (NC_001132.2); swinepox virus (NC_003389.1); deerpox virus (NC_006966.1); sheeppox virus (NC_004002.1); lumpy skin disease virus (NC_003027.1); fowlpox virus (NC_002188.1); rabbit fibroma virus (NC_001266.1), molluscum contagiosum virus (NC_001731.1).

Amino acid sequences of single proteins (DNA polymerase and DNA topoisomerase I) were aligned using the CLUSTAL W multiple alignment tool, CLC Main workbench, version 7.6.4. For phylogenetic analyses, genomes were trimmed manually and neighbor-joining trees were calculated using nucleotide distance measurement Jukes-Cantor parameters. Bootstrap analysis was performed with 1000 iterations.

Following an approach described before³⁸ phylogenetic analysis of a concatenated protein sequences was performed by maximum-likelihood tree construction using PHYML3³⁹ (Figure 3). The sequences were obtained by identifying 47 proteins conserved in SePPV and 14 representative genomes (AY386264; NC_001132.2; NC_001266; NC_001611.1; NC_002188.1; NC_003027.1; NC_003389.1; NC_004002.1; NC_005337.1; NC_006966.1; NC_006998.1; NC_013804.1; NC_025963.1; NC_001731.1) of the subfamily Chordopoxvirinae. Protein sequences were considered being conserved, if the corresponding sequence of SePPV yielded BLASTP alignments over at least 90% of the SePPV protein sequence length with sequences of all 13 representative genomes.

References

- Haller, S. L., Peng, C., McFadden, G. & Rothenburg, S. Poxviruses and the evolution of host range and virulence. *Infect Genet Evol* **21**, 15–40, doi:10.1016/j.meegid.2013.10.014 (2014).
- Falk, E. S. Parapoxvirus infections of reindeer and musk ox associated with unusual human infections. *Br J Dermatol* **99**, 647–654 (1978).
- Klein, J. & Tryland, M. Characterisation of parapoxviruses isolated from Norwegian semi-domesticated reindeer (*Rangifer tarandus tarandus*). *Virology* **79**, doi:10.1186/1743-422X-2-79 (2005).
- Lederman, E. *et al.* Zoonotic parapoxviruses detected in symptomatic cattle in Bangladesh. *BMC Res Notes* **7**, 816, doi:10.1186/1756-0500-7-816 (2014).
- Becher, P., König, M., Müller, G., Siebert, U. & Thiel, H. J. Characterization of sealpox virus, a separate member of the parapoxviruses. *Archives of virology* **147**, 1133–1140, doi:10.1007/s00705-002-0804-8 (2002).
- Airas, N. *et al.* Infection with Possible Novel Parapoxvirus in Horse, Finland, 2013. *Emerg Infect Dis* **22**, 1242–1245, doi:10.3201/eid2207.151636 (2016).
- Clark, C., McIntyre, P. G., Evans, A., McInnes, C. J. & Lewis-Jones, S. Human sealpox resulting from a seal bite: confirmation that sealpox virus is zoonotic. *Br J Dermatol* **152**, 791–793, doi:10.1111/j.1365-2133.2005.06451.x (2005).
- Hicks, B. D. & Worthy, G. A. Sealpox in captive grey seals (*Halichoerus grypus*) and their handlers. *J Wildl Dis* **23**, 1–6 (1987).
- Burek, K. A. *et al.* Poxvirus infection of Steller sea lions (*Eumetopias jubatus*) in Alaska. *J Wildl Dis* **41**, 745–752, doi:10.7589/0090-3558-41.4.745 (2005).
- Nettleton, P. F. *et al.* Isolation of a parapoxvirus from a grey seal (*Halichoerus grypus*). *The Veterinary record* **137**, 562–564 (1995).
- Nollens, H. H. *et al.* Seroepidemiology of parapoxvirus infections in captive and free-ranging California sea lions *Zalophus californianus*. *Dis Aquat Organ* **69**, 153–161, doi:10.3354/dao069153 (2006).
- Nollens, H. H. *et al.* Parapoxviruses of seals and sea lions make up a distinct subclade within the genus *Parapoxvirus*. *Virology* **349**, 316–324, doi:10.1016/j.virol.2006.01.020 (2006).
- Nollens, H. H. *et al.* Pathology and preliminary characterization of a parapoxvirus isolated from a California sea lion (*Zalophus californianus*). *J Wildl Dis* **42**, 23–32, doi:10.7589/0090-3558-42.1.23 (2006).
- Ohno, Y., Inoshima, Y., Maeda, K. & Ishiguro, N. Molecular analysis of parapoxvirus from a spotted seal *Phoca largha* in Japan. *Dis Aquat Organ* **97**, 11–16, doi:10.3354/dao02405 (2011).
- Toplu, N., Aydoğan, A. & Oguzoglu, T. C. Visceral leishmaniasis and parapoxvirus infection in a Mediterranean monk seal (*Monachus monachus*). *J Comp Pathol* **136**, 283–287, doi:10.1016/j.jcpa.2007.02.005 (2007).
- Tryland, M., Klein, J., Nordoy, E. S. & Blix, A. S. Isolation and partial characterization of a parapoxvirus isolated from a skin lesion of a Weddell seal. *Virus research* **108**, 83–87, doi:10.1016/j.virusres.2004.08.005 (2005).
- Bracht, A. J. *et al.* Genetic identification of novel poxviruses of cetaceans and pinnipeds. *Archives of virology* **151**, 423–438, doi:10.1007/s00705-005-0679-6 (2006).
- Cottone, R. *et al.* Analysis of genomic rearrangement and subsequent gene deletion of the attenuated Orf virus strain D1701. *Virus research* **56**, 53–67 (1998).
- Müller, G. *et al.* Parapoxvirus infection in harbor seals (*Phoca vitulina*) from the German North Sea. *Vet Pathol* **40**, 445–454, doi:10.1354/vp.40-4-445 (2003).

20. Ip, C. L. *et al.* MinION Analysis and Reference Consortium: Phase 1 data release and analysis. *F1000Res* **4**, 1075, doi:[10.12688/f1000research.7201.1](https://doi.org/10.12688/f1000research.7201.1) (2015).
21. Hautaniemi, M. *et al.* The genome of pseudocowpoxvirus: comparison of a reindeer isolate and a reference strain. *The Journal of general virology* **91**, 1560–1576, doi:[10.1099/vir.0.018374-0](https://doi.org/10.1099/vir.0.018374-0) (2010).
22. Fleming, S. B. *et al.* Infection with recombinant orf viruses demonstrates that the viral interleukin-10 is a virulence factor. *The Journal of general virology* **88**, 1922–1927, doi:[10.1099/vir.0.82833-0](https://doi.org/10.1099/vir.0.82833-0) (2007).
23. Fleming, S. B., McCaughan, C. A., Andrews, A. E., Nash, A. D. & Mercer, A. A. A homolog of interleukin-10 is encoded by the poxvirus orf virus. *Journal of virology* **71**, 4857–4861 (1997).
24. Fleming, S. B., Wise, L. M. & Mercer, A. A. Molecular genetic analysis of orf virus: a poxvirus that has adapted to skin. *Viruses* **7**, 1505–1539, doi:[10.3390/v7031505](https://doi.org/10.3390/v7031505) (2015).
25. Haig, D. M. *et al.* The orf virus OV20.0L gene product is involved in interferon resistance and inhibits an interferon-inducible, double-stranded RNA-dependent kinase. *Immunology* **93**, 335–340 (1998).
26. Diel, D. G., Delhon, G., Luo, S., Flores, E. F. & Rock, D. L. A novel inhibitor of the NF- κ B signaling pathway encoded by the parapoxvirus orf virus. *Journal of virology* **84**, 3962–3973, doi:[10.1128/JVI.02291-09](https://doi.org/10.1128/JVI.02291-09) (2010).
27. Diel, D. G. *et al.* Orf virus ORFV121 encodes a novel inhibitor of NF- κ B that contributes to virus virulence. *Journal of virology* **85**, 2037–2049, doi:[10.1128/JVI.02236-10](https://doi.org/10.1128/JVI.02236-10) (2011).
28. Delhon, G. *et al.* Genomes of the parapoxviruses ORF virus and bovine papular stomatitis virus. *Journal of virology* **78**, 168–177 (2004).
29. Inder, M. K., Ueda, N., Mercer, A. A., Fleming, S. B. & Wise, L. M. Bovine papular stomatitis virus encodes a functionally distinct VEGF that binds both VEGFR-1 and VEGFR-2. *The Journal of general virology* **88**, 781–791, doi:[10.1099/vir.0.82582-0](https://doi.org/10.1099/vir.0.82582-0) (2007).
30. Ueda, N., Inder, M. K., Wise, L. M., Fleming, S. B. & Mercer, A. A. Parapoxvirus of red deer in New Zealand encodes a variant of viral vascular endothelial growth factor. *Virus research* **124**, 50–58, doi:[10.1016/j.virusres.2006.09.012](https://doi.org/10.1016/j.virusres.2006.09.012) (2007).
31. Ueda, N., Wise, L. M., Stacker, S. A., Fleming, S. B. & Mercer, A. A. Pseudocowpox virus encodes a homolog of vascular endothelial growth factor. *Virology* **305**, 298–309 (2003).
32. Lyttle, D. J., Fraser, K. M., Fleming, S. B., Mercer, A. A. & Robinson, A. J. Homologs of vascular endothelial growth factor are encoded by the poxvirus orf virus. *Journal of virology* **68**, 84–92 (1994).
33. Deane, D. *et al.* Conservation and variation of the parapoxvirus GM-CSF-inhibitory factor (GIF) proteins. *The Journal of general virology* **90**, 970–977, doi:[10.1099/vir.0.006692-0](https://doi.org/10.1099/vir.0.006692-0) (2009).
34. Seet, B. T. *et al.* Analysis of an orf virus chemokine-binding protein: Shifting ligand specificities among a family of poxvirus viroreceptors. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 15137–15142, doi:[10.1073/pnas.2336648100](https://doi.org/10.1073/pnas.2336648100) (2003).
35. Hall, T. A. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp Ser* **41**, 95–98 (1995).
36. Lehmbecker, A., Rittinghausen, S., Rohn, K., Baumgartner, W. & Schaudien, D. Nanoparticles and pop-off technique for electron microscopy: a known technique for a new purpose. *Toxicol Pathol* **42**, 1041–1046, doi:[10.1177/0192623313509906](https://doi.org/10.1177/0192623313509906) (2014).
37. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology: a journal of computational molecular cell biology* **19**, 455–477, doi:[10.1089/cmb.2012.0021](https://doi.org/10.1089/cmb.2012.0021) (2012).
38. Friederichs, S., Krebs, S., Blum, H., Lang, H. & Buttner, M. Parapoxvirus (PPV) of red deer reveals subclinical infection and confirms a unique species. *The Journal of general virology* **96**, 1446–1462, doi:[10.1099/vir.0.000080](https://doi.org/10.1099/vir.0.000080) (2015).
39. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic biology* **59**, 307–321, doi:[10.1093/sysbio/syq010](https://doi.org/10.1093/sysbio/syq010) (2010).

Acknowledgements

The project is funded by the German Center for Infection Research, project funding given to NF, AG and PB. The authors are grateful to Daniela Indenbirken, Lia Burkhart and Kerstin Reumann for technical assistance with Illumina library preparation and sequencing; we also thank Kerstin Rohn for assistance with the pop-off electron microscopic technique. We thank Oxford Nanopore Technologies for the flow cells and reagents used in these experiments.

Author Contributions

P.B., N.F. and A.G. designed the project. J.M. collected the sample; T.G. performed the sequencing experiments; P.W. performed the histological analysis, electron microscopy and *in situ* hybridization; L.H. performed RT-PCR and Sanger Sequencing; T.G., M.A., A.G. and N.F. analyzed the sequencing data; N.F. A.G. and L.H. wrote the manuscript.

Additional Information

Supplementary information accompanies this paper at doi:[10.1038/s41598-017-03997-y](https://doi.org/10.1038/s41598-017-03997-y)

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2017

Novel poly-uridine insertion in the 3'UTR and E2 amino acid substitutions in a low virulent classical swine fever virus

Liani Coronado^{1,2#}, Matthias Liniger^{3#}, Sara Muñoz-González², Alexander Postel⁴, Lester Josue Pérez¹, Marta Pérez-Simó², Carmen Laura Perera¹, Maria Teresa Frías¹, Rosa Rosell^{2,5}, Adam Grundhoff⁶, Daniela Indenbirken⁶, Malik Alawi^{6,7}, Nicole Fischer⁸, Paul Becher⁴, Nicolas Ruggli³, Lillianne Ganges^{2*}

¹Centro Nacional de Sanidad Agropecuaria (CENSA), La Habana, Cuba

²IRTA, Centre de Recerca en Sanitat Animal (CReSA, IRTA-UAB), Campus de la Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain

³ Institute of Virology and immunology IVI, Mittelhäusern, Switzerland

⁴EU and OIE Reference Laboratory for Classical Swine Fever, Institute of Virology, Department of Infectious Diseases, University of Veterinary Medicine, Hannover, Germany

⁵Departament d'Agricultura, Ramaderia, Pesca, Alimentació i Medi Natural, (DAAM), Generalitat de Catalunya, Spain

⁶Heinrich Pette Institute, Leibniz Institute for Experimental Virology, Research Group Virus Genomics, Hamburg, Germany

⁷Bioinformatics Service Facility, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

⁸Institute for Medical Microbiology, Virology and Hygiene, University Medical Center Hamburg- Eppendorf, Hamburg, Germany

contributed equally to this work

*Corresponding Author: Lillianne Ganges (LG)

Lillianne.ganges@irta.cat

Abstract

In this study, we compared the virulence in weaner pigs of the Pinar del Rio isolate and the virulent Margarita strain. The latter caused the Cuban classical swine fever (CSF) outbreak of 1993. Our results showed that the Pinar del Rio virus isolated during an endemic phase is clearly of low virulence. We analysed the complete nucleotide sequence of the Pinar del Rio virus isolated after persistence in newborn piglets, as well as the genome sequence of the inoculum. The consensus genome sequence of the Pinar del Rio virus remained completely unchanged after 28 days of persistent infection in swine. More importantly, a unique poly-uridine tract was discovered in the 3'UTR of the Pinar del Rio virus, which was not found in the Margarita virus or any other known CSFV sequences. Based on RNA secondary structure prediction, the poly-uridine tract results in a long single-stranded intervening sequence (SS) between the stem-loops I and II of the 3'UTR, without major changes in the stem-loop structures when compared to the Margarita virus. The possible implications of this novel insertion on persistence and attenuation remain to be investigated. In addition, comparison of the amino acid sequence of the viral proteins E^{ms}, E1, E2 and p7 of the Margarita and Pinar del Rio viruses showed that all non-conservative amino acid substitutions acquired by the Pinar del Rio isolate clustered in E2, with two of them being located within the B/C domain. Immunisation and cross-neutralisation experiments in pigs and rabbits suggest differences between these two viruses, which may be attributable to the amino acid differences observed in E2. Altogether, these data provide fresh insights into viral molecular features which might be associated with the attenuation and adaptation of CSFV for persistence in the field.

Keywords: CSFV, B/C Domain, E2 glycoprotein, poly-U insertion, secondary structure, amino acid differences, antigenic differences, immunogenicity, pigs, rabbits

1. Introduction

Classical swine fever (CSF) is a highly contagious viral disease affecting domestic pigs and wild boar which results in major losses in stock farming, particularly in developing countries (Moennig et al., 2003). The disease is caused by the CSF virus (CSFV), a member of the genus *Pestivirus* within the family *Flaviviridae* (Simmonds et al., 2012). CSFV comprises three different genotypes 1, 2, and 3, with three to four subgenotypes each (Postel et al., 2013; Beer et al., 2015). The virions consist of a lipid envelope, a capsid, and a single plus-strand RNA genome carrying a single large open reading frame (ORF) flanked by two untranslated regions (UTRs). The ORF encodes a polyprotein of typically 3898 amino acids, which is processed by cellular and viral proteases in the four structural proteins—C, E^{ns}, E1, E2, and in the eight non-structural proteins: N^{pro}; p7; NS2; NS3; NS4A; NS4B; NS5A; and NS5B (Tautz et al., 2015).

The disease is endemic in several South and Central American countries, in the Caribbean region, in Asia, and in some Eastern European countries (Pérez et al., 2012; Postel et al., 2013). Live attenuated CSFV vaccines are highly efficacious, resulting in complete protection from CSF even before neutralising antibodies can be detected, and nearly 100% protection against strains from different genotypes (Aynaud and Launais, 1978; Ganges et al., 2008; Graham et al., 2012). This type of vaccine was applied in numerous countries because of its high efficacy and safety, and the Chinese vaccine strain was the most widely used (C-strain, genotype 1.1) (Liu et al., 2011). Nevertheless, the disease is still endemic in several countries despite intensive vaccination programmes implemented over years. This has been attributed to failures in the responses to vaccination caused by a combination of distinct problems, for example, quality and availability of the vaccine, and gaps in the cold chain between vaccine production and application, among others (Diaz de Arce et al., 2005; Ganges et al., 2008; Shen et al., 2011; Pérez et al., 2012). In endemic areas, a whole range of disease severity from acute haemorrhagic to chronic and subclinical disease has been observed (Van Oirschot et al., 1983; Ganges et al., 2008). A previous study revealed that the increased rate of non-synonymous substitutions in E2 of CSFV among the isolates is associated with this trend (Diaz de Arce et al., 2005). Towards this background, it is worth noting the Cuban CSFV epizootic which started in 1993 after 20 years of epidemiological silence, manifested in several outbreaks each year until today, despite regular vaccination with a lapinised, live, attenuated vaccine C-strain. The molecular

studies within this epidemiological situation showed that viruses currently circulating belong to the 1.4 subgenotype (Postel et al., 2013) and are closely related to those viral strains isolated during the 1993–1997 epizootic, including the CSFV virulent Margarita strain (Díaz de Arce et al., 1999; Díaz de Arce et al., 2005; Ganges et al., 2005; Pérez et al., 2012). Therefore, the most likely origin of the Cuban outbreaks in 1993 was due to the reintroduction of the Margarita strain to the field (Díaz de Arce et al., 1999; Díaz de Arce et al., 2005; Pérez et al., 2012). Similarly, these studies rule out the external reintroduction from another CSFV genogroup.

