Virus Evolution, 2018, 4(1): vey007

doi: 10.1093/ve/vey007
Resources

Easy and accurate reconstruction of whole HIV genomes from short-read sequence data with `shiver`

Chris Wymant,[1,2,*,†] François Blanquart,[2] Tanya Golubchik,[1,3] Astrid Gall,[4,5] Margreet Bakker,[6] Daniela Bezemer,[7] Nicholas J. Croucher,[2] Matthew Hall,[1,2] Mariska Hillebregt,[7] Swee Hoe Ong,[5,‡] Oliver Ratmann,[2,8] Jan Albert,[9,10] Norbert Bannert,[11] Jacques Fellay,[12,13] Katrien Fransen,[14] Annabelle Gourlay,[15,16] M. Kate Grabowski,[17] Barbara Gunsenheimer-Bartmeyer,[18] Huldrych F. Günthard,[19,20] Pia Kivelä,[21] Roger Kouyos,[19,20] Oliver Laeyendecker,[22] Kirsi Liitsola,[21] Laurence Meyer,[23] Kholoud Porter,[15] Matti Ristola,[21] Ard van Sighem,[7] Ben Berkhout,[6,§] Marion Cornelissen,[6] Paul Kellam,[24,25] Peter Reiss,[7,26] and Christophe Fraser,[1,2,**] on Behalf of the BEEHIVE Collaboration[††]

[1]Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Medicine, University of Oxford, Oxford, UK, [2]Medical Research Council Centre for Outbreak Analysis and Modelling, Department of Infectious Disease Epidemiology, Imperial College London, London, UK, [3]Wellcome Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, Oxford, UK, [4]Department of Veterinary Medicine, University of Cambridge, Cambridge, UK, [5]Virus Genomics, Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK, [6]Laboratory of Experimental Virology, Department of Medical Microbiology, Center for Infection and Immunity Amsterdam (CINIMA), Academic Medical Center of the University of Amsterdam, Amsterdam, The Netherlands, [7]Stichting HIV Monitoring, Amsterdam, The Netherlands, [8]Department of Mathematics, Imperial College London, London, UK, [9]Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet, Stockholm, Sweden, [10]Department of Clinical Microbiology, Karolinska University Hospital, Stockholm, Sweden, [11]Division for HIV and Other Retroviruses, Robert Koch Institute, Berlin, Germany, [12]School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, [13]Swiss Institute of Bioinformatics, Lausanne, Switzerland, [14]HIV/STI Reference Laboratory, Department of Clinical Science, WHO Collaborating Centre, Institute of Tropical Medicine, Antwerpen, Belgium, [15]Institute for Global Health, University College London, London, UK, [16]Department of Population Health, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK, [17]Department of Pathology, John Hopkins University, Baltimore, MD, USA, [18]Department of Infectious Disease Epidemiology, Robert Koch-Institute, Berlin, Germany, [19]Division of Infectious Diseases and Hospital Epidemiology, University Hospital Zurich, Zurich, Switzerland, [20]Institute of Medical Virology, University of Zurich, Zurich, Switzerland, [21]Department of Infectious Diseases, Helsinki University Hospital, Helsinki, Finland, [22]Division of Intramural Research, NIAID, NIH, Baltimore, MD, USA,

[23]INSERM CESP U1018, Université Paris Sud, Université Paris Saclay, APHP, Service de Santé Publique, Hôpital de Bicêtre, Le Kremlin-Bicêtre, France, [24]Kymab Ltd, Cambridge, UK, [25]Division of Infectious Diseases, Department of Medicine, Imperial College London, London, UK and [26]Department of Global Health, Academic Medical Center and Amsterdam Institute for Global Health and Development, Amsterdam, The Netherlands

*Corresponding author: E-mail: chris.wymant@bdi.ox.ac.uk

†http://orcid.org/0000-0002-9847-8226

‡http://orcid.org/0000-0002-3629-5387

§http://orcid.org/0000-0002-1905-8486

**http://orcid.org/0000-0003-2399-9657

††See Supplementary Material for a complete list.

Abstract

Studying the evolution of viruses and their molecular epidemiology relies on accurate viral sequence data, so that small differences between similar viruses can be meaningfully interpreted. Despite its higher throughput and more detailed minority variant data, next-generation sequencing has yet to be widely adopted for HIV. The difficulty of accurately reconstructing the consensus sequence of a quasispecies from reads (short fragments of DNA) in the presence of large between- and within-host diversity, including frequent indels, may have presented a barrier. In particular, mapping (aligning) reads to a reference sequence leads to biased loss of information; this bias can distort epidemiological and evolutionary conclusions. *De novo* assembly avoids this bias by aligning the reads to themselves, producing a set of sequences called contigs. However contigs provide only a partial summary of the reads, misassembly may result in their having an incorrect structure, and no information is available at parts of the genome where contigs could not be assembled. To address these problems we developed the tool `shiver` to pre-process reads for quality and contamination, then map them to a reference tailored to the sample using corrected contigs supplemented with the user's choice of existing reference sequences. Run with two commands per sample, it can easily be used for large heterogeneous data sets. We used `shiver` to reconstruct the consensus sequence and minority variant information from paired-end short-read whole-genome data produced with the Illumina platform, for sixty-five existing publicly available samples and fifty new samples. We show the systematic superiority of mapping to `shiver`'s constructed reference compared with mapping the same reads to the closest of 3,249 real references: median values of 13 bases called differently and more accurately, 0 bases called differently and less accurately, and 205 bases of missing sequence recovered. We also successfully applied `shiver` to whole-genome samples of Hepatitis C Virus and Respiratory Syncytial Virus. `shiver` is publicly available from https://github.com/ChrisHIV/shiver.

Key words: bioinformatics; next-generation sequencing; HIV; diversity; genome assembly; mapping.

1. Introduction

The genetic sequences of pathogens are a rich data source for studying their epidemiology and evolution, and provide information for vaccine and therapeutic design. In the past decade, next-generation sequencing (NGS) has transformed genomics, with decreasing costs and enormous increases in the amount of data available. Despite the success of NGS in other fields, sequencing of human immunodeficiency virus (HIV) is still largely based on the older method of Sanger sequencing. For example, on the comprehensive Los Alamos National Laboratory HIV database (http://www.hiv.lanl.gov/ accessed 11 October 2017), of the 147,751 samples with platform information, 90.8% were generated by Sanger sequencing, 6.9% with the Roche 454 platform, 2.2% with Illumina platforms, and 0.02% with the IonTorrent platform. Breakdowns of these numbers by date and sequence length are in Supplementary Section S1.

More broadly, NGS has been hugely successful both for sequencing samples with no within-sample diversity, and at the opposite end of the spectrum, for metagenomic studies. In the first case, any apparent within-sample diversity is attributable to sequencing error; in the latter case, there is no presumption that different fragments of sequence in the same sample have the same origin, and so each fragment is checked against large databases to catalogue these diverse origins (Kunin et al. 2008; Thomas, Gilbert, and Meyer, 2012).

HIV is an intermediate case: the long duration of chronic infection coupled with high rates of replication and mutation mean that a single infection, and hence a single sample, will contain a diverse collection of related viral particles, frequently called a quasispecies. The long generation time for HIV transmission, together with continual within-host evolution, results in large, star-like phylogenies at the between-host level (Grenfell et al. 2004), i.e. each individual's quasispecies is quite distinct from the quasispecies of others. Reconstructing different aspects of these diverse quasispecies from *reads* (fragments of sequence; see Fig. 1) has proven technically challenging (Beerenwinkel et al. 2012) and may have hindered the widespread adoption of NGS for HIV. The complications of working with reads derived from a quasispecies can be bypassed with single genome amplification (SGA): in SGA, by limiting dilution, samples are reduced to single-virion aliquots that are sequenced separately (Simmonds et al. 1990; Palmer et al. 2005; Keele et al. 2008). However, the costs of using SGA for large population studies may be prohibitively high.

Here, we present the user-friendly programme `shiver` for working with HIV NGS data. Note that a variety of NGS platforms exist, which can be broadly classified into short-read-

