Use molecular barcodes to study emerging infectious diseases

Heng-Chang Chen[1]

[1]Institut de Génétique Humaine

May 18, 2022

Abstract

The appearance for viruses that evolve to adapt to a new living niche often reflect on viral sequence changes. Fixation of these changes may require a long time through repeated transmission, thereby rendering a reduced size of an effective population harboring dominant alterations in their sequence spaces. Those approaches, with which we can closely monitor and survey the transient changes of viral sequences over the longer timescales, thus become a requisite to better understand the evolution of viral pathogenicity. Molecular barcodes are a powerful and practical molecular tool to individually label sequences, allowing for correcting sequencing errors and identifying true mutants of interest with a single nucleotide resolution. Molecular barcoding has also been implemented as a useful approach to study several zoonotic viruses. In this review, the emphasis will not only be limited to summarize current studies focusing on viral pathogenesis and fitness; we will also propose ideas that molecular barcodes can be used to execute surveillance of changes of viral sequences. We believe that this review will be helpful for the readers to better understand the rationale and the usage of molecular barcodes and the perspectives of what molecular barcodes can do for fighting upcoming emerging infectious diseases.

Introduction

While looking back to our history, outbreaks of infectious diseases always bring tragedy for humans. Even though we have been able to conquer most of them, several remain to circulate in human populations and emerge from time to time. Recent outbreak of novel coronavirus diseases 2019 (COVID-19)[1–3]caused by coronaviruses has made us more alert to the emergence of infectious diseases originating from animal reservoirs and transmitting between animals and people (so called zoonotic diseases). It is known that nearly two-thirds of emerging infectious diseases (EIDs) have their origins in animals[4–6]. In the U.S. zoonotic diseases of most concern include zoonotic influenza, salmonellosis, West Nile virus, plague, emerging coronaviruses, rabies, brucellosis and Lyme disease[4–6]. Other EIDs, such as human immunodeficiency virus type 1 (HIV-1) infections, *Escherichia* coli O157:H7, hantavirus, dengue fever and the Zika virus are also a significant burden on public health and global economies at present. Therefore, the way how to closely surveil and to efficiently control EIDs for pandemic prevention are urgent to acquire. In this review article, we will focus on EIDs caused by viruses.

Relative to DNA viruses, RNA viruses have high rates of mutation[7]due to, in part, the high error-prone and low-fidelity of the RNA-dependent RNA polymerases that replicate their genomes[8], subsequently attributing to viral sequence changes. These changes are somehow essential and necessary for viruses to maintain their fitness, especially allowing viruses to frequently undergo host switching under different selection pressures[9]. On a per-site level, it is known that virus sequence change is often dominated by synonymous nucleotide substitutions in coding regions; the protein sequence is thus unaffected[10]. In contrast, nonsynonymous substitutions that change protein sequences frequently result in changing physicochemical properties of amino acids, thereby bringing a much greater effect on an individual. Take severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) as an example, the most common type of nonsynonymous mutation observed is alanine to valine (Ala - Val)[11]. Either way, substituted nucleotides that retrain after repeated circulation within a population are sort of *imprints* , reflecting how viruses adapt as a host niche changes throughout

evolutionary timescales. It is important to stress the point that these changes in the viral sequence space can however be not straightforward; some substituted nucleotides may disappear during the period of this evolutionarily transient process; others may remain until the viruses can fully adapt to new hosts. Therefore, in the second part of this review we will focus on the potential of the molecular barcoding technology, a systematic and quantitative approach, with which we will be able to experimentally follow up sequential changes of the viral genomic sequences at a single-sequence level, being indispensable to dissect the molecular basis of any present and upcoming EID caused by emerging- and newly discovered viruses.

Molecular barcoding has been invented as a useful tool to investigate population diversity. The molecular barcoding strategy has first been proposed to solve the problems of PCR duplications and to improve the accuracy of next generation sequencing quantification[12–15]. In the past, molecular barcodes have been given various names, such as unique identifier, unique molecular identifier (UMI)[16], primer ID[17]and duplex barcodes. Molecular barcodes are commonly in the string form of random nucleotides, partially degenerate nucleotides, or defined nucleotides. The concept of molecular barcodes is that each original DNA or RNA fragment, within the same pool of the samples, is tagged with a unique sequence of molecular barcodes[18]. Sequence reads that contain different molecular barcodes illustrate different origins of molecules, whereas sequence reads with the same molecular barcodes are the result of PCR duplication from the same original molecule[18]. The length of molecular barcodes can vary (normally 4 - 20 base pairs): with a longer sequence of molecular barcodes we have a lower probability of identical barcodes present between two or more sequence reads. By employing molecular barcodes, we can thus possibly distinguish PCR artifacts from sequence variants present in different original molecules[13,19].

In the past ten years, technological progress of the molecular barcoding strategy has been made to reach the resolution at a single-molecule level[12,13,16,17,20–22]and detect low-frequency and subclonal variations[23]. This strategy can now be applied to study viruses in many aspects, for example, viral transmission[24], transcriptomics analyses of viruses[25], evolutionary dynamics[26], diagnostics of infectious diseases[27], viral capsid functions[28]as well as the analysis of a viral gene[17]. In this review we summarize several examples of studies, in which molecular barcodes are used to understand the molecular bases and viral fitness[29–31]of zoonotic viruses with the emphasis placed on SARS-CoV-2, HIV-1, influenza virus and Zika virus followed by elucidating our ideas about how molecular barcodes can be applied to closely survey and predict evolutionary of dynamic sequence changes of other emerging- and new discovered viruses *in vitro* .

Main text

Principle and design of molecular barcodes tagged on the viral genome

The molecular barcoding strategy has been successfully employed in "omics" techniques for virus studies (summarized in Table 1). The principle of molecular barcoding is to tag the viral genome with a string of random nucleotides in order to distinguish one virus from another. Molecular barcodes can be introduced into a template by using different approaches. First, molecular barcodes can be embedded into sequencing adaptors while constructing sequencing libraries (Figure 1A). A classical example was given by Schmitt et al. (2012)[19]. They constructed sequencing libraries with two independent molecular barcodes associated at both ends of each library to correct sequencing errors present on every sequencing read. It is noteworthy to point out the difference between "molecular barcodes" and "sample barcodes (sample indexes)" here. Although both are common to be used in the same sequencing reads, molecular barcodes aim to correct sequencing errors and increase sequencing accuracy or define true mutations; whereas sample barcodes enable multiple samples to be sequenced together, so called multiplexed into the same lane of flow cells (Figure 1B). The second approach is to employ molecular inversion probes (MIP) carrying molecular barcodes. A good example of this approach is the single molecule Molecular Inversion Probes method[23](Figure 1C), which combines the MIP strategy for targeted capture[32–34]at a single-molecule level[12,13,16,17,21–22]. Third, molecular barcodes can also be introduced on a template by PCR amplification with target-specific primers (Figure 1D).

Estimation barcode collision

Molecular barcode sequences can be altered during barcode synthesis, primer ligation, PCR amplification and sequencing errors, leading to incorrect sample identification. These errors can be either nucleotide substitutions or small insertions and deletions[35]. It is thus critical to take barcode complexity into account while generating a barcoded library in order to diminish the negative impact on the accuracy of readouts when such circumstances happen. In general, barcode complexity determines the quality of sequencing libraries: the higher barcode complexity of a library is, the better experimental outcomes can reflect natural phenomena. It is also recommended to implement error correcting algorithms and codes in every analytical pipeline while decoding barcode sequences. Hamming codes[36,37]and Levenshtein codes[38] are two popular methods for error-correcting. Furthermore, evaluation of the influence of molecular barcodes on the capacity of a virus to fulfill its life cycle, so-called viral fitness[29–31]should be considered as well. Below we will detail approaches used to embed molecular barcodes in SARS-CoV-2, HIV-1 and Simian immunodeficiency virus (SIV), Influenza virus and Zika virus.

Molecular barcodes designed for virus studies

SARS-CoV-2

SARS-CoV-2 continues to threaten our lives after more than two years of the COVID-19 outbreak. Up to date, more than 500 millions cases of COVID-19 have been reported, including more than six millions deaths worldwide, numbers that increase daily[39]. One of the major challenges to suppress local spread of SARS-CoV-2 is due to the fact that 33% of people with SARS-CoV-2 infection are estimated to be asymptomatic[40]. Population-scale testing thus appears to be very important for us to effectively detect people with SARS-CoV-2 infections and rapidly place them in quarantine[41–43]at the first place. The success of population-scale testing relies on how many individual samples can be synchronously tested. To tackle this issue, several groups have applied the barcode strategy on the high-throughput sequencing platform to make their testing capacity as maximal as possible. Ludwig and colleagues reported a high-throughput technology named LAMP-seq to sequence ten of millions of individual samples at the same time[27]. LAMP-seq is derived from Reverse-Transcription Loop-mediated Isothermal Amplification (RT-LAMP)[44,45]by employing molecular barcodes specific for each sample. LAMP-seq can be started with an unpurified swab sample[46]followed by a single heating step, extensive sample pooling, massively parallel RT-LAMP, and standard computational analysis[27]to identify infected people. Six different LAMP primers [F3, B3, forward inner primer (FIP), backward inner primer (BIP), LF, LB] were used in a RT-LAMP reaction to target the SARS-CoV-2 N gene; a sequence of 10 base pairs DNA barcode was incorporated in the FIP primer (primer sequence is TCTGGCCCAGTTCCTAGGTAGTNNNNNNNNNNCCAGACGAATTCGTGGTGG, where Ns represent a unique barcode sequence)[27,47,48](Figure 2A). The authors further suggested using the combination of unique barcodes per sample to form a compressed barcode space, where up to five barcodes can be present in order to enlarge the capacity for sample pooling as long as a small fraction of samples is expected to be positive during population scale testing. The characterization of the sequencing library is illustrated in Figure 2A. The authors calculated a minimum Levenshtein distance to ensure that one to two insertion, deletion or substitution errors between any pair of barcodes were detectable in barcode sets, indicating that the molecular barcode system employed in LAMP-seq was robust. It is important to note that the complexity of molecular barcodes (1,000 - 10,000 barcode set) in LAMP-seq is sufficient to cover the dynamic range of input viral loads and the presence of molecular barcodes do not affect LAMP sensitivity, product amounts, or downstream PCR amplification. Finally, this molecular barcoding system was validated by employing a modified Bloom filter[49]based on a pool of 10,000 barcodes to estimate the probability of false-positive and false-negative generated by LAMP-seq. After computational simulation the authors concluded that when using 5 barcodes per sample, 3 barcodes are detectable.

Bloom and colleagues proposed another high-throughput method named Swab-Seq by employing molecular barcodes for a population-scale testing[50,51]. Two sets of 1,536 unique barcodes were placed adjacent to the P5 and P7 adaptors in Illumina sequencing primers (so-called i5 and i7 sample barcodes), respectively, rendering the presence of two independent barcodes in each amplicon (Figure 2B). Barcoded primers were designed to amplify the SARS-CoV-2 S gene. Every barcode is 10 base pairs in length and does not contain

homopolymer repeats greater than 2 nucleotides. At least three nucleotides (a minimum Levenshtein distance of 3)[52]are different between two barcodes present in the same amplicon, allowing for demultiplexing even in the face of sequencing errors. Swab-Seq has been validated on the bench by employing purified RNA nasopharyngeal samples and extraction-free saliva specimens, showing that it has extremely high sensitivity and specificity for the detection of viral RNA[50,53]. Swab-Seq has been scaled up since 2021 to support asymptomatic screening. To date, over 80,000 tests have been performed by applying this high-throughput protocol, at a scale of ~10,000 testing samples per week. More methods[53–55]that employ the similar strategy of molecular barcoding for sample multiplexing, rendering the feasibility of initiating a population-scale testing have been expected to come up one after another. At present, the molecular barcoding strategy used for SARS-CoV-2 studies mainly focus on scaling up the testing capacity efficiently. How to use molecular barcodes to gain more insight into the SARS-CoV-2 Virology will be considerable for further investigation.

HIV and SIV

Retroviral zoonoses have received public health attention due to the origin of HIV-1 and human immunodeficiency virus type 2 linked to cross-species transmission of SIV from naturally infected primates, with a prevalence ranging up to 36%[56,57]. Even though viruses have been adapted in human society over thirty years, up to 2020, the estimated number of people living with HIV was 37.7 million[58,59], indicating research on HIV/AIDS (Acquired immunodeficiency syndrome) remains a top priority worldwide. Jabara and colleagues employed the molecular barcoding strategy (so called Primer ID in their article) to reveal HIV-1 dynamic changes in genetic variation[17]. A string of eight nucleotides of a molecular barcode followed by a string of three nucleotides of a sample barcode were incorporated in a reversed primer which anneals downstream of the HIV-1 protease gene (Figure 3A). Both barcodes can be followed through the PCR amplification and deep sequencing and be used to calibrate artificial errors generated by PCR amplification and high-throughput sequencing and then created a consensus sequence for individual viral templates. After calibration by using molecular barcodes, 80% of the unique sequence polymorphisms were removed in this dataset[17]. Of note, the distribution of identical molecular barcodes did not form a Gaussian distribution, meaning that each viral genome was not equally amplified. The authors assumed that this phenomenon was because the number of cDNA templates amplified in each RNA cycle varies[17].

In 2017 the laboratory of Dr. Filion published a high-throughput method named Barcoded HIV Ensembles (B-HIVE) to monitor the expression of thousands of proviruses in a cell population[25]. The principle of B-HIVE is to tag the HIV genome with a molecular barcode of 20 random nucleotides prior to generating infectious viral particles (Figure 3B)[25,60]. Advantages of molecular barcodes implemented in B-HIVE are two-folds. First, molecular barcodes were used to map HIV integration sites by generating a high-throughput sequencing library, in which every chimeric DNA is composed of the host local genomic region where provirus DNA inserts. Second, molecular barcodes were used to quantify HIV transcription by performing RT-PCR on viral RNA with the primers flanking the region of a barcode[25,60]. The authors verified the insert-specific provirus expression measured by B-HIVE by using another method called T7-PCR: a positive correlation between RNA barcode counts measured by B-HIVE and RT-qPCR values measured by T7-PCR for the selected barcodes were detected[25,60]. It is worthy noting that all the possible 20-mers represent enough complexity (4^{20} distinct barcode combinations) so that the probability of two viruses having the same barcode is negligible. The authors also observed that the presence of the barcodes did not interfere with the efficiency of HIV (HIV-based vector) infection; side by side comparisons of infection rates between barcoded- and non-barcoded viral particles showed that barcoded viral particles were only ~5% less infectious[25].