The role of vaccination in viral evolution has been reported for different diseases affecting both animals and humans. When the immunity is not sterilising, wild strains are able to circulate. Numerous examples of viruses that adapted to this new scenario through immune-escape are available (Cecchinato et al., 2010; Woźniakowski and Samorek-Salamonowicz, 2014; Vera et al., 2015). Recent molecular epidemiology studies from some CSF endemic countries suggest that the viruses circulating in the field have evolved under the positive selection pressure exerted by immune responses to the vaccine, among other factors, and this could favour the generation of new attenuated viral variants that induce milder forms of CSF (Pérez et al., 2012; Ji et al., 2014). The selection pressure analysis previously reported highlighted some viral isolates that showed the G761R replacement into the E2-B/C-domain, which suggests the possible association of this substitution in the virulence and pathogenesis of the virus in the field (Pérez et al., 2012; Ji et al., 2014; Hu et al., 2016). Interestingly, one of these latter Cuban CSFV field isolates, Pinar del Rio (PR-11/10-3), carrying the G761R substitution in the B/C-domain of E2 (Pérez et al., 2012), was able to induce persistence in piglets upon neonatal infection (Muñoz-González et al., 2015a). Therefore, the present study was aimed at determining the virulence and immunogenicity of the Pinar del Rio isolate in comparison to the virulent Margarita strain in weaner pigs. Additionally, E2 sequences and the antigenic reactivity of Pinar del Rio and Margarita were compared. Analysis of the complete nucleotide sequence of Pinar del Rio before and after persistence showed that the consensus genome sequence remained completely unchanged during persistence. Importantly, a novel poly-uridine sequence was discovered in the 3'UTR of the Pinar del Rio isolate.

2. Materials and Methods

2.1. Cells and viruses

We amplified and titrated viral stocks in the porcine kidney cell line PK-15, (ATCC CCL 33) cultured in Dulbecco's modified Eagle medium, (DMEM) supplemented with a 10 % pestivirus-free foetal bovine serum (FBS) at 37 °C in 5 % CO₂. The titrations were performed by end-point dilution using a peroxidase-linked assay, (PLA) and the titres were calculated according to Reed and Muench (Reed and Muench, 1938).

The CSFV Pinar del Rio field isolate originates from the Cuban CSF epizootic of 2010 (PR-11/10-3) (Pérez et al., 2012) and was classified within the new CSFV subgenotype 1.4 (Postel et al., 2013). This virus was isolated by transfection of RNA in PK-15 cells (Meyer et al., 2015) and kept at the European Union Reference Laboratory for Classical Swine Fever (EURL-CSF), in Hannover, Germany, with the reference nomenclature CSF1058. The virulent CSFV Margarita strain which also belongs to subgenotype 1.4 (Díaz de Arce et al., 1999; Postel et al., 2013) was isolated in Cuba in 1958 and has been used since 1965 for vaccine potency testing (Ganges et al., 2005). The CSFV Alfort/187 strain was kindly provided by the CSFV EU Reference Laboratory (EURL), in Hannover, Germany. Finally, the lapinised live attenuated vaccine Labiofam strain belongs to subgenotype 1.1 and has been used in Cuba since 1965 for prophylactic vaccination against CSF (Díaz de Arce et al., 1999; Pérez et al., 2012).

2.2. The experimental infection of the pigs

The pathogenicity of the Pinar del Rio field strain (PR-11/10-3) (Pérez et al., 2012) was evaluated in healthy six-week-old Landrace × Large White × Duroc pigs. The animals were free of porcine circovirus type 2, porcine reproductive and respiratory syndrome virus, and pestiviruses. The animals were distributed in two groups: (1) Pinar del Rio (n=5, numbered 1 to 5) and (2) Margarita (n=3, numbered 6 to 8). Both groups were inoculated intranasally with 2 ml of DMEM containing 10⁵ TCID₅₀ Pinar del Rio field isolate or 10² TCID₅₀ Margarita strain, respectively.

A trained veterinarian recorded the rectal temperature and clinical symptoms of disease daily in a blinded manner. The clinical status of the animals was determined as previously described (Tarradas et al., 2014). The status was scored from 0 to 7 as follows: 0: no signs; 1: mild pyrexia; 2: pyrexia plus mild clinical signs; 3: mild-moderate clinical signs but no nervous disorders; 4: slight nervous disorders and rest of clinical signs moderate; 5: moderate nervous disorders and moderate–severe other clinical signs; 6: severe clinical signs (including nervous disorders); and 7: death. For

ethical reasons, the animals were euthanised either when the clinical score reached 5 or higher, or when they exhibited a fall of the hindquarter, inability to get up to drink, prostration, or when they exhibited moderate nervous disorders. After euthanasia, an exhaustive necropsy was conducted in order to evaluate the presence of pathological symptoms in different organs and tissues. International standards of animal welfare were used, following the regulations for the Institute of Veterinary Medicine (IMV), and Ministry of Agriculture (MINAGRI) of the Republic of Cuba. The protocol was approved by the ethics committee of the MINAGRI of the Republic of Cuba and all efforts were made to minimise any suffering of the animals. The IMV is the official regulatory body of the Republic of Cuba; therefore, additional permits were not required. After the infection, serum samples and nasal and rectal swabs were collected on the first day, then on days 4, 7, 11, 14, 18 and 21 post infection (dpi), and at the time of necropsy (for the Margarita group), or at 21 dpi (for the Pinar del Rio and control groups). Tissue samples of the tonsils, spleen, lymph nodes and ileum were collected. Serum samples were tested with neutralisation peroxidase-linked assay (NPLA) (Terpstra et al., 1984), and the titres obtained were expressed as the reciprocal dilution of serum that neutralised 100 TCID₅₀ of the Alfort 187 strain in 50% of the culture replicates.

2.3. RNA extraction and qRT-PCR for the detection of CSFV RNA

Viral RNA was extracted from 140 µL of the sample by using a QIAamp Viral RNA Mini Kit (Qiagen GmbH) in accordance to the manufacturer's directions. The synthesis of cDNA was performed by random priming, using M-MLV reverse transcriptase as described previously (Díaz de Arce et al., 2009). CSFV RNA in the serum and nasal and rectal swabs, as well as tonsil, spleen, lymph node and ileum samples, was detected using the qRT-PCR assay previously described (Pérez et al., 2011). This test has been used for inter-laboratory comparisons of CSFV diagnoses, organised by the EURL-CSF. Positive results were considered to be threshold cycle values (CT) ≤35. Samples in which fluorescence was undetectable were considered negative. All qRT-PCR analyses were run on a LightCycler 2.0 instrument (Roche Applied Science, Mannheim, Germany).

2.4. Complete genome sequence of the CSFV Pinar del Rio field strain

The complete genome sequence of CSFV Pinar del Rio was determined first, from the original CSF1058 rescued by transfection of RNA in PK-15 cells at EURL-CSF in Hannover, Germany, and second, from the serum of an experimentally and persistently infected piglet collected 28 days post infection with the original CSF1058 stock (pig number seven, Muñoz-González et al., 2015). The total RNA from the lysate CSF1058-infected PK-15 cells and from the serum of the persistently infected pig number seven was prepared using the Viral Amp RNA purification kit (Qiagen) and TRIzol solution (Life Technologies), respectively. To sequence the nucleotides of the original Pinar del Rio CSF1058 viral RNA, Illumina libraries were generated from 15 ng of RNA using a modified protocol of the SCRIPT SEQ version 2 RNA Seq kit (Epicentre Biotechnologies) as described previously (Fischer et al., 2014). The diluted library (2 nM) was sequenced on an Illumina MiSeq instrument (2x250bp paired end; 4,078,934 reads). De novo assembly and taxonomic classification of the assembled contig were performed as described (Becher et al., 2014; Neill et al., 2014).

To determine the complete genome sequence of the CSFV Pinar der Rio recovered from the persistently infected piglet at 28 dpi (Muñoz-González et al., 2015a), we used reverse transcription (RT) with Superscript III reverse transcriptase (Life Technologies) and PCR amplification with Phusion Hot Start II DNA Polymerase (Thermo Scientific) to generate four overlapping long DNA fragments covering the complete genome except the 5' and 3' extremities. The amplicons from four independent RT-PCR assays of each of the long fragments were inserted in pJet1.2 (CloneJET PCR Cloning Kit). The details of the oligonucleotides used for RT-PCR can be obtained on request. The plasmid DNA was sequenced bi-directionally with forward and reverse primers using standard dideoxy-chain terminator sequencing with the BigDye Terminator v3.1 Cycle Sequencing Kit, (Life Technologies) and an Applied Biosystems 3130 Genetic Analyser (Thermo Fisher Scientific). The DNA sequences were assembled with DNA baser software (Heracle BioSoft SRL) using the quality files and default settings to produce a consensus sequence. The 5' and 3' termini of the viral genome were determined using the 5' and 3' RACE System for Rapid Amplification of cDNA Ends (Life Technologies). Twenty-six clones of RACE fragments were sequenced on both strands to establish the consensus.

The nucleotide and deduced amino acid (aa) sequences were analysed with MAFFT and Clustal Omega software (EMBL-EBI). The nucleotide sequence alignment and the

analysis of the Pinar del Rio sequence was performed with other CSFV sequences obtained from GenBank: Alfort/187 (GenBank accession no. X87939), SXCDK (GQ923951), SXYL2006 (GQ122383), RUCSFPLUM (AY578688), BRESSCIAXPLUM (AY578687), NNQianA (AY663656), 96TD (AY554397), 0406CH01TWN (AY568569), GXWZ02 (AY367767), Riems IVI (AY259122), HCLV (AF531433), 39 (AF407339), LPC (AF352565), cF114 (AF333000), Eystrup IVI (AF326963), CS (AF099102), Alfort Tuebingen (J04358), Shimen (AF092448), HCLV (AF091507), Brescia IVI (AF091661), Alfort A19 (U90951), Glentorf (U45478), Riems Giessen (U45477), LOM (EU789580), flcLOM (UE915211), zj0801 (FJ529205), Sp01 (FJ265020), India (EU857642), Thiverval (EU490425), JL1 (EU497410), Shimen HVRI (AY775158), 944IL94TW (AY646427), CHVRI (AY805221), SWH (DQ127910), Brescia (M31768), CAP (X96550), C-strain (AY382481), ALD (D49532), GPE- (D49533), Paderborn (GQ902941), Roesrath (GU233734), Hennef (GU233733), Euskirchen (GU233732), Borken (GU233731), C-ZJ-2008 (HM175885) HEBZ (GU592790) and Koslov (HM237795).

2.5. Nucleotide sequence of the 3'UTR-terminal region of CSFV Margarita and Labiofam strains

To determine the sequence of the 3'UTR-terminal region, we retrieved 15 representative sequences of the three CSFV genotypes, including the vaccine and low and high virulence previously reported CSFV strains, from GenBank. The sequences were aligned using BioEdit (Hall, 1999). Two primers were designed into the 3' end of the CSFV as follows: forward primer (12010-12029), 5'-TCAACATAGTGTTAAGGAGG-3'; reverse primer (12251-12301), 5'-CCGTTAGGAAATTACCTTAG -3'. The nucleotide positions were based on the genome sequence of the Pinar del Rio field strain determined in this study. The one-step RT-PCR protocol was undertaken using the commercially available One-Step RT-PCR Kit (Qiagen). The temperature profile was 30 minutes at 50 °C (reverse transcription), 10 minutes at 95 °C (inactivation reverse transcriptase/activation Taq polymerase), followed by 35 cycles of 30 seconds at 94 °C (denaturation), 35 seconds at 52 °C (annealing) and 30 seconds at 72 °C (elongation). The amplification products were checked by electrophoresis on 2% agarose gel and were directly cleaned with a Wizard® PCR Preps DNA Purification System (Promega, Madison, Wisconsin, USA). The sequencing reactions were conducted under BigDye™ terminator-cycling

conditions using an ABI 3130XL. Forward and reverse sequences obtained from each amplicon were assembled using the Contig Express application in Vector NTI software, version 11 (Invitrogen). Finally, the 3'-UTR RNA secondary structure from the 3'-UTR start nucleotide position (12075) to the genome final (12301) of the CSFV Pinar del Rio field isolate and Margarita strain were predicted using the mfold Web Server software (The RNA Institute college of Arts and Science University at Albany, New York, USA).

2.6. Antibody production and cross-neutralisation assays

Six New Zealand White rabbits weighing 2.5 kg were separated into three groups of two each. Group 1 was immunised with the CSFV Pinar del Rio field isolate derived from the Cuban CSF epizootic of 2010 (PR-11/10-3) (Pérez et al., 2012); group 2 was inoculated with Margarita strain, and group 3 with PBS. The rabbits were subcutaneously immunised three times at 3-week intervals, at six deposits in the back (Huang et al., 2012). The virus solution or PBS was emulsified V/V with Montanide 888 adjuvant (Seppic, France). Each dose of emulsion consisted of $10^{7.2}$ TCID₅₀ of Pinar del Rio or Margarita strains, respectively. Before each inoculation, serum samples were collected from the rabbits via the jugular vein. After euthanising the animals (2 weeks after the last inoculation), sera pools for each inoculated group were prepared. The antibody titres were monitored using the NPLA assay against the homologous and, in parallel, a cross-neutralisation assay against the heterologous virus, respectively (Terpstra et al., 1984). The titres obtained were expressed as the reciprocal dilution of serum that neutralised 100 TCID₅₀ of Pinar del Rio field strain or Margarita strains, respectively in 50% of the culture replicates. Similarly, the neutralising titers were calculated using Kärber's method (Kärber, 1931). Lastly, the mean value of neutralising antibody titres were determined for each virus and experimental group.

In addition, we had previously prepared and stocked pig hyperimmune serum against the CSFV Labiofam vaccine strain. The pig antiserum was generated with two doses at 4-week intervals in 6-week-old CSFV-free domestic pigs. The protocols were approved by the ethics committee of the MINAGRI of the Republic of Cuba. Each dose contained 100 protective doses (PD) administered by intramuscular injection in the neck (Muñoz-González et al., 2015b). The sera were collected at different times post vaccination, and the samples collected 78 days post immunisation were used in the neutralisation test. The titres obtained were expressed as the reciprocal dilution of serum that neutralised

100 TCID₅₀ of Alfort/187 strain, in 50% of the culture replicates. In addition, a neutralisation test was conducted (Terpstra et al., 1984) in order to evaluate the neutralisation capacity of the anti-Labiofam vaccine pig hyperimmune sera against the CSFV Pinar del Rio field isolate originated from the Cuban CSF epizootic of 2010 (PR-11/10-3) (Pérez et al., 2012) and the Margarita strain, respectively. To this end, triplicate heat-inactivated sera with a neutralisation titre of 1:200 against CSFV Alfort/187 strain were mixed with equal volumes of 10^{7.2} TCID₅₀ of Margarita or Pinar del Rio virus suspensions, respectively. The reactions were incubated at 37°C for 1 h and subsequently transferred to 96-well plates with PK-15 cells. Peroxidase-linked assay (PLA) (Wensvoort et al., 1986) was used for viral titration following the statistical methods described by Reed and Muench (Reed and Muench, 1938).

3. Results

3.1. The CSFV Pinar del Rio field isolate is low virulent in weaner pigs

In order to confirm the low pathogenicity of the CSFV Pinar del Rio field isolate that originated from the Cuban CSF epizootic of 2010 (PR-11/10-3) (Pérez et al., 2012) in comparison with Margarita strain, the virulence of these two viruses were assessed side by side in weaner pigs. All the Margarita-infected pigs developed pyrexia (rectal temperature higher than 40 °C) starting at day 4 and lasting until 18 dpi, with peaks greater than 41.5 °C. Moderate and severe clinical symptoms (≥ 4 points in score value), such as anorexia, conjunctivitis, diarrhoea, constipation, abdominal petechiae and nervous disorders, were observed particularly from 12 dpi onwards. Pig number 6 suddenly died at 15 dpi. The other two animals were euthanised for ethical reasons at 17 or 18 dpi (Fig. 1). This contrasted with Pinar del Rio that induced only mild symptoms and slightly elevated body temperatures between 4 and 9 dpi. The clinical scores of the two groups were significantly different from day 6 until the end of the experiment (Fig. 1A). Neutralising antibodies were detected in all the Pinar del Rio-infected pigs as early as 11 dpi and the titres increased up to 800-3200 reciprocal 50% neutralising dilution on 21 dpi, while the titres remained very low in pigs infected with the Margarita strain (Fig. 1B). Compared to Margarita, the replication of Pinar del Rio was delayed by 4 days, and the viral RNA content of the serum and nasal swabs was an average 1000 times lower (Fig. 1C and E). In the rectal swabs, the viral RNA content was low for both viruses and the differences were less pronounced (Fig. 1D). A clearly lower viral RNA content of the same order of magnitude was also detected in the tissue samples,

that is, the tonsils, spleen, lymph nodes and ileum of Pinar del Rio- versus Margarita-infected pigs (Fig. 1F).

3.2. Complete nucleotide sequence analysis of CSFV Pinar del Rio revealed a novel poly-uridine tract in the 3'UTR

In order to determine whether the Pinar del Rio virus did change genetically during its persistence in piglets, the complete genome sequence of the virus was determined from serum collected after 28 days of persistence and from the original Pinar del Rio virus stock (CSF1058). The complete sequence of the Pinar del Rio virus was determined from both the inoculum and the virus recovered from the persistently infected piglets. However, 11 nucleotides from the 5' end and 53 from the 3' end were not possible to sequence from the Pinar del Rio inoculum by the mass sequencing Illumina method. Eventually, the 3' end was fully amplified and sequenced by using the Sanger Methods as previously described. Remarkably, the two consensus nucleotide sequences from the Pinar del Rio virus stock (CSF1058) and the Pinar del Rio virus collected after 28 days of persistence in a pig were 100% identical. This sequence, the first obtained from the CSFV 1.4 genogroup, was deposited in GenBank with the accession number KX576461.