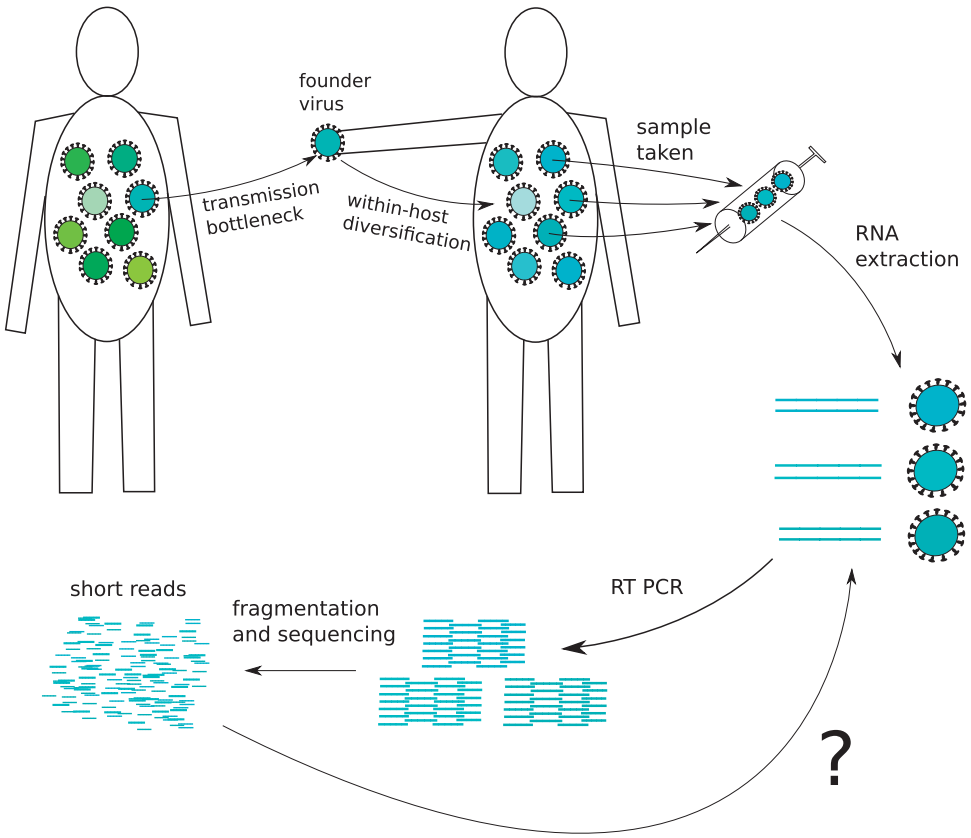

Figure 1. Interpreting NGS data for HIV. The question mark shows our question here: how best to discover the viral genotype from NGS data despite the high diversity of HIV between and within hosts?

low-error platforms and long-read-high-error platforms (see e.g. Goodwin, McPherson, and McCombie, 2016); here we focus on the former. Our programme was developed as part of the BEEHIVE project (*Bridging the Evolution and Epidemiology of HIV in Europe*) in which samples from over 3,000 individuals with known date of HIV infection are being sequenced to investigate the viral-molecular basis of virulence (Fraser et al. 2014). The power of genome-wide association studies (GWASs), and of epidemiological analyses e.g. identifying transmission risk factors, is enhanced by focussing resources on the widest possible population coverage (and so use of SGA is not a priority). We explain the need for `shiver` in the following subsection.

1.1 Mapping reads: problems and solutions

The quasispecies in one infected individual can be summarised by the consensus sequence—the 'average' sequence of those virions sampled, as represented in the reads. Determining the most common base at each position in the genome, and which other bases are present and at what frequencies, requires the reads to be *mapped* (aligned) to a reference sequence. To what should they be mapped? Mapping to a reference too far from the quasispecies' true consensus leads to biased loss of information (Archer et al. 2010; Henn et al. 2012; Iqbal et al. 2012; McElroy, Thomas, and Luciani, 2014). Like any form of sequence alignment, mapping relies upon sequence similarity; the more a read differs from its reference, the less likely it is to be aligned correctly or at all. This tends to hide differences between the sample and the reference, giving a consensus genome erroneously similar to the reference chosen.

The implications of this problem for downstream sequence analysis are worrying. Using the same reference for multiple infected individuals will tend to make their consensuses artefactually similar, overestimating proximity in a transmission network and distorting epidemiological conclusions. Using old reference sequences to construct new ones biases the new to resemble the old, which could distort our picture of evolution and hinder monitoring of emerging virulent or resistant variants. As an example, in a survey of *env* gene diversity in currently circulating viruses for vaccine design, it would be highly undesirable to artificially bias the reconstructed sequence towards similarity with the standard HXB2 reference virus isolated in 1983.

An example of this biased data loss is shown in Fig. 2, in which an insertion in the sample is lost because it is missing in the reference to which the reads were mapped. Reads containing insertions/deletions (indels) are particularly difficult to map correctly (Li, Ruan, and Durbin, 2008; Ye et al. 2009; McKenna et al. 2010; Albers et al. 2011). Inaccurate mapping at the sites of indels does not only result in missing the indel, as here, but can also prevent any reads from being mapped, or cause bases to be called incorrectly due to misalignment. This is an important point: even if the bases in an insertion are considered uninformative and are excluded from a particular comparative analysis, for example phylogenetic inference, it is undesirable that the insertion should cause missing or incorrect bases at neighbouring sites. Indels are known to be very common in HIV (Wood et al. 2009), especially in the *env* gene (Starcich et al. 1986). To quantify this further, we calculated indel size and position distributions in 3,249 whole genomes from the Los Alamos National Laboratory HIV database, shown in Fig. 3.

A How the reads should have been aligned to the reference:

```
Reference: ...ATATTTGGATGGCCTACTGTAAGGGAAAGAATG----------------------------AGACGAGCTGAGCCAGCAGCAGATGGGGTGGGAGCAGC...
Read 1:    ...GGGGGTGGATGGTCTGCTGTACGGGAAAGAATGAGGCGAGCTGCACCAACAGCAGAGGGGGTGAGGCGAGCTGAACCAGCAGCA
Read 2:                                AAGAATGAGGCGAGCTGCACCAACAGCAGAGGGGGTGAGGCGAGCTGAACCAGCAGCAGAGGGGGTGGGAGCAGC...
```

B How the reads were aligned to the reference:

```
Reference: ...ATATTTGGATGGCCTACTGTAAGGGAAAGAATGAGACGAGCTGAGCCAGCAGCAGATGGGGTGGGAGCAGC...
Read 1:    ...GGGGGTGGATGGTCTGCTGTACGGGAAAGAATGAGGCGAGCTGCACCAACAGCAGAGGGGGTGAGGCGAGCTGAACCAGCAGCA
Read 2:    AAGAATGAGGCGAGCTGCACCAACAGCAGAGGGGGTGAGGCGAGCTGAACCAGCAGCAGAGGGGGTGGGAGCAGC...
```

Figure 2. An example of biased loss of information encountered in our data when mapping to an existing reference. The reads contain a 30 bp insertion relative to the reference. Correct alignment, shown in the upper panel, would have inserted a 30 bp gap into the reference to accommodate this. What the mapper actually did (lower panel) was to align part of each read correctly either to the left of the insertion or to the right of it, and discard the rest of the read. 'Read 1' and 'Read 2' each represent roughly 2,000 similar reads; their consensus is therefore well supported but misses the insertion. This bias occurred despite the reference having been identified as the closest of 3,249 to this set of reads. Similar errors were made by the mapper's smalt, BWA, and bowtie, resulting in the same erroneous consensus being called in each case. Bases in the reads that differ from the reference are shown in blue; the ends of the reads that were discarded during mapping (i.e. not aligned) are shown in grey with strikethrough. This figure corresponds to Position 8450 in Fig. 5.

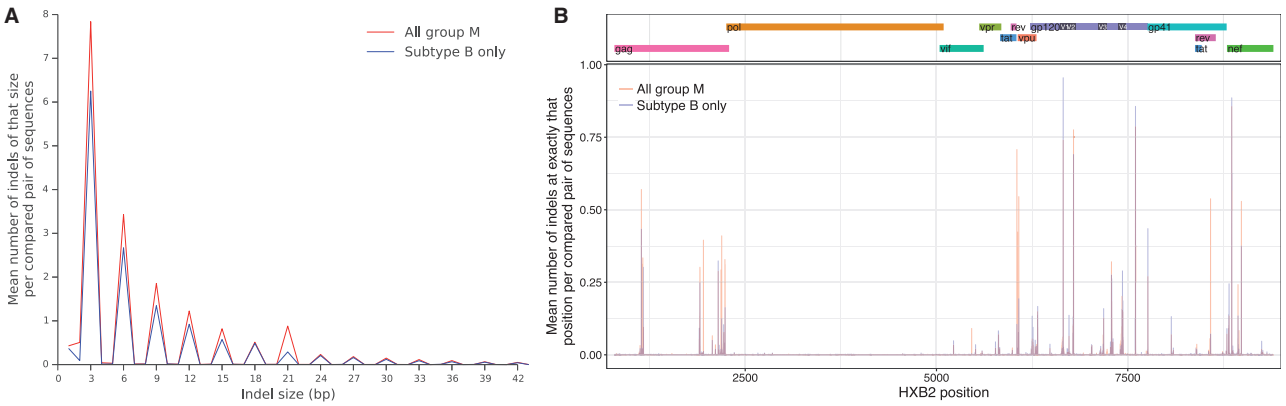

Figure 3. Quantifying indels in 3,249 whole genomes—those in the 2016 'all genome' group M alignment from the Los Alamos National Laboratory HIV database. We trimmed both 3' and 5' ends of the alignment where sequences align poorly, then considered each of the roughly 5.3 million possible pairs of references therein. For each pair we calculated the size and position of their relative indels (i.e. taking their relative alignment from the overall alignment, ignoring positions at which both have a gap). We also considered just the subset of 1,019 subtype B sequences, which is less diverse than group M as a whole but shows similar indel patterns. Left panel: the distribution of indel sizes. The striking bias towards frame-preserving indels could be biological (frame-shifting indels will generally have a large fitness cost), artefactual (removal of frame-shifting indels from sequences during analysis before public release, on the assumption that this is sequencing or bioinformatic error), or a combination of both. Right panel: where in the genome the indels tend to occur. The observed pattern is consistent with purifying selection in pol and diversifying selection in env.