Marsden and colleagues have successfully tagged molecular barcodes on the genome of nearly full-length and replication-competent HIV[61]. This HIV construct is derived from NL4-3, which has been modified to express an NL-hemagglutinin (HA) epitope in place of Vpr for the purpose of enriching HIV-infected cells[62]. Molecular barcodes encompassing 21 nucleotides in length were introduced in the region right downstream of the HA tag (Figure 3C); this region is not responsible for any gene expression and lack for any cis elements as well as splice sites (Figure 3C). It is important to note that a thymidine was designed to be placed in the sequence of barcodes with an interval of 2 nucleotides to avoid random strings of consecutive GC bases,

subsequently resulting in the complexity of barcodes down to 4^{14} combinations (Figure 3C). The authors verified the complexity of barcodes by employing Illumina HiSeq sequencing, showing that around 18,000 unique sequences of barcodes were homogeneous present in the original plasmid preparation. The average number of different nucleotides between any two barcodes was 11 base pairs, showing an appreciable diversity of barcodes. Moreover, the frequency of barcodes present in virions correlated with the frequency of barcodes present in the original plasmid preparation, meaning that the complexity of barcodes is able to be recapitulated in infectious particles[61]. Viral infectivity and replication were validated in GHOST CXCR4+CCR5+ cells and primary human peripheral blood mononuclear cells, showing that barcoded viruses are capable of infecting and replicating in primary cells without causing any bias in the presence of barcodes[61].

Apart from HIV, Fennessey and colleagues first established a model of rhesus macaques infected with barcoded SIV to study the dynamics of viral reservoir establishment and viral rebound over time[63]. Molecular barcodes encompassing 10 random nucleotides in length were inserted between the stop codon of *vpx* and the start codon of*vpr* in the SIVmac239 plasmid (Figure 3D). Since only one restriction enzyme, MluI, was chosen for cloning barcodes, insertion of each barcode can thus be bidirectional, consequently doubling the complexity of the barcodes in viral stocks. In order to accurately estimate the complexity of input barcodes named viral clonotypes in the text, the authors reverse transcribed viral RNA into cDNA, which was diluted down to 5,000 viral copies in total 168 aliquots in order to avoid any sequencing error. The authors were able to nicely separate probability distributions between barcodes generated from PCR-induced errors and true sequences of barcodes (9,336 sequences)[63]. Every barcode contains an average of seven distinct nucleotides. Furthermore, the authors also showed that barcoded SIV did not cause any effect on viral infectivity and replication.

Influenza virus

In addition to HIV, Influenza virus A is another classic example that the viruses have zoonotic origins with associated public health concern of host adaptation, permanently establishing in the human population[64]. Varble and colleagues[24]first barcoded influenza A virus, with which the authors attempted to understand the probability and stochastic nature resulting from different transmission routes. Varble and colleagues[24]employed molecular barcodes on the genome of a model influenza A virus to track viral transmission. The challenge to insert molecular barcodes in the genome of influenza A virus is because the eight single-stranded RNA segments are all required for virion production; therefore there is little space for molecular manipulation. Previously Varble and colleagues[65]have chosen to engineer segment 8, encoding the nonstructural protein 1 (NS1) and the nuclear export protein (NEP, also referred to as NS2) because of the shorter of two viral transcripts, facilitating to give any genetic material. Since NS1 and NEP use different reading frames with the overlap between the C-terminal transcript of NS1 and the N-terminal transcript of NEP, the authors thus disrupted the endogenous splice acceptor site and placed it after the stop codon of NS1[65](Figure 4A). This modified segment 8 thus harbored a noncoding intergenic region where a string of 22 nucleotides of a molecular barcode was inserted (Figure 4A). Barcodes were amplified from the 30 arm of a shRNA library[66], along with 100 base pairs of common flanking sequence (Figure 4A). The authors confirmed that inserted molecular barcodes did not affect viral replication in human lung epithelial cells.

Russel and colleagues[67]utilized three different types of barcodes, cell barcodes, UMI and viral barcodes to distinguish every single mRNA transcript based on the single-cell droplet-based platform[67,68](Figure 4B). Cell barcodes were used to distinguish one cell from another; UMIs incorporated in primers were designed to separate each mRNA molecule during the process of reverse transcription, allowing to quantify the number of molecules of each mRNA that have been captured for each cell. The challenge here is how to separate viral- and cellular mRNA generated in the same cell. The authors thus engineered the viral genomic sequence where two synonymous mutations were given in the vicinity of the 3' end of each transcript (Figure 4B) for this discrimination. The transcription of the viral hemagglutinin from the viruses carrying these synonymous mutations was compatible with the wild-type viruses[67], indicating that synonymous mutations did not affect viral fitness. The authors successfully used these viral barcodes to distinguish the viral transcripts carrying mutations from the transcripts produced by the wild-type viruses[67]. By employing this three-

barcodes approach, the authors established a threshold determining truly infected cells and quantitatively investigated stochastic gene expression of influenza viruses among individual infected cells[67].

Zika virus

ZIKV is an arbovirus that consists of an envelope, positive-sense, single-stranded RNA. Its genome encodes three structural proteins (capsid, precursor of membrane and envelope) and seven non-structural proteins (NS1, NS2A, NS2B, NS3, NS4A, NS4B and NS5) responsible for RNA replication[69]. Aliota and colleagues introduced a string of eight degenerate codons where the third nucleotide of each codon is degenerated into a region of the NS2A protein to generate as many synonymous mutations as possible (Figure 5). This barcoded Zika virus was named ZIKV-BC-1.0[26]. A multiplex-PCR approach was applied to sequence the entire coding genome of the ZIKV-BC-1.0 library. Compared with its original strain, ZIKV-IC, three single nucleotide mutations (site 1964, 8488 and 9581) were detected outside of the barcode region in ZIKV-BC-1.0[26].

Aliota and colleagues further used three different approaches to define trustable barcodes referred to as authentic barcodes in their study. Among these three approaches, the third approach (Approach 'C'), with which the authors acquired the highest threshold frequency (0.57%) was placed to be the minimum threshold to identify authentic barcodes. By using this threshold, 20 sequences in the sequencing output list were considered to be authentic[26]. Intriguingly, the first three barcodes (Zika_BC01, Zika_BC02 and Zika_BC03) were more dominant than the rest of 17 barcodes when limited input viral RNA was used for amplification[26]; however at these stage it is still quite challenging to explain whether barcode heterogeneity comes from a potential bias introduced by a barcode itself or the evolutionary selection. The authors also claimed that the absence of barcodes in sequencing reads could be due to the fact that either barcodes were not present in certain biological samples or the copy number of the barcodes were not sufficient to be detected by sequencing, meaning that the presence of each barcode in ZIKV-infected animals was not homogeneous. Viral infectivity and replication of ZIKV-BC-1.0 showed no difference *in vitro* compared with Zika infectious clone-derived- and wild-type viruses[26], indicating that *in vitro* additional degenerate nucleotides in the viral genome did not cause a significant effect on viral fitness.*In vivo* three rhesus macaques were inoculated with ZIKV-BC-1.0 to validate its replication capacity. Viruses were detectable in plasma one day post inoculation (dpi) and viral load elevating between two and four dpi was comparable to ZIKV-IC and wild-type viruses[26], indicating that viral infectivity and replication was not damaged by molecular barcodes present in ZIKV-BC-1.0 *in vivo* , neither.

What have we learned from molecular barcodes in virus studies?

Measure single virus transcriptomics

It is known that transcription measured by standard methods from a bulk sample of materials is typically lost upto 20% variants present in a population[70,71], thereby causing a strong bias for investigating HIV gene expression which is known for its stochastic phenotype of transcription[72,73]. Barcoded HIV thus provides us a power matter to quantitatively measure insert-specific provirus transcription at a single-virus level. Taking the B-HIVE technology as an example, the authors first found that the transcriptional phenotype of stochastic HIV expression falls under the concept of 'position effects' first discovered by Joseph Müller in 1930[74]when he observed that the translocation of the *Drosophila* white gene to the centromere caused its expression to fluctuate. This phenomenon remained elusive until the discovery of histone modifications. On the basis of the concept of position effects, it turns out that proviruses that integrate in the proximity of active enhancers (defined by histone H3 lysine 27 acetylation, H3K27ac) displayed a better transcriptional level compared with those present far from active enhancers[25]. This observation was supported by the previous study, showing that HIV that integrated in different chromosomal regions displayed different transcription levels in Jurkat clonal cell lines[75]and the study conducted by Vansant and colleagues[76], showing that integration sites retargeted by 'LEDGINs'[77,78]harbored low HIV transcription measured by B-HIVE. Another example was given by Russel and colleagues[67], who attempted to study how well influenza virus genes are expressed in hundreds of individual mammalian cells in a single-cell scale. By employing the three-barcodes approach described in the previous section, the authors statistically quantified viral transcription from every infected

cell by using the method called the Gini coefficient[79], a measure of the difference between a given distribution of variables to represent an inequality within a group. The authors calculated the Gini coefficient value that was at least 0.64 from viral mRNA per infected cell, indicating a detectable variation present in viral gene expression and concluded that such transcriptional variation was caused by both cellular factors and inherent stochasticity in addition to viral transcription itself[80–82], which only partially explained this phenotype.

Monitor behavior of latent proviruses in the frame of antiretroviral therapy

In many cases of viruses, the status of viral transcription shows a strong correlation to their pathogenesis. Viruses, such as hepatitis B virus, herpes virus and HIV[83–85]require a transcriptionally silent state, so-called latency throughout their life cycles, allowing viruses to persist in host cells for a long time and establish chronic infections. In the case of HIV, viral rebound frequently occurs when the antiretroviral therapy is interrupted, subsequently leading to the AIDS symptoms resurface. It thus becomes an obvious obstacle for curing HIV-1 infection. In the previous section we talked about molecular barcodes used to quantitatively measure single virus transcription. Here we are going to give several examples that illustrate how molecular barcodes help us to interfere with translation research involving antiretroviral therapy in addition to the measurement of viral transcription.

The "shock and kill" antiretroviral therapy has been the most attractive strategy for purging HIV latent infection as it was first proposed in 2012[86]. The concept of this strategy is to give HIV-1-infected individuals latency reversing agents (LRAs) such as histone deacetylase inhibitors to deliberately reactivate proviral transcription in latently infected cells, rendering reactivated cells destroyed by the host immune system. The success of this strategy relies on the drugs used to reactivate latent proviruses, yet several exploratory clinical trials have already shown limited evidence of efficacy[87]. Since molecular barcodes allow us to quantitative measure HIV transcription at a single-virus level, Chen and colleagues thus attempted to use them for estimating the spectrum of vorinostat, an histone deacetylase inhibitor used in clinical trials[88]as a proof of concept. Comparing with phytohemagglutinin, a strong activator for T cells, it turned out that the same provirus can either respond more strongly to phytohemagglutinin than vorinostat, or*vice versa* , indicating that vorinostat fully reactivate only a subset of the proviruses in a cell population. Proviruses that were prone to be reactivated by vorinostat were more frequently in the vicinity of active regulatory elements (H3K27ac and histone H3 trimethylated at K4 and monomethylated at K4)[25]. This observation was later confirmed by Battivelli and colleagues[89]using a second generation of full-length dual-fluorescence reporter HIV, HIV_{GKO}, to investigate the reactivation potential of various LRAs, including panobinostat[90], JQ1[91–95]and bryostatin[96,97]in pure latent populations. They showed that less than 5% of the latently infected primary CD4+ T cells were able to be reactivated by LRAs and confirmed that the proviruses refractory to be reactivated appeared to be present closer to heterochromatin (histone H3 trimethylated at K27 and at K9) and non-accessible regions (DNase hyposensitivity)[89].

In vivo barcoded HIV has also been used to track viral rebounds in animal models. Marsden and colleagues[61]infected bone marrow-liver-thymus mice, an animal model that can represent the human immune system[98], with barcoded nearly full-length and replication-competent HIV and measured viral rebound after giving a synthetic bryostatin analog, SUW133[99], which is a PKC inhibitor[100]. It was observed that mice treated with SUW133 showed a delay of viral rebound after antiretroviral treatment was stopped compared with vehicle control. In addition, barcode diversity present in reactivated viruses showed a decrease compared with the mice treated with vehicle control, meaning that the administration of SUW133 can eliminate a subset of latently infected cells in this *in vivo* model.

Nonhuman primates are another useful animal model to study retroviruses. SIV-infected nonhuman primates are similar in their pathophysiology and viral dynamics to humans infected with HIV-1[101–105]and therefore make ideal models for HIV studies in many aspects. Fennessey and colleagues[63]successfully employed the model of rhesus macaques infected with barcoded SIV to identify and quantify the dynamics of the establishment of viral reservoir and viral rebound after combination antiretroviral therapy administration and interruption, demonstrating that the time of initiation and duration of therapy administration in nonhuman primates can alter the size of the reservoir. Soon afterwards Khanal and colleagues[106]generated a derivative

based on this barcoded SIV. At present barcoded SIV continues to be applied on animal models of rhesus macaques to tackle several important questions in the field of HIV/AIDS in many aspects[107–111].