The complete 3'-UTR sequences of the Margarita and Labiofam strains were also determined and deposited in GenBank with the accession number (LT601427) and (LT601428), respectively. All the Pinar del Rio sequences analysed showed a novel uninterrupted polyuridine (poly-U) sequence of an average length of 36 uridines in the 3'-UTR starting at nucleotide 12225, a position where all other CSFV strains sequenced to date including the Margarita strain have typically only 4 to 5 uridines (Fig. 2). This poly-U sequence was also found in the Pinar del Rio virus stock used to inoculate the newborn piglets. The length of the poly-U sequence was heterogeneous, varying from 19 to 47 nucleotides in a total of 26 independent cDNA clones analysed. In some individual clones, the poly-U was interrupted with one or two cytidines and/or adenosines. It is worth noting that in the Labiofam vaccine strain, a poly-U tract was found starting at position 12137 within the 3'UTR, that is, nearly 90 nucleotides upstream of the position of the poly-U stretch discovered in the genome of Pinar del Rio (Fig. 2).

3.3. The secondary structure prediction of the 3'-UTR end

The effect of the novel poly-U sequence on the predicted RNA secondary structure was compared in the Pinar del Rio and Margarita strains using the mFold program. A secondary structure containing four stem-loops (SL-I to SL-IV) was obtained with the 3'-UTR of both CSFV Margarita and Pinar del Rio (Fig. 3). The poly-U insertion in Pinar del Rio forms a long single-stranded intervening sequence (SS) that increases the distance between SL-I and SL-II. The 3'-UTR of CSFV Pinar del Rio and Margarita strain have the same predicted RNA secondary structure when the poly-U sequence of Pinar del Rio is removed.

3.4. The amino acid differences in the envelope proteins of the CSFV Pinar del Rio and Margarita virus clusters in E2

The envelope protein sequences of E^{ms}, E1, E2 and p7 were compared by aligning the sequence data available for CSFV Pinar del Rio (Genbank accession nos. JX028204 and KX576461) and Margarita (Genbank accession nos. JX028201 and AJ704817.1). Three conservative K385R, V674I, and D1097E substitutions were found in E^{ms}, E1 and p7, respectively. Six substitutions were found in E2, two of them, G761R and L763S, being located within the B/C domain (Fig 4). The four other changes were the conservative I780V and the non-conservative D898V, L956S, and T1028A mutations. Of note, a single non-synonymous change was found between the newly determined Pinar del Rio sequence (KX576461) and the partial sequence of Pinar del Rio published previously (JX028204), resulting in a H852Y substitution in E2 of the newly determined Pinar del Rio sequence (KX576461). At this position, the Margarita virus carries H852 (Fig. 4) as Pinar del Rio JX028204.

3.5. Differences in the antigenic reactivity and immunogenicity between Pinar del Rio and Margarita strains

In order to determine whether the amino acid differences found in E2 modify the antigenic properties of Pinar del Rio and Margarita, we first compared the susceptibility of the two viruses to neutralisation by the porcine hyperimmune serum against the CSFV Labiofam vaccine strain. A porcine hyperimmune serum against Labiofam vaccine strain neutralised the Margarita virus completely from 10^7 TCID₅₀/mL to undetectable levels, while the infectivity of the Pinar del Rio virus was reduced by three log₁₀ only, from $10^{7.2}$ TCID₅₀/mL to $10^{4.3}$ TCID₅₀/mL.

Finally, we compared the immunogenicity of Pinar del Rio and Margarita in rabbits. Rabbits were used for this purpose because the virulent Margarita strain causes severe clinical disease and even death in infected pigs before the induction of an effective humoral response. Neutralising antibody titres were determined against both the homologous and the heterologous virus in a cross-neutralisation assay. Higher neutralising antibody titres were found in the Margarita-immunised rabbits with titres of 7499 and 2239 against the homologous and heterologous viruses, respectively. In contrast, 13 and 8 times lower neutralisation titres were detected in sera from the Pinar del Rio-immunised animals, with titres of 562 and 282 against the homologous (Pinar del Rio field isolate) and heterologous (Margarita strain) virus strains, respectively.

4. Discussion

Despite the importance of low virulent CSFV strains, available data on the molecular biology of CSFV attenuation and pathogenesis induced by this type of viral strain are scarce. This study describes the full-length nucleotide sequence of the Pinar del Rio isolate PR-11/10-3 (Pérez et al., 2012). This field strain was recovered in the west of Cuba in 2010 with the G761R replacement into the E2-B/C-domain, being previously suggested its implication in the virulence and the pathogenesis of CSFV in the field (Shen et al., 2011; Pérez et al., 2012; Hu et al., 2016). Accordingly, molecular and antigenic properties of the E2 protein from Pinar del Rio virus and the virulent Margarita strain, the most likely origin for the Cuban outbreaks in 1993, were compared. The virulence of the Pinar del Rio isolate and Margarita strain was also evaluated side by side in weaner pigs. In accordance with previous data (Postel et al., 2015), the field isolate Pinar del Rio strain induced only very mild clinical signs in the infected piglets, along with only short-term low viral RNA levels in the blood, organs and body secretions. Considering the level of CSFV replication *in vivo* as a major correlate of virulence (Leifer et al., 2013; von Rosen et al., 2013; Tamura et al., 2014; Tarradas et al., 2014), these data confirmed the low virulence of the Pinar del Rio field isolate, being in line with previous studies from endemic countries (Pérez et al., 2012; Ji et al., 2014; Hu et al., 2016).

During experimental infection, neutralising antibody responses were mainly detected in the Pinar del Rio-infected piglets, while the virulent Margarita strain was unable to induce an effective humoral response. This confirmed the ineffectiveness of the virulent CSFV strains to induce neutralising antibodies during severe CSF disease (Ganges et

al., 2005 and Tarradas et al., 2010). Pigs with higher neutralising antibody titres had lower CSFV RNA loads, especially at 21 dpi when two out of five pigs had cleared the virus. Nevertheless, viral RNA was detected in 55% of the tissues samples at necropsy, principally in the tonsils and spleen of the Pinar del Rio-infected pigs (Fig. 1F), corroborating the role of these lymphoid organs in the spread and persistence of CSFV (Ganges et al., 2008; Tarradas et al., 2014; Muñoz-González et al., 2015a).

We have previously reported that the early post-natal infection of piglets with the Pinar del Rio isolate resulted in persistent infection lasting for at least 4 weeks in the absence of any detectable antibody response (Muñoz-González et al., 2015a). In this study, we determined the complete genome sequence of the Pinar del Rio virus used to infect the newborn piglets and the virus re-isolated from the serum after 28 days of persistence. Importantly, this full-length consensus sequence represents the first complete genome sequence of a CSFV subgenotype 1.4. It is noteworthy that the consensus sequences obtained before and after persistence were completely identical. As a result, the CSFV genome sequence of low virulent Pinar del Rio field strain remains unaltered after 28 days of persistent infection in piglets (Muñoz-González et al., 2015a). It was recently determined that viral samples from pigs infected with highly virulent CSFV isolates “Koslov” or “Brescia” showed higher quasispecies diversity and more nucleotide variability, compared to samples of pigs infected with low and moderately virulent isolates (Töpfer et al., 2013). In addition, a previous work analysing the Cuban CSFV population dynamics should be noted. This study revealed a decrease in the genetic diversity of the viral population from the epizootic beginning, when the acute disease was prevalent, to recent years (Pérez et al., 2012). The high stability showed by this low virulent CSFV field strain may be linked to a potential adaptive advantage, which may favour the prevalence of the low virulent strain circulating in the field. Considering the role of the mechanism showing the suppression of superinfection (SIE) in CSFV persistently infected pigs, the possible outcome of the generation of the new low pathogenicity CSFV strains circulating in an endemic situation and the impact of the SIE on disease control cannot be underrated (Muñoz-González et al., 2016).

A unique uninterrupted poly-U insertion of an average length of 36 nucleotides was found in the 3'-UTR the Pinar del Rio genome. Such a poly-U insertion has never been described before at this position in the 3'UTR of pestiviruses according to GenBank. Also of interest is a similar 6- to 32-nucleotide-long U-rich insertion that can be found

in the 3'-UTR of several attenuated CSF vaccine virus strains, such as HCLV, C-strain, Porcivac, Rovac, Russian LK and Thiverval, nearly 90 nucleotides upstream compared to the Pinar del Rio poly-U (Wu et al., 2001; Fan et al., 2008). This latter U-rich sequence was also found in the Labiofam vaccine strain used in Cuba to control CSFV over decades, as well as and in the lapinised Harbin vaccine strain (Genbank accession number AY805221.1) (Fig. 2). Although the possible functions of the U-rich insertions in the 3'-UTR of the attenuated vaccine strains are still unknown, it has been speculated that the U-rich sequence in the C-strain vaccine was acquired during the adaptation of the virus to the rabbit host (Li et al., 2014). Accordingly, it is of interest to speculate that the U-rich sequences of the vaccine strains and the poly-U insertion in the genome of Pinar del Rio contribute to attenuation. Of note, the long poly-U/UC tract located in the 3'UTR of HCV was found to trigger innate immune induction via RIG-I, which was dependent on the composition and length of the poly-U stretch (Schnell et al., 2012; Kell et al., 2015). The 3'-UTR of pestiviruses contains *cis*-active elements that are indispensable for viral replication and translation (Austermann-Busch and Becher, 2012; Li et al., 2014). This structure also plays a regulatory role in IRES-mediated translation (Huang et al., 2012). Previous studies showed that the insertion and deletion of nucleotides in the 3'-UTR of pestiviruses may lead to significant changes in viral RNA synthesis (Xiao et al., 2004; Pankraz et al., 2005). Although the implication and functionality of the poly-U insertion in the Pinar del Rio field strain remain to be determined, it could be associated with decreased virulence and a resulting potential advantage of this viral strain within the endemic situation which may favour the establishment of low virulent CSFV. Previously described RNA secondary structure models of CSFV 3'-UTR predict four consecutive stem-loop structures, SL-I, SL-II, SL-III and SL-IV (Fan et al., 2008; Huang et al., 2012). These can also be predicted for the Margarita strain. The poly-U insertion in the genome of Pinar del Rio leads to a long single-stranded intervening sequence (SS) between SL-I and SL-II in the predicted RNA secondary structure (Fig 3). Previous studies revealed that the SL-I and the SS region between SL-I and SL-II are essential for viral replication (Pankraz et al., 2005; Huang et al., 2012). Pestivirus replication depends on the binding of microRNA-17 (miR-17) to a highly conserved sequence between SL-I and SL-II (Scheel et al., 2016). This miR-17 binding site is located 3 nucleotides downstream of the large poly-U insertion identified in the Pinar del Rio genome. The binding of miR-17 by the

pestivirus genome enhances translation and RNA stability. Whether and how the poly-U stretch and the miR-17 binding may be linked functionally is a matter of future studies.

In our comparison of the sequences of the Pinar del Rio and Margarita strains, we also noticed the fact that the amino acid differences in the envelope proteins clustered within E2. Among these was the previously described G761R mutation within the B/C domain of E2 (Shen et al., 2011; Pérez et al., 2012; Hu et al., 2016). Previous studies have suggested that the TAVSPTTLR motif and an additional 13 amino acids in the carboxy-terminal part of E2 may be significantly implicated in CSFV virulence (Risatti et al., 2006; Risatti et al., 2007; Shen et al., 2011; Tamura et al., 2012). However, these amino acid residues are conserved in Margarita and Pinar del Rio. In the p7 protein, which is a viroporin possibly implicated in CSFV virulence (Gladue et al., 2012), a conservative D (Margarita) to E (Pinar del Rio) substitution was found at position 1096. This position connects two transmembrane helices of p7. For the related hepatitis C virus (HCV) it has been reported that changes around this site (aa 1095 to 1101) prevent HCV replication in chimpanzees (Sakai et al., 2003).

Finally, we found differences in the antigenic properties of E2 of the Pinar del Rio and Margarita virus. The sera from pigs vaccinated with the Labiofam vaccine were less efficient at completely neutralising the Pinar del Rio field strain than the Margarita strain. In addition, when using sera from rabbits infected with either of the two viruses, we found that the neutralising antibody titres against both the homologous and heterologous viruses were considerably lower in the Pinar del Rio-infected rabbits compared to rabbits infected with Margarita. Glycosylation of viral envelope proteins may influence the immunogenicity and the sensitivity of the virus to neutralising antibodies by maintaining the appropriate conformation of proteins (Li et al., 2008). Accordingly, mutations in the E2 B/C domain including the amino acid residues 761 and 763 might lead to variations in virus neutralisation (Chen et al., 2010; Chang et al., 2012; Hu et al., 2016). Additionally, previous studies showed that alteration into the B/C domain, mainly from amino acids 690 to 773, reduced the binding of pig serum raised against the C strain (Tong et al., 2015). Thus, the changes found in E2 of the attenuated CSFV strain Pinar del Rio may be associated with a possible viral escape from neutralising antibodies, which may be related to the viral persistence and low virulence of Pinar del Rio.

In conclusion, the evolutionarily related low and high virulent CSFV strains of Pinar del Rio and Margarita represent a valuable virus pair for studying CSFV virulence factors. Future experimental approaches employing reverse genetics are required to clarify the issues identified in the present study.

Acknowledgements

The research in CReSA, Spain, was supported by grant AGL2015-66907 from the Spanish government. L.C. had a 2015/16 scholarship from the MAEC–AECID Program of the Spanish Government. S. M. had a pre-doctoral fellowship FI-DGR 2014 from AGAUR, Generalitat de Catalunya. M.L. was supported by grant #310030-141045 from the Swiss National Science Foundation to N.R.

Competing interests

The authors declare that they have no competing interests.

References

- Austermann-Busch, S., Becher, P., 2012. RNA structural elements determine frequency and sites of nonhomologous recombination in an animal plus-strand RNA virus. *J. Virol.* 86, 7393–402. doi:10.1128/JVI.00864-12
- Aynaoud, J.M., Launais, M., 1978. Hog cholera: immunization of young pigs with the Thiverval strain vaccine in the presence of colostral immunity. *Dev. Biol. Stand.* 41, 381–7.
- Becher, P., Fischer, N., Grundhoff, A., Stalder, H., Schweizer, M., Postel, A. 2014. Complete genome sequence of bovine pestivirus strain PG-2, a second member of the tentative Pestivirus species giraffe. *Genome Announc.* 15, pii: e00376-14. doi: 10.1128/genomeA.00376-14.
- Beer, M., Goller, K., Staubach, C., Blome S., 2015. Genetic variability and distribution of Classical swine fever virus. *Anim. Health Res Rev.* 16,33-9.
- Cecchinato, M., Catelli, E., Lupini, C., Ricchizzi, E., Clubbe, J., Battilani, M., Naylor, C.J., 2010. Avian metapneumovirus (AMPV) attachment protein involvement in probable virus evolution concurrent with mass live vaccine introduction. *Vet. Microbiol.* 146, 24–34. doi:10.1016/j.vetmic.2010.04.014

- Chang, C.Y., Huang, C.C., Deng, M.C., Huang, Y.L., Lin, Y.J., Liu, H.M., Lin, Y.L., Wang, F.I., 2012. Antigenic mimicking with cysteine-based cyclized peptides reveals a previously unknown antigenic determinant on E2 glycoprotein of classical swine fever virus. *Virus Res.* 163, 190–6. doi:10.1016/j.virusres.2011.09.019
- Chen, N., Tong, C., Li, D., Wan, J., Yuan, X., Li, X., Peng, J., Fang, W., 2010. Antigenic analysis of classical swine fever virus E2 glycoprotein using pig antibodies identifies residues contributing to antigenic variation of the vaccine C-strain and group 2 strains circulating in China. *Virol. J.* 7, 378. doi:10.1186/1743-422X-7-378
- Díaz de Arce, H., Núñez, J.I., Ganges, L., Barreras, M., Teresa Frías, M., Sobrino, F., 1999. Molecular epidemiology of classical swine fever in Cuba. *Virus Res.* 64, 61–7.
- Díaz de Arce, H., Pérez, L.J., Frías, M.T., Rosell, R., Tarradas, J., Núñez, J.I., Ganges, L., 2009. A multiplex RT-PCR assay for the rapid and differential diagnosis of classical swine fever and other pestivirus infections. *Vet. Microbiol.* 139, 245–52. doi:10.1016/j.vetmic.2009.06.004
- Díaz de Arce, H., Ganges, L., Barrera, M., Naranjo, D., Sobrino, F., Frías, M.T., Núñez, J.I., 2005. Origin and evolution of viruses causing classical swine fever in Cuba. *Virus Res.* 112, 123–31. doi:10.1016/j.virusres.2005.03.018
- Fan, Y., Zhao, Q., Zhao, Y., Wang, Q., Ning, Y., Zhang, Z., 2008. Complete genome sequence of attenuated low-temperature Thiverval strain of classical swine fever virus. *Virus Genes* 36, 531–8. doi:10.1007/s11262-008-0229-x
- Ganges, L., Barrera, M., Núñez, J.I., Blanco, I., Frías, M.T., Rodríguez, F., Sobrino, F., 2005. A DNA vaccine expressing the E2 protein of classical swine fever virus elicits T cell responses that can prime for rapid antibody production and confer total protection upon viral challenge. *Vaccine* 23, 3741–52. doi:10.1016/j.vaccine.2005.01.153
- Ganges, L., Núñez, J.I., Sobrino, F., Borrego, B., Fernández-Borges, N., Frías-Lepoureau, M.T., Rodríguez, F., 2008. Recent advances in the development of recombinant vaccines against classical swine fever virus: cellular responses also play a role in protection. *Vet. J.* 177, 169–77. doi:10.1016/j.tvjl.2007.01.030
- Gladue, D.P., Holinka, L.G., Largo, E., Fernandez Sainz, I., Carrillo, C., O'Donnell, V., Baker-Branstetter, R., Lu, Z., Ambroggio, X., Risatti, G.R., Nieva, J.L., Borca, M.V.,

2012. Classical swine fever virus p7 protein is a viroporin involved in virulence in swine. *J. Virol.* 86, 6778–91. doi:10.1128/JVI.00560-12

Graham, S.P., Haines, F.J., Johns, H.L., Sosan, O., La Rocca, S.A., Lamp, B., Rüménapf, T., Everett, H.E., Crooke, H.R., 2012. Characterisation of vaccine-induced, broadly cross-reactive IFN- γ secreting T cell responses that correlate with rapid protection against classical swine fever virus. *Vaccine* 30, 2742–8. doi:10.1016/j.vaccine.2012.02.029

Hall, T.A., 1999. BioEdit: A user-friendly biological sequence alignment program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.* 41, 95–98.