The loss of reads during mapping has been shown to be roughly proportional to the divergence between the true consensus and the reference used (Archer et al. 2010). The bias in the loss of reads (and the loss of accuracy in their alignment) occurs at different scales. Data are more likely to be lost in (1) those samples in a dataset that differ more greatly from the reference used for their mapping; (2) those parts of the genome, in a single sample, where the sample and reference are most different; and (3) a subset of genotypes, in a single diverse sample, that are more different from the reference than the other genotypes.

This problem means the simplest mapping strategy—using as a reference some existing, standard genome, even if chosen specifically for each sample based on the reads—has much room for improvement. For example one could map once to a standard reference, call the consensus, then use this as the reference for one or more rounds of remapping (Willerth et al. 2010; Gibson et al. 2014; McElroy, Thomas, and Luciani, 2014; Verbist et al. 2014; Ode et al. 2015). Remapping is expected to be more accurate, because the consensus initially called is expected to be closer to the true consensus than the standard reference is. For this to be the case all along the genome however, reads must map correctly all along the genome in the first step. If the sample has an indel not present in the reference, inaccurate mapping at the site of the indel may cause it to be missed when the consensus is called, as in Fig. 2. Remapping is then doomed to repeat the same error.

To correct for this, between initial mapping to the standard reference and calling the first consensus, multiple sequence alignment can be performed with the reads (Archer et al. 2010; Zanini et al. 2015). This removes some of the bias imposed by the initial mapping, because while mapping aligns each read to the reference sequentially and independently, multiple sequence alignment with the reads considers how the reads align to each other. It is then less important that reads map *correctly* all along the genome, since realignment may correct misalignment around indels, but the reads do still need to *map* all along the genome. If biased data loss leads to a failure of reads to map at a given point, the missing reads will not shape the initial consensus and remapping to that consensus will not recover them. For the variable loop regions of HIV's *env* gene in particular, reads from one virus can easily fail to map to another; many examples of this can be seen in Supplementary Sections S4 and S5, visible as parts of the genome where reads do map to a reference tailored to the sample, but not to the closest identified real reference, resulting in missing sequence in the latter case.

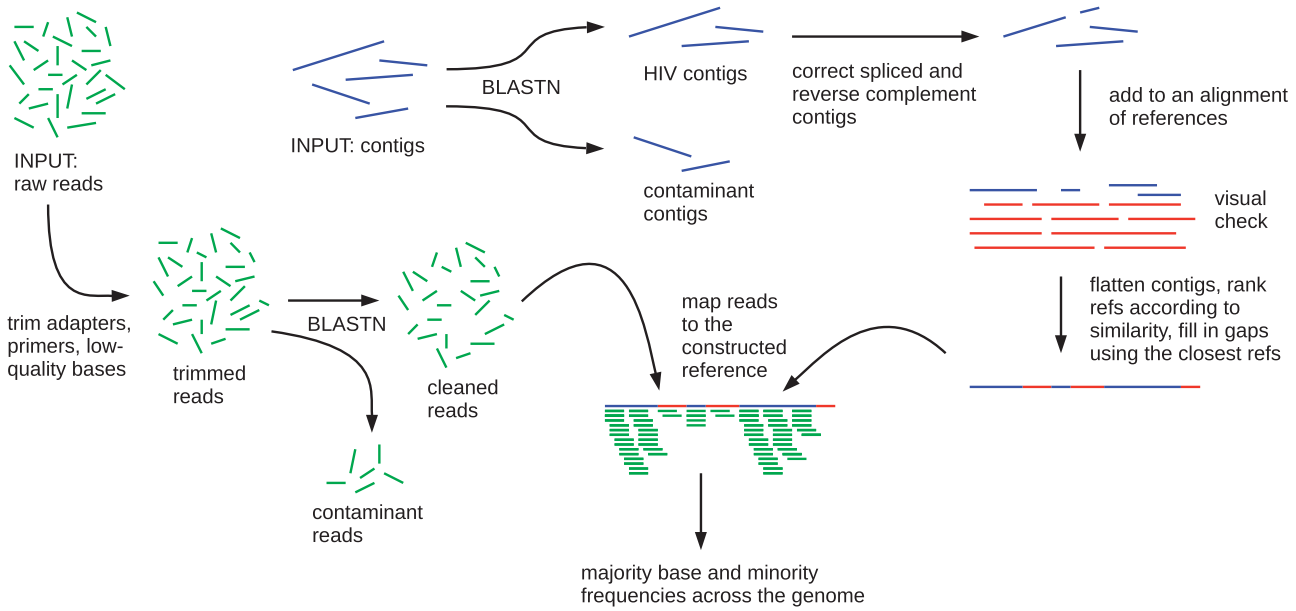

Figure 4. A summary of the steps in our method `shiver`.

These problems motivate *de novo* assembly (hereafter just assembly). Roughly, this consists of aligning overlapping reads to each other, tolerating some pre-set level of disagreement between them to allow for some within-sample diversity or sequencing error, iteratively extending using reads overhanging the edges, finally resulting in a set of sequences called contigs (see e.g. http://en.wikipedia.org/wiki/Sequence_assembly). Remapping to contigs (Henn et al. 2012; Yang et al. 2012; Malboeuf et al. 2013; McElroy, Thomas, and Luciani, 2014; Ode et al. 2015) settles ambiguity at positions spanned by multiple contigs which disagree, corrects positions where assembly did not call the most common base, provides minority variant information, and allows greater use to be made of base quality information than is typically done during assembly.

However, contigs may differ from the true consensus by more than just a few SNPs that can be corrected by mapping. Misassembly may occur, giving contigs supported by a high depth of reads but whose structure is very different from the known genome. This can arise *in silico* (McElroy, Thomas, and Luciani, 2014), i.e. by misassembly of correct reads; or as a result of chimeric reads produced during sequencing, due to recombination during library preparation (Meyerhans, Vartanian, and Wain-Hobson, 1990; Judo, Wedel, and Wilson, 1998; McElroy, Thomas, and Luciani, 2014), concatemerisation/ligation (Croucher et al. 2009), or stem loops of RNA secondary structure (Malboeuf et al. 2013).

Furthermore, the set of contigs resulting from assembly may not fully cover the genome. Gaps between contigs can be due to a total absence of reads there, following sequencing failure or only a partial genome present in the sample. They can also be due to the reads being too few (though non-zero), or too diverse, for successful assembly; in this case, mapping can recover consensus sequence not present in assembly output.

Finally, as the set of reads will generally contain contamination, so will the set of contigs. These contigs should be identified and discarded.

To address these problems we developed the tool `shiver`—*Sequences from HIV Easily Reconstructed*—to preprocess and map reads from each sample to a custom reference, tailored to be as close as possible to the expected consensus, constructed by correcting contigs and filling in gaps between them with the closest identified existing reference sequences. We wrote it to be easy to use, suitable for simple scripted application to large heterogeneous data sets, in our population genomics study and elsewhere.

2. Methods

2.1 A summary of the `shiver` method

The steps in `shiver` are shown in Fig. 4; see Supplementary Section S2 for more details.

In summary: paired-end short reads and contigs assembled from those reads are required as input for each sample; also required is a set of existing reference genomes, chosen by the user. Contigs are compared with the existing references using BLASTN (Altschul et al. 1990), then partitioned into those judged to be HIV and those judged to be contamination. HIV contigs are corrected as follows. First, spliced contigs—those concatenating two separated regions of the genome into a single sequence—are cut. The motivation for this cutting of contigs is the assumption that HIV does not exhibit major structural variation, e.g. variation in gene presence/absence or gene order, which is supported by sequence compendiums to date (http://www.hiv.lanl.gov/). Second, parts of contigs that did not have a blast hit to any existing reference are removed. Third, any contig (or part of a contig) found to be in the opposite orientation to the existing references is reverse-complemented. The contigs are added to the alignment of existing references using MAFFT (Katoh et al. 2002), and contigs found to have an overly large internal deletion are split into two separate contigs at that point.

At this point `shiver` stops to allow a visual check of the alignment of contigs and existing references. Once it is checked, `shiver` continues (all remaining steps in the programme are performed by the second of two commands needed for full processing). From this alignment, the closest existing reference is identified by comparison with all of the contigs. This is expected to be a more accurate identification of the closest existing reference than, for example, finding which existing

reference most reads match most closely, which gives undue weighting to regions of the genome where more rounds of amplification resulted in an exponentially greater number of reads. shiver creates a reference for mapping by using contig sequence where available, and the closest existing reference to fill in any gaps between contigs (at parts of the genome where assembly failed). Before mapping, reads are trimmed for low-quality bases, adapter and primer sequences using Trimmomatic (Bolger, Lohse, and Usadel, 2014) and fastaq (https://github.com/sanger-pathogens/Fastaq); contaminant read pairs are diagnosed as those matching contaminant contigs more closely than the tailored reference, and are removed. The remaining reads are mapped to the tailored reference. By default we map using smalt with a minimum read identity (the fractional agreement between a read and the reference to be considered mapped) of 70%, independent mapping of mates in a pair, a maximum insert size of 2,000 bp, and discarding improperly paired reads. Optionally, BWA (Li and Durbin 2010) and bowtie (Langmead et al. 2009) can be used instead of smalt.