Examine evolutionary bottlenecks to depict viral fitness landscapes

An evolutionary bottleneck is the event that reduces the size of a population, limiting the genetic diversity of the species. It occurs when viruses transmit from one host to another, rendering the functionally dominant populations remaining in the circulation among hosts. Molecular barcodes thus become a powerful tool, with which we can simply trace transmission paths of a population of the viruses over different hosts at a single-virus level. Influenza virus is one of the viruses that has subsisted in the human community over a long history. The most remarkable pandemics caused by influenza virus is the so-called Spanish flu, or the 1918 H1N1 pandemic, infecting ~30% of the population[112,113]. Influenza virus A is endemic in waterfowl and is usually nonpathogenic in these birds[114]until the emergence of H5N1 outbreaks in Asia[115–117]. Soon afterwards H5N1 influenza virus is known to be zoonoses, which can transmit from its natural host to humans[118–122]. To date, in addition to avian influenza virus A(H5N1), several zoonotic influenza viruses, such as avian influenza virus A(H7N9)[123,124]and A(H5N6)[125–127]as well as swine influenza virus A(H1N1)[128–131]and A(H3N2)[132,133]and its variant (H3N2v)[134]have been reported. Rapid antigenic variation that enables the viruses to escape from neutralizing antibodies[135]is frequently observed in influenza viruses, rendering vaccines not fully effective. It thus a requisite to understand which variations of the viruses could go through the pressure of selection bottlenecks in different cross-species transmission routes before developing a better strategy against influenza virus infection.

With barcoded influenza A virus in hand, Varble and colleagues examined viral transmission bottlenecks under the selection forces present in four different experimental systems, including Madin-Darby canine kidney cells, embryonated eggs, direct-contact transmission in guinea pigs and ferrets and respiratory-droplet transmission in ferrets[24]. In the context of population dynamics and evolution, the bottleneck concept is often used in reference to events that limit population size: only individuals which successfully pass through the bottleneck retrain in a population in a new environment. The authors detected that bottleneck selection occurred only when embryonated-chicken eggs were infected with barcoded viruses: 5 - 13 viral clones harboring unique barcodes were successfully amplified in each egg[24]. Intriguingly, such bottlenecks observed in individual recipient animals can be independent from viral genome sequences. Furthermore, the authors observed that viral transmission routes affect bottleneck stringency: survival of barcoded viruses transmitting via contact- and airborne infection showed a dramatic decrease compared with direct inoculation[24]. As such, the authors concluded that viruses like H5N1 have to overcome two distinct bottlenecks before the emergence of pandemic. The first one lies on the level of the viral genome sequence. Viruses acquire genetic mutations and/or reassortments to maintain replication fitness in recipient hosts[136]. Another is dependent on viral transmission routes. This barcoded influenza A virus has further been applied by Muñoz-Moreno and colleagues to reveal the evolutionary relationship between influenza NS1 genotypic variants and host tropism[137].

Zika virus is another example of zoonotic diseases that is through mosquito-borne transmission. It was first discovered in a Rhesus monkey and in *Aedes africanus* mosquitoes during a yellow fever study in the Zika forest of Uganda in 1947 and 1948[138,139]. Through the past fifteen years several outbreaks of Zika virus were seen, bringing this virus to the world's attention. Even though the World Health Organization declared the end of the Zika virus epidemic in 2016[140], it still represents a highly significant and long-term problem. The virus induces a mild flu-like disease in about one in five infected individuals[141–144]; in other words, approximately 80% of Zika infections are asymptomatic. Pathologically speaking, Zika virus can also infect the placenta, transmitting from pregnant women to fetuses[145]in addition to blood-brain barrier endothelial cells, neurons and neural stem cells[146,147]. Even though the frequency of vertical transmission is not clear, current data suggest that it may be a very common case, especially if infection occurs in the period of the first trimester[148]. At least 5% of babies contaminated by Zika virus during maternal pregnancy suffered a collection of neurological, visual, auditory, and developmental birth defects, so called congenital Zika syndrome[145]. It is now also known that Zika virus infection is closely associated with diseases. In Brazil

the outbreak of Zika virus correlated with an increase in cases of infant microcephaly[149,150], and in other parts of the Amercias there has been a significant increase in Guillain-Barré syndrome coinciding with Zika infections[151].

In order to better understand the *in vivo* replication and evolutionary dynamics of Zika virus infection, Aliota and colleagues[26]employed barcoded Zika viruses to track and monitor individual viral lineages during acute infection in pregnant and nonpregnant macaques. The authors first confirmed that viral RNA from barcoded Zika virus was detectable in placentae. Fetal histology also showed several pathological symptoms relevant to Zika virus infections[26], meaning that this animal model with barcoded Zika virus infection can most likely bring us the picture representing Zika virus infections in humans. Although no significant changes in the frequency of barcode distribution was observed in nonpregnant macaques, a bottleneck effect appeared after seven dpi in pregnant macaques[26]: the prevalence of the dominant viral lineages identified by unique barcodes changes mainly after day 8 post infection. Aliota and colleagues further used mosquitos fed with *Aedes aegypti* vector competence for barcoded Zika virus to infect the pregnant macaque. Only one viral lineage, Zika_BC02, representing the second most dominant barcode present in the early period of infection was successfully transmitted via mosquitoes.

How molecular barcodes can bring us one step closer for better understanding and preventing from EIDs

As mentioned in the previous section that evolutionary bottlenecks filter out those functionally recessive viruses; the viruses that are able to pass through the selection pressure of bottlenecks usually emerge certain genotypic changes, to be more specific, nucleotide substitutions, allowing the viruses to adapt to a new host. Virus sequence changes caused by evolutionary constraints are critical for zoonotic viruses to jump across the barrier between different species; thereby owning experimental evolution models (Figure 6), with which we can track dynamic changes of viral sequences in the short term followed by testing pathogenic phenotypes associated with their genetic robustness, visualizing fitness landscapes and predicting their long-term evolutionary paths at a single-sequence level becomes irreplaceable to better understand pathogenesis of emerging viruses and to set up antiviral strategies in advance. Below we will share two of our ideas about what molecular barcodes can bring us one step closer to dissecting emerging infectious diseases.

Unveil critical mutations responsible for antiviral drug resistance, contributing to precision medicine

Previously Jabara and colleagues have employed molecular barcodes to create a consensus sequence for individual viral templates amplified from HIV-1infected individuals and sought for the emergence of multiple single-nucleotide polymorphisms (SNPs) responsible for drug-resistance[17]. In such a way only SNPs present in the time as materials were collected can be identified; whether these SNPs have been present from the beginning of the primary infection or they were the SNPs derived from several rounds of nucleotide substitutions over the period of the treatment is however not traceable. Indeed, some substituted nucleotides in virus sequences may not be essential for the final phenotype and could be replaced due to the effects from different constraints during the selection process. A genotype-traceable model for monitoring the dynamics of virus sequence changes is thus important to recapitulate evolutionary trajectories at a single-sequence level.

Our idea is that, by using barcoded viruses to infect host cells present in the selection pressure of antiviral drugs, after several rounds of passages we will then look for the variants that emerge mutations, rendering the viruses capable of surviving in the presence of antiviral drugs. Similar ideas have been applied on mosquito Infection models used to recapitulate the emerging variant of chikungunya virus that occurred in the Indian Ocean islands[152]. At that stage, molecular barcodes, accompanying high-throughput sequencing, allow us to segregate every dominant variant at a single-sequence level and decode corresponding genotypes. This approach can be done with different sorts of combinations of antiviral drugs to build up a database, which systematically associates different antiviral drug-resistant variants with corresponding mutations. Since viruses used in this approach are barcoded, we can also construct phylogenetic trees for individual antiviral drug-resistant variants, allowing us to monitor whether or not any intermediate nucleotide substitutions are

required during the course of the evolutionary process to result in final antiviral drugs-resistant phenotypes. If this is the case, such information could be used as a guideline for clinical diagnostics. In other words, after unveiling the genotypes of viruses present in patients, based on experimental outputs released from this genotype-traceable model, medical doctors can foresee that an indicated patient might exhibit antiviral drug tolerance to certain drugs and adjust antiviral regimens for treatments to satisfy the need for every individual.

Unveil fitness landscapes of newly discovered zoonotic viruses, contributing to preventive medicine

Having this experimental evolution model mentioned above in hand, we can further use this system, especially while dealing with emerging- and newly discovered zoonotic viruses, to *in vitro* generate mutations by reiterating several times of infection in natural host cells/animals followed by testing viral fitness landscapes of variant strains in human cells, thereby helping us to understand and predict the possible impact of variant strains derived from these emerging- and newly discovered zoonotic viruses on humans. Retrospective studies of constructing evolutionary trajectories based on phylogenetic analyses should be performed; in such a way, we can closely monitor the sequential transition of virus sequence changes reshaped by evolutionary constraints throughout every passage of cells or every filial generation in animal models at a single-sequence level and estimate the evolutionary timescale that requires viruses to reprogram their tropisms. This experimental evolution model is inspired by the evolutionary approach named retrovolution[153], with which the authors studied the human deoxycytidine kinase[153]. Here under the framework of this approach with modifications, we import the idea of incorporating molecular barcodes embedded in our viruses of interest, thereby bringing our investigation down to the resolution of a single virus sequence. Most importantly, in line with the procedure designed in this model we will be able to immediately assay viral fitness enhanced by emerging mutations that can be traced through phylogenetic trees. In other words, this barcoding evolution model enables us to associate genotypic changes occurring in natural host cells/animals with predictive phenotypes in human cells in every single evolutionary event in order to accelerate our understanding of newly discovered zoonotic viruses and benefit to develop antiviral strategies in advance. It is worthy to stress the point that in addition to different experimental approaches can be used for predicting the likelihood of predictive mutations[152,154–156], we also propose to use *in silico* approaches for this purpose: we suggest that algorithms, like generative adversarial networks[157], an architecture for training generative models, can be for instance applied for predicting evolutionary paths when the number of input sequences is sufficient. Eventually virus sequences with predictive mutations will also be necessary to test for fitness in human cells.

Perspectives of molecular barcoding for virus research

Undoubtedly, molecular barcodes should strengthen their power in the context of molecule quantification. Together with precise instruments, such as the microfluidic system to delicately detect individual molecules and deep learning algorithms, dedicated to accurately interpret, calculate and classify each sequence of barcodes, we can foresee that molecular barcodes will continue to be one of the most attractive methods applied for quantitative biology. Additional advantage of the molecular barcoding approach is that we will be able to track any intermediate product relevant to the origin of genomic DNA, such as RNA isoforms and antisense RNA or the functional unit which is absent following mRNA processing; few examples have already shown the feasibility of molecular barcoding for tackling these questions[28,158,159]. Furthermore, no matter barcoding DNA- or RNA viruses, molecular barcodes are presently embedded in genomic DNA or cDNA. However welcome, it is known that many zoonotic viruses causing EIDs belong to RNA viruses. Direct barcoding and sequencing native viral RNA molecules without conversion to cDNA will be expected to bring us a more informative picture concerning chemical RNA modifications present in the native RNA molecules and the whole transcriptome. Example given by Smith and colleagues[160]demonstrated the potential for barcoding directly on native RNA molecules by using the platform offered by Oxford Nanopore Technologies.

Conclusions

It is said "we all live in the past." It is probably true. If so, how we can use the past to foresee the future could then be our task at present.

While preparing this manuscript, SARS-CoV-2 still continues circulating in human populations and threatening public health. This pandemic has further stressed the importance of closely monitoring the emergence of infectious diseases originating from animal reservoirs and transmitted between animals and people. It is obvious that this pandemics will not be the last; a powerful experimental evolution approach allowing us to closely survey the dynamics of genetic mutations in a population of circulating zoonotic viruses in the short term and possibly predict upcoming variants is therefore pivotal to improve the scientific, medical and societal response to future zoonoses.

Molecular barcoding has a long history in yeast genetics[161]. And to date, it has made great progress in investigating EIDs. Concerning RNA viruses, as mentioned in the article, molecular barcodes have been used as a tracer to follow the consequence of individual viruses encountering evolutionary selection forces *in vivo* , rendering us better understanding viral transmission fitness and evolutionary dynamics. Molecular barcoding on SARS-CoV-2 are focusing on sample multiplexing thus far; application of molecular barcodes on coronaviruses to gain better understanding of cross-species transmission fitness over time will be extremely important. Whereas molecular barcodes in retroviruses (SIV and HIV)[17,25,61,63]have been applied to identify different polymorphisms on the viral gene, study single provirus transcriptomics and track viral rebound in animal models. Collectively, molecular barcodes provide us a good opportunity for virus studies in a quantitative manner.

The key to using molecular barcodes as universal identifiers for either identifying single molecules or tracing individual viruses is to have a very large initial collection of barcodes. Good complexity of barcodes diminish the probability that two independent molecules harbor the same barcodes. Several approaches used to validate the quality of barcoded libraries, such as calculation of a minimum Levenshtein distance[38] or establishment of a threshold for sorting out trustable barcodes[26]have been mentioned previously. In general, barcode duplications are recommended to be kept at a frequency below 0.1%[60]. It is also important to note that imbalance in the representation of the barcodes in the library has been reported; a thorough investigation into a bias of barcode representation will be required in the future. Eventually, the impact of molecular barcodes on viral infectivity and replication should be always taken into account; fitness of barcoded viruses is recommended to be verified for the usage of every new prepared barcoded library.

Acknowledgements

I would like to thank Dr. Benkirane (L'institut de Génétique Humaine, Montpellier, France) and Dr. Negroni (Institut de biologie moléculaire et cellulaire, Strasbourg, France) for his frequent exchanges and constructive discussions. I would like to thank Dr. Horiuchi (Yokohama City University, Yokohama, Japan) and Dr. Bechara (Center for Human Technology, Istituto Italiano di Tecnologia, Genoa, Italy) for their critical feedback on the manuscript. This manuscript was financially supported by the Łukasiewicz–PORT Polish Center for Technology Development (startup funding package).

Data availability statement: Data sharing is not applicable to this article as no new data were created or analyzed.

Conflict-of-interest statement

The authors declare that they have no conflict-of-interest.