Hu, D., Lv, L., Gu, J., Chen, T., Xiao, Y., Liu, S., 2016. Genetic diversity and positive selection analysis of Classical swine fever virus envelope protein gene E2 in east China under C-Strain vaccination. *Front. Microbiol.* 7, 85. doi:10.3389/fmicb.2016.00085

Huang, S.W., Chan, M.Y., Hsu, W.L., Huang, C.C., Tsai, C.H., 2012. The 3'-terminal hexamer sequence of classical swine fever virus RNA plays a role in negatively regulating the IRES-mediated translation. *PLoS One* 7, e33764. doi:10.1371/journal.pone.0033764

Ji, W., Niu, D.D., Si, H.L., Ding, N.Z., He, C.Q., 2014. Vaccination influences the evolution of classical swine fever virus. *Infect. Genet. Evol.* 25, 69–77. doi:10.1016/j.meegid.2014.04.008

Kärber, G., 1931. Beitrag zur kollektiven Behandlung pharmakologischer Reihenversuche. *Naunyn. Schmiedebergs. Arch. Exp. Pathol. Pharmacol.* 162, 480–483. doi:10.1007/BF01863914

Kell A., Stoddard M., Li, H., Marcotrigiano, J., Shaw G.M, Gale M Jr., 2015. Pathogen-associated molecular pattern recognition of Hepatitis C virus transmitted/founder Variants by RIG-I is dependent on U-Core length. *J Virol.* 89, 11056-68. doi: 10.1128/JVI.01964-15.

Leifer, I., Ruggli, N., Blome, S., 2013. Approaches to define the viral genetic basis of classical swine fever virus virulence. *Virology* 438, 51–5. doi:10.1016/j.virol.2013.01.013

- Li, C., Li, Y., Shen, L., Huang, J., Sun, Y., Luo, Y., Zhao, B., Wang, C., Yuan, J., Qiu, H.J., 2014. The role of noncoding regions of classical swine fever virus C-strain in its adaptation to the rabbit. *Virus Res.* 183, 117–22. doi:10.1016/j.virusres.2014.02.003
- Li, H., Chien, P.C., Tuen, M., Visciano, M.L., Cohen, S., Blais, S., Xu, C.F., Zhang, H.T., Hioe, C.E., 2008. Identification of an N-linked glycosylation in the C4 region of HIV-1 envelope gp120 that is critical for recognition of neighboring CD4 T cell epitopes. *J. Immunol.* 180, 4011–21.
- Liu, L., Xia, H., Everett, H., Sosan, O., Crooke, H., Meindl-Böhmer, A., Qiu, H.J., Moennig, V., Belák, S., Widén, F., 2011. A generic real-time TaqMan assay for specific detection of lapinized Chinese vaccines against classical swine fever. *J. Virol. Methods* 175, 170–4. doi:10.1016/j.jviromet.2011.05.003
- Meyer, D., Schmeiser, S., Postel, A., Becher, P., 2015. Transfection of RNA from organ samples of infected animals represents a highly sensitive method for virus detection and recovery of classical swine fever virus. *PLoS One* 10, e0126806. doi:10.1371/journal.pone.0126806
- Moennig, V., Floegel-Niesmann, G., Greiser-Wilke, I., 2003. Clinical signs and epidemiology of classical swine fever: a review of new knowledge. *Vet. J.* 165, 11–20.
- Muñoz-González, S., Pérez-Simó, M., Colom-Cadena, A., Cabezón, O., Bohórquez, J.A., Rosell, R., Pérez, L.J., Marco, I., Lavín, S., Domingo, M., Ganges, L., 2016. Classical Swine Fever Virus vs. Classical Swine Fever Virus: The Superinfection Exclusion Phenomenon in Experimentally Infected Wild Boar. *PLoS One* 11, e0149469. doi:10.1371/journal.pone.0149469
- Muñoz-González S, Perez-Simó M, Muñoz M, Bohorquez JA, Rosell R, Summerfield A, Domingo M, Ruggli N, Ganges L., 2015. Efficacy of a live attenuated vaccine in classical swine fever virus postnatally persistently infected pigs. *Vet Res.* 46:78. doi: 10.1186/s13567-015-0209-9.
- Muñoz-González, S., Ruggli, N., Rosell, R., Pérez, L.J., Frías-Leuporeau, M.T., Fraile, L., Montoya, M., Cordoba, L., Domingo, M., Ehrensperger, F., Summerfield, A., Ganges, L., 2015. Postnatal persistent infection with classical Swine Fever virus and its immunological implications. *PLoS One* 10, e0125692. doi:10.1371/journal.pone.0125692

- Neill, J.D., Ridpath, J.F., Fischer, N., Grundhoff, A., Postel, A., Becher, P., 2014. Complete genome sequence of pronghorn virus, a pestivirus. *Genome Announc.* 12;2(3). pii: e00575-14. doi: 10.1128/genomeA.00575-14.
- Pankraz, A., Thiel, H.J., Becher, P., 2005. Essential and nonessential elements in the 3' nontranslated region of Bovine viral diarrhea virus. *J. Virol.* 79, 9119–27. doi:10.1128/JVI.79.14.9119-9127.2005
- Pérez LJ, Díaz de Arce H, Tarradas J, Rosell R, Perera CL, Muñoz M, Frías MT, Nuñez JI, Ganges L., 2011. Development and validation of a novel SYBR Green real-time RT-PCR assay for the detection of classical swine fever virus evaluated on different real-time PCR platforms. *J Virol Methods.* 174:53-9. doi: 10.1016/j.jviromet.2011.03.022.
- Pérez, L.J., Díaz de Arce, H., Perera, C.L., Rosell, R., Frías, M.T., Percedo, M.I., Tarradas, J., Dominguez, P., Núñez, J.I., Ganges, L., 2012. Positive selection pressure on the B/C domains of the E2-gene of classical swine fever virus in endemic areas under C-strain vaccination. *Infect. Genet. Evol.* 12, 1405–12. doi:10.1016/j.meegid.2012.04.030
- Postel, A., Pérez, L.J., Perera, C.L., Schmeiser, S., Meyer, D., Meindl-Boehmer, A., Rios, L., Austermann-Busch, S., Frias-Lepoureau, M.T., Becher, P., 2015. Development of a new LAMP assay for the detection of CSFV strains from Cuba: a proof-of-concept study. *Arch. Virol.* 160, 1435–48. doi:10.1007/s00705-015-2407-1
- Postel, A., Schmeiser, S., Perera, C.L., Rodríguez, L.J., Frias-Lepoureau, M.T., Becher, P., 2013. Classical swine fever virus isolates from Cuba form a new subgenotype 1.4. *Vet. Microbiol.* 161, 334–8. doi:10.1016/j.vetmic.2012.07.045
- Reed, L.J., Muench, H., 1938. A simple method of estimating fifty per cent endpoints. *Am. J. Epidemiol.* 27, 493–497.
- Risatti, G.R., Holinka, L.G., Carrillo, C., Kutish, G.F., Lu, Z., Tulman, E.R., Sainz, I.F., Borca, M.V., 2006. Identification of a novel virulence determinant within the E2 structural glycoprotein of classical swine fever virus. *Virology* 355, 94–101. doi:10.1016/j.virol.2006.07.005
- Risatti, G.R., Holinka, L.G., Fernandez Sainz, I., Carrillo, C., Kutish, G.F., Lu, Z., Zhu, J., Rock, D.L., Borca, M.V., 2007. Mutations in the carboxyl terminal region of E2

glycoprotein of classical swine fever virus are responsible for viral attenuation in swine. *Virology* 364, 371–82. doi:10.1016/j.virol.2007.02.025

Sakai, A., Claire, M.S., Faulk, K., Govindarajan, S., Emerson, S.U., Purcell, R.H., Bukh, J., 2003. The p7 polypeptide of hepatitis C virus is critical for infectivity and contains functionally important genotype-specific sequences. *Proc. Natl. Acad. Sci. U. S. A.* 100, 11646–51. doi:10.1073/pnas.1834545100

Scheel, T.K., Luna, J.M., Liniger, M., Nishiuchi, E., Rozen-Gagnon, K., Shlomai, A., Auray, G., Gerber, M., Fak, J., Keller, I., Bruggmann, R., Darnell, R.B., Ruggli, N., Rice, C.M., 2016. A broad RNA virus survey reveals both miRNA dependence and functional sequestration. *Cell Host Microbe.* 19, 409-23. doi: 10.1016/j.chom.2016.02.007.

Schnell, G., Loo, Y.M., Marcotrigiano, J., Gale, M. Jr. 2012. Uridine composition of the poly-U/UC tract of HCV RNA defines non-self recognition by RIG-I. *PLoS Pathog.* 8, e1002839

Shen, H., Pei, J., Bai, J., Zhao, M., Ju, C., Yi, L., Kang, Y., Zhang, X., Chen, L., Li, Y., Wang, J., Chen, J., 2011. Genetic diversity and positive selection analysis of classical swine fever virus isolates in south China. *Virus Genes* 43, 234–42. doi:10.1007/s11262-011-0625-5

Tamura, T., Nagashima, N., Ruggli, N., Summerfield, A., Kida, H., Sakoda, Y., 2014. Npro of classical swine fever virus contributes to pathogenicity in pigs by preventing type I interferon induction at local replication sites. *Vet. Res.* 45, 47-57. doi:10.1186/1297-9716-45-47

Tamura, T., Sakoda, Y., Yoshino, F., Nomura, T., Yamamoto, N., Sato, Y., Okamatsu, M., Ruggli, N., Kida, H., 2012. Selection of classical swine fever virus with enhanced pathogenicity reveals synergistic virulence determinants in E2 and NS4B. *J. Virol.* 86, 8602–13. doi:10.1128/JVI.00551-12

Tarradas, J., de la Torre, M.E., Rosell, R., Pérez, L.J., Pujols, J., Muñoz, M., Muñoz, I., Muñoz, S., Abad, X., Domingo, M., Fraile, L., Ganges, L., 2014. The impact of CSFV on the immune response to control infection. *Virus Res.* 185, 82–91. doi:10.1016/j.virusres.2014.03.004

- Tarradas J, Argilagué JM, Rosell R, Nofrarías M, Crisci E, Córdoba L, Pérez-Martín E, Díaz I, Rodríguez F, Domingo M, Montoya M, Ganges L., 2010. Interferon-gamma induction correlates with protection by DNA vaccine expressing E2 glycoprotein against classical swine fever virus infection in domestic pigs. *Vet Microbiol.* 142:51-8. doi: 10.1016/j.vetmic.2009.09.043.
- Tautz, N., Tews, B.A., Meyers, G., 2015. The Molecular Biology of Pestiviruses. *Adv. Virus Res.* 93, 47-160. doi: 10.1016/bs.aivir.2015.03.002. Epub 2015 Apr 29.
- Terpstra, C., Bloemraad, M., Gielkens, A.L., 1984. The neutralizing peroxidase-linked assay for detection of antibody against swine fever virus. *Vet. Microbiol.* 9, 113–20.
- Simmonds, P., Becher, P., Collett, M.S., Gould, E.A., Heinz, F.X., Meyers, G., Monath, T., Pletnev, A., Rice, C.M., Stiasny, K., Thiel, H.J., Weiner, A., Bukh, J., 2012. Family Flaviviridae. In: King, A.M.Q., Adams, M.J., Carstens, E.B., Lefkowitz, E.J. (Eds.), *Virus Taxonomy. Ninth Report of the International Committee on Taxonomy of Viruses*, Elsevier Academic Press, San Diego, USA, pp. 1004-20
- Tong, C., Chen, N., Liao, X., Xie, W., Li, D., Li, X., Fang, W., 2015. The epitope recognized by monoclonal antibody 2B6 in the B/C domains of Classical swine fever virus glycoprotein E2 affects viral binding to hyperimmune sera and replication. *J. Microbiol. Biotechnol.* 25, 537–46.
- Töpfer, A., Höper, D., Blome, S., Beer, M., Beerenwinkel, N., Ruggli, N., Leifer, I., 2013. Sequencing approach to analyze the role of quasispecies for classical swine fever. *Virology* 438, 14–9. doi:10.1016/j.virol.2012.11.020
- Van Oirschot, J.T., De Jong, D., Huffels, N.D., 1983. Effect of infections with swine fever virus on immune functions. II. Lymphocyte response to mitogens and enumeration of lymphocyte subpopulations. *Vet. Microbiol.* 8, 81–95.
- Vera, F., Craig, M.I., Olivera, V., Rojas, F., König, G., Pereda, A., Vagnozzi, A., 2015. Molecular characterization of infectious bursal disease virus (IBDV) isolated in Argentina indicates a regional lineage. *Arch. Virol.* 160, 1909–21. doi:10.1007/s00705-015-2449-4
- von Rosen, T., Lohse, L., Nielsen, J., Uttenthal, Å., 2013. Classical swine fever virus infection modulates serum levels of INF- α , IL-8 and TNF- α in 6-month-old pigs. *Res. Vet. Sci.* 95, 1262–7. doi:10.1016/j.rvsc.2013.09.011

Wensvoort, G., Terpstra, C., Boonstra, J., Bloemraad, M., Van Zaane, D., 1986. Production of monoclonal antibodies against swine fever virus and their use in laboratory diagnosis. *Vet Microbiol.* 12, 101–8. doi: 10.1016/0378-1135(86)90072-6

Woźniakowski, G., Samorek-Salamonowicz, E., 2014. Molecular evolution of Marek's disease virus (MDV) field strains in a 40-year time period. *Avian Dis.* 58, 550–7. doi:10.1637/10812-030614-Reg.1

Wu, H.X., Wang, J.F., Zhang, C.Y., Fu, L.Z., Pan, Z.S., Wang, N., Zhang, P.W., Zhao, W.G., 2001. Attenuated lapinized chinese strain of classical swine fever virus: complete nucleotide sequence and character of 3'-noncoding region. *Virus Genes* 23, 69–76.

Xiao, M., Gao, J., Wang, Y., Wang, X., Lu, W., Zhen, Y., Chen, J., Li, B., 2004. Influence of a 12-nt insertion present in the 3' untranslated region of classical swine fever virus HCLV strain genome on RNA synthesis. *Virus Res.* 102, 191–8. doi:10.1016/j.virusres.2004.01.029

Figure Legends

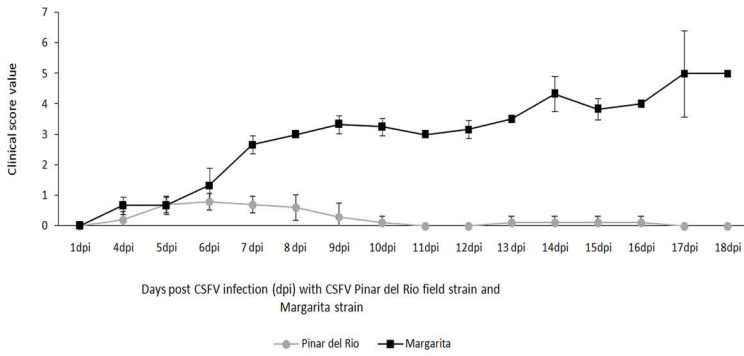
Figure 1. Assessment of the virulence of the CSFV strains Pinar del Rio and Margarita in weaner pigs. The clinical scores (A), neutralising antibody titres (B), and viral RNA load in nasal swabs (C), rectal swabs (D) and serum (E) were monitored at the indicated days post infection (dpi) for the pigs infected with CSFV Pinar del Rio (grey lines and symbols, pigs # 1 to 5) or Margarita (black lines and symbols, pigs # 6 to 8). For the clinical scores, the mean values are shown with error bars representing the standard deviation. At necropsy, the viral RNA load was determined in the tonsils, spleen, lymph node and ileum (F). Viral RNA was quantified by qRT-PCR (B to F) and represented with the Ct values. Ct values equal to (dotted line) or smaller than 35 were considered positive. *Pig number 6 died at 15 dpi and it was not possible to collect tissue samples from this animal.

Figure 2. Comparison of the 3'UTR sequences of the CSFV Pinar del Rio isolate, Margarita strain, the Labiofam vaccine strain and other selected strains. The 3'UTRs were aligned from nucleotide position 12074 of Pinar del Rio to the 3' end. The Genbank accession numbers are indicated with the virus nomenclature, and the poly-uridine tracts are boxed.

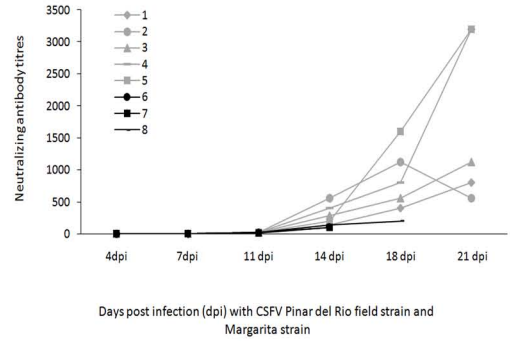
Figure 3. Secondary structure prediction of the 3'-UTR of the CSFV Pinar del Rio and Margarita strains. The basic profile contains SL-I, SL-II, SL-III, SL-IV and region SS. (A) Secondary structure prediction of the 3'-UTR of CSFV Margarita strain 226 bp, free energy = -51.29. (B) Secondary structure prediction of the 3'-UTR of CSFV Pinar del Rio field strain 257 bp, free energy = -53.59.

Figure 4. Amino acid sequence alignment of the glycoprotein E2 region. The last 23 residues of E1, the complete E2, and 47 first residues of p7 of CSFV Margarita (AJ704817) and Pinar del Rio (KX576461) are shown. The amino acid differences between the two strains are highlighted in boldface. Arrows indicate the first amino acid of E2 and p7.