Following mapping, each position in the genome is considered in turn using SAMtools (Li et al. 2009), to find the frequencies of different bases. At positions where some reads have deletions relative to the mapping reference, we count the frequency of the gap character together with actual bases. At positions where some reads have insertions relative to the mapping reference, for the consensus we use the most common insertion size (which may be 0, i.e. no insertion). By default the most common base is called to give the consensus; optionally ambiguity codes can be used more readily, when the frequency of the most common base(s) is below a threshold. A consensus base is only called if the coverage equals or exceeds a minimum threshold specified by the user, to protect against the effect of residual low-coverage contaminant reads in genomic regions lacking genuine HIV reads. By default this is 15, but this is likely to need adjusting for different datasets. A tool contained in shiver helps the user to explore appropriate values (see the discussion of LinkIdentityToCoverage.py in Supplementary Section S3).

By default, once the consensus is called, the cleaned reads are re-mapped to it (with any missing coverage in the consensus filled in with the corresponding part of the original tailored reference) for a second iteration of calling the base frequencies and the consensus. (This is why the shiver reference does not match the contigs exactly in Fig. 5 and the figures of Supplementary Sections S4 and S5).

shiver also produces a 'global alignment' of all consensuses it generates by coordinate translation, without need for an alignment algorithm.

2.2 Running shiver fully automatically

Alternatively shiver can be run from beginning to end without the break in the middle described above, for applications where visually checking the contigs is impractical. This is only possible for samples not requiring contig correction, and does not produce the global alignment of all samples' consensuses together. The different alignment strategy used in this case, and our recommendation that the contigs be checked instead, are discussed further in Supplementary Section S2.5.

2.3 Using the shiver code

shiver and its documentation are available at https://github.com/ChrisHIV/shiver. It was designed to be run in Linux-like environments, including Mac OS. Once dependent packages are installed, shiver itself requires no installation: it is a set of executable scripts. The Genomic Virtual Laboratory (Afgan et al. 2015), provided for example on the UK Medical Research Centre's Cloud Infrastructure for Microbial Bioinformatics (MRC CLIMB) (Connor et al. 2016), contains all dependencies (except smalt, which is loaded on MRC CLIMB with the single command `brew install smalt`, and otherwise available at http://www.sanger.ac.uk/science/tools/smalt-0), allowing shiver to be run immediately. The GitHub repository also contains a platform-independent virtual machine containing shiver with all of its dependencies pre-installed.

Before processing with shiver, short reads must be assembled into contigs. This important step, though difficult technically, is not onerous for the user: our chosen assembler IVA assembles contigs from reads with a single command from the command line, and can be run on a virtual machine provided by the Sanger pathogens group (http://sanger-pathogens.github.io/pathogens-vm/). The user can use any assembler; others are available in the Genomic Virtual Laboratory, including SPAdes, Velvet and MIRA, though currently none designed specifically for viral data.

shiver is run from the command line using three commands. Firstly, a one-off initialisation command:

```
shiver_init.sh MyInitDir config.sh MyReferences.fasta \
MyAdapters.fasta MyPrimers.fasta
```

(the slash indicating that one command is here being split over multiple lines), which sets up an initialisation directory of files for shiver based on the user's choice of existing references, and adapter and primer sequences to remove. Subsequently, for each sample to be processed, one command blasts, corrects and aligns the contigs:

```
shiver_align_contigs.sh MyInitDir config.sh \
MyContigs.fasta MyID
```

where MyID is used for labelling output. After inspection of the corrected contigs aligned to the existing references, a second command constructs a tailored reference for mapping, preprocesses the reads, maps them and calls the consensus:

```
shiver_map_reads.sh MyInitDir config.sh \
MyContigs.fasta MyID MyID.blast \
MyAlignedContigs.fasta MyForwardReads.fastq \
MyReverseReads.fastq
```

This produces, for each sample,

- the mapped reads in BAM format;
- a plain text file with the counts of the different bases at each position, also including HXB2 coordinates (by default; not relevant for non-HIV samples);
- the consensus;
- a coordinate-translated version of the consensus for a global alignment; and
- the insert-size distribution.

The global alignment of consensuses produced from all samples is constructed simply by combining the coordinate-translated consensus files from all samples into one file, e.g. running from the command line

```
cat file1 file2 [...] > MyGlobalAlignment.fasta
```

For our data, shiver typically look less than an hour to process each Miseq sample, and up to ten hours for each Hiseq sample (the latter containing roughly ten times as many reads), on a single core of the Imperial College London High-Performance Cluster (which is a mixture of computational resources with different specifications).

All bioinformatic parameters can be changed in the configuration file (`config.sh` above), allowing customisation of how reads are trimmed, how they are mapped, and how the consensus is called as a function of coverage and diversity. `shiver` also includes simple command-line tools for partial reprocessing (modifying sample output without rerunning the whole pipeline), and for analysis—see Supplementary Section S3.

2.4 Example data and its processing by `shiver`

We used two datasets as examples for processing with `shiver`. The first was sixty-eight publicly available Miseq samples: those sequenced and released with the IVA publication (Hunt et al. 2015), namely accession numbers ERR732065–ERR732132 on the European Nucleotide Archive. The samples have different origins; six are from a longitudinally sampled transmission pair studied by Brener et al. 2015. ERR732065–ERR732072 were sequenced with 150 bp reads, ERR732073–ERR732132 with 250 bp reads. Only forty-two of these sixty-eight samples were assembled by Hunt et al. 2015: the rest failed quality control checks designed to pre-select robust whole-genome samples. After downloading the short reads from the European Nucleotide Archive, we reassembled all sixty-eight samples with IVA for processing with `shiver`, as by design our method can be run in exactly the same way for those samples devoid of genuine sequence, those with partial genomes and those with whole genomes.

The second dataset was fifty Hiseq samples newly generated for the BEEHIVE project, from confirmed seroconverters from Europe. RNA was extracted manually from blood samples following the procedure of Cornelissen et al. 2016. This was amplified using universal primers that define four overlapping amplicons spanning the whole genome, following the procedure of Gall et al. 2012. Specifically, 5 μl of Amplicon 1 (the shortest and most successfully amplified amplicon) was pooled with 10 μl each of Amplicons 2–4. Multiple samples were pooled during library preparation, using one of 192 multiplex adaptors for each sample. The library was sequenced in 'rapid run mode' on both lanes of a HiSeq2500 instrument with read lengths of 2×250 bp, resulting in two lanes of short reads per sample. Automatic processing at the Wellcome Trust Sanger Institute used IVA to generate contigs for each lane, i.e. two sets of contigs per sample. We combined the two sets to allow comparison of the assembly output resulting from two technical replicates of short reads. For the large majority of cases the contigs were nearly identical, but stochastic differences in the read populations between lanes mean the resulting contigs occasionally differ.

The fifty Hiseq samples were chosen from a larger dataset currently being collected and sequenced for the BEEHIVE project's primary aim of investigating the viral-molecular basis of virulence. Selection criteria for inclusion in the project include a known date of infection, either by negative and positive tests separated by less than a year, or by clinical signs of acute infection at diagnosis; and a sample obtained for sequencing between 6 and 24 months after diagnosis, before beginning antiretroviral treatment and before progression to AIDS. The fifty samples processed here were chosen as follows. (1) One sample chosen with a large difference in the fraction of the genome assembled between the two Hiseq lanes, as an example of the variability of assembly output. (2) Nine samples chosen with misassembled contigs for one or both Hiseq lanes, to illustrate the necessity of `shiver`'s contig correction. (3) From each of the Dutch, French, German and Swiss cohorts, ten samples with contigs spanning the whole genome: five subtype B and five non-B samples (subtype was determined with the COMET software (Struck et al. 2014)).

The existing reference set we used was the 2016 'compendium' group M genome alignment from the Los Alamos National Laboratory, with a small amount of sequence trimmed from both edges of the alignment to match the region of the genome amplified by the sequencing protocol used for all data here (Gall et al. 2012), which partially excludes the flanking long terminal repeat regions.

For comparison with `shiver`'s constructed mapping reference, for each sample we used kallisto (Bray et al. 2016) to pseudo-align all the reads, using an index constructed from 3,249 whole genome references from the Los Alamos National Laboratory HIV database (those in the 2016 'all genome' alignment) together with the whole human genome (as an attractor for human contaminant reads). We defined the closest existing/real reference sequence for that sample as the one with the highest transcript per million score.

For this analysis, we set the minimum coverage threshold (the number of mapped reads required to call the base at each position) to be 10 throughout, since the assembler we used—IVA—requires at least ten reads to extend a contig, and we compare the consensus to the contigs.

To illustrate application of `shiver` outside of HIV, we used it to process Illumina paired reads from a whole-genome Hepatitis C Virus (HCV) sample: accession number DRR000928 on the European Nucleotide Archive. We assembled the reads into contigs using SPAdes (Bankevich et al. 2012), and for the existing reference set required as `shiver` input we used the 2008 'all genome' alignment of 471 references from the Los Alamos National Laboratory HCV database (Kuiken et al. 2005). We also ran `shiver` on Illumina paired reads from a whole-genome Respiratory Syncytial Virus (RSV) sample: accession number ERR438932 on the European Nucleotide Archive. We assembled the reads into contigs using SPAdes, and for the existing reference set we used the sixty-three whole genomes sequenced by Bose et al. 2015 from four continents to help capture global RSV diversity. For both the RSV and HCV reads we used kallisto to identify the closest sequence in the existing reference set, in the same manner as described above for the HIV dataset.