References

1. C. Drosten, S. Günther, W. Preiser, S. van der Werf, H.-R. Brodt, S. Becker, H. Rabenau, M. Panning, L. Kolesnikova, R. A. M. Fouchier, A. Berger, A.-M. Burguière, J. Cinatl, M. Eickmann, N. Escriou, K. Grywna, S. Kramme, J.-C. Manuguerra, S. Müller, V. Rickerts, M. Stürmer, S. Vieth, H.-D. Klenk, A. D. M. E. Osterhaus, H. Schmitz, H. W. Doerr,*N. Engl. J. Med.* **2003** , *348* , Issue 20, pp. 1967–1976*https://doi.org/10.1056/nejmoa030747*

2. T. G. Ksiazek, D. Erdman, C. S. S. R. Goldsmith, Zaki, T. Peret, S. Emery, S. Tong, C. Urbani, J. A. Comer, W. Lim, P. E. Rollin, S. F. Dowell, A.-E. Ling, C. D. Humphrey, W.-J. Shieh, J. Guar-

ner, C. D. Paddock, P. Rota, B. Fields, J. DeRisi, J. Y. Yang, N. Cox, J. M. Hughes, J. W. LeDuc, W. J. Bellini, L. J. Anderson, SARS Working Group,*N. Engl. J. Med.* **2003** , *348* , Issue 20, pp. 1953–1966*https://doi.org/10.1056/nejmoa030781*

3. A. M. Zaki, S. van Boheemen, T. M. Bestebroer, A. D. M. Osterhaus, R. A. M. Fouchier, *N. Engl. J. Med.* **2012** , *367* , Issue 19, pp. 1814–1820*https://doi.org/10.1056/nejmoa1211721*

4. L. H. Taylor, S. M. Latham, M. E. J. Woolhouse, *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **2001** , *356* , 983–9*http://dx.doi.org/10.1098/rstb.2001.0888*

5. M. E. J. Woolhouse, S. Gowtage-Sequeria, *Emerging Infectious Diseases* , **2005** , *11* , 1842–1847*http://doi.org/10.3201/eid1112.050997*

6. K. E. Jones, N. G. Patel, M. A. Levy, A. Storeygard, D. Balk, J. L. Gittleman, P. Daszak, *Nature* **2008** , *451* , 990–993*https://doi.org/10.1038/nature06536*

7. R. Sanjuán, P. Domingo-Calap, *Cell. Mol. Life Sci* .**2016** , *73* , Issue 23, pp. 4433–4448*https://doi.org/10.1007/s00018-016-2299-6*

8. K. M. Peck, A. S. Lauring, *J. Virol.* 2018, 92*https://doi.org/10.1128/JVI.01031-17*

9. M. Vignuzzi, J. K. Stone, J. J. Arnold, C. E. Cameron, R. Andino,*Nature* **2006** , *439* , Issue 7074, pp. 344–348*https://doi.org/10.1038/nature04388*

10. P. Simmonds, P. Aiewsakun, A. Katzourakis, *Nat. Rev. Microbiol* . **2019** , *17* , 321–328*https://doi.org/10.1038/s41579-018-0120-2*

11. K. Yi, S. Y. Kim, T. Bleazard, T. Kim, J. Youk, Y. S. Ju, *Exp. Mol. Med.* **2021** , *53* , 1229–1237*https://doi.org/10.1038/s12276-021-00658-z*

12. J. A. Casbon, R. J. Osborne, S. Brenner, C. P. Lichtenstein,*Nucleic Acids Res.* **2011** , *39* , e81*https://doi.org/10.1093/nar/gkr217*

13. I. Kinde, J. Wu, N. Papadopoulos, K. W. Kinzler, B. Vogelstein,*Proc. Natl. Acad. Sci. U. S. A.* **2011** , *108* , Issue 23, pp. 9530–9535*https://doi.org/10.1073/pnas.1105422108*

14. T. M. Blomquist, E. L. Crawford, J. L. Lovett, J. Yeo, L. M. Stanoszek, A. Levin, J. Li, M. Lu, L. Shi, K. Muldrew, J. C. Willey,*PLoS ONE* **2013** , *8* , Issue 11, p. e79120*https://doi.org/10.1371/journal.pone.0079120*

15. G. K. Fu, W. Xu, J. Wilhelmy, M. N. Mindrinos, R. W. Davis, W. Xiao, S. P. A. Fodor, *Proc. Natl. Acad. Sci. U. S. A.* **2014** ,*111* , Issue 5, pp. 1891–1896*https://doi.org/10.1073/pnas.1323732111*

16. T. Kivioja, A. Vähärautio, K. Karlsson, M. Bonke, M. Enge, S. Linnarsson, J. Taipale, *Nat. Methods* **2011** , *9* , 72–74*https://doi.org/10.1038/nmeth.1778*

17. C. B. Jabara, C. D. Jones, J. Roach, J. A. Anderson, R. Swanstrom,*Proc. Natl. Acad. Sci. U. S. A.* **2011** , *108* , 20166–20171*https://doi.org/10.1073/pnas.1110064108*

18. Q. Peng, R. Vijaya Satya, M. Lewis, P. Randad, Y. Wang, *BMC Genomics* **2015** , *16* , 589*https://doi.org/10.1186/s12864-015-1806-8*

19. M. W. Schmitt, S. R. Kennedy, J. J. Salk, E. J. Fox, J. B. Hiatt, L. A. Loeb, *Proc. Natl. Acad. Sci . U. S. A.* **2012** ,*09* , Issue 36, pp. 14508–14513*https://doi.org/10.1073/pnas.1208715109*

20. J. B. Hiatt, R. P. Patwardhan, E. H. Turner, C. Lee, J. Shendure,*Nat. Methods* **2010** , *7* , Issue 2, pp. 119–122*https://doi.org/10.1038/nmeth.1416*

21. G. K. Fu, J. Hu, P.-H. Wang, S. P. A. Fodor, *Proc. Natl. Acad. Sci. U. S. A.* **2011** , *108* , Issue 22, pp. 9026–9031*https://doi.org/10.1073/pnas.1017621108*

22. K. Shiroguchi, T. Z. Jia, P. A. Sims, X. S. Xie, *Proc. Natl. Acad. Sci. U. S. A.* **2012** , *109* , Issue 4, pp. 1347–1352*https://doi.org/10.1073/pnas.1118018109*

23. J. B. Hiatt, C. C. Pritchard, S. J. Salipante, B. J. O'Roak, J. Shendure, *Genome Res.* **2013** , *23* , Issue 5, pp. 843–854*https://doi.org/10.1101/gr.147686.112*

24. A. Varble, R. A. Albrecht, S. Backes, M. Crumiller, N. M. Bouvier, D. Sachs, A. García-Sastre, A., B. R. tenOever, *Cell Host & Microbe* **2014** , *16* , 691–700*https://doi.org/10.1016/j.chom.2014.09.020*

25. H.-C. Chen, J. P. Martinez, E. Zorita, A. Meyerhans, G. J. Filion,*Nat. Struct. Mol. Biol.* **2017** , *24* , Issue 1, pp. 47–54*https://doi.org/10.1038/nsmb.3328*

26. M. T. Aliota, D. M. Dudley, C. M. Newman, J. Weger-Lucarelli, L. M. Stewart, M. R. Koenig, M. E. Breitbach, A. M. Weiler, M. R. Semler, G. L. Barry, K. R. Zarbock, A. K. Haj, R. V. Moriarty, M. S. Mohns, E. L. Mohr, V. Venturi, N. Schultz-Darken, E. Peterson, W. Newton, M. L. Schotzko, H. A. Simmons, A. Mejia, J. M. Hayes, S. 3rd Capuano, M. P. Davenport, T. C. Friedrich, G. D. Ebel, S. L. O'Connor, D. H. O'Connor,*PLoS Pathog.* **2018** , *14* , e1006964*https://doi.org/10.1371/journal.ppat.1006964*

27. K. U. Ludwig, R. M. Schmithausen, D. Li, M. L. Jacobs, R. Hollstein, K. Blumenstock, J. Liebing, M. Słabicki, A. Ben-Shmuel, O. Israeli, S. Weiss, T. S. Ebert, N. Paran, W. Rüdiger, G. Wilbring, D. Feldman, B. Lippke, N. Ishorst, L. M. Hochfeld, E. C. Beins, I. H. Kaltheuner, M. Schmitz, A. Wöhler, M. Döhla, E. Sib, M. Jentzsch, J. D. Borrajo, J. Strecker, J. Reinhardt, B. Cleary, M. Geyer, M. Hölzel, R. Macrae, M. M. Nöthen, P. Hoffmann, M. Exner, A. Regev, F. Zhang, J. L. Schmid-Burgk,*Nat. Biotechnol.* **2012** , *39* **,** 1556–1562*https://doi.org/10.1038/s41587-021-00966-9*

28. K. Adachi, T. Enoki, Y. Kawano, M. Veraz, H. Nakai, *Nat. Commun.* **2014** , *5* , 3075*https://doi.org/10.1038/ncomms4075*

29. E. Domingo, *Vet. Res* . **2010** , *41* , 38.*https://doi.org/10.1051/vetres/2010010*

30. E. Domingo, J. J. Holland, *Annu. Rev. Microbiol.***1997** , *51* , 151–178*https://doi.org/10.1146/annurev.micro.51.1.151*

31. A. R. Wargo, G. Kurath, *Curr. Opin. Virol.* **2012** ,*2* , 538–545*https://doi.org/10.1016/j.coviro.2012.07.007*

32. E. H. Turner, C. Lee, S. B. Ng, D. A. Nickerson, J. Shendure,*Nat. Methods* **2009** , *6* , Issue 5, pp. 315–316*https://doi.org/10.1038/nmeth.f.248*

33. P. Shen, W. Wang, S. Krishnakumar, C. Palm, A.-K. Chi, G. M. Enns, R. W. Davis, T. P. Speed, M. N. Mindrinos, C. Scharfe, *Proc. Natl. Acad. Sci. U. S. A.* **2011** , *108* , 6549–6554*https://doi.org/10.1073/pnas.1018981108*

34. B. J. O'Roak, L. Vives, W. Fu, J. D. Egertson, I. B. Stanaway, I. G. Phelps, G. Carvill, A. Kumar, C. Lee, K. Ankenman, J. Munson, J. B. Hiatt, E. H. Turner, R. Levy, D. R. O'Day, N. Krumm, B. P. Coe, B. K. Martin, E. Borenstein, D. A. Nickerson, H. C. Mefford, D. Doherty, J. M. Akey, R. Bernier, E. E. Eichler, J. Shendure, *Science***2012** , *338* , 1619–1622*https://doi.org/10.1126/science.1227764*

35. P. Nguyen, J. Ma, D. Pei, C. Obert, C. Cheng, T. L. Geiger,*BMC Genomics* **2011** , *12* , Issue 1*https://doi.org/10.1186/1471-2164-12-106*

36. R. W. Hamming, *The Bell System Technical Journal***1950** , *29* , Issue 2, pp. 147–160*https://doi.org/10.1002/j.1538-7305.1950.tb00463.x*

37. M. Hamady, J. J. Walker, J. Kirk Harris, N. J. Gold, R. Knight,*Nat. Methods* **2008** , *5* , Issue 3, pp. 235–237).*https://doi.org/10.1038/nmeth.1184*

38. V. I. Levenshtein, *Sov. Phys.-Dokl.* **1966** *, 10* , no. 8, pp. 707–710

39. COVID-19 situation update worldwide, as of week 18, updated 12 May 2022. (n.d.). Retrieved May 14, 2022, from*https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases*

40. D. P. Oran, E. J. Topol, *Ann. Intern. Med.* **2021** ,*174* , 1344-1345*https://doi.org/10.7326/M20-6976*

41. D. Berger, K. Herkenhoff, S. Mongey, *NBER Working Paper No. 26901* 2020*https://doi.org/10.3386/w26901*

42. J. Taipale, P. Romer, S. Linnarsson, *medRxiv* **2020***https://doi.org/10.1101/2020.04.27.20078329*

43. C. M. Peak, R. Kahn, Y. H. Grad, L. M. Childs, R. Li, M. Lipsitch, C. O. Buckee, *Lancet Infect. Dis.* **2020** , *20* , Issue 9, pp. 1025–1033*https://doi.org/10.1016/s1473-3099(20)30361-3*

44. T. Notomi, H. Okayama, H. Masubuchi, T. Yonekawa, K. Watanabe, N. Amino, T. Hase, *Nucleic Acids Res.* 2020, 28, E63*https://doi.org/10.1093/nar/28.12.e63*

45. K. Nagamine, T. Hase, T. Notomi, *Mol. Cell. Probes***2002** , *16* , 223–229*https://doi.org/10.1006/mcpr.2002.0415*

46. P. F. N. Estrela, G. de Melo Mendes, K. G. de Oliveira, A. M. Bailão, C. M. de Almeida Soares, N. A. Assunção, G. R. M. Duarte,*J. Virol. Methods* **2019** , *271* , p. 113675*https://doi.org/10.1016/j.jviromet.2019.113675*

47. M. W. Schmitt, S. R. Kennedy,J. J. Salk, E. J. Fox, J. B. Hiatt, L. A. Loeb, *Proc. Natl. Acad. Sci. U. S. A.* **2012** ,*109* , Issue 36, pp. 14508–14513*https://doi.org/10.1073/pnas.1208715109*

48. D. Thompson, Y. Lei, *Sens. Actuators Rep.* **2020** ,*2* , Issue 1, p. 100017*https://doi.org/10.1016/j.snr.2020.100017*

49. B. H. Bloom, *Commun. ACM* **1970** , *13* , Issue 7, pp. 422–426*https://doi.org/10.1145/362686.362692*

50. J. S. Bloom, E. M. Jones, M. Gasperini, N. B. Lubock, L. Sathe, C. Munugala, A. S. Booeshaghi, O. F. Brandenberg, L. Guo, J. Boocock, S. W. Simpkins, I. Lin, N. LaPierre, D. Hong, Y. Zhang, G. Oland, B. J. Choe, S. Chandrasekaran, E. E. Hilt, M. J. Butte, R. Damoiseaux, C. Kravit, A. R. Cooper, Y. Yin, L. Pachter, O. B. Garner, J. Flint, E. Eskin, C. Luo, S. Kosuri, L. Kruglyak, V. A. Arboleda, *medRxiv* , **2020** ,*https://doi.org/10.1101/2020.08.04.20167874*