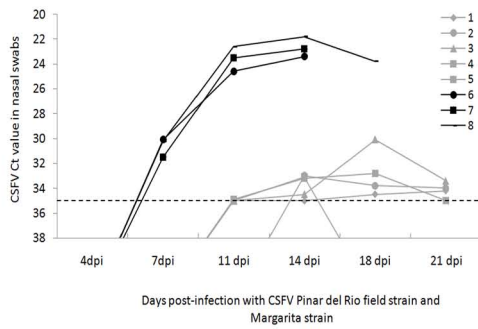
A



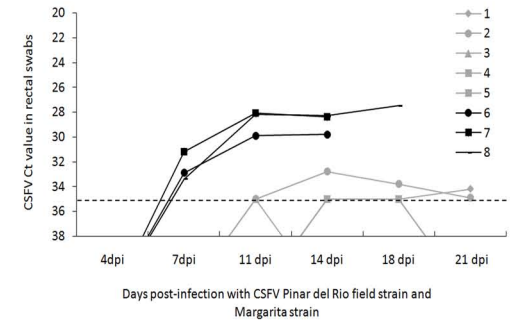
B



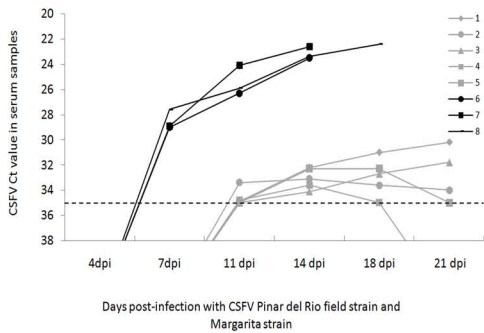
C



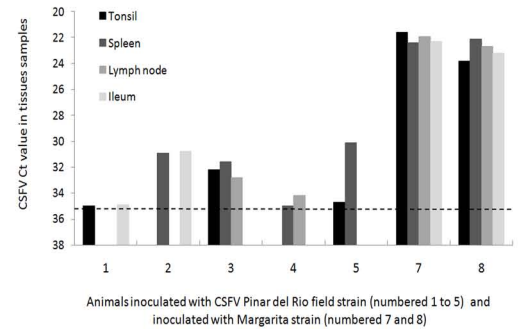
D



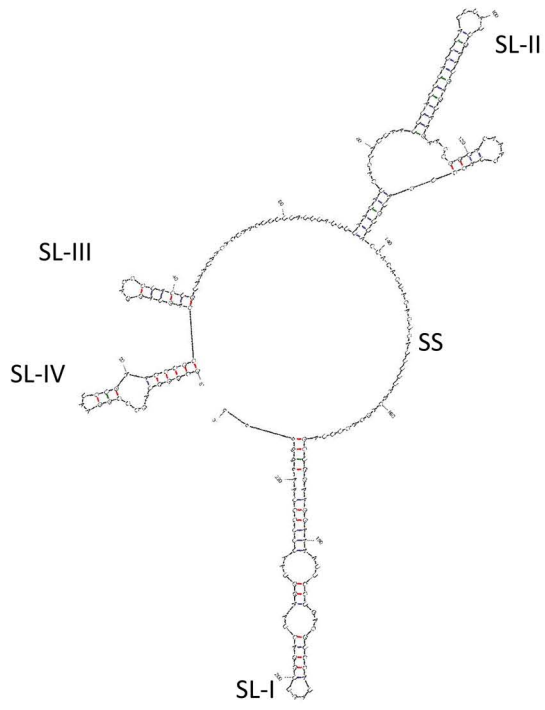
E



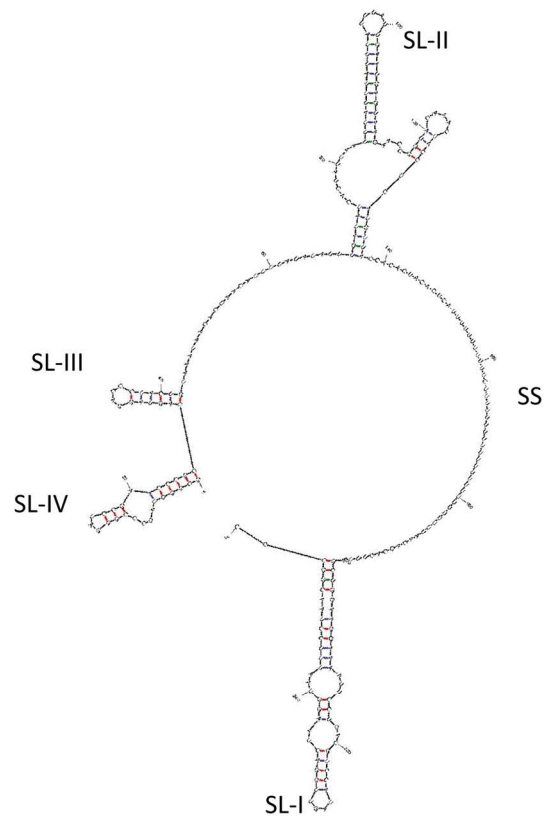
F



A



B



	E1	↓	E2
Margarita PdR	IKVLRGQ V VQGVIVLLLVLTGAQGR LACKEDFRYAISSTNEIGLLGAELTTTWKDYDHNL		IKVLRGQ I VQGVIVLLLVLTGAQGR LACKEDFRYAISSTNEIGLLGAELTTTWKDYDHNL *****:*****
Margarita PdR	QLDDGTIKAICTAGSFKVIALNVVSRRYLASLHK G ALP TSVTFELFDGTSPS I EMGDD		QLDDGTIKAICTAGSFKVIALNVVSRRYLASLHK R ASP TSVTFELFDGTSPS V EMGDD ***** * *****:*****
Margarita PdR	FGFGLCPFDTS PVVKGRYNTLLNGSAFYLVCPIGWTGVIECTAVSPTTLRTEVVKTFRR		FGFGLCPFDTS PVVKGRYNTLLNGSAFYLVCPIGWTGVIECTAVSPTTLRTEVVKTFRR *****
Margarita PdR	EKPPF H RKDCVTTTVENEDLFYCR LGGNWT CVKGEPIYTGGLVKQCRWCGF D FNEPDGL		EKPPF V RKDCVTTTVENEDLFYCR LGGNWT CVKGEPIYTGGLVKQCRWCGF V FNEPDGL *****:*****
Margarita PdR	PHYPIGKCILANETGYRIVDSTDCNRNGVVI STEGSHECLIGNTSVK V HALDERLGPMP		PHYPIGKCILANETGYRIVDSTDCNRNGVVI STEGSHECLIGNTSVK V HASDERLGPMP ***** *
Margarita PdR	RPKEIVSSEGPVRKTSCTFN YTKTLRNKYEP RDSYFQQYMLKGEYQYWFDDLVDTHHSD		RPKEIVSSEGPVRKTSCTFN YTKTLRNKYEP RDSYFQQYMLKGEYQYWFDDLVDTHHSD *****
Margarita PdR	YF T EFLVLVVVALLGGRYVLWLVITYVVLTEQLAAGLQLGQGEVVLIGNLIHTDIEVVV	↓	YF A EFLVLVVVALLGGRYVLWLVITYVVLTEQLAAGLQLGQGEVVLIGNLIHTDIEVVV **:* *****
Margarita PdR	YFLLLYLIMR D DP IKKWILLLFH		YFLLLYLIMR E DP IKKWILLLFH *****:*****

SCIENTIFIC REPORTS

OPEN

Presence of atypical porcine pestivirus (APPV) genomes in newborn piglets correlates with congenital tremor

Received: 01 February 2016

Accepted: 24 May 2016

Published: 13 June 2016

Alexander Postel^{1,2,*}, Florian Hansmann^{3,4,*}, Christine Baechlein², Nicole Fischer⁵, Malik Alawi⁶, Adam Grundhoff⁶, Sarah Derking⁷, Jörg Tenhündfeld⁷, Vanessa Maria Pfankuche^{3,4}, Vanessa Herder^{3,4}, Wolfgang Baumgärtner^{3,4}, Michael Wendt⁸ & Paul Becher^{1,2}

Pestiviruses are highly variable RNA viruses belonging to the continuously growing family *Flaviviridae*. A genetically very distinct pestivirus was recently discovered in the USA, designated atypical porcine pestivirus (APPV). Here, a screening of 369 sera from apparently healthy adult pigs demonstrated the existence of APPV in Germany with an estimated individual prevalence of 2.4% and ~10% at farm level. Additionally, APPV genomes were detected in newborn piglets affected by congenital tremor (CT), but genomes were absent in unaffected piglets. High loads of genomes were identified in glandular epithelial cells, follicular centers of lymphoid organs, the inner granular cell layer of the cerebellum, as well as in the trigeminal and spinal ganglia. Retrospective analysis of cerebellum samples from 2007 demonstrated that APPV can be found in piglets with CT of unsolved aetiology. Determination of the first European APPV complete polyprotein coding sequence revealed 88.2% nucleotide identity to the APPV sequence from the USA. APPV sequences derived from different regions in Germany demonstrated to be highly variable. Taken together, the results of this study strongly suggest that the presence of APPV genomes in newborn piglets correlates with CT, while no association with clinical disease could be observed in viremic adult pigs.

Pestiviruses are enveloped highly variable RNA viruses with a genome of about 12.3 kb belonging to the family *Flaviviridae*¹. Besides the classical pestiviruses bovine viral diarrhoea virus-1 (BVDV-1), BVDV-2, border disease virus (BDV) and classical swine fever virus (CSFV), a growing number of additional tentative pestivirus species was discovered in the last few years^{2–4}. Until recent discovery of very distantly related pestivirus sequences in bats and rats, it was commonly accepted that pestivirus infections are limited to ungulate hosts^{2,3}. Infections with the classical pestivirus species are worldwide of utmost socioeconomic relevance. In consequence, many countries have implemented compulsory eradication programs for bovine viral diarrhoea (BVD) and classical swine fever (CSF), the latter is also reportable to the World Organization for Animal Health (OIE). Besides dramatic clinical signs and high mortality rates particularly observed in acute infections with highly virulent CSFV strains, fetopathogenicity is a common feature of intrauterine pestivirus infections and has significant consequences for productivity of animal breeding⁵.

¹University of Veterinary Medicine, Department of Infectious Diseases, Institute of Virology, EU and OIE Reference Laboratory for Classical Swine Fever, Hannover, 30559, Germany. ²University of Veterinary Medicine, Department of Infectious Diseases, Institute of Virology, Hannover, 30559, Germany. ³University of Veterinary Medicine, Department of Pathology, Hannover, 30559, Germany. ⁴University of Veterinary Medicine, Center for Systems Neuroscience, Hannover, 30559, Germany. ⁵University Medical Center Hamburg- Eppendorf, Institute for Medical Microbiology, Virology and Hygiene, Hamburg, 20246, Germany. ⁶Leibniz Institute for Experimental Virology, Heinrich Pette Institute, Hamburg, 20251, Germany. ⁷Veterinary practice Vetland[®] Dr. Tenhündfeld & Kollegen, Vreden, 48691, Germany. ⁸University of Veterinary Medicine, Clinic for Swine, Small Ruminants, Forensic Medicine and Ambulatory Service, Hannover, 30173, Germany. *These authors contributed equally to this work. Correspondence and requests for materials should be addressed to P.B. (email: Paul.Becher@tiho-hannover.de)

Piglet [ID]	Congenital tremor	APPV genome detection in qRT-PCR [Cq values]			
		Serum	CSF ^a	CNS pool ^b	Cerebellum
51	–	–	–	–	–
52	+	29.7	28.1	27.0	25.8
53	–	–	–	–	–
54	+	29.3	27.0	26.6	27.5
55	+	29.9	27.5	30.6	21.9
56	+	27.4	27.2	28.0	24.8
57	+	27.3	28.4	29.8	24.9
59	+	27.4	26.0	22.3	21.0

Table 1. Association of congenital tremor with atypical porcine pestivirus (APPV) genome detection.
^aCerebrospinal fluid. ^bCentral nervous system comprising cerebrum, cerebellum, spinal cord.

Recently a novel, genetically very distinct pestivirus, tentatively designated “atypical porcine pestivirus” (APPV), was discovered in pigs from the USA by high throughput sequencing⁶. So far, the clinical relevance of APPV infections remained elusive. In this study, we report the detection of APPV genomes in serum of apparently healthy pigs from two different Federal states in Germany. In addition, APPV genomes were identified in different tissues including cerebellum and peripheral nerves of piglets with congenital tremor (CT, *Myoclonia congenita*) but not in healthy piglets from the same herd providing evidence for a so far unknown association with CT in newborn piglets.

Results

Identification of APPV in healthy adult pigs and newborn piglets with CT. Two different SYBR-Green based APPV RT-PCRs targeting the NS3 and NS4B encoding regions of APPV were developed based on the only available APPV sequence from the USA and taking into account other atypical pestivirus sequences from rat and bat. A total number of 369 serum samples from clinically unsuspecting sows and fattening pigs originating from South Germany (20 farms, 200 sera) and North Germany (9 farms, 169 sera) were screened with these RT-PCRs. Both assays identified APPV genomes in three sows (two herds) from Bavaria and six finishing pigs (one herd) from Lower Saxony resulting in an individual genome prevalence of 2.4% and a prevalence of ~10% at farm level.

Several farms located in the western part of Germany experienced cases of CT becoming evident by unintended shivering of newborn piglets, but no clinical signs in the sows or other adult pigs were observed. The etiology of this clinical syndrome is elusive, but an infectious etiology was suspected due to the regional accumulation of cases and epidemiological links. To exclude CSFV infection, eight piglets (six affected and two unaffected animals) from three affected litters were sacrificed for pathological and virological investigations. As no viral genomes of CSFV and other established pestivirus species were detectable by generic Pan-Pestivirus RT-PCR, the two SYBR-Green based APPV-specific PCRs were applied. All of the six clinically affected piglets showed APPV genomes in serum, cerebrospinal fluid and pooled central nervous system samples (comprising cerebrum, cerebellum, and spinal cord), while both piglets without tremor were tested negative (Table 1). Sera (n = 23) obtained from sows of the affected farm with and without affected litters were negative for APPV genomes.

Epidemiology. Outbreaks of CT were observed in several farms located in Western Germany (North Rhine-Westphalia) in 2015. With the exception of one litter (out of 50 affected litters in total) only litters of newly introduced gilts from the same multiplier herd were affected. Clinical cases of CT appeared in the investigated sow herd (720 “DanBred” hybrid sows) since August 2015 after changing the source of breeding stock by placing of new gilts. On average 30–40% piglets of the affected litters showed typical symptoms of CT. Most of them recovered within two to three weeks, total piglet mortality in the herds remained unchanged.

Post mortem examination. Necropsies were performed on six diseased and two clinically unaffected piglets (two days of age). Gross lesions were restricted to a local facial dermatitis in seven out of eight piglets. Histologically, a mild suppurative omphalitis was detected in four out of eight piglets while central and peripheral nervous system as well as skeletal muscle were without significant findings. Luxol fast blue staining revealed a mild reduced staining intensity accentuated in the lateral white matter of the spinal cord in four out of six diseased animals while the two age-matched, clinically unaffected animals showed a regular myelination (Fig. 1).

Tissue tropism of APPV. 26 organ samples from one affected piglet (no. 59) were analyzed by quantitative RT-PCR (qRT-PCR) with a specific TaqMan probe to determine the tissue tropism of APPV (Fig. 2). Highest genome loads were found in glands of the *Arcus palatoglossus* and the *Lymphonodus mandibularis* (quantification cycles, Cq values: 24.3 and 24.4). Other tissues that typically contain high genome loads in case of CSFV infection, like kidney and spleen, gave a much weaker signal corresponding to approximately 1000 fold less APPV genome equivalents. Cerebellum, trigeminal and spinal ganglia revealed to contain high amounts of APPV genomes, but also peripheral nerves were tested positive. In addition, all cerebellar samples obtained from diseased piglets contained high loads of APPV genomes, while both clinically unaffected piglets were APPV negative (Table 1). Fluorescent *in-situ* hybridization (FISH) substantiated the results of the TaqMan qRT-PCR with exception of a positive PCR result obtained for the thymus that could not be confirmed by *in situ* hybridisation (Fig. 2). Three

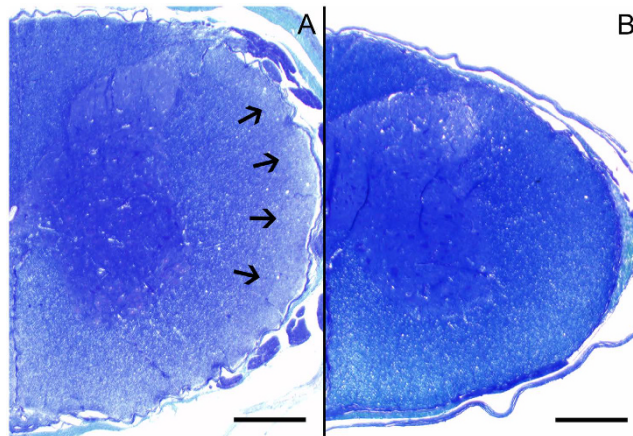


Figure 1. Histochemical visualization of myelin using Luxol fast blue cresyl fast violet staining in the spinal cord of two days old piglets. Congenital tremor affected piglet (no. 52) showed a mildly reduced myelin staining intensity accentuated in the lateral white matter (A, indicated by arrows) compared to an unaffected piglet (piglet no. 53) with regular myelination (B). Scale bars = 500 μm.