3. Results

We ran `shiver` on the paired-end short read HIV data described earlier—sixty-eight Illumina Miseq samples and fifty Illumina Hiseq samples. Only sixty-five of the Miseq samples had at least one contig that returned a BLASTN hit to a sequence in our chosen set of existing references; these and all fifty Hiseq samples were fully processed, giving whole or partial genomes. We produced consensus sequences, together with summary minority-variant information (base frequencies at each position) and detailed minority-variant information (all reads aligned to their correct position in the genome).

For comparison, for each sample we also mapped to the closest existing reference sequence identified using kallisto, instead of the `shiver` reference. We used the same mapping parameters, mapped the same set of reads (following `shiver`'s removal of adapters, primers, low-quality bases and contaminant read pairs), and called the consensus of the mapped reads in the same way (still using `shiver`), i.e. we changed only the reference sequence used for mapping.

Supplementary Sections S4 and S5 contain figures showing, for each sample, the genes of HIV in their reading frames, a set

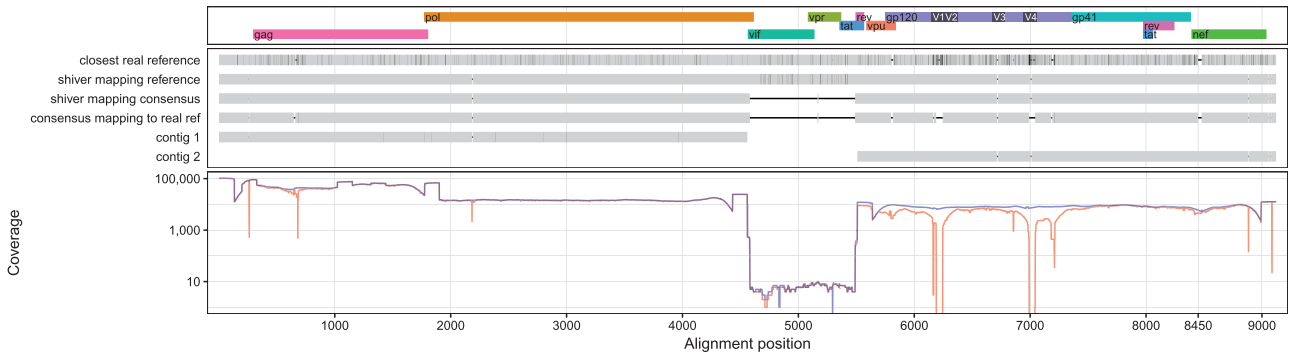

Figure 5. Top panel: HIV genes in their reading frames. Middle panel: sequences for the Miseq sample ERR732065. From top to bottom these are the closest identified real reference (see main text), the reference created and used for mapping by `shiver`, the consensus of reads mapped to `shiver` reference, the consensus of the same reads mapped to the real reference, and the contigs generated by *de novo* assembly. Vertical black lines inside sequences in the alignment denote single nucleotide polymorphisms (SNPs), defined here relative to the most common base among these sequences. Horizontal black lines indicate a lack of bases, i.e. a deletion relative to another sequence in the alignment or, for the two consensuses, simply missing sequence due to insufficient coverage. Bottom panel: the coverage (number of mapped reads) for the `shiver` reference in blue, and for the real reference in red. Mapping problems at Position 8450 are shown in detail in Fig. 2. Where the real reference and the sample differ by many close SNPs or an indel, differences often arise between the `shiver` consensus and the consensuses mapping to the real reference. The coverage plot beneath the sequences shows that at such points, the coverage mapping to the real reference almost always drops below the coverage mapping to the `shiver` reference; given that the same reads are being mapped to the same part of the genome with the same mapping parameters, this strongly suggests that the `shiver` consensus is more accurate. This is the case at Position 8450 in this figure, in the *nef* gene; the problem mapping to the real reference here was shown in detail in Fig. 2. Though the coverage here drops due to the problem aligning the reads, it is still more than 4,000, showing that a large absolute number of reads is no guarantee of accuracy. Mapping to the `shiver` reference on the other hand, coverage remains locally smooth. Similar errors mapping to the real reference in this figure can be seen in *gag* and in five different places in gp120.

of sequences connected to this sample, and the coverage (number of reads mapped at each position) along the genome. We reproduce the figure for the first Miseq sample here (Fig. 5)—as an example for discussion. We see that there is no sequence data in the region around the *vif* and *vpr* genes, which is the part of Amplicon 3 in this sequencing protocol that is not overlapped by neighbouring Amplicons 2 or 4. Evidently Amplicon 3 failed to amplify for this sample. There is no contig sequence in this region, a coverage less than the threshold of 10, and so consensus sequence was not called. (The information contained in the few reads that did map to this region is retained in the minority-variant files produced by `shiver`; consensus sequence could be called here, if one chose to lower the minimum coverage threshold parameter below 10.)

Comparisons of these sequences are quantified for each sample in Supplementary Table S1, and in summary in Tables 1 and 2. For example Table 1 shows that mapping a sample's reads to the `shiver` reference instead of the real reference, the median number of bases called differently and supported by higher coverage is 13; the median number of bases called differently but with equal or lower coverage is 0. Interpreting higher coverage as more accurate mapping, mapping to the `shiver` reference instead of the real reference typically corrects thirteen false SNPs per sample. For this comparison we only considered positions where a base was called in both consensuses, but the base differed. As in the case of Fig. 2, inaccurate mapping may also result in a stretch of sequence being missed from the consensus. The median increase in the consensus sequence length when mapping a sample's reads to the `shiver` reference instead of the real reference is 205 bp.

Table 2 shows that, for more than half of the samples, the `shiver` consensus is no longer than the set of contigs (the median length increase is zero). However it is occasionally much longer—see the relevant column of Supplementary Table S1—due to assembly failure. The median number of bases in the `shiver` consensus that differ from all contigs at that point is 7. (Where the contigs disagree amongst themselves but one agrees with the consensus, we count this as agreement.) As the consensus is derived by mapping to the contig sequence at such points and calling the most common base, such positions of disagreement are probably improvements. Seven corrected SNPs is a highly conservative estimate of the improvement over the contigs, however, as the comparison was made after `shiver` performed contig correction (including both structural correction and trimming of contig ends where they have no BLASTN hit). This is because a base-by-base comparison of two sequences requires them to be aligned, and aligning the *spliced* or partially reverse-complemented contigs that `shiver` corrects (see Section 2.1) would give a nonsensical alignment. In addition, deriving the consensus from mapping instead of relying solely on *de novo* assembly means that minority-variant information is available.

As mentioned in Section 2, nine of the Hiseq samples were chosen as illustrations of misassembled contigs, and twenty-three of the Miseq samples with HIV contigs (twenty-six including those without HIV contigs) were not considered in the IVA publication due to failing sample quality control checks. These samples are identified in Supplementary Table S1. The statistics for `shiver`'s performance for these nine Hiseq samples are not worse than those for the all the data, e.g. a median of thirty-one bases called differently with higher coverage in the shiver consensus, and 0 bases called differently with higher coverage mapping to the real reference. This illustrates that problematic contigs do not mean that mapping to an existing reference becomes preferable, thanks to `shiver`'s contig correction. The IVA QC failures are mostly partial genomes; statistics for these samples are scaled down from their values for the whole data set due to these being shorter sequences. An exception is the increase in the consensus sequence length over the length of the contigs, whose median value is 0 for the whole dataset but thirty-two for the QC failures. It is not surprising that contigs should be shorter than mapping-derived consensuses for problematic samples previously excluded from consideration for assembly.

These improvements from using `shiver` are small compared with the length of the HIV genome—roughly 9,000 bases. However the aim of sequencing a known pathogen is not to produce a roughly correct picture of the known genome, but to

Table 1. Comparing the consensus from mapping to the reference constructed by `shiver` with the consensus from mapping to the closest identified real reference.

Number of bases called differently, with higher coverage when mapping to the `shiver` reference than to the real reference	Min	0
	Median	13
	Mean	16.8
	Max	57
Number of bases called differently, with higher (or equal) coverage mapping to the real reference than to the `shiver` reference	Min	0
	Median	0
	Mean	1.2
	Max	24
Extra length of the `shiver` consensuses compared with the real reference's consensus (in number of bases)	Min	−54
	Median	205
	Mean	239.4
	Max	1,262

Minima, medians, means, and maxima are over the combined set of sixty-five Miseq and fifty Hiseq samples processed. Means are rounded to one decimal place.

Table 2. Comparing the consensus from mapping to the reference constructed by `shiver` with the contigs (after correction of the contigs by `shiver`).

Length of sequence present in the contigs but missing from the consensus	Min	0
	Median	0
	Mean	0
	Max	0
Length of sequence present in the consensus but missing from the contigs	Min	0
	Median	0
	Mean	114.1
	Max	2,443
Number of positions where all corrected contigs disagree with the consensus	Min	0
	Median	7
	Mean	13.7
	Max	106

Minima, medians, means, and maxima are over the combined set of sixty-five Miseq and fifty Hiseq samples processed. Means are rounded to one decimal place.

obtain each sample's sequence as accurately as possible, so that small numbers of differences between similar samples can be meaningfully interpreted.