51. J. S. Bloom, L. Sathe, C. Munugala, E. M. Jones, M. Gasperini, N. B. Lubock, F. Yarza, E. M. Thompson, K. M. Kovary, J. Park, D. Marquette, S. Kay, M. Lucas, T. Love, A. Sina Booeshaghi, O. F. Brandenberg, L. Guo, J. Boocock, M. Hochman, S. W. Simpkins, I. Lin, N. LaPierre, D. Hong, Y. Zhang, G. Oland, B. J. Choe, S. Chandrasekaran, E. E. Hilt, M. J. Butte, R. Damoiseaux, C. Kravit, A. R. Cooper, Y. Yin, L. Pachter, O. B. Garner, J. Flint, E. Eskin, C. Luo, S. Kosuri, L. Kruglyak, V. A. Arboleda, *Nat. Biomed. Eng.* **2021** , *5* , 657–665*https://doi.org/10.1038/s41551-021-00754-5*

52. L. Yujian, L. Bo, *IEEE Trans. Pattern Anal. Mach. Intell.***2007** , *29* , Issue 6, pp. 1091–1095*https://doi.org/10.1109/tpami.2007.1078*

53. N. E. Freed, M. Vlková, M. B. Faisal, O. K. Silander, *Biol. Methods Protoc.* **2020** , *5https://doi.org/10.1093/biomethods/bpaa014*

54. P. James, D. Stoddart, E. D. Harrington, J. Beaulaurier, L. Ly, S. W. Reid, D. J. urner, S. Juul, *medRxiv* **2020***https://doi.org/10.1101/2020.08.07.20161737*

55. M.-M. Aynaud, J. J. Hernandez, S. Barutcu, U. Braunschweig, K. Chan, J. D. Pearson, D. Trcka, S. L. Prosser, J. Kim, M. Barrios-Rodiles, M. Jen, S. Song, J. Shen, C. Bruce, B. Hazlett, S. Poutanen, L. Attisano, R. Bremner, B. J. Blencowe, T. Mazzulli, H. Han, L. Pelletier, J. L. Wrana, *Nat. Commun.* **2021** , *12* , 1405*https://doi.org/10.1038/s41467-021-21653-y*

56. P. M. Sharp, B. H. Hahn, *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **2010** , *365* , Issue 1552, pp. 2487–2494*https://doi.org/10.1098/rstb.2010.0031*

57. J. W. B. van der Giessen, L. D. Isken, E. W. Tiemersma, *The National Institute for Public Health and the Environmen* **2004***https://www.rivm.nl/bibliotheek/rapporten/330200002.pdf*

58. HIV/AIDS. (n.d.). Retrieved May 15, 2022, from*https://www.who.int/data/gho/data/themes/hiv-aids*

59. Global HIV & AIDS statistics — 2020 fact sheet. (n.d.). Retrieved May 15, 2022, from*https://www.unaids.org/en/resources/fact-sheet*

60. H.-C. Chen, E. Zorita, E., G. J. Filion, *Curr. Protoc. Mol. Biol.* **2018** , *122* , Issue 1*https://doi.org/10.1002/cpmb.56*

61. M. D. Marsden, T.-H. Zhang, Y. Du, M. Dimapasoc, M. S. A. Soliman, X. Wu, J. T. Kim, A. Shimizu, A. Schrier, P. A. Wender, R. Sun, J. A. Zack, *Cell Rep. Med.* **2020** , *1* , 100162*https://doi.org/10.1016/j.xcrm.2020.100162*

62. A. Ali, O. O. Yang, *J. Virol. Methods* **2006** ,*133* , Issue 1, pp. 41–47*https://doi.org/10.1016/j.jviromet.2005.10.016*

63. C. M. Fennessey, M. Pinkevych, T. T. Immonen, A. Reynaldi, V. Venturi, P. Nadella, C. Reid, L. Newman, L. Lipkey, K. Oswald, W. J. Bosche, M. T. Trivett, C. Ohlen, D. E. Ott, J. D. Estes, G. Q. Del Prete, J. D. Lifson,M. P. Davenport, B. F. Keele, *PLoS Pathog* .**2017** , *13* , e1006359*https://doi.org/10.1371/journal.ppat.1006359*

64. M. Woolhouse, E. Gaunt, *Crit. Rev. Microbiol.* **2007** ,*33* , 231–242*https://doi.org/10.1080/10408410701647560*

65. A. Varble, M. A. Chua, J. T. Perez, B. Manicassamy, A. Garcia-Sastre, B. R. tenOever, *Proc. Natl. Acad. Sci. U. S. A.***2010** , *107* , Issue 25, pp. 11519–11524*https://doi.org/10.1073/pnas.1003115107*

66. J. M. Silva, M. Z. Li, K. Chang, W. Ge, M. C. Golding, R. J. Rickles, D. Siolas, G. Hu, P. J. Paddison, M. R. Schlabach, N. Sheth, J. Bradshaw, J. Burchard, A. Kulkarni, G. Cavet, R. Sachidanandam, W. R. McCombie, M. A. Cleary, S. J. Elledge, G. J. Hannon, *Nat. Genet.***2005** , *37* , 1281–1288*https://doi.org/10.1038/ng1650*

67. A. B. Russell, C. Trapnell, J. D. Bloom, *eLife* **2018** ,*7https://doi.org/10.7554/eLife.32303*

68. A. M. Klein, L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D. A. Weitz, M. W. Kirschner, *Cell* **2015** ,*61* , Issue 5, pp. 1187–1201*https://doi.org/10.1016/j.cell.2015.04.044*

69. G. Kuno, G.-J. J. Chang, *Arch. Virol.* **2007** ,*152* , Issue 4, pp. 687–696*https://doi.org/10.1007/s00705-006-0903-z*

70. H. F. Günthard, J. K. Wong, C. C. Ignacio, D. V. Havlir, D. D. Richman, *AIDS Res. Hum. Retroviruses* **1998** ,*14* , Issue 10, pp. 869–876*https://doi.org/10.1089/aid.1998.14.869*

71. K. Van Laethem, K. Van Vaerenbergh, J.-C. Schmit, S. Sprecher, P. Hermans, V. De Vroey, R. Schuurman, T. Harrer, M. Witvrouw, E. Van Wijngaerden, L. Stuyver, M. Van Ranst, J. Desmyter, E. De Clercq, A.-M. Vandamme, *J. Acquir. Immune Defic. Syndr.* **1999** ,*22* , Issue 2, p. 107*https://doi.org/10.1097/00126334-199910010-00001*

72. L. Weinberger, R. Dar, M. Simpson, *Nat. Genet* . **2008** ,*40* , 466-70*https://doi.org/10.1038/ng.116*

73. B. S. Razooky, A. Pai, K. Aull, I. M. Rouzine, L. S. Weinberger,*Cell* **2015** , *160* , 990–1001*https://doi.org/10.1016/j.cell.2015.02.009*

74. H. J. Müller, *Journ. of Gen.* **1930** , *22* , Issue 3, pp. 299–334*https://doi.org/10.1007/bf02984195*

75. A. Jordan, *EMBO J.* **2001** , *20* , Issue 7, pp. 1726–1738*https://doi.org/10.1093/emboj/20.7.1726*

76. G. Vansant, H.-C. Chen, E. Zorita, K. Trejbalová, D. Miklík, G. Filion, Z. Debyser, *Nucleic Acids Res.* **2020** , *48* , Issue 14, pp. 7801–7817*https://doi.org/10.1093/nar/gkaa536*

77. F. Christ, A. Voet, A. Marchand, S. Nicolet, B. A. Desimmie, D. Marchand, D. Bardiot, N. J. Van der Veken, B. Van Remoortel, S. V. Strelkov, M. De Maeyer, P. Chaltin, Z. Debyser, *Nat. Chem. Biol.***2010** , *6* , 442–448*https://doi.org/10.1038/nchembio.370*

78. J. Demeulemeester, P. Chaltin, A. Marchand, M. De Maeyer, Z. Debyser, F. Christ, *Expert Opin. Ther. Pat.* **2014** ,*24* , Issue 6, pp. 609–632*https://doi.org/10.1517/13543776.2014.898753*

79. C. Gini, Concentration and dependency ratios (in Italian). English translation in *Rivista di Politica Economica* **1997** *, 87* , 769–789 (1909)

80. C. B. Brooke, W. L. Ince, J. Wei, J. R. Bennink, J. W. Yewdell,*Proc. Natl. Acad. Sci. U. S. A.* **2014** , *111* , 16854–16859*https://doi.org/10.1073/pnas.1415396111*

81. F. S. Heldt, S. Y. Kupke, S. Dorl, U. Reichl, T. Frensing,*Nat. Commun* . **2015** , *6* , Issue 1*https://doi.org/10.1038/ncomms9938*

82. D. Dou, I. Hernández-Neuta, H. Wang, H. Östbye, X. Qian, S. Thiele, P. Resa-Infante, N. M. Kouassi, V. Sender, K. Hentrich, P. Mellroth, B. Henriques-Normark, G. Gabriel, M. Nilsson, R. Daniels, *Cell Rep* .**2017** , *20* , Issue 1, pp. 251–263*https://doi.org/10.1016/j.celrep.2017.06.021*

83. T.-W. Chun,L. Stuyver, S. B. Mizell, L. A. Ehler, J. A. M. Mican, M. Baseler, A. L. Lloyd, M. A. Nowak, A. S. Fauci, *Proc. Natl. Acad. Sci. U. S. A.* **1997** , *94* , Issue 24, pp. 13193–13197*https://doi.org/10.1073/pnas.94.24.13193*

84. T.-W. Chun, R. T. Davey Jr, D. Engel, H.-C. Lane, A. S. Fauci,*Nature* **1999** , *401* , 874–875*https://doi.org/10.1038/44755*

85. D. Finzi, J. Blankson, J. D. Siliciano, J. B. Margolick, K. Chadwick, T. Pierson, K. Smith, J. Lisziewicz, F. Lori, C. Flexner, T. C. Quinn, R. E. Chaisson, E. Rosenberg, B. Walker, S. Gange, J. Gallant, R. F. Siliciano, *Nat. Med.* **1999** , *5* , Issue 5, pp. 512–517*https://doi.org/10.1038/8394*

86. N. M. Archin, A. L. Liberty, A. D. Kashuba, S. K. Choudhary, J. D. Kuruc, A. M. Crooks, D. C. Parker, E. M. Anderson, M. F. Kearney, M. C. Strain, D. D. Richman, M. G. Hudgens, R. J. Bosch, J. M. Coffin, J. J. Eron, D. J. Hazuda, D. M. Margolis, *Nature* **2012** ,*487* , 482–485*https://doi.org/10.1038/nature11286*

87. A. R. Cillo, M. D. Sobolewski, R. J. Bosch, E. Fyne, M. Piatak, J. M. Coffin, J. W. Mellors, *Proc. Natl. Acad. Sci. U. S. A.***2014** , *111* , Issue 19, pp. 7078–7083*https://doi.org/10.1073/pnas.1402873111*

88. C. K. Bullen, G. M. Laird, C. M. Durand, J. D. Siliciano, R. F. Siliciano, *Nat. Med.* **2014** , *20* , 425–429*https://doi.org/10.1038/nm.3489*

89. E. Battivelli, M. S. Dahabieh, M. Abdel-Mohsen, J. Peter Svensson, I. T. Da Silva, L. B. Cohn, A. Gramatica, S. Deeks, W. C. Greene, S. K. Pillai, E. Verdin, *eLife* **2018** , *7 https://doi.org/10.7554/elife.34655*

90. T. Rasmussen, M. Tolstrup, A. Winckelmann, L. Østergaard, O. S. Søgaard, *Hum. Vaccin. Immunother.* **2013** , *9* , Issue 4, pp. 790–799*https://doi.org/10.4161/hv.23202*

91. C. Banerjee, N. Archin, D. Michaels, A. C. Belkina, G. V. Denis, J. Bradner, P. Sebastiani, D. M. Margolis, M. Montano, *J. Leukoc. Biol.* **2012** , *92* , Issue 6, pp. 1147–1154*https://doi.org/10.1189/jlb.0312165*

92. D. Boehm, V. Calvanese, R. D. Dar, S. Xing, S. Schroeder, L. Martins, K. Aull, P.-C. Li, V. Planelles, J. E. Bradner, M.-M. Zhou, R. F. Siliciano, L. Weinberger, E. Verdin, M. Ott, *Cell Cycle***2013** , *12* , Issue 3, pp. 452–462*https://doi.org/10.4161/cc.23309*

93. P. Filippakopoulos, J. Qi, S. Picaud, Y. Shen, W. B. Smith, O. Fedorov, E. M. Morse, T. Keates, T. T. Hickman, I. Felletar, M. Philpott, S. Munro, M. R. McKeown, Y. Wang, A. L. Christie, N. West, M. J. Cameron, B. Schwartz, T. D. Heightman, N. La Thangue, C. A. French, O. Wiest, A. L. Kung, S. Knapp, J. E. Bradner, *Nature***2010** , *468* , Issue 7327, pp. 1067–1073*https://doi.org/10.1038/nature09504*

94. Z. Li, J. Guo, Y. Wu, Q. Zhou, *Nucleic Acids Res.***2013** , *41* , Issue 1, pp. 277–287*https://doi.org/10.1093/nar/gks976*

95. J. Zhu, G. D. Gaiha, S. P. John, T. Pertel, C. R. Chin, G. Gao, H. Qu, B. D. Walker, S. J. Elledge, A. L. Brass, *Cell Rep.***2012** , *2* , Issue 4, pp. 807–816*https://doi.org/10.1016/j.celrep.2012.09.008*

96. R. Mehla, S. Bivalkar-Mehla, R. Zhang, I. Handy, H. Albrecht, S. Giri, P. Nagarkatti, M. Nagarkatti, A. Chauhan, *PLoS ONE***2010** , *5* , Issue 6, p. e11160*https://doi.org/10.1371/journal.pone.0011160*

97. G. del Real, S. Jimenez-Baranda, E. Mira, R. A. Lacalle, P. Lucas, C. Gomez-Mouton, M. Alegret, J. M. Pena, M. Rodriguez-Zapata, M. Alvarez-Mon, C. Martinez-A., S. Manes, *J. Exp. Med.***2004** , *200* , Issue 4, pp. 541–547*https://doi.org/10.1084/jem.20040061*