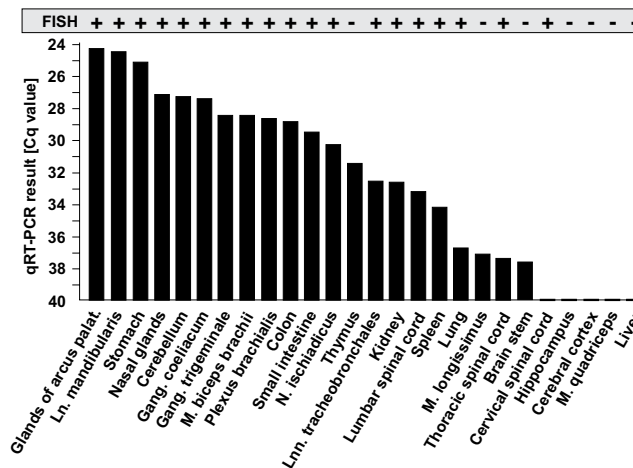


Figure 2. Tissue tropism of atypical porcine pestivirus (APPV). Detection of APPV genomes in different tissues of a two day old piglet with congenital tremor by fluorescent *in-situ* hybridization (FISH) and quantitative reverse transcription polymerase chain reaction (qRT-PCR). Organ distribution of APPV genome as detected by FISH is shown as present (+) or absent (-). Furthermore, for each organ the respective quantification cycle (Cq) values are given. Ln = Lymphonodus, Gang = Ganglion, M = Musculus, N = Nervus, Lnn = Lymphonodi.

tissues containing very low genome loads (*M. longissimus*, brain stem, cervical spinal cord) gave inconsistent results probably due to the detection limits of both assays and uneven distribution of APPV positive cells in the tissues. In the central nervous system, virus genome was located in the inner granular cell layer of the cerebellum (Fig. 3A) as well as in spinal (Fig. 3B) and trigeminal ganglia. Furthermore, glandular epithelial cells in the *Arcus palatoglossus* (Fig. 3C) and Brunner’s glands (duodenum) as well as lymphoid organs showed a strong positive signal which was most prominent in the follicular centres (Fig. 3D).

Evidence for a common association of APPV genomes in cerebellum samples and CT.

Comparable to the recent outbreak, several sow farms were affected by CT in 2007, after placing of new gilts from the same multiplier herd (“JSR Hybrid” sows). For retrospective analyses, formalin-fixed and paraffin-embedded cerebellum samples archived from this earlier outbreak were analyzed by real-time PCR and *in situ* hybridization. Two out of eleven cerebellar samples showed a positive result in both APPV PCRs (NS3 and NS4B genomic regions). FISH of the cerebellum revealed a strong APPV specific signal mainly located in the inner granular layer of the cerebellum in both diseased piglets (data not shown).

Molecular characterization of APPV genomes. One serum (S5/9), obtained from a sow in Bavaria that revealed to contain highest APPV genome loads, was used to determine the first complete polyprotein coding

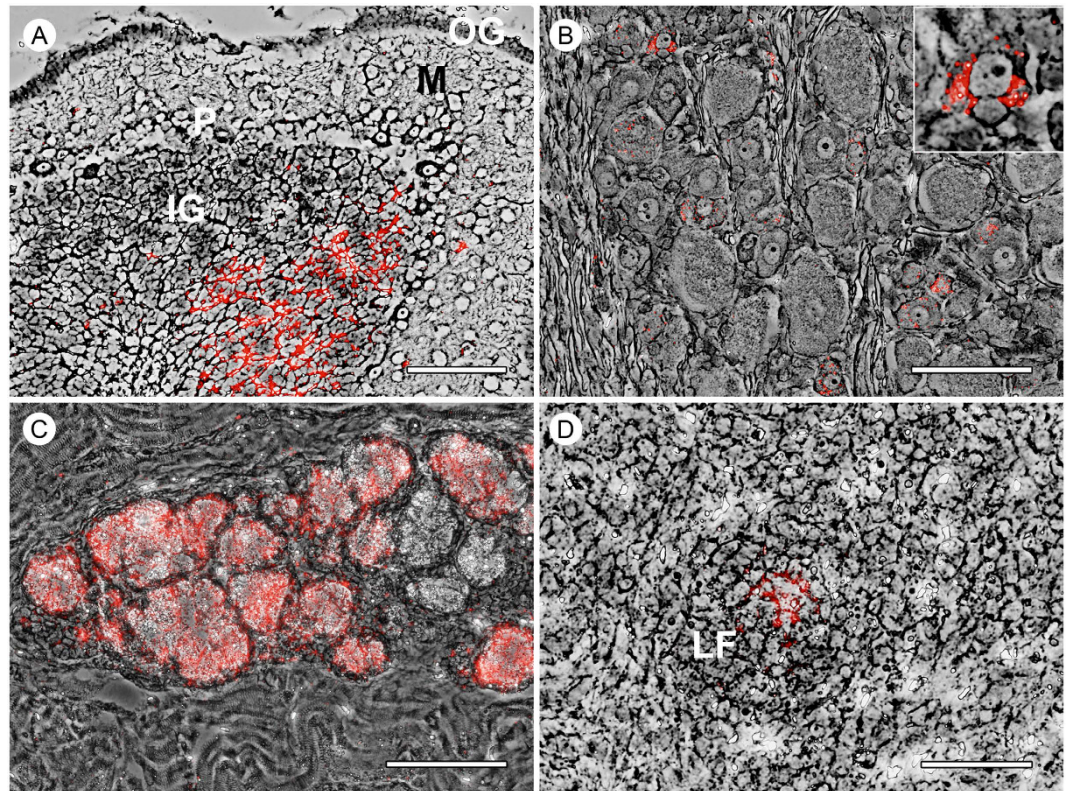


Figure 3. Fluorescent *in-situ* hybridization showing the organ tropism of atypical porcine pestivirus (APPV) in a two day old piglet with congenital tremor. APPV genome (red) was detected in the inner granular cell layer of the cerebellum (A), cytoplasm of spinal ganglia neurons (B), glandular epithelial cells of the *Arcus palatoglossus* (C) and in follicular centers of the mandibular lymph node (D). Insert in B shows a higher magnification of an APPV positive neuron. OG = outer granular cell layer, P = Purkinje cell layer, IG = inner granular cell layer of cerebellum; M = molecular layer; LF = lymphoid follicle; scale bars in A–C = 100 μm and D = 50 μm .

sequence of European APPV by NGS. The ORF of S5/9 showed 88.2% nucleotide identity to the only so far known APPV sequence from the USA (5) encoding for a polyprotein of the same length (3,635 amino acids). The complete polyprotein sequence of this European APPV revealed an identity of 94.3% to the polyprotein encoded by the American APPV, but only 37.3% identity to the polyprotein of CSFV strain Alfort-Tuebingen⁷. Amino acid variability between the APPV polyprotein sequences from Germany and the USA showed the same range and a very similar pattern of variability as observed among the three established CSFV genotypes (Fig. 4). Similar to the three genotypes of CSFV, the two APPV sequences showed a high degree of conservation in the amino acid composition of nonstructural proteins NS3 and NS5B, but also a significant conservation in NS3 and NS5B when compared to other pestiviruses (Fig. 4). Phylogenetic analyses as well as amino acid scan revealed only a very distant relatedness of APPV to other pestiviruses including CSFV, the atypical porcine pestivirus Bungowannah, a broad spectrum of ruminant pestiviruses, and the recently described pestivirus from Norway rat (Figs 4 and 5). Remarkably, the partial sequence of a pestivirus obtained from a bat (*Rhinolophus affinis*) has an intermediate position between APPV and other pestiviruses, nevertheless showing the typical pattern of conserved and variable regions (Fig. 4). The distances of APPV to the bat pestivirus in the NS2-3 region were 33% on nucleotide level and 26% in the amino acid composition. Analyses of the three partial NS2-3 encoding APPV sequences obtained from Germany displayed similar genetic distances (8.9–10.2%) among each other and to the APPV sequence from the USA (Fig. 5), while the pairwise distances of the respective deduced amino acid sequences were below 2.1% (1.1–2.1%), indicating highly conserved protein functions. In contrast, differences between 51% and 52% were observed on amino acid level to the different CSFV sequences representing the three established CSFV genotypes (Figs 4 and 5).

Discussion

Pestiviruses are highly variable RNA viruses causing economically relevant diseases in swine, cattle, sheep, and goats. In the last two decades a growing number of novel pestiviruses has been discovered in various domestic and wild ruminant species as well as pigs⁴. The recent identification of highly distinct pestivirus sequences from bats and rats was of particular interest as these discoveries provided the first evidence of pestivirus infections in non-*Artiodactyla* hosts^{2,3}. An association of these atypical pestiviruses with disease is not known so far. In addition to these findings in wild animals, genomes of another highly distinct pestivirus, tentatively designated “atypical porcine pestivirus” (APPV), were identified in apparently healthy domestic pigs in the USA⁶. The discovery

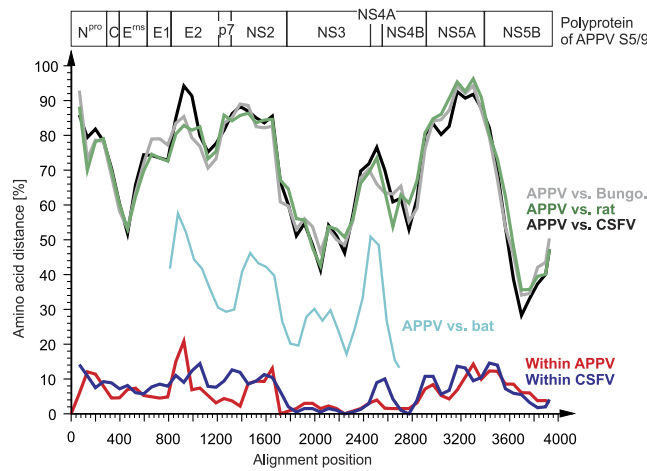


Figure 4. Variability in the polyprotein sequences of atypical porcine pestivirus (APPV) and other pestiviruses. Shown is an amino acid scan performed with the p distance algorithm of the SSE software platform applying a sliding window of 400 and an increment of 200 residues³⁴. The novel APPV polyprotein sequence from Germany (S5/9) and the APPV sequence from the USA (GenBank KR011347) were compared to the complete polyprotein sequences of porcine pestivirus Bungowannah (GenBank NC023176), three different CSFV strains, representing genotype 1 (Alfort/187, GenBank X87939), genotype 2 (Alfort-Tuebingen, GenBank J04358) and genotype 3 (94.4/TWN, GenBank AY646427) as well as the complete polyprotein sequence of a pestivirus from a rat (GenBank KJ950914) and a partial pestivirus polyprotein sequence obtained from a bat (GenBank JQ814854). The organization of the APPV S5/9 polyprotein is indicated above the amino acid scan.

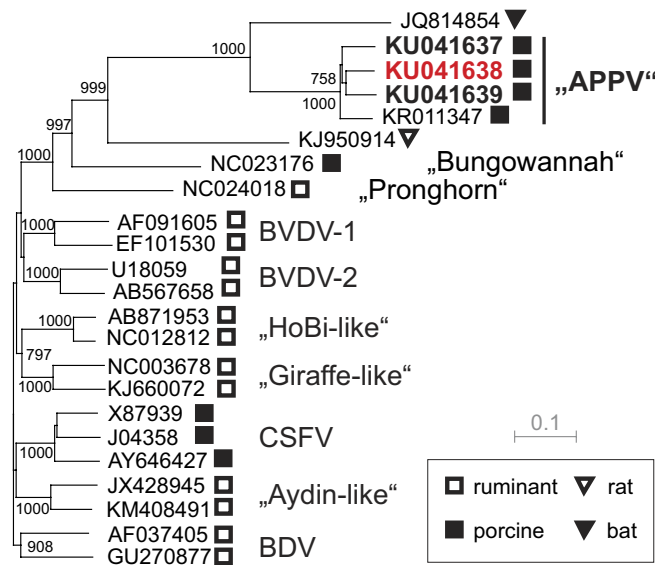


Figure 5. Phylogenetic tree displaying relatedness of atypical porcine pestiviruses (APPV) from three different regions in Germany. Genetic distances were calculated by the Kimura 2-parameter substitution model and phylogenetic analyses by applying the neighbor-joining method as described for CSFV phylogeny³³. Phylogenetic analysis of partial NS2-3 encoding APPV sequences (1581 nt) from three different regions in Germany (bold) together with respective sequences of established and tentative pestivirus species were analyzed. The GenBank accession number and the corresponding host species (symbols) are indicated for each individual sequence. APPV sequences from Lower Saxony [KU041637] and from Bavaria [KU041639] were obtained from apparently healthy adult pigs. APPV sequence from North Rhine Westphalia [KU041638] was obtained from a congenital tremor diseased piglet (highlighted in red). Bootstrap values were calculated for 1000 iterations. Only significant bootstrap values (≥ 700) of major nodes are given in the tree. Trees were displayed by dendroscope³⁸.

and further characterization of this novel APPV is of particular interest, as other porcine pestiviruses like CSFV and the Australian pestivirus Bungowannah are causative agents of severe diseases in pigs^{5,8}. An association of APPV with disease remained elusive for many decades. The results of the present study show the association of APPV genomes in the cerebellum and other nervous tissues with the occurrence of CT in newborn piglets. At the

time this manuscript was in review, another research group demonstrated an association of APPV and CT in US swine herds by performing an animal experiment with clinical sample material⁹.

CT can be differentiated by the presence (types AI–AV) or absence (type B) of morphological lesions in the brain and spinal cord. Etiologically, CT is associated with infection (AI = CSFV, AII = unknown infectious agent), genetic background (AIII = Landrace; sex-linked recessive), AIV = Saddleback (autosomal recessive; Landrace/Saddleback) and intoxication (AV = metrifonate, trichlorfon)¹⁰. So far, the etiology of CT type AII remained elusive, although this disease of newborn piglets is known since many decades^{11,12}. An association of porcine circovirus type 2 (PCV2) with CT type AII was proposed, but several contrary studies could not support this finding^{13,14}. A recent study reported the presence of astrovirus genomes in the brain of piglets suffering from CT, but RNA loads in the brain appeared to be rather low as a nested-PCR approach was necessary to amplify the viral genomes¹⁵. Aside from CT caused by CSFV (type AI), a similar clinical presentation is known to occur also in other host species after infection with ruminant pestiviruses like BVDV and BDV (“hairy shaker lambs”). Clinical signs of type AII CT had been experimentally induced in newborn piglets by intramuscular or intrauterine inoculation of pregnant sows with brain suspensions from CT-affected piglets^{16,17}. In a very recent study, transferability of APPV with serum from a healthy but genome-positive (viremic) sow was demonstrated and inoculation into the fetal amniotic vesicles of a gravid sow resulted in the birth of CT-affected piglets⁹. Experimental infections of pregnant sows with a virus isolate will be required to finally prove that APPV solely can cause CT in newborn piglets and to understand the mechanisms of pathogenesis. However, it was not possible to isolate and propagate the virus so far⁶. Attempts of virus isolation and virus recovery by RNA transfection using various porcine cell lines were not successful and hampered by limited amounts of sample material obtained from the piglets investigated in this study. Therefore, an animal experiment with a virus isolate could not be performed so far. Nevertheless, the detection of high APPV genome loads in the cerebellum, different ganglia and other tissues by qRT-PCR and FISH, including retrospective analyses, together with the recently published data by Arruda *et al.* (2016) strongly suggest that APPV represents a previously unrecognized virus which is associated with the occurrence of CT in piglets.

Intrauterine infections with classical pestiviruses like CSFV, BDV or BVDV are known to induce hypomyelination and cerebellar hypoplasia^{18–20}. Cerebellar hypoplasia is due to a primary infection of the outer granular layer of the cerebellum with consecutive death of the affected neurons²¹. The lack of the outer granular cell layer results in cerebellar hypoplasia with persistent neurological signs since no neurons from the outer granular cell layer can migrate into the inner granular cell layer²¹. Remarkably, in the presented study APPV genomes are most abundantly detected in the cerebellar inner granular cell layer, but not in the outer granular cell layer like observed in infections with classical pestiviruses. This finding may explain the transient clinical signs (recovery of the affected piglets) observed in the present cases since a loss of inner granular cells may be compensated by an immigration of cells from the outer granular cell layer during the first weeks *post natum*. Additionally, the infection of spinal ganglia may contribute to the observed clinical signs in infected piglets. However, since no inflammation was detected within the central and peripheral nervous system of piglets with CT the mechanism of virus clearance remains elusive so far. Distribution of virus within herds may occur via the orofecal route since a significant amount of APPV genome was present in salivary glands, duodenum, pancreas and colon. Future studies will address the transmission routes of this novel virus.

Postmortem investigation of CT type II diseased piglets revealed varying degrees of hypomyelination of brain and spinal cord²². In the present study, a mild reduction of myelin in the spinal cord was observed in four out of six affected piglets. As two APPV positive piglets showed no convincing hypomyelination despite shivering, this morphologic finding probably results from a transiently delayed myelination. A transient course of infection and subsequent completion of myelogenesis is likely to occur in the piglets as the majority of affected piglets recover after some time (2–3 weeks). In addition, the lack of APPV genomes observed in sows with affected litter gives strong evidence for a transient and obviously clinically inapparent infection during gestation. This is in line with a previous study reporting that an infection of adult sows with clinical sample material did not result in a clinical manifestation except for the production of CT-affected piglets showing varying degrees of hypomyelination of brain and spinal cord²². Screening for APPV genomes in a larger number of gilts from the affected farms will help to identify acute infections during gestation and will provide information with respect to the clinical signs in sows under field conditions. In addition, serial bleeding of affected piglets will be performed in the future to answer the question whether abrogation of clinical signs coincidence with clearance of viremia or whether recovered piglets remain persistently infected. First results show that some of the CT affected piglets still have high genome loads (C_q values 21–24) in the serum at the age of 31 days, albeit showing eased clinical signs. Based on the knowledge of other pestiviruses it can be speculated that chronically or persistently infected, but clinically healthy pigs shedding the virus are the source of infection for serologically naïve sows, which are newly introduced in a herd with these viremic animals and subsequently experience a transient infection with significant consequences for the unborn piglets in case of pregnancy. This hypothesis – which is supported by the current knowledge in CT type AII²³ – is further strengthened by the epidemiological investigation of the APPV associated outbreak of CT described here. A sero-negative APPV status of the farm delivering the gilts to the CT-affected farms (with putatively high seroprevalence of APPV specific antibodies) would strongly support this hypothesis. A novel serological assay for APPV is in development to address the epidemiology and the seroprevalence of APPV.