The problems arising from mapping to a reference that differs from the sample in question do not arise simply from an inappropriate choice of mapper. To illustrate this, for the Miseq dataset we also mapped the reads to their closest real reference sequence using BWA and bowtie in both its 'local' and 'end-to-end' modes (for both mappers we used their default settings except for specifying a maximum insert size of 2,000 for bowtie, retaining only properly paired reads as we did with smalt). Figure 6A shows the resulting coverage along the genome for the same sample shown in Fig. 5. Localised drops in coverage indicate the same problems described previously. This was common across all of the samples; Fig. 6B shows a more extreme example, for which mapping to the closest real reference using any of the mappers performs very poorly.

Among the reads mapped by `shiver`, interesting within-host diversity is maintained, capable of revealing structure in the quasispecies. Figure 7 shows an example for our Hiseq sample 17796_3_29. The reads are from the boundary between p2 and p7 in the *gag* gene; roughly a third of them have a 21-bp insertion relative to the others. This insertion is not seen in any other sequence in the Los Alamos National Laboratory alignment 'HIV1_ALL_2015_gag_DNA' of 7,903 *gag* sequences (http://www.hiv.lanl.gov/). Though not a duplication at the nucleotide level, it duplicates the GATAMMQ amino acid motif. Mutations at the p2/p7 boundary (Ho et al. 2008) and insertions at other *gag* cleavage sites (Tamiya et al. 2004) have been implicated in restoring replicative capacity in viruses treated with protease inhibitors.

For the HCV sample, compared with mapping to the closest existing reference identified from the reads, `shiver` called nine bases differently, all supported by higher coverage. `shiver` also recovered a 15-bp stretch of sequence that was missing from the consensus after mapping to the closest existing reference. These nine different base calls and 15 bp of sequence were close together at the start of the E2 gene. There were no indels between the sample and the closest existing reference here, but a very high density of SNPs which prevented accurate mapping of the sample's reads to the closest existing reference.

For the RSV sample, compared with mapping to the closest existing reference identified from the reads, `shiver` called only one base differently, supported by higher coverage. Clearly, examining only a single sample does not allow us to draw any conclusions; however, this much more modest improvement in using a constructed reference over an existing reference for RSV is not surprising. The smaller amount of diversity in RSV (especially within each of its two distinct subgroups, A and B) compared with HIV or HCV should make it easier to find an existing reference with a very high degree of similarity to the sample in consideration. On the other hand, for viruses exhibiting less diversity, each erroneous base call will have greater impact on comparative analyses; `shiver` may therefore still be useful in these cases.

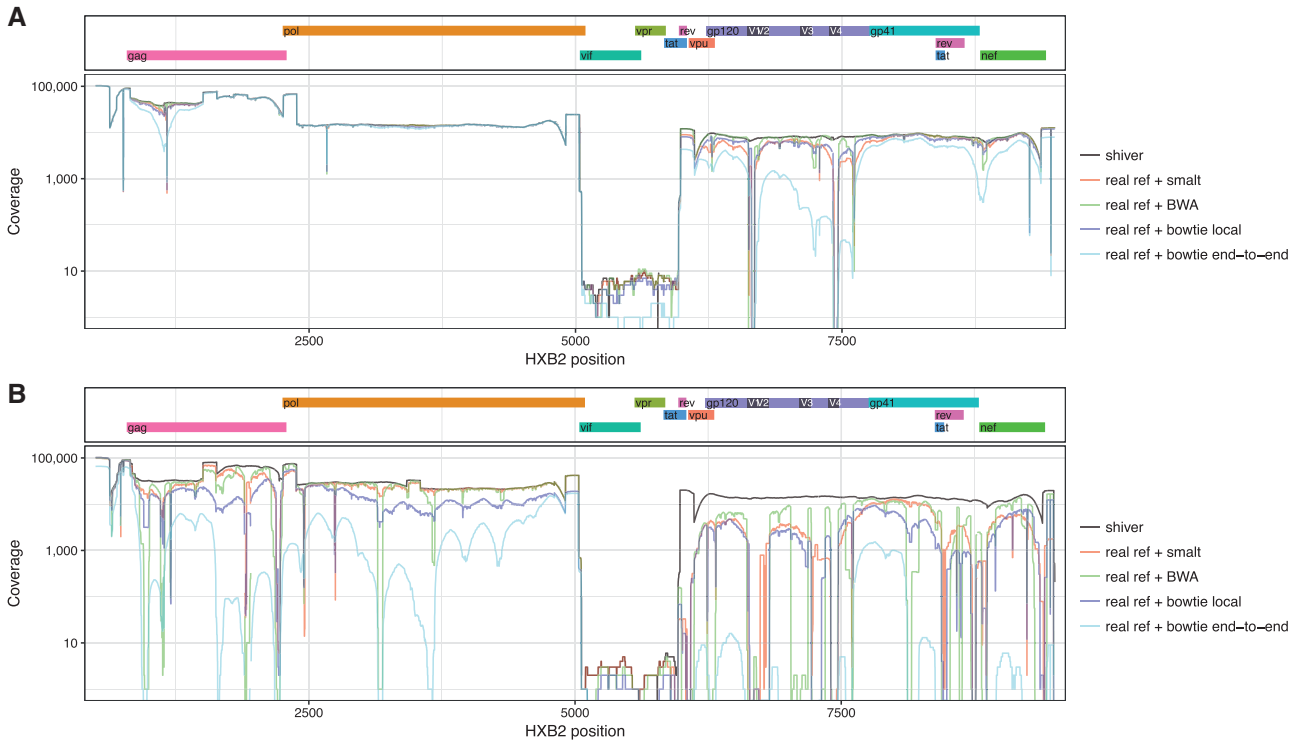

Figure 6. Coverage over the genome following different mapping strategies. The line marked '`shiver`' is for mapping to the reference constructed for the sample using contigs, with smalt. The other lines are for mapping to the closest identified real reference using the indicated mapping algorithm. (A) is the same sample shown in Fig. 5; (B) is Miseq sample ERR732071, a more extreme example of mapping failure using the closest real reference.

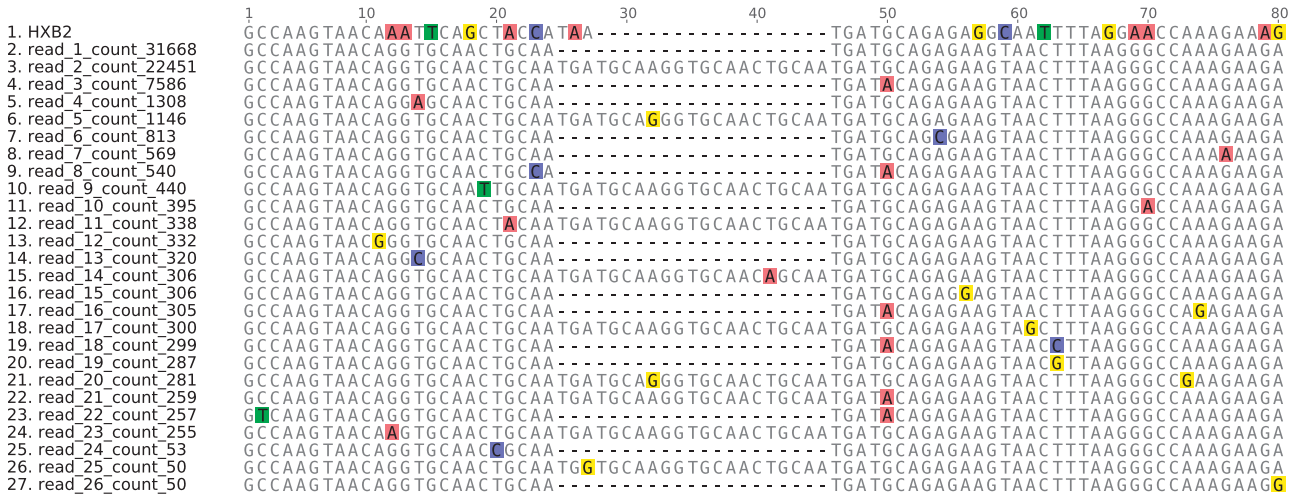

Figure 7. Within-host indel polymorphism in our Hiseq sample 17796_3_29: a 21 bp insertion in roughly a third of the reads duplicates the GATAMMQ amino acid motif at the boundary between p2 and p7 in *gag*. The value following '_count ' in the sequence name is the number of times that exact sequence was found in the reads here following mapping with `shiver`; only sequences found at least fifty times are shown. HXB2 is included for comparison. Coloured squares highlight bases differing from the consensus; bases without a coloured square agree with the consensus base at that position (ignoring gaps).

In this methods article we do not present conclusions drawn from analyses of sequences generated with `shiver`. Two such analyses published so far are by Blanquart et al. 2017 and Ratmann et al. 2017. Blanquart et al. 2017 determined that the fraction of variability in HIV set-point viral load that is explained by viral genetic factors was around one third, using 1,373 European whole genomes. Ratmann et al. 2017 found predictors of HIV sequencing success or failure for 3,985 African whole genomes, and studied the effect of the observed amplification failure patterns on phylogenetic inference. Analysis of both genomic datasets is ongoing.