98. M. D. Marsden, J. A. Zack, *Annu. Rev. Virol.* **2017** ,*4* , Issue 1, pp. 393–412*https://doi.org/10.1146/annurev-virology-101416-041703*

99. M. D. Marsden, B. A. Loy, X. Wu, C. M. Ramirez, A. J. Schrier, D. Murray, A. Shimizu, S. M. Ryckbosch, K. E. Near, T.-W. Chun, P. A. Wender, J. A. Zack, *PLoS Pathog* . **2017** , *13* , e1006575*https://doi.org/10.1371/journal.ppat.1006575*

100. G. M. Laird, C. Korin Bullen, D. I. S. Rosenbloom, A. R. Martin, A. L. Hill, C. M. Durand, J. D. Siliciano, R. F. Siliciano, *J. Clin. Invest.* **2015** , *125* , Issue 5, pp. 1901–1912*https://doi.org/10.1172/jci80142*

101. A. Shen, M. Christine Zink, J. L. Mankowski, K. Chadwick, J. B. Margolick, L. M. Carruth, M. Li, J. E. Clements, R. F. Siliciano,*J. Virol* . **2003** , *77* , Issue 8, pp. 4938–4949*https://doi.org/10.1128/jvi.77.8.4938-4949.2003*

102. J. B. Dinoso, S. Alireza Rabi, J. N. Blankson, L. Gama, J. L. Mankowski, R. F. Siliciano, M. Christine Zink, J. E. Clements, *J. Virol* . **2009** , *83* , Issue 18, pp. 9247–9257*https://doi.org/10.1128/jvi.00840-09*

103. J. B. Whitney, A. L. Hill, S. Sanisetty, P. Penaloza-MacMaster, J. Liu, M. Shetty, L. Parenteau, C. Cabral, J. Shields, S. Blackmore, J. Y. Smith, A. L. Brinkman, L. E. Peter, S. I. Mathew, K. M. Smith, E. N. Borducchi, D. I. S. Rosenbloom, M. G. Lewis, J. Hattersley, B. Li, J. Hesselgesser, R. Geleziunas, M. L. Robb, J. H. Kim, N. L. Michael, D. H. Barouch, *Nature* **2014** , *512* , Issue 7512, pp. 74–77*https://doi.org/10.1038/nature13594*

104. E. N. Borducchi, C. Cabral, K. E. Stephenson, J. Liu, P. Abbink, D. Ng'ang'a, J. P. Nkolola, A. L. Brinkman, L. Peter, B. C. Lee, J. Jimenez, D. Jetton, J. Mondesir, S. Mojta, A. Chandrashekar, K. Molloy, G. Alter, J. M. Gerold, A. L. Hill, M. G. Lewis, M. G. Pau, H. Schuitemaker, J. Hesselgesser, R. Geleziunas, J. H. Kim, M. L. Robb, N. L. Michael, D. H. Barouch, *Nature* **2016** , *540* , Issue 7632, pp. 284–287*https://doi.org/10.1038/nature20583*

105. G. Q. D. Prete, G. Q. Del Prete, J. D. Lifson, *Curr. Top. Microbiol. Immunol* . **2017** , *417* , pp. 69–109*https://doi.org/10.1007/82_2017_73*

106. S. Khanal, C. M. Fennessey, S. P. O'Brien, A. Thorpe, C. Reid, T. T. Immonen, R. Smith, J. W. Bess, A. E. Swanstrom, G. Q. Del Prete, M. P. Davenport, A. A. Okoye, L. J. Picker, J. D. Lifson, B. F. Keele,*J. Virol* . **2019** , *94* , Issue 1*https://doi.org/10.1128/jvi.01420-19*

107. T. T. Immonen, C. Camus, C. Reid, C. M. Fennessey, G. Q. Del Prete, M. P. Davenport, J. D. Lifson, B. F. Keele, *Proc. Natl. Acad. Sci. U. S. A.* **2020** , **117** , 494–502*https://doi.org/10.1073/pnas.1914967117*

108. M. Pinkevych, C. M. Fennessey, D. Cromer, C. Reid, C. M. Trubey, J. D. Lifson, B. F. Keele, M. P. Davenport, *eLife* **2019** ,*8https://doi.org/10.7554/eLife.49022*

109. S. Long, C. M. Fennessey, L. Newman, C. Reid, S. P. O'Brien, Y. Li, G. Q. Del Prete, J. D. Lifson, R. J. Gorelick, B. F. Keele, *J. Virol.* **2019** , *94https://doi.org/10.1128/JVI.01308-19*

110. T. T. Immonen, C. M. Fennessey, L. Lipkey, A. Thorpe, G. Q. Del Prete, J. D. Lifson, M. P. Davenport, B. F. Keele, *PLoS Pathog* .**2021** , *17* , e1009686*https://doi.org/10.1371/journal.ppat.1009686*

111. A. E. Swanstrom, T. T. Immonen, K. Oswald, C. Pyle, J. A. Thomas, W. J. Bosche, L. Silipino, M. Hull, L. Newman, V. Coalter, A. Wiles, R. Wiles, J. Kiser, D. R. Morcock, R. Shoemaker, R. Fast, M. W. Breed, J. Kramer, D. Donohue, T. Malys, C. M. Fennessey, C. M. Trubey, C. Deleage, J. D. Estes, J. D. Lifson, B. F. Keele, G. Q. Del Prete, *J. Clin. Invest.* **2021** , *131https://doi.org/10.1172/JCI142421*

112. W. H. Frost, *Public Health Reports (1896-1970)***1920** , *35* , Issue 11, p. 584*https://doi.org/10.2307/4575511*

113. N. P. A. S. Johnson, N. P. A. Johnson, J. Mueller, *Bull. Hist. Med* . **2002** , *76* , Issue 1, pp. 105–115*https://doi.org/10.1353/bhm.2002.0022*

114. R. G. Webster, W. J. Bean, O. T. Gorman, T. M. Chambers, Y. Kawaoka, *Microbiol. Rev.* **1992** , *56* , Issue 1, pp. 152–179*https://doi.org/10.1128/mr.56.1.152-179.1992*

115. K. F. Shortridge, N. N. Zhou, Y. Guan, P. Gao, T. Ito, Y. Kawaoka, S. Kodihalli, S. Krauss, D. Markwell, K. Gopal Murti, M. Norwood, D. Senne, L. Sims, A. Takada, R. G. Webster, *Virology* **1998** ,*252* , Issue 2, pp. 331–342*https://doi.org/10.1006/viro.1998.9488*

116. Y. Guan, J. S. M. Peiris, A. S. Lipatov, T. M. Ellis, K. C. Dyrting, S. Krauss, L. J. Zhang, R. G. Webster, K. F. Shortridge,*Proc. Natl. Acad. Sci. U. S. A* . **2002** , *99* , Issue 13, pp. 8950–8955*https://doi.org/10.1073/pnas.132268999*

117.L. D. Sims, T. M. Ellis, K. K. Liu, K. Dyrting, H. Wong, M. Peiris, Y. Guan, K. F. Shortridge, *Avian Dis* . **2003** , *47* , Issue s3, pp. 832–838*https://doi.org/10.1637/0005-2086-47.s3.832*

118. E. C. J. Claas, A. D. Osterhaus, R. van Beek, J. C. De Jong, G. F. Rimmelzwaan, D. A. Senne, S. Krauss, K. F. Shortridge, R. G. Webster,*Lancet* **1998** , *351* , Issue 9101, pp. 472–477*https://doi.org/10.1016/s0140-6736(97)11212-0*

119.K. Subbarao, *Science* **1998** , *279* , Issue 5349, pp. 393–396*https://doi.org/10.1126/science.279.5349.393*

120. Y. Guan, L. L. M. Poon, C. Y. Cheung, T. M. Ellis, W. Lim, A. S. Lipatov, K. H. Chan, K. M. Sturm-Ramirez, C. L. Cheung, Y. H. C. Leung, K. Y. Yuen, R. G. Webster, J. S. M. Peiris, *Proc. Natl. Acad. Sci. U. S. A* . **2004** , *101* , Issue 21, pp. 8156–8161*https://doi.org/10.1073/pnas.0402443101*

121. J. S. M. Peiris,W. C. Yu, C. W. Leung, C. Y. Cheung, W. F. Ng, J. M. Nicholls, T. K. Ng, K. H. Chan, S. T. Lai, W. L. Lim, K. Y. Yuen, Y. Guan, *Lancet* **2004** , *363* , Issue 9409, pp. 617–619*https://doi.org/10.1016/s0140-6736(04)15595-5*

122. T. H. Tran, T. L. Nguyen, T. D. Nguyen, T. S. Luong, P. M. Pham, van V. C. Nguyen, T. S. Pham, C. D. Vo, T. Q. M. Le, T. T. Ngo, B. K. Dao, P. P. Le, T. T. Nguyen, T. L. Hoang, V. T. Cao, T. G. Le, D. T. Nguyen, H. N. Le, K. T. Nguyen, H. S. Le, V. T. Le, D. Christiane, T. T. Tran, de J. Menno, C. Schultsz, P. Cheng, W. Lim, P. Horby, J. Farrar, World Health Organization International Avian Influenza Investigative Team, *N. Engl. J. Med* . **2004** *350* , 1179–1188*https://doi.org/10.1056/NEJMoa040419*

123. J. Parry, *BMJ* **2013** , *346* , f2222*https://doi.org/10.1136/bmj.f2222*

124. J. Parry, *BMJ* **2013** , *346* , Issue apr04 2, pp. f2151–f2151*https://doi.org/10.1136/bmj.f2151*

125. M. Pan, R. Gao, Q. Lv, S. Huang, Z. Zhou, L. Yang, X. Li, X. Zhao, X. Zou, W. Tong, S. Mao, S. Zou, H. Bo, X. Zhu, L. Liu, H. Yuan, M. Zhang, D. Wang, Z. Li, W. Zhao, M. Ma, Y. Li, T. Li, H. Yang, J. Xu, L. Zhou, X. Zhou, W. Tang, Y. Song, T. Chen, T. Bai, J. Zhou, D. Wang, G. Wu, D. Li, Z. Feng, G. F. Gao, Y. Wang, S. He, Y. Shu, *J. Infect* .**2016** , *72* , 52–59*https://doi.org/10.1016/j.jinf.2015.06.009*

126. Y. Bi, Q. Chen, Q. Wang, J. Chen, T. Jin, G. Wong, C. Quan, J. Liu, J. Wu, R. Yin, L. Zhao, M. Li, Z. Ding, R. Zou, W. Xu, H. Li, H. Wang, K. Tian, G. Fu, Y. Huang, A. Shestopalov, S. Li, B. Xu, H. Yu,

T. Luo, L. Lu, X. Xu, Y. Luo, Y. Liu, W. Shi, D. Liu, G. F. Gao, *Cell Host Microbe* **2016** , *20* , Issue 6, pp. 810–821*https://doi.org/10.1016/j.chom.2016.10.022*

127. L. Yang, X. Zhao, X. Li, H. Bo, D. Li, J. Liu, D. Wang,*Biosafety and Health* **2020** , *2* , Issue 1, pp. 49–52*https://doi.org/10.1016/j.bsheal.2020.02.003*

128. Y. Zheng, Y. Yang, W. Zhao, H. Wang, *Am. J. Trop. Med. Hyg* .**2010** , *82* , Issue 4, pp. 743–745*https://doi.org/10.4269/ajtmh.2010.09-0666*

129. A. Puvanalingam, C. Rajendiran, K. Sivasubramanian, S. Ragunanthanan, S. Suresh, S. Gopalakrishnan, *J. Assoc .Physicians India* **2011** , *59* , 14–16, 18

130. H. Sun, Y. Xiao, J. Liu, D. Wang, F. Li, C. Wang, C. Li, J. Zhu, J. Song, H. Sun, Z. Jiang, L. Liu, X. Zhang, K. Wei, D. Hou, J. Pu, Y. Sun, Q. Tong, Y. Bi, K.-C. Chang, S. Liu, G. F. Gao, J. Liu, *Proc. Natl. Acad. Sci. U. S. A* . **2020** , *117* , 17204–17210*https://doi.org/10.1073/pnas.1921186117*

131. Correction to Supporting Information for H. Sun, Y. Xiao, J. Liu, D. Wang, F. Li, C. Wang, C. Li, J. Zhu, J. Song, H. Sun, Z. Jiang, L. Liu, X. Zhang, K. Wei, D. Hou, J. Pu, Y. Sun, Q. Tong, Y. Bi, K.-C. Chang, S. Liu, G. F. Gao, J. Liu, *Proc. Natl. Acad. Sci. U. S. A* .**2020** , 117, 23194*https://doi.org/10.1073/pnas.2017619117*

132. A. L. Vincent, W. Ma, K. M. Lager, B. H. Janke, J. A. Richt,*Adv. Virus Res* . **2008** , *72* , 127–154*https://doi.org/10.1016/S0065-3527(08)00403-X*

133. Centers for Disease Control and Prevention (CDC), Swine-origin influenza A (H3N2) virus infection in two children–Indiana and Pennsylvania, July-August 2011. *MMWR. Morb. Mortal. Wkly. Rep.***2011** , *60* , 1213–1215*https://www.cdc.gov/mmwr/preview/mmwrhtml/mm6035a6.htm*

134. M. A. Jhung, S. Epperson, M. Biggerstaff, D. Allen, A. Balish, N. Barnes, A. Beaudoin, L. Berman, S. Bidol, L. Blanton, D. Blythe, L. Brammer, T. D'Mello, R. Danila, W. Davis, S. de Fijter, M. Diorio, L. O. Durand, S. Emery, B. Fowler, R. Garten, Y. Grant, A. Greenbaum, L. Gubareva, F. Havers, T. Haupt, J. House, S. Ibrahim, V. Jiang, S. Jain, D. Jernigan, J. Kazmierczak, A. Klimov, S. Lindstrom, A. Longenberger, P. Lucas, R. Lynfield, M. McMorrow, M. Moll, C. Morin, S. Ostroff, S. L. Page, S. Y. Park, S. Peters, C. Quinn, C. Reed, S. Richards, J. Scheftel, O. Simwale, B. Shu, K. Soyemi, J. Stauffer, C. Steffens, S. Su, L. Torso, T. M. Uyeki, S. Vetter, J. Villanueva, K. K. Wong, M. Shaw, J. S. Bresee, N. Cox, L. Finelli, *Clin. Infect. Dis.***2013** , *57* , 1703–1712*https://doi.org/10.1093/cid/cit649*