Despite all differences of APPV to classical pestiviruses, the APPV genomes from Europe and the USA encode for proteins being unique for members of the genus *Pestivirus*, namely the N-terminal protease N^{pro} and the secreted glycoprotein E^{gns} containing also the conserved motif required for RNase activity. The different APPV sequences characterized in this study revealed significant differences to the reference sequence from the USA, but also between the APPVs originating from three different regions in Germany. This finding points towards geographically isolated virus populations in Germany, which must have evolved over a longer period of time and are epidemiologically not linked with each other. In addition, the observed genetic variability and the distribution

Primer/Probe	Sequence (5'-3')	Target	Purpose
APPV_5587-fw	CAGAGRAAAGGKCGAGTGGG	NS3	PCR 1, qRT-PCR
APPV_5703-rev	ACCATAYTCTTGGGCCTGSAG		PCR 1, qRT-PCR, sequencing PCR A
APPV_CT-59 probe	[6FAM] ACTACTATCCTTCGGGGGTAGTACCGA [BHQ1]		qRT-PCR
APPV_6869-fw	CTTTCATGGARTCWGGCGGTG	NS4B	PCR 2
APPV_6950-rev	AGACTCCTRTTCTGCATGTT		PCR 2
APPV_5087-fw	GAAAGTGTCTGCCGCTTCATG	NS3	sequencing PCR A
APPV_4186-fw	GTGCGGCCTCCCAACTGTAG	NS2	sequencing PCR B
APPV_4273-fw	TGGGGACCTCACCAGTGATG	NS2	sequencing PCR C
APPV_5169-rev	ACGTCACCTCTTCCGCTC	NS3	sequencing PCR B/C

Table 2. Atypical porcine pestivirus (APPV)-specific primers and probes used in the study.

of variable and conserved regions in the polyprotein of the characterized APPV sequences are comparable to the variability among different CSFV genotypes (Fig. 4). Considering the high variability of the characterized APPV genomes and their estimated prevalence in domestic pigs, it appears likely that infections with these newly discovered pestiviruses frequently occur, but without knowledge of its association with CT in newborn piglets remained clinically unrecognized in the past.

In the presented study, we identified APPV genomes in the cerebellum and ganglia of new-born piglets suffering from CT type AII. This strongly suggests that APPV infection contributes to the induction of CT in piglets. In addition, APPV sequences can be detected in clinically healthy adult pigs. Future studies will address the biology of this atypical pestivirus and help to reduce losses in pig production by tailored herd management and prevention strategies.

Materials and Methods

Sample material. 369 serum samples from clinically unsuspecting sows and fattening pigs originating from Bavaria (20 farms, 200 sera), Lower Saxony and Schleswig-Holstein (8 farms, 158 sera) were obtained in the framework of veterinary microbiological diagnostics or the Salmonella monitoring program in accordance to German legislation (SchwSalmoV §2) and residual volumes of these samples were provided for use in the present study. Therefore no ethical approval was required for the use of these samples. Eleven samples were residual volumes obtained from a previous experimental study conducted in Schleswig-Holstein, which was notified and approved by the local authorities (Ministry of Energy, Agriculture, the Environment and Rural Areas, Schleswig-Holstein: reference number V244-7224.121.9-34). Samples and piglets from the CT-affected farms were sent to the University of Veterinary Medicine, Hannover, for diagnostic reasons, especially for the exclusion of CSF, which is in accordance to German legislation (SchHaltHygV §8) and does not require further approval by the authorities. None of the animals included in this study was infected experimentally.

Post mortem examination and sample collection. Serum, cerebrospinal fluid and tissue samples from different organs were taken from CT-affected (n = 6) and unaffected (n = 2) piglets of different litters at the age of two days. For histology, all tissues were routinely processed in paraffin wax, cut at 2 µm thickness, and stained with hematoxylin and eosin. In addition, cross-sections of spinal cord and cerebellum were stained with Luxol fast blue for the investigation of myelination as previously described^{24,25}.

RNA isolation. RNA from liquids (serum, cerebrospinal fluid) was prepared with the ViralAmp Kit (Qiagen, Hilden) and RNA from formalin-fixed archived cerebellum samples was extracted with the RNeasy FFPE kit according to the recommendations of the manufacturer (Qiagen, Hilden). RNA from tissues was isolated by phenol-chloroform precipitation or with the Nucleospin RNA kit (Macherey-Nagel, Düren).

Reverse transcription PCR (RT-PCR). The presence of CSFV and genomes of other established pestiviruses was excluded by applying the accredited methods of the EU and OIE Reference Laboratory for CSF, Hannover, using primers described previously^{26,27}. Based on the only available APPV sequence and the sequences of atypical pestiviruses from bat (*Rhinolophus affinis*) and rat (*Rattus norvegicus*), primers targeting conserved regions in the NS3 (PCR 1) and the NS4B (PCR 2) encoding regions were designed for APPV screening (Table 2). These APPV screening PCRs were performed using the QuantiTect SYBR-Green kit (Qiagen, Hilden) with subsequent visualization of the PCR products by agarose gel electrophoresis. Quantitative TaqMan-PCR was performed based on the established real-time PCR protocol of the EURL with the QuantiTect Probe RT-PCR kit (Qiagen, Hilden) containing 1.5 pmol of each primer used in APPV-specific screening PCR 1 and 0.25 pmol probe (specific for the APPV sequence obtained from CT-affected piglet no. 59) per 20 µl mastermix. For all PCR reactions five microliters of RNA were added to the mastermix and amplification was performed in a one-step PCR reaction with 50 °C, 30 min; 95 °C, 15 min and 40 cycles comprising 95 °C, 30 sec; 56 °C, 30 sec; 72 °C, 30 sec.

APPV fluorescent *in-situ* hybridization (FISH). A broad spectrum of formalin-fixed, paraffin-embedded tissues from the piglet (no. 59) with the highest amount of APPV genome, cerebellum from two non-shivering piglets as well as cerebellum from two shivering piglets necropsied in 2007 were analyzed using FISH, as previously described²⁸, with a probe targeting a fragment of the NS3 encoding sequence of APPV (GenBank KU041638) and a probe specific for porcine ubiquitin as positive control with the following modifications:

Pretreatment of tissue sections included boiling (85–90 °C) in pretreatment solution (Affymetrix-Panomics, Santa Clara, CA) for 20 minutes, followed by protease QF (Affymetrix-Panomics) digestion for 10 minutes at 40 °C. Hybridization, pre-amplification, amplification, and detection were performed according to the manufacturers' instructions. Images were acquired with a color video camera (DP72, 12.8 megapixel CCD; Olympus, Hamburg, Germany) mounted on an IX50 microscope (Olympus) using the cellF Software (version 3.3; Olympus, Hamburg, Germany).

Determination of nucleotide sequences. The complete polyprotein encoding sequence of one APPV-positive porcine serum sample from Bavaria (sample S5/9) was determined by next generation sequencing on an Illumina HiSeq (2500 2 × 150 bp paired end run, sequencing depth: 6 Mio reads) as recently described^{29,30}. The complete genome was assembled using the IDBA-UD algorithm with a minimum coverage of 36.8³¹.

For amplification and determination of partial NS2-3 encoding sequences complementary DNA was transcribed using Superscript II reverse transcriptase and random hexamers. Amplification was performed in three different PCRs using the AccuPrime polymerase (LifeTechnologies, Darmstadt) and primers indicated in Table 2. The nucleotide sequences (1581 nucleotides) of the obtained PCR products were determined by conventional Sanger sequencing (LGC genomics, Berlin).

Sequence analyses. To genetically characterize the identified APPV, the partial NS2-3 encoding sequence (1581 nucleotides) obtained from a CT-affected piglet in 2015 (North Rhine Westphalia) was compared to APPV sequences obtained from sera of one clinically unsuspecting fattening pig from Lower Saxony and one unsuspecting sow originating from Bavaria. Multiple sequence alignments were generated with ClustalW of the Multiple Sequence Comparison by Log-Expectation (MUSCLE) tool provided by EMBL-EBI³². Genetic distances were calculated with the Kimura 2-parameter substitution model and phylogenetic analysis was performed by the Neighbour-joining method as previously reported for pestiviruses^{4,33}. Amino acid scan of polyprotein sequences was performed with the sequence distance calculation tool of the SSE software platform using the p distance method and a sliding sequence window of 400 residues with 200 residues increment³⁴.

Virus isolation. Attempts of virus isolation were performed with tissue homogenate (undiluted and a 1:10 dilution) of the diseased and highly APPV genome positive piglet #59 (*Arcus palatoglossus*; Cq-value 17) on cell lines susceptible for established pestiviruses (PK-15, SK6 and STE cells) according to the accredited protocol of the EU and OIE Reference Laboratory for CSF. In addition, the tissue homogenate as well as APPV genome positive pig sera from Bavaria and Lower Saxony were incubated on porcine lymphoma cells 38A1D and porcine endothelial cells PEDSV.15^{35,36}. RNA preparations of the tissue homogenate and of a viremic pig serum (S5/9, Bavaria) were used to transfect SK6 cells as described previously³⁷. At least three serial blind passages were performed followed by qRT-PCR screening of cells and supernatants for the presence of APPV genomes.

References

1. Simmonds, P. *et al.* Family Flaviviridae. in *Virus Taxonomy. Ninth Report of the International Committee on Taxonomy of Viruses* (San Diego, USA, 2012).
2. Firth, C. *et al.* Detection of zoonotic pathogens and characterization of novel viruses carried by commensal *Rattus norvegicus* in New York City. *MBio* **5**, e01933–01914 (2014).
3. Wu, Z. *et al.* Virome analysis for identification of novel mammalian viruses in bat species from Chinese provinces. *J Virol* **86**, 10999–11012 (2012).
4. Postel, A. *et al.* Close relationship of ruminant pestiviruses and classical Swine Fever virus. *Emerg Infect Dis* **21**, 668–672 (2015).
5. Moennig, V. & Becher, P. Pestivirus control programs: how far have we come and where are we going? *Anim Health Res Rev* **16**, 83–87 (2015).
6. Hause, B. *et al.* Discovery of a novel putative atypical porcine pestivirus in pigs in the United States. *J Gen Virol* **96**, 2994–2998 (2015).
7. Rumenapf, T., Meyers, G., Stark, R. & Thiel, H. J. Hog cholera virus—characterization of specific antiserum and identification of cDNA clones. *Virology* **171**, 18–27 (1989).
8. Kirkland, P. D., Read, A. J., Frost, M. J. & Finlaison, D. S. Bungovannah virus—a probable new species of pestivirus—what have we found in the last 10 years? *Anim Health Res Rev* **16**, 60–63 (2015).
9. Arruda, B. L. *et al.* Identification of a Divergent Lineage Porcine Pestivirus in Nursing Piglets with Congenital Tremors and Reproduction of Disease following Experimental Inoculation. *PLoS One* **11**, e0150104 (2016).
10. Done, J. T. & Harding, J. D. [Congenital tremor in pigs (trembling disease of piglets): lesions and causes]. *Dtsch Tierarztl Wochenschr* **74**, 333–336 (1967).
11. Kinsley, A. Dancing pigs? *Vet Med* **17**, 123 (1922).
12. Gustafson, D. & Kanitz, C. Experimental transmission of congenital tremor in swine. *Proc Ann Meet USAHA* **78**, 338–345 (1974).
13. Stevenson, G. W. *et al.* Tissue distribution and genetic typing of porcine circoviruses in pigs with naturally occurring congenital tremors. *J. Vet. Diagn. Invest.* **13**, 57–62 (2001).
14. Chae, C. A review of porcine circovirus 2-associated syndromes and diseases. *Vet J* **169**, 326–336 (2005).
15. Blomstrom, A. L., Ley, C. & Jacobson, M. Astrovirus as a possible cause of congenital tremor type AII in piglets? *Acta Vet Scand* **56**, 82 (2014).
16. Done, J. T., Woolley, J., Upcott, D. H. & Hebert, C. N. Porcine congenital tremor type AII: spinal cord morphometry. *Br Vet J* **142**, 145–150 (1986).
17. Vandekerckhove, P., Maenhout, D., Curvers, P., Hoorens, J. & Ducatelle, R. Type A2 congenital tremor in piglets. *Zentralbl Veterinarmed A* **36**, 763–771 (1989).
18. Emerson, J. L. & Delez, A. L. Cerebellar hypoplasia, hypomyelinogenesis, and congenital tremors of pigs, associated with prenatal hog cholera vaccination of sows. *J. Am. Vet. Med. Assoc.* **147**, 47–54 (1965).
19. Brown, T. T., DeLahunta, A., Bistner, S. I., Scott, F. W. & McEntee, K. Pathogenetic studies of infection of the bovine fetus with bovine viral diarrhoea virus. I. Cerebellar atrophy. *Vet. Pathol.* **11**, 486–505 (1974).
20. Barlow, R. M. & Dickinson, A. G. On the Pathology and Histochemistry of the Central Nervous System in Border Disease of Sheep. *Res Vet Sci* **6**, 230–237 (1965).
21. Zachary, J. F. Nervous System. In *Pathologic Basis of Veterinary Disease* (eds McGavin, M. D. & Zachary, J. F.) 771–870 (Elsevier, St. Louis, Missouri, USA, 2012).

22. Patterson, D. S., Done, J. T., Foulkes, J. A. & Sweasey, D. Neurochemistry of the spinal cord in congenital tremor of piglets (type AII), a spinal dysmyelination of infectious origin. *J Neurochem* **26**, 481–485 (1976).
23. Bolin, S. R. Congenital tremors virus. In *Diseases of swine* (eds. Leman, A. D., Straw, B. E., Mengeling, W. L., D'Allaire, S. & Taylor, D. J.) 247–249 (Iowa State University Press, Ames, 1992).
24. Herder, V., Wohlsein, P., Peters, M., Hansmann, F. & Baumgärtner, W. Salient lesions in domestic ruminants infected with the emerging so-called Schmallenberg virus in Germany. *Vet Pathol* **49**, 588–591 (2012).
25. Riedelsheimer, B. & Welsch, U. Romeis Mikroskopische Technik. In *Färbungen* (eds. Mulisch, M. & Welsch, U.) 198–297 (Spektrum Akademischer Verlag, Heidelberg, 2010).
26. Hoffmann, B., Beer, M., Schelp, C., Schirrmeyer, H. & Depner, K. Validation of a real-time RT-PCR assay for sensitive and specific detection of classical swine fever. *J Virol Methods* **130**, 36–44 (2005).
27. Vilcek, S. *et al.* Genetic variability of classical swine fever virus. *Virus Res* **43**, 137–147 (1996).
28. Pfaender, S. *et al.* Clinical course of infection and viral tissue tropism of hepatitis C virus-like nonprimate hepaciviruses in horses. *Hepatology* **61**, 448–459 (2015).
29. Fischer, N. *et al.* Rapid metagenomic diagnostics for suspected outbreak of severe pneumonia. *Emerg Infect Dis* **20**, 1072–1075 (2014).
30. Fischer, N. *et al.* Evaluation of Unbiased Next-Generation Sequencing of RNA (RNA-seq) as a Diagnostic Method in Influenza Virus-Positive Respiratory Samples. *J Clin Microbiol* **53**, 2238–2250 (2015).
31. Peng, Y., Leung, H. C., Yiu, S. M. & Chin, F. Y. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420–1428 (2012).
32. McWilliam, H. *et al.* Analysis Tool Web Services from the EMBL-EBI. *Nucleic Acids Res* **41**, W597–600 (2013).
33. Postel, A. *et al.* Improved strategy for phylogenetic analysis of classical swine fever virus based on full-length E2 encoding sequences. *Vet Res* **43**, 50 (2012).
34. Simmonds, P. SSE: a nucleotide and amino acid sequence analysis platform. *BMC Res Notes* **5**, 50 (2012).
35. Strandstrom, H. *et al.* C-type particles produced by a permanent cell line from a leukemic pig. I. Origin and properties of the host cells and some evidence for the occurrence of C-type-like particles. *Virology* **57**, 175–178 (1974).
36. Seebach, J. D. *et al.* Immortalized bone-marrow derived pig endothelial cells. *Xenotransplantation* **8**, 48–61 (2001).
37. Meyer, D., Schmeiser, S., Postel, A. & Becher, P. Transfection of RNA from organ samples of infected animals represents a highly sensitive method for virus detection and recovery of classical swine fever virus. *PLoS One* **10**, e0126806 (2015).
38. Huson, D. H. & Scornavacca, C. Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks. *Syst Biol* **61**, 1061–1067 (2012).

Acknowledgements

We want to thank our colleagues Dr. Thomas große Beilage (Veterinary practice Essen/Oldenburg) and Dr. Jens Böttcher (Bavarian Animal Health Service) for providing pig sera. This work was supported in part by DG SANCO of the European Commission, Niedersachsen-Research Network on Neuroinfectiology (N-RENNT) of the Ministry of Science and Culture of Lower Saxony, Germany and by the European Union's Horizon 2020 research and innovation program under grant agreement No. 643476 (COMPARE). We are grateful to Inga Grotha, Caroline Schütz and Bettina Buck for excellent technical assistance. We thank Daniela Indenbirken for technical support in preparing the Illumina NGS libraries.

Author Contributions

A.P. and F.H. contributed equally to this study. Identification, quantification of APPV by PCR and subsequent molecular characterization was performed and interpreted by A.P., C.B. and P.B. Necropsies, histological investigation and FISH were performed and interpreted by F.H., V.H., V.M.P. and W.B. NGS and subsequent contig assembly was conducted by N.F., M.A. and A.G. Sampling of the pigs and epidemiological investigation were conducted by S.D., J.T. and M.W. All authors contributed to the manuscript and reviewed the final version.

Additional Information

Accession codes: The obtained consensus sequences were deposited in GenBank (KU041637- KU041639).

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Postel, A. *et al.* Presence of atypical porcine pestivirus (APPV) genomes in newborn piglets correlates with congenital tremor. *Sci. Rep.* **6**, 27735; doi: 10.1038/srep27735 (2016).