4. Discussion

We developed the tool `shiver` to preprocess and map reads from each sample to a custom reference, constructed using *de novo* assembled contigs supplemented by existing reference genomes. Tailoring the reference to be as close as possible to the

expected consensus before mapping maximises the accuracy of the mapping, and therefore of the resulting consensus. `shiver`'s identification, ranking, and use of the closest existing references to fill in gaps between contigs boosts data recovery for samples with amplification failure or assembly failure. Such partial-genome samples, which are inevitable in large diverse data sets, are processed with exactly the same two commands; this simplifies scripted application of `shiver` to all samples in a data set. `shiver` also produces a global alignment containing all of the consensuses separately generated for each sample, which is usually required for comparative analysis of the sequences such as for phylogenetics or GWASs.

Mapping to `shiver`'s constructed reference instead of mapping the same reads to the closest identified real reference gives a median increase in consensus sequence length of 205 bp, with thirteen of the original bases called differently and more accurately. This shows the importance of tailoring the reference to the sample before mapping. `shiver`'s consensus, obtained by mapping reads to a reference constructed from the contigs, has a median of 7 bases called differently from the contigs even after correcting structural problems in the contigs and trimming suspicious sequence from their ends. This illustrates the need for mapping in addition to assembly.

A limitation of the method is that after reads have been successfully mapped (which imposes requirements on base quality and good alignment to the reference), we consider each read to carry equal weight in determining the consensus and the frequency of variant bases. The frequency of a variant in the reads and its frequency in the sampled virions may differ due to PCR bias—amplification of some virions more than others. A proper reconciliation of these frequencies would require modelling the number of virions in the sample, their diversity, the process generating PCR bias, and sequencing error, which is beyond the scope of this work. Included in `shiver` is the option to *deduplicate* mapped reads based on their position: from each set of paired reads with identical mapped coordinates, retaining only one pair and discarding the rest as suspected PCR duplicates (using Picard). This is turned off by default, as decreasing the coverage and discarding some diversity in the reads may not be appropriate for every sequencing protocol. We do not include an option for removal of duplicate reads before mapping based on exact sequence matches, as this preferentially retains reads with sequencing error. Instead of addressing the problem of PCR bias at the analysis stage, it can be addressed with the sequencing protocol: primer IDs (Jabara et al. 2011) can associate every read to its template, allowing identification of all PCR duplicates (as well as permitting separate reconstruction of all haplotypes). As with SGA however, higher costs for each sample currently limit applicability to large population studies.

Another limitation is that no mapping of diverse reads can guarantee perfect accuracy at every position in every sample, as perfect sequence alignment is an unsolved problem. In particular where samples contain indel polymorphisms, or where localised misassembly results in an indel not present in the reads, mapping may misalign reads in a way that is not cured by remapping to their own consensus, since the misalignment gives an error in the consensus. As with all automatic sequence alignment, there is scope for improvement by manual inspection. `shiver`'s performance is also linked to that of the assembler used to produce the input contigs. For a sample with parts of the genome where assembly failed to produce contigs, `shiver`'s reference is constructed using the closest identified reference in lieu of the missing contigs. For such samples the bias of mapping to an existing reference is still present to some degree, though mitigated by `shiver`'s option to map a second time to its initial consensus.

For sequences that are recombinants of a type not seen in existing reference sets, `shiver` will nevertheless construct an appropriate reference for mapping provided contigs were fully assembled from the available reads, i.e. either the contigs span the whole genome, or they are missing only where reads are missing. As `shiver` fills in gaps between contigs using the single closest existing reference (supplemented by further existing references only at the ends, i.e. if the closest reference is shorter than some others), in the event of partial assembly failure for a novel recombinant this might not produce a mapping reference as well tailored to the sample as some process of mixing different existing references at different parts of the genome to locally match the available contigs. However `shiver`'s second round of mapping to the first round's consensus will partially mitigate this, and as novel recombinant samples with partial assembly failure are expected to be rare (noting that the success of *de novo* assembly is independent of subtype or recombination), we prefer not to mix existing references throughout the genome, for simplicity and robustness to reference misalignment.

A design choice is that `shiver` does not take into account translation to amino acids, and in particular does not bias towards maintaining reading frames. Deliberately including this bias would be clearly justified for many organisms, but the case is arguable for HIV due to overlapping reading frames, frameshifting polymorphisms, and possibly antisense expression (Miller 1988; Cassan et al. 2016). Other tools exist to extract in-frame gene sequences from `shiver` consensuses, such as Gene Cutter (https://www.hiv.lanl.gov/content/sequence/GENE_CUTTER/cutter.html).

Individuals who are dually infected—hosting two distinct quasispecies, whether by two distinct founder viruses establishing productive infections, or by superinfection—are known to be special cases clinically, and perhaps for evolution, because of the opportunity for recombination. It is important to note that they are also special cases for bioinformatic processing (Giallonardo et al. 2014). If one of the two quasispecies is more highly represented in the reads at every position in the genome, the consensus sequence for the infected individual will be simply the consensus of the more abundant quasispecies. However if one quasispecies has more reads at part of the genome and the other has more reads elsewhere in the genome, the consensus will be a recombinant of both quasispecies; a recombinant which may never have existed *in vivo*, and which may invalidate phylogenies in which it is included. Clearly, care must be taken in identifying such individuals as their dually infected status may not be known.

Our focus here has been reconstruction of the consensus sequence that summarises a quasispecies. The process of doing this from diverse reads—from different virions in the quasispecies—retains rich information on within-host diversity. Our separate tool phyloscanner (Wymant et al. 2017) allows easy extraction, processing, alignment and parallel phylogenetic analysis of the short reads from many genomic windows of many mapped read files, for example those produced by `shiver`. Examination of within-host and between-host diversity together, at every position along the genome, allows identification of dual infections, transmission, recombination and contamination. These more detailed pictures of quasispecies and the relationships between them, in addition to their summaries as consensus sequences, further motivate the valuable role NGS has to play in our understanding of HIV.

Data availability

The Miseq short reads processed here are publicly available on the European Nucleotide Archive: accession numbers ERR732065–ERR732132. The newly generated Hiseq short reads processed here will be made available subject to a data access request, to ensure patient confidentiality is protected.

Supplementary data

Supplementary data are available at *Virus Evolution* online.

Conflict of interest: A.J.G. participated in an advisory board meeting for ViiV Healthcare in July 2016. K.P. is a member of the Viiv Dolutegravir' Advisory Board and Viiv Data and Insights: Standardisation in Measuring and Collecting Care Continuum Data Advisory Board. H.G. reports receipt of grants from the Swiss National Science Foundation, Swiss HIV Cohort Study, University of Zurich, Yvonne Jacob Foundation, and Gilead Sciences; fees for data and safety monitoring board membership from Merck; consulting/advisory board membership fees from Gilead Sciences; and travel reimbursement from Gilead, Bristol-Myers Squibb, and Janssen. P.R. through his institution has received independent scientific grant support from Gilead Sciences, Janssen Pharmaceuticals Inc, Merck & Co, Bristol-Myers Squibb, and ViiV Healthcare; he has served on scientific advisory boards for Gilead Sciences and ViiV Healthcare and on a data safety monitoring committee for Janssen Pharmaceuticals Inc, for which his institution has received remuneration.

Acknowledgements

Thanks to Martin Hunt and Dan Frampton for helpful discussions and to Simon Burbidge and Matt Harvey for help with Imperial College London High Performance Cluster computing. This work was funded by ERC Advanced Grant (PBDR-339251). This work used the computing resources of the UK MEDical BIOinformatics partnership—aggregation, integration, visualization, and analysis of large, complex data (UK MED-BIO), which is supported by the Medical Research Council [grant number MR/L01632X/1].

References

Afgan, E. et al. (2015) 'Genomics Virtual Laboratory: A Practical Bioinformatics Workbench for the Cloud', *Plos One*, 10: e0140829–0.

Albers, C. A. et al. (2011) 'Dindel: Accurate Indel Calls from Short-Read Data', *Genome Research*, 21: 961–73. DOI: 10.1101/gr.112326.110.

Altschul, S. F. et al. (1990) 'Basic Local Alignment Search Tool', *Journal of Molecular Biology*, 215: 403–10.

Archer, J. et al. (2010) 'The Evolutionary Analysis of Emerging Low Frequency HIV-1 CXCR4 Using Variants through Time—an Ultra-Deep Approach', *PLoS Computational Biology*, 6: e1001022–11.

Bankevich, A. et al. (2012) 'SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing', *Journal of Computational Biology*, 19: 455–77.

Beerenwinkel, N. et al. (2012) 'Challenges and Opportunities in Estimating Viral Genetic Diversity from Next-Generation Sequencing Data', *Frontiers in Microbiology*, 3: 329.

Blanquart, F. et al. (2017) 'Viral Genetic Variation Accounts for a Third of Variability in HIV-1 Set-Point Viral Load in Europe. (R. Sanjuán, Ed.)', *PLoS Biology*, 15: e2001855.

Bolger, A. M., Lohse, M., and Usadel, B. (2014) 'Trimmomatic: A Flexible Trimmer for Illumina Sequence Data', *Bioinformatics (Oxford, England)*, 30: 2114–20.

Bose, M. E. et al. (2015) 'Sequencing and Analysis of Globally Obtained Human Respiratory Syncytial Virus a and B Genomes', *PLoS One*, 10: /e0120098–22.

Bray, N. L. et al. (2016). Near-optimal probabilistic RNA-seq quantification, 34: 525EP–. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved. SN. DOI: 10.1038/nbt.3519.