135. W. M. Fitch, J. M. Leiter, X. Q. Li, P. Palese, *Proc. Natl. Acad. Sci. U. S. A* . **1991** , *88* , 4270–4274*https://doi.org/10.1073/pnas.88.10.4270*

136. C. A. Russell, J. M. Fonville, A. E. X. Brown, D. F. Burke, D. L. Smith, S. L. James, S. Herfst, S. van Boheemen, M. Linster, E. J. Schrauwen, L. Katzelnick, A. Mosterin, T. Kuiken, E. Maher, G. Neumann, A. D. M. Osterhaus, Y. Kawaoka, R. A. M. Fouchier, D. J. Smith,*Science* **2012** , *336* , Issue 6088, pp. 1541–1547*https://doi.org/10.1126/science.1222526*

137. R. Muñoz-Moreno, C. Martínez-Romero, D. Blanco-Melo, C. V. Forst, R. Nachbagauer, A. A. Benitez, I. Mena, S. Aslam, V. Balasubramaniam, I. Lee, M. Panis, J. Ayllón, D. Sachs, M.-S. Park, F. Krammer, B. R. tenOever, A. García-Sastre, *Cell Rep* . **2019** , *29* , 3997–4009.e5*https://doi.org/10.1016/j.celrep.2019.11.070*

138. G. W. A. Dick, *Trans. R. Soc. Trop. Med. Hyg* . **1952** ,*46* , Issue 5, pp. 521–534*https://doi.org/10.1016/0035-9203(52)90043-6*

139. G. W. A. Dick, S. F. Kitchen, A. J. Haddow, *Trans. R. Soc. Trop. Med. Hyg* . **1952** , 46, Issue 5, pp. 509–520*https://doi.org/10.1016/0035-9203(52)90042-4*

140. World Health Organization: WHO. Zika virus. *World Health Organization: WHO* **2018** , July 20*https://www.who.int/news-room/fact-sheets/detail/zika-virus*

141. W. G. C. Bearcroft, *Trans. R. Soc. Trop. Med. Hyg* .**1956** , 50, Issue 5, pp. 438–441*https://doi.org/10.1016/0035-9203(56)90090-6*

142. D. I. H. Simpson, *Trans. R. Soc. Trop. Med. Hyg* .**1964** , 58, Issue 4, pp. 339–348*https://doi.org/10.1016/0035-9203(64)90201-9*

143. A. H. Fagbami, *J. Hyg. (Lond)* . **1979** , *83* , Issue 2, pp. 213–219*https://doi.org/10.1017/s0022172400025997*

144. J. G. Olson, T. G. Ksiazek, Suhandiman, Triwibowo, *Trans. R. Soc. Trop. Med. Hyg* . **1981** , 75, Issue 3, pp. 389–393*https://doi.org/10.1016/0035-9203(81)90100-0*

145. M. R. Reynolds, *MMWR. Morb. Mortal. Wkly. Rep.***2017** , *66https://doi.org/10.15585/mmwr.mm6613e1*

146. A. J. Rolfe, D. B. Bosco, J. Wang, R. S. Nowakowski, J. Fan, Y. Ren, *Cell Biosci* . **2016** , 6, 42*https://doi.org/10.1186/s13578-016-0110-x*

147. K. Himmelsbach, E. Hildt, *World J. Virol* . **2018** ,*7* , Issue 1, pp. 10–20*https://doi.org/10.5501/wjv.v7.i1.10*

148. S. M. Nguyen, K. M., Antony, D. M., Dudley, S. Kohn, H. A. Simmons, B. Wolfe, M. Shahriar Salamat, L. B. C. Teixeira, G. J. Wiepz, T. H. Thoong, M. T. Aliota, A. M. Weiler, G. L. Barry, K. L. Weisgrau, L. J. Vosler, M. S. Mohns, M. E. Breitbach, L. M. Stewart, M. N. Rasheed, C. M. Newman, M. E. Graham, O. E. Wieben, P. A. Turski, K. M. Johnson, J. Post, J. M. Hayes, N. Schultz-Darken, M. L. Schotzko, J. A. Eudailey, S. R. Permar, E. G. Rakasz, E. L. Mohr, S. Capuano 3rd, A. F. Tarantal, J. E. Osorio, S. L. O'Connor, T. C. Friedrich, D. H. O'Connor, T. G. Colos,*PLoS Pathog* . **2017** , *13* , Issue 5, p. e1006378*https://doi.org/10.1371/journal.ppat.1006378*

149. G. Calvet, R. S. Aguiar, A. S. O. Melo, S. A. Sampaio, I. de Filippis, A. Fabri, E. S. M. Araujo, P. C. de Sequeira, M. C. L. de Mendonça, L. de Oliveira, D. A. Tschoeke, C. G. Schrago, F. L. Thompson, P. Brasil, F. B. dos Santos, R. M. R. Nogueira, A. Tanuri, A. M. B. de Filippis, *Lancet Infect. Dis* . **2016** , *16* , Issue 6, pp. 653–660*https://doi.org/10.1016/s1473-3099(16)00095-5*

150. N. S. D. Carvalho, N. S. De Carvalho, B. F. De Carvalho, C. A. Fugaça, B. Dóris, E. S. Biscaia, *Braz. J. Infect. Dis* .**2016** , *20* , Issue 3, pp. 282–289*https://doi.org/10.1016/j.bjid.2016.02.006*

151. J. Mlakar, M. Korva, N. Tul, M. Popović, M. Poljšak-Prijatelj, J. Mraz, M. Kolenc, K. R. Rus, T. V. Vipotnik, V. F. Vodušek, A. Vizjak, J. Pižem, M. Petrovec, T. A. Županc, *N. Engl. J. Med.* **2016** ,*374* , Issue 10, pp. 951–958*https://doi.org/10.1056/nejmoa1600651*

152. K. A. Stapleford, L. L. Coffey, S. Lay, A. V. Bordería, V. Duong, O. Isakov, K. Rozen-Gagnon, C. Arias-Goeta, H. Blanc, S. Beaucourt, T. Haliloğlu, C. Schmitt, I. Bonne, N. Ben-Tal, N. Shomron, A.-B. Failloux, P. Buchy, M. Vignuzzi, *Cell Host Microbe* **2014** ,*15* , 706–716*https://doi.org/10.1016/j.chom.2014.05.008*

153. P. Rossolillo, F. Winter, E. Simon-Loriere, S. Gallois-Montbrun, M. Negroni, *PLoS Genet* . **2012** , *8* , Issue 8, p. e1002904*https://doi.org/10.1371/journal.pgen.1002904*

154. S. Herfst, E. J. A. Schrauwen, M. Linster, S. Chutinimitkul, E. de Wit, V. J. Munster, E. M. Sorrell, T. M. Bestebroer, D. F. Burke, D. J. Smith, G. F. Rimmelzwaan, A. D. M. E. Osterhaus, R. A. M. Fouchier,*Science* **2012** , *336* , 1534–1541*https://doi.org/10.1126/science.1213362*

155. M. Imai, T. Watanabe, M. Hatta, S. C. Das, M. Ozawa, K. Shinya, G. Zhong, A. Hanson, H. Katsura, S. Watanabe, C. Li, E. Kawakami, S. Yamada, M. Kiso, Y. Suzuki, E. A. Maher, G. Neumann, Y. Kawaoka,*Nature* **2012** , *486* , 420–428*https://doi.org/10.1038/nature10831*

156. A. B. Allison, D. J. Kohler, A. Ortega, E. A. Hoover, D. M. Grove, E. C. Holmes, C. R. Parrish, *PLoS Pathog* . **2014** ,*10* , e1004475*https://doi.org/10.1371/journal.ppat.1004475*

157. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, *Advances in neural information processing systems* **2014** .

158. R. P. Patwardhan, C. Lee, O. Litvin, D. L. Young, D. Pe'er, J. Shendure, *Nat. Biotechnol* . **2009** , *27* , Issue 12, pp. 1173–1175*https://doi.org/10.1038/nbt.1589*

159. M. Davidsson, P. Díaz-Fernández, M. Torroba, O. D. Schwich, P. Aldrin-Kirk, L. Quintino, A. Heuer, G. Wang, C. Lundberg, T. Björklund,*RNA* **2018** , *24* , Issue 5, pp. 673–687*https://doi.org/10.1261/rna.063925.117*

160. M. A. Smith, T. Ersavas, J. M. Ferguson, H. Liu, M. C. Lucas, O. Begik, L. Bojarski, K. Barton, E. M. Novoa, *Genome Res* .**2020** , *30* , 1345–1353*https://doi.org/10.1101/gr.260836.120*

161. A. Goffeau, FEBS Lett. 2000, 480(1), 37–41*https://doi.org/10.1016/s0014-5793(00)01775-0*

162. M. Davidsson, P. Diaz-Fernandez, O. D. Schwich, M. Torroba, G. Wang, T. Björklund, *Sci. Rep* . **2016** , *6* , 37563*https://doi.org/10.1038/srep37563*

163. J. Weinmann, S. Weis, J. Sippel, W. Tulalamba, A. Remes, J. El Andari, A.-K. Herrmann, Q. H. Pham, C. Borowski, S. Hille, T. Schönberger, N. Frey, M. Lenter, T. VandenDriessche, O. J. Müller, M. K. Chuah, T. Lamla, D. Grimm, *Nat. Commun* . **2020** , *11* , 5432*https://doi.org/10.1038/s41467-020-19230-w*

164. G. Tetz, V. Tetz, *Microorganisms* **2018** , *6* , Issue 2, p. 54*https://doi.org/10.3390/microorganisms6020054*

165. G. Tetz, V. Tetz, *Gut Pathog* . **2016** , *8* , Issue 1*https://doi.org/10.1186/s13099-016-0109-1*

166. G. V. Tetz, K. V. Ruggles, H. Zhou, A. Heguy, A. Tsirigos, V. Tetz,*Sci. Rep.* **2017** , *7* , Issue 1*https://doi.org/10.1038/s41598-017-07278-6*

167.M. Łusiak-Szelachowska, B. Weber-Dąbrowska, E. Jończyk-Matysiak, R. Wojciechowska, A. Górski, *Gut Pathog.* **2017** , *9* , 44*https://doi.org/10.1186/s13099-017-0196-7*

168. B. T. McCune, M. R. Lanahan, B. R. tenOever, J. K. Pfeiffer,*J. Virol.* **2020** , *94https://doi.org/10.1128/JVI.01590-19*

169. S. K. Kuss, C. A. Etheredge, J. K. Pfeiffer, *PLoS Pathog* .**2008** , *4* , e1000082*https://doi.org/10.1371/journal.ppat.1000082*

170. A. S. Lauring, R. Andino, *J. Virol* . **2011** ,*85* , 3780–3791*https://doi.org/10.1128/JVI.02217-10*

171. A. K. Erickson, J. K. Pfeiffer, *J. Virol* . **2013** ,*87* , 12392–12397*https://doi.org/10.1128/JVI.02149-13*

172. J. Weger-Lucarelli, S. M. Garcia, C. Rückert, A. Byas, S. L. O'Connor, M. T. Aliota, T. C. Friedrich, D. H. O'Connor, G. D. Ebel,*Virology* **2018** , *521* , 138–148*https://doi.org/10.1016/j.virol.2018.06.004*

Tables

Table 1. Virus studies employed molecular barcodes

Virus	Viral genome	Disease	Zoonotic diseases	Main reservoirs	Usual mode of transmission to humans	Studies employed molecular barcodes
Adeno-associated viruses (AAV)	Linear, single-stranded DNA of about 48,000 nucleotides	AAV are not currently known to cause disease	No	Humans	Direct contact	[28,159,162,163]
Bacteriophage[a]	Linear, double-stranded DNA of about 44,000 nucleotides (SP6), 38,000 nucleotides (T3) and 40,000 nucleotides (T7)	The concept that bacteriophages as human pathogens has been proposed[164] Bacterio-phages may trigger and worsen a number of human diseases[165–167]	No	Bacteria	Direct contact	[158]
Human enterovirus (HEV)[b]	Linear, single-stranded RNA of about 7,400 nucleotides (coxsackievirus B3)	Human myocarditis or dilated cardiomyopathy	No	Humans	The oral-fecal route	[168]
HIV	Linear, single-stranded RNA of about 9,700 nucleotides	AIDS	Yes	Humans, chim-panzees, gorillas[c]	Direct contact with certain body fluids (blood, semen and pre-seminal fluid, rectal fluids, vaginal fluids, breast milk)	[17,25,61]

Virus	Viral genome	Disease	Zoonotic diseases	Main reservoirs	Usual mode of transmission to humans	Studies employed molecular barcodes
Influenza A virus	Linear, single-stranded RNA of about 14,000 nucleotides (Influenza A virus)	Influenza	Yes	Wild birds, poultry, swine	Aerosols, faeces	[24,67]
Poliovirus	Single-stranded RNA of about 7,500 nucleotides	Poliomyelitis, infantile paralysis	No	Humans	The oral-fecal route	[169,170]d
SARS-CoV-2	Single-stranded RNA of about 30,000 nucleotides	SARS	Yes	Bats	Direct, indirect, or close contact with infected people Airborne transmission	[27,51,53–55]
SIV	Linear, single-stranded RNA of about 9,500 nucleotides	-	No	Chimpanzees, sooty mangabeys	Bites and scratches	[63,107,108,110,111]
Yellow fever virus (YFV)	Linear, single-stranded RNA of about 11,000 nucleotides	Yellow fever, yellow jack, yellow plague	Yes	Humans, monkeys[e]	Mosquitoes	[171]f
Zika virus	Linear, single-stranded RNA of about 11,000 nucleotides	Mild fever and rash, with possible links to microcephaly and Guillain Barré syndrome	Yes	Humans, other primates	Mosquitoes	[26,172]

[a)] This indicated study, in which the molecular barcoding strategy was employed, focused on the bacteriophage promoters SP6 (SP6p32), T3 (class 3, phi13) and T7 (class 3, phi10)[158].