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

References

1. van Leeuwen M, Williams MM, Koraka P, Simon JH, Smits SL, Osterhaus AD. Human picobornaviruses identified by molecular screening of diarrhea samples. *J Clin Microbiol*. 2010;48:1787–94. <http://dx.doi.org/10.1128/JCM.02452-09>
2. Schürch AC, Schipper D, Bijl MA, Dau J, Beckmen KB, Schapendonk CM, et al. Metagenomic survey for viruses in Western Arctic caribou, Alaska, through iterative assembly of taxonomic units. *PLoS One*. 2014;9:e105227. <http://dx.doi.org/10.1371/journal.pone.0105227>
3. Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and parallel computing. *Nat Methods*. 2012;9:772. <http://dx.doi.org/10.1038/nmeth.2109>
4. Tamura K, Stecher G, Peterson D, Filipinski A, Kumar S. MEGA6: Molecular evolutionary genetics analysis version 6.0. *Mol Biol Evol*. 2013;30:2725–9. <http://dx.doi.org/10.1093/molbev/mst197>
5. Zhou F, Sun H, Wang Y. Porcine bocavirus: achievements in the past five years. *Viruses*. 2014;6:4946–60. <http://dx.doi.org/10.3390/v6124946>
6. Bodewes R, Lapp S, Hahn K, Habierski A, Förster C, König M, et al. Novel canine bocavirus strain associated with severe enteritis in a dog litter. *Vet Microbiol*. 2014;174:1–8. <http://dx.doi.org/10.1016/j.vetmic.2014.08.025>
7. Chen D, Wei Y, Huang L, Wang Y, Sun J, Du W, et al. Synergistic pathogenicity in sequential coinfection with *Mycoplasma hyorhinis* and porcine circovirus type 2. *Vet Microbiol*. 2016;182:123–30. <http://dx.doi.org/10.1016/j.vetmic.2015.11.003>
8. Mori D, Ranawaka U, Yamada K, Rajindrajith S, Miya K, Perera HK, et al. Human bocavirus in patients with encephalitis, Sri Lanka, 2009–2010. *Emerg Infect Dis*. 2013;19:1859–62. <http://dx.doi.org/10.3201/eid1911.121548>
9. Benjamin LA, Lewthwaite P, Vasanthapuram R, Zhao G, Sharp C, Simmonds P, et al. Human parvovirus 4 as potential cause of encephalitis in children, India. *Emerg Infect Dis*. 2011;17:1484–7.
10. Barah F, Whiteside S, Batista S, Morris J. Neurological aspects of human parvovirus B19 infection: a systematic review. *Rev Med Virol*. 2014;24:154–68. <http://dx.doi.org/10.1002/rmv.1782>

Address for correspondence: Wolfgang Baumgärtner, Department of Pathology, University of Veterinary Medicine, Bünteweg 17 D-30559 Hannover, Germany; email: wolfgang.baumgaertner@tiho-hannover.de

Pegivirus Infection in Domestic Pigs, Germany

Christine Baechlein,¹ Adam Grundhoff,¹ Nicole Fischer, Malik Alawi, Doris Hoeltig, Karl-Heinz Waldmann, Paul Becher

Author affiliations: University of Veterinary Medicine Hannover, Hannover, Germany (C. Baechlein, D. Hoeltig, K.-H. Waldmann, P. Becher); German Center for Infection Research Partner Site Hannover–Braunschweig, Hannover (C. Baechlein, P. Becher); German Center for Infection Research Partner Site Hamburg–Lübeck–Borstel, Hamburg (N. Fischer, A. Grundhoff); Heinrich Pette Institute, Hamburg, Germany (A. Grundhoff, M. Alawi); University Medical Center Hamburg–Eppendorf, Hamburg (N. Fischer, M. Alawi)

DOI: <http://dx.doi.org/10.3201/eid2207.160024>

¹These authors contributed equally to this article.

To the Editor: The family *Flaviviridae* includes many human and animal virus pathogens. Recently, in addition to the genera *Flavivirus*, *Hepacivirus*, and *Pestivirus*, a fourth genus, *Pegivirus*, has been identified (1). In addition to human pegiviruses, a range of phylogenetic, highly divergent pegiviral sequences have been identified in various animal species, including primates, bats, rodents, and horses (2). We report the detection of a porcine pegivirus (PPgV) in serum samples from pigs.

Initially, we investigated pooled serum samples by using high-throughput sequencing methods and isolated RNA from individual porcine serum samples by using the QIAmp Viral RNA Mini Kit (QIAGEN, Hilden, Germany). We prepared libraries compatible with Illumina (San Diego, CA, USA) sequencing from pooled samples and individual serum samples by using the ScriptSeq version 2 RNA-Seq Library Preparation Kit (Epicenter, Madison, WI, USA) and sequenced them by using a HiSeq 2500 (2 × 150 cycles paired-end; Illumina) for pooled samples and MiSeq (2 × 250 cycles paired-end; Illumina) for individual samples (3).

We conducted quantitative reverse transcription PCR (RT-PCR) by using a Quantitect-SYBR Green Assay (QIAGEN) and primers PPgV_fwd: 5'-CTGTCTATGCTGGTCAC-GGA-3' and PPgV_rev: 5'-GCCATAGAACGGGAAGTC-GC-3'. By using high-throughput sequencing of the pooled serum sample library (23,167,090 reads), we identified 1 contig (4,582 bp) that had distant nucleotide sequence similarity to bat pegivirus (69% and 4% sequence coverage) and 2 contigs (2,683 bp and 665 bp) that had 73% sequence coverage, thereby covering 8% and 37% of the identified sequence. RT-PCR with primers designed on basis of recovered sequences identified the sample containing pegivirus sequences. Subsequent MiSeq analysis (7,085,595 reads) of an RNA library prepared from a sample from 1 animal identified 1 contig (9,145 nt) with sequence similarity to pegivirus sequences.

We performed 3' end completion of the viral genome by rapid amplification of cDNA ends and identified the entire open reading frame of PPgV_903 encoding 2,972 aa (GenBank accession no. KU351669). Analysis of the pegivirus 5' untranslated region identified a highly structured internal ribosome entry site motif (online Technical Appendix, <http://wwwnc.cdc.gov/EID/article/22/7/16-0024-Techapp1.pdf>), which was similar in structure to previously described 5' untranslated region structures of other pegiviruses (4,5).

Pegiviruses do not encode a protein homologous to the capsid protein of other viruses of the family *Flaviviridae*, another common feature of pegiviruses (6). The presence of cleavage sites for cellular signal peptidases and viral proteases indicates that, similar to polyproteins of other pegiviruses and members of the genus *Hepacivirus*, the pegivirus polyprotein NH₂-E1-E2-Px-NS2-NS3-NS4A-NS4B-NS5A-NS5B-COOH (E [envelope], NS [nonstructural],

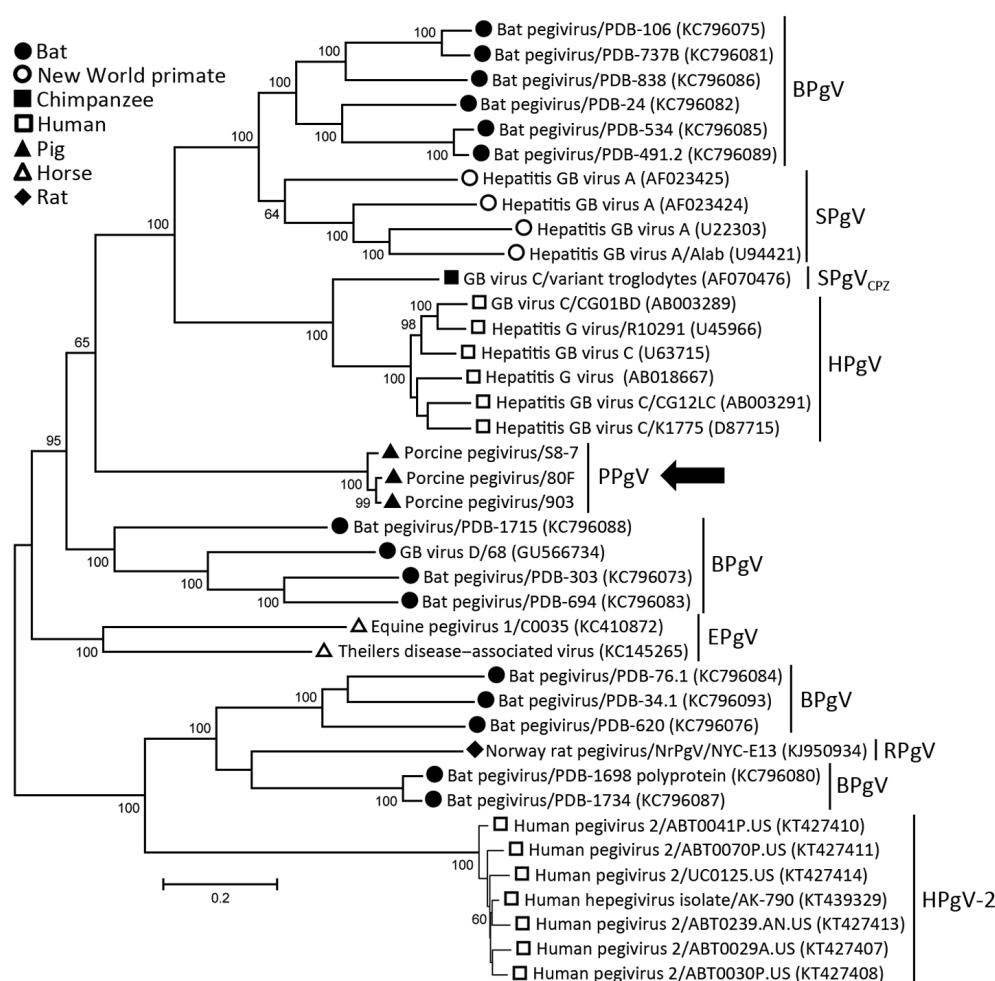


Figure. Phylogenetic analysis of human and animal pegiviruses. We constructed a maximum-likelihood tree on the basis of the complete coding region and used the general time reversible model for modeling of substitutions. Bootstrap analysis was performed with 200 replicates. Numbers along branches are percentage bootstrap values. GenBank accession numbers are in parentheses. Arrow indicates viruses isolated in this study. Scale bar indicates nucleotide substitutions per site. BPgV, bat pegivirus; SPgV, simian pegivirus; SPgV_{CPZ}, simian pegivirus (chimpanzee); HPgV, human pegivirus; PPgV, porcine pegivirus; EPgV, equine pegivirus; RPgV, rodent pegivirus. GB viruses have recently been reclassified as pegiviruses.

and Px [protein X]) is cleaved co-translationally and posttranslationally.

We tested 3 additional animals from the same breeding cohort for virus RNA at irregular intervals for 22 months. One animal was positive for pegivirus RNA for 7 months, and the other 2 animals had pegivirus RNA in serum for 16 and 22 months. None of these animals showed obvious clinical signs attributable to virus infection. Follow-up investigation of 455 serum samples from 37 swine holdings from Germany identified 10 (2.2%) samples from 6 pig holdings that contained pegivirus RNA. We obtained 2 additional near full-length genomic sequences (PPgV_80F and PPgV_S8-7) from 2 animals in different herds by high-throughput sequencing, RT-PCR, and Sanger sequencing (GenBank accession nos. KU351670 and KU351671).

Phylogenetic analyses of complete coding regions showed the close relationship of the 3 pegivirus sequences from Germany. These 3 sequences formed a separate clade within the genus *Pegivirus* (Figure). Pairwise comparison between PPgV_903 and the other 2 pegivirus sequences showed strong nucleotide identities (96.0%–98.4%). A distance scan over the entire polyprotein showed genetic distance to other pegiviruses and demonstrated that NS3 and

NS5B contain the most conserved regions among pegivirus polyproteins (online Technical Appendix).

In horses, 2 distinct pegiviruses that had different potentials to cause clinical disease in infected animals have been described (4,7). No obvious clinical effects were observed in pegivirus-infected animals during our study. However, potential consequences of viral infection for animal health and food production need to be explored more closely under field and experimental conditions. Pegiviruses can interact with the immune system of the host. Coinfection with human pegivirus and HIV can have beneficial effects, which result in decreased retroviral loads and delayed disease progression (8).

It will be useful to investigate whether co-infections with pegiviruses can influence clinical manifestations of infectious diseases of swine, including multifactorial diseases such as postweaning multisystemic wasting syndrome, in which unknown immune modulating virus infections have been suggested to influence the degree of clinical illness (9). RNA viruses have considerable potential to adapt to new environmental conditions and to overcome host restrictions (10). Until now, the host tropism of PPgV has not been investigated in detail. Therefore,

additional studies will be required to elucidate whether the spectrum of potential hosts might include other farm or companion animals, and whether the virus might be able to infect humans.

Acknowledgments

We thank Jens Böttcher, Thomas Große Beilage, Diana Meemken, Alexandra von Altröck, and Cornelia Schwennen for collecting serum samples; Polina Parfentev for providing excellent technical assistance; and Daniela Indenbirken for providing support in preparation of the RNA library.

This study was supported by the German Center for Infection Research/Thematic Translational Unit Emerging Infections.

References

1. Stapleton JT, Bukh J, Muerhoff AS, Fong S, Simmonds P. Assignment of human, simian and bat pegiviruses (previously described as GBV-A, GBV-C, and GBV-D) as members of a new genus (*Pegivirus*) within the *Flaviviridae* [cited 2015 Oct 21]. <http://www.ictvonline.org/proposals/2012.011a-dV.A.v2.Pegivirus.pdf>
2. Thézé J, Lowes S, Parker J, Pybus OG. Evolutionary and phylogenetic analysis of the hepaciviruses and pegiviruses. *Genome Biol Evol.* 2015;7:2996–3008. <http://dx.doi.org/10.1093/gbe/evv202>
3. Baechlein C, Fischer N, Grundhoff A, Alawi M, Indenbirken D, Postel A, et al. Identification of a novel hepacivirus in domestic cattle from Germany. *J Virol.* 2015;89:7007–15. <http://dx.doi.org/10.1128/JVI.00534-15>
4. Kapoor A, Simmonds P, Cullen JM, Scheel TK, Medina JL, Giannitti F, et al. Identification of a pegivirus (GB virus-like virus) that infects horses. *J Virol.* 2013;87:7185–90. <http://dx.doi.org/10.1128/JVI.00324-13>
5. Simons JN, Desai SM, Schultz DE, Lemon SM, Mushahwar IK. Translation initiation in GB viruses A and C: evidence for internal ribosome entry and implications for genome organization. *J Virol.* 1996;70:6126–35.
6. Stapleton JT, Fong S, Muerhoff AS, Bukh J, Simmonds P. The GB viruses: a review and proposed classification of GBV-A, GBV-C (HGV), and GBV-D in genus *Pegivirus* within the family *Flaviviridae*. *J Gen Virol.* 2011;92:233–46. <http://dx.doi.org/10.1099/vir.0.027490-0>
7. Chandriani S, Skewes-Cox P, Zhong W, Ganem DE, Divers TJ, Blaricum AJ, et al. Identification of a previously undescribed divergent virus from the *Flaviviridae* family in an outbreak of equine serum hepatitis. *Proc Natl Acad Sci U S A.* 2013;110:E1407–15. <http://dx.doi.org/10.1073/pnas.1219217110>
8. Schwarze-Zander C, Blackard JT, Rockstroh JK. Role of GB virus C in modulating HIV disease. *Expert Rev Anti Infect Ther.* 2012;10:563–72. <http://dx.doi.org/10.1586/eri.12.37>
9. Grau-Roma L, Fraile L, Segalés J. Recent advances in the epidemiology, diagnosis and control of diseases caused by porcine circovirus type 2. *Vet J.* 2011;187:23–32. <http://dx.doi.org/10.1016/j.tvjl.2010.01.018>
10. Rosenberg R. Detecting the emergence of novel, zoonotic viruses pathogenic to humans. *Cell Mol Life Sci.* 2015;72:1115–25. <http://dx.doi.org/10.1007/s00018-014-1785-y>

Address for correspondence: Paul Becher, Institute of Virology, Department of Infectious Diseases, University of Veterinary Medicine, Buenteweg 17, 30559 Hannover, Germany; email: paul.becher@tiho-hannover.de

New Chimeric Porcine Coronavirus in Swine Feces, Germany, 2012

Valerij Akimkin,¹ Martin Beer,¹ Sandra Blome,¹ Dennis Hanke,¹ Dirk Höper,¹ Maria Jenckel,¹ Anne Pohlmann¹

Author affiliations: Chemical and Veterinary Investigations Office Stuttgart, Fellbach, Germany (V. Akimkin); Friedrich-Loeffler-Institut, Greifswald–Insel Riems, Germany (M. Beer, S. Blome, D. Hanke, D. Höper, M. Jenckel, A. Pohlmann)

DOI: <http://dx.doi.org/10.3201/eid2207.160179>

To the Editor: Porcine epidemic diarrhea virus (PEDV) and transmissible gastroenteritis virus (TGEV) can cause severe enteritis in pigs accompanied by diarrhea, vomiting, and dehydration. Clinical signs are most prominent in young suckling pigs, in which high mortality rates are common. As seen in recent porcine epidemic diarrhea outbreaks in the United States and Asia, the effect on the pig industry can be tremendous.

Recently, Boniotti et al. (1) reported detection and genetic characterization of swine enteric coronaviruses (CoVs) circulating in Italy during 2007–2014. Characterization was based on sequencing and phylogenetic analyses of spike genes of TGEV and PEDV isolates. This study also reported a new recombinant CoV strain with a TGEV backbone and a PEDV spike gene (SeCoV/Italy/213306/2009; KR061459), which was identified as a swine enteric CoV (SeCoV). This chimeric virus presumably resulted from a recombination event.

Accompanying a study of recent porcine epidemic diarrhea cases in Germany caused by a new PEDV Indel strain (2), we retrospectively analyzed fecal samples from pigs that showed typical clinical symptoms of a PEDV infection. The sample set included fecal material collected from a farm in southern Germany on which an episode of diarrhea among pigs occurred in 2012. This material was shown by electron microscopy to contain CoV-like particles (Figure), but showed negative results by reverse transcription PCRs specific for the PEDV nucleocapsid gene.

Subsequent metagenomic analyses resulted in the full-genome sequence of a swine enteric CoV (SeCoV/GER/L00930/2012). We found a sequence showing high similarity (99.5% identity) with the TGEV/PEDV recombinant reported by Boniotti et al. (1). Network analysis of complete genome sequences of similar CoVs underline the chimeric nature of the genome between TGEV and PEDV genome sequences (online Technical Appendix Figure,

¹All authors contributed equally to this article.

Eidesstattliche Versicherung

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

A handwritten signature in blue ink, appearing to read 'M. Aland', is positioned to the right of the text.