Brener, J. et al. (2015) 'Disease Progression despite Protective HLA Expression in an HIV-Infected Transmission Pair', *Retrovirology*, 12: 1–13.

Cassan, E. et al. (2016) 'Concomitant Emergence of the Antisense Protein Gene of HIV-1 and of the Pandemic', *Proceedings of the National Academy of Sciences of the United States of America*, 113: 11537–42.

Connor, T. R. et al. (2016) 'CLIMB (the Cloud Infrastructure for Microbial Bioinformatics): An Online Resource for the Medical Microbiology Community', *Microbial Genomics*, 2: e000086. DOI: 10.1101/064451.

Cornelissen, M. et al. (2016) 'From Clinical Sample to Complete Genome: Comparing Methods for the Extraction of HIV-1 RNA for High-Throughput Deep Sequencing', *Virus Research*, 239: 10–16. DOI: http://dx.doi.org/10.1016/j.virusres.2016.08.004.

Croucher, N. J. et al. (2009) 'A Simple Method for Directional Transcriptome Sequencing Using Illumina Technology', *Nucleic Acids Research*, 37: e148.

Fraser, C. et al. (2014) 'Virulence and Pathogenesis of HIV-1 Infection: An Evolutionary Perspective', *Science*, 343: 1243727.

Gall, A. et al. (2012) 'Universal Amplification, Next-Generation Sequencing, and Assembly of HIV-1 Genomes', *Journal of Clinical Microbiology*, 50: 3838–44.

Giallonardo, F. D. et al. (2014) 'Full-Length Haplotype Reconstruction to Infer the Structure of Heterogeneous Virus Populations', *Nucleic Acids Research*, 42: e115.

Gibson, R. M. et al. (2014) 'Sensitive Deep-Sequencing-Based HIV-1 Genotyping Assay to Simultaneously Determine Susceptibility to Protease, Reverse Transcriptase, Integrase, and Maturation Inhibitors, as Well as HIV-1 Coreceptor Tropism', *Antimicrobial Agents and Chemotherapy*, 58: 2167–85.

Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016) 'Coming of Age: Ten Years of Next-Generation Sequencing Technologies', *Nature Reviews Genetics*, 17: 333–51.

Grenfell, B. T. et al. (2004) 'Unifying the Epidemiological and Evolutionary Dynamics of Pathogens', *Science*, 303: 327–32.

Henn, M. R. et al. (2012) 'Whole Genome Deep Sequencing of HIV-1 Reveals the Impact of Early Minor Variants upon Immune Recognition during Acute Infection', *PLoS Pathogens*, 8: e1002529–14.

Ho, S. K. et al. (2008) 'Drug-Associated Changes in Amino Acid Residues in Gag p2, p7NC, and p6Gag/p6Pol in Human Immunodeficiency Virus Type 1 (HIV-1) Display a Dominant Effect on Replicative Fitness and Drug Response', *Virology*, 378: 272–81.

Hunt, M. et al. (2015) 'IVA: Accurate De Novo Assembly of RNA Virus Genomes', *Bioinformatics*, 31: 2374–6. DOI: 10.1093/bioinformatics/btv120.

Iqbal, Z. et al. (2012) 'De Novo Assembly and Genotyping of Variants Using Colored De Bruijn Graphs', *Nature Genetics*, 44: 226–32.

Jabara, C. B. et al. (2011) 'Accurate Sampling and Deep Sequencing of the HIV-1 Protease Gene Using a Primer ID', *Proceedings of the National Academy of Sciences of the United States of America*, 108: 20166–71.

Judo, M. S. B., Wedel, A. B., and Wilson, C. (1998) 'Stimulation and Suppression of PCR-Mediated Recombination', *Nucleic Acids Research*, 26: 1819–25.

Katoh, K. et al. (2002) 'MAFFT: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform', *Nucleic Acids Research*, 30: 3059–66.

Keele, B. F. et al. (2008) 'Identification and Characterization of Transmitted and Early Founder Virus Envelopes in Primary HIV-1 Infection', *Proceedings of the National Academy of Sciences of the United States of America*, 105: 7552–7.

Kuiken, C. et al. (2005) 'The Los Alamos Hepatitis C Sequence Database', *Bioinformatics (Oxford, England)*, 21: 379–84.

Kunin, V. et al. (2008) 'A Bioinformatician's Guide to Metagenomics', *Microbiology and Molecular Biology Reviews*, 72: 557–78.

Langmead, B. et al. (2009) 'Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome', *Genome Biology*, 10: R25.

Li, H., and Durbin, R. (2010) 'Fast and Accurate Long-Read Alignment with Burrows–Wheeler Transform', *Bioinformatics (Oxford, England)*, 26: 589–95.

—— et al. (2009) 'The Sequence Alignment/Map (SAM) Format and SAMtools', *Bioinformatics*, 25: 2078–9. DOI: 10.1093/bioinformatics/btp352.

——, Ruan, J., and Durbin, R. (2008) 'Mapping Short DNA Sequencing Reads and Calling Variants Using Mapping Quality Scores', *Genome Research*, 18: 1851–8.

Malboeuf, C. M. et al. (2013) 'Complete Viral RNA Genome Sequencing of Ultra-Low Copy Samples by Sequence-Independent Amplification', *Nucleic Acids Research*, 41: e13.

McElroy, K., Thomas, T., and Luciani, F. (2014) 'Deep Sequencing of Evolving Pathogen Populations: Applications, Errors, and Bioinformatic Solutions', *Microbial Informatics and Experimentation*, 4: 1–14.

McKenna, A. et al. (2010) 'The Genome Analysis Toolkit: A MapReduce Framework for Analyzing Next-Generation DNA Sequencing Data', *Genome Research*, 20: 1297–303. DOI: 10.1101/gr.107524.110.

Meyerhans, A., Vartanian, J.-P., and Wain-Hobson, S. (1990) 'DNA Recombination during PCR', *Nucleic Acids Research*, 18: 1687–91.

Miller, R. H. (1988) 'Human Immunodeficiency Virus May Encode a Novel Protein on the Genomic DNA plus Strand', *Science*, 239: 1420–2.

Ode, H. et al. (2015) 'Quasispecies Analyses of the HIV-1 near-Full-Length Genome with Illumina MiSeq', *Frontiers in Microbiology*, 6: 1258. DOI: 10.3389/fmicb.2015.01258.

Palmer, S. et al. (2005) 'Multiple, Linked Human Immunodeficiency Virus Type 1 Drug Resistance Mutations in Treatment-Experienced Patients Are Missed by Standard Genotype Analysis', *Journal of Clinical Microbiology*, 43: 406–13.

Ratmann, O. et al. (2017) 'HIV-1 Full-Genome Phylogenetics of Generalized Epidemics in Sub-Saharan Africa: Impact of Missing Nucleotide Characters in Next-Generation Sequences', *AIDS Research and Human Retroviruses*, 33: 1083–98.

Simmonds, P. et al. (1990) 'Analysis of Sequence Diversity in Hypervariable Regions of the External Glycoprotein of Human Immunodeficiency Virus Type 1', *Journal of Virology*, 64: 5840–50.

Starcich, B. R. et al. (1986) 'Identification and Characterization of Conserved and Variable Regions in the Envelope Gene of HTLV-III/LAV, the Retrovirus of AIDS', *Cell*, 45: 637–48.

Struck, D. et al. (2014) 'COMET: Adaptive Context-Based Modeling for Ultrafast HIV-1 Subtype Identification', *Nucleic Acids Research*, 42: e144.

Tamiya, S. et al. (2004) 'Amino Acid Insertions near Gag Cleavage Sites Restore the Otherwise Compromised Replication of Human Immunodeficiency Virus Type 1 Variants Resistant to Protease Inhibitors', *Journal of Virology*, 78: 12030–40.

Thomas, T., Gilbert, J., and Meyer, F. (2012) 'Metagenomics - A Guide from Sampling to Data Analysis', *Microbial Informatics and Experimentation*, 2: 3.

Verbist, B. M. P. et al. (2014) 'VirVarSeq: A Low-Frequency Virus Variant Detection Pipeline for Illumina Sequencing Using Adaptive Base-Calling Accuracy Filtering', *Bioinformatics*, 31: 94–101. DOI: 10.1093/bioinformatics/btu587.

Willerth, S. M. et al. (2010) 'Development of a Low Bias Method for Characterizing Viral Populations Using Next Generation Sequencing Technology', *PLoS One*, 5: e13564–9.

Wood, N. et al. (2009) 'HIV Evolution in Early Infection: Selection Pressures, Patterns of Insertion and Deletion, and the Impact of APOBEC', *PLoS Pathogens*, 5: e1000414–6.

Wymant, C. et al. (2017) 'PHYLOSCANNER: Inferring Transmission from within- and between-Host Pathogen Genetic Diversity', *Molecular Biology and Evolution*, 35: 719–33. DOI: 10.1093/molbev/msx304.

Yang, X. et al. (2012) 'De Novo Assembly of Highly Diverse Viral Populations', *BMC Genomics*, 13: 1–13.

Ye, K. et al. (2009) 'Pindel: A Pattern Growth Approach to Detect Break Points of Large Deletions and Medium Sized Insertions from Paired-End Short Reads', *Bioinformatics*, 25: 2865–71.

Zanini, F. et al. (2015) 'Population Genomics