[b)]This indicated study, in which the molecular barcoding strategy was employed, focused on coxsackievirus B3 (CVB3)[168].

[c)]Chimpanzees and gorillas are the original but humans are now the primary reservoir.

[d)]The barcoding strategy used by Kuss et al. (2008)[169]relied on site-directed mutagenesis (point mutations) on the viral genome to distinguish each of the ten viruses from one another.

[e)]Monkeys are the original but humans are now the primary reservoir.

[f)]The barcoding strategy used by Erickson and Pfeiffer (2013)[171]relied on site-directed mutagenesis (point mutations) on the viral genome to distinguish each of the ten viruses from one another.

Figure Legends

Figure 1. Schematic representation of three principal approaches to prepare barcoded libraries. (A, C and D) Molecular barcodes can be introduced to a template by ligating sequencing adaptors (A) or hybridizing molecular inversion probes (C) or directly given by PCR amplification with target-specific primers (D). Of note, both forward- and reverse adaptors were synthesized followed by adaptor extension and A tailing. (B) Schematic representation of the different applications between the usage of molecular barcodes and sample barcodes. Molecular barcodes used in pooled sample 1 aims to correct sequencing errors: a misreading nucleotide, guanosine (G) for example, is corrected in final consensus sequences. Molecular barcodes used in pooled sample 2 aims to identify true mutations (marked by red triangles); a mistaken mutation (marked by a blue triangle) is eventually removed in final consensus sequences. Panel (A) is modified based on Figure 1 in Schmitt et al. (2012)[19]; panel (C) is modified based on Figure 1 in Hiatt et al. (2013)[23].

Figure 2. Schematic representation of the molecular barcoding strategies used for a population-scale testing to screen SARS-CoV-2-infected individuals. (A) A sequence of 10 base pairs DNA barcodes named LAMP barcodes was incorporated in the FIP primer while performing RT-PCR. PCR barcodes adjacent to Illumina P5 and P7 sequences flanked the both ends of the library. Annotated amplicon sequence is modified based on Figure 1b in Ludwig et al. (2021)[27]. (B) Two sets of unique barcodes named i5 and i7 sample barcodes were placed adjacent to the P5 and P7 adaptors in Illumina sequence primers in the stage of PCR amplification. Illustration is modified based on Figure 1b in Bloom et al. (2021)[51].

Figure 3. Schematic representation of the molecular barcoding strategies applied in HIV and SIV. (A) A swarm of cDNA synthesis primers containing a string of eight degenerate nucleotides named Primer ID and a three nucleotides sample barcode were used to PCR amplify the HIV-1 protease (*pro*) gene. (B) A sequence of 20-nucleotides molecular barcodes was used to tag the region downstream the HIV 5' long terminal repeat in the HIV-based vector. After infection, viral DNA containing molecular barcodes were inserted in the host genome. Inverse PCR performed on genome DNA isolated from the infected cells identifies provirus insertion sites; RT-PCR performed mRNA of barcoded proviruses measures viral transcription driven by the HIV 5' long terminal repeat. (C) The genomic characterization of nearly full-length HIV and the composition of the molecular barcode sequence. A 21 nucleotides barcode sequence was inserted in a non-expressed region upstream of the HA tag overlapped with the HIV *vpr* gene in a nearly full-length and replication-competent HIV genome. The original sequence in the barcode region was given in this illustration. Each third nucleotide was replaced by a thymidine in the sequence of barcodes. Pink stick marks the region where a molecular barcode is inserted. Illustration is modified based on Figure 1A in Marsden et al. (2020)[61]. (D) The genomic characterization of SIV and the composition of the molecular barcode sequence. Molecular barcodes encompassing 10 random nucleotides in length were inserted between the stop codon of the SIV*vpx* gene and the start codon of the SIV *vpr* gene in the SIVmac239 plasmid. Illustration is modified based on Figure 1A in Fennessey et al. (2017)[63].

Figure 4. Schematic representation of the molecular barcoding strategies applied in Influenza A virus. (A) A string of 22 nucleotides molecular barcodes were carried by a shRNA library

with amplification[66]. Amplified products encompassing molecular barcodes were inserted between the Influenza A virus genes encoding NS1 and NEP, which have been manipulated in their previous work[65]. Illustration is modified based on Figure 1A in Varble et al. (2014)[24]. (B) Three sorts of barcodes, including cell barcodes, UMI and viral barcodes were applied on viral mRNA to measure single mRNA transcript in cells infected with Influenza A viruses. Annotated amplicon was illustrated based on Figure 1C in Russell et al. (2018)[67].

Figure 5. Schematic representation of the genomic characterization of Zika virus and the composition of the molecular barcode sequence. Molecular barcode consisting of eight degenerate codons (24 nucleotides) was embedded into the gene encoding the NS2A protein. Pink stick marks the region where a molecular barcode is inserted. Nucleotides written in pink color are referred to as degenerate nucleotides.

Figure 6. Schematic representation of the rationale design of experimental evolution models proposed in this review article. Proposed experimental evolution models are composed of two parts: the reservoir of natural host cells/animals used to experimentally generate a swarm of laboratory-produced variant strains and the validation stage, in which we will predict phenotypes that are caused by genotypic changes and survey their impacts on human cells at a single-sequence level.

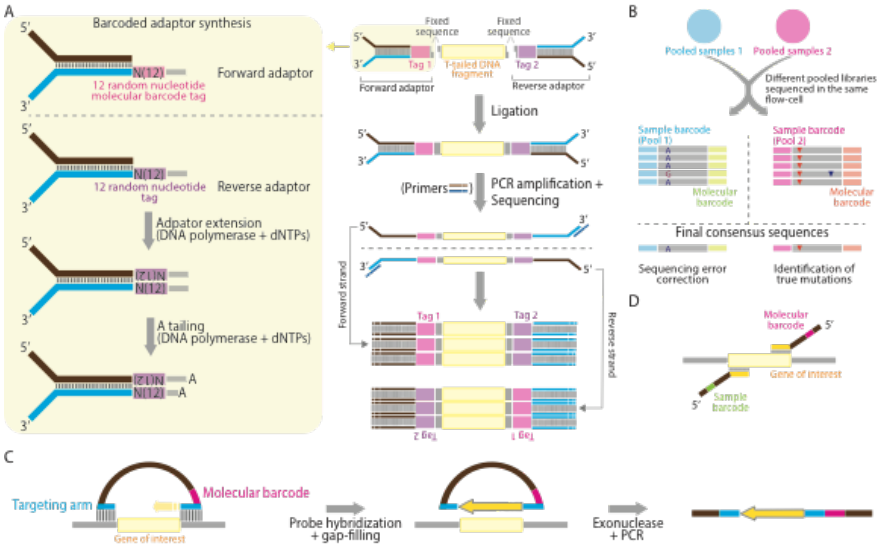

Figure 1

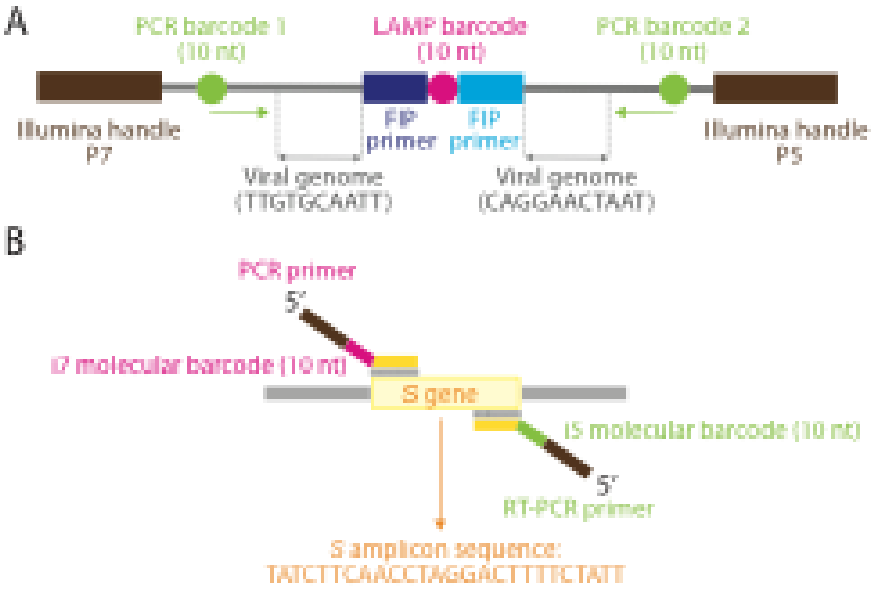

Figure 2

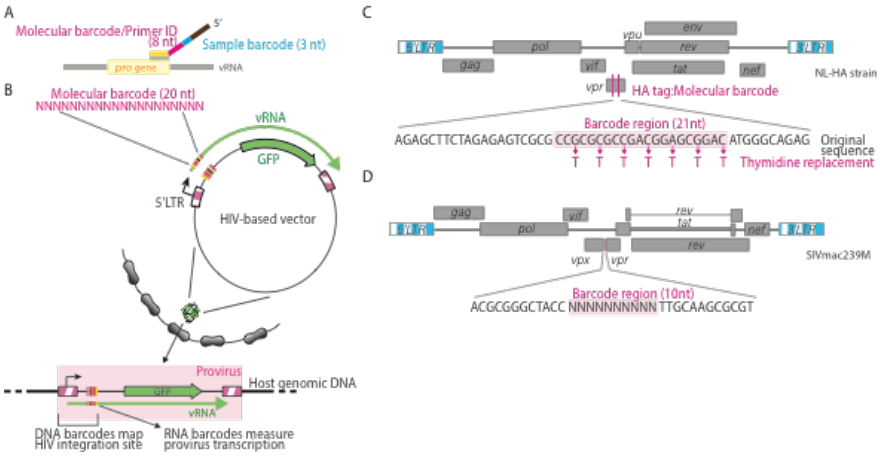

Figure 3

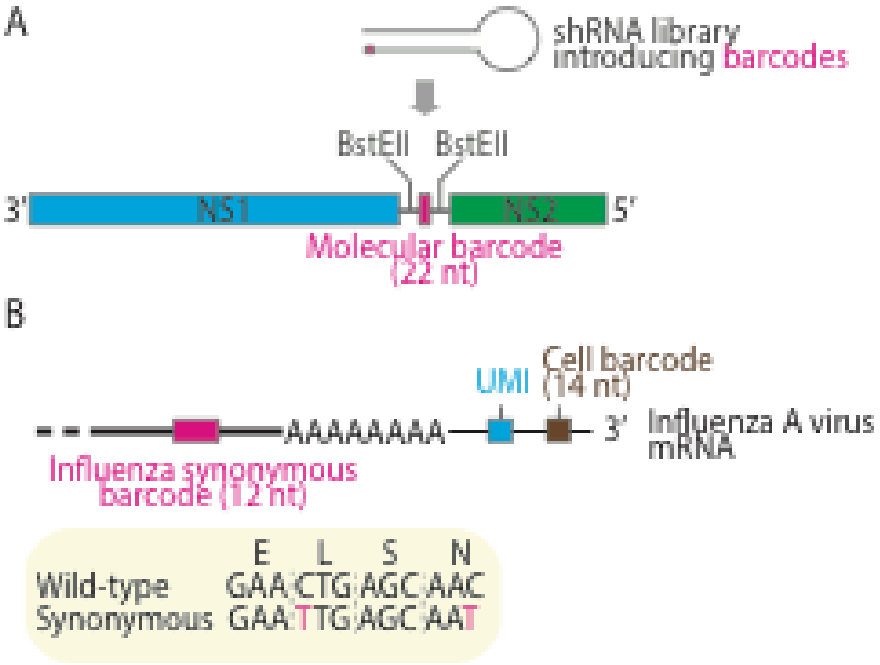

Figure 4

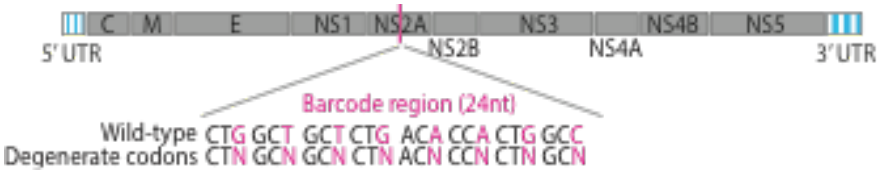

Figure 5

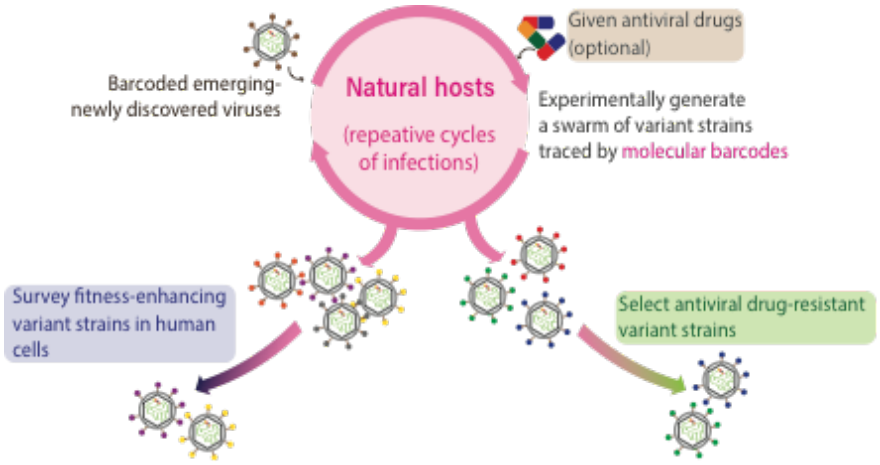

Figure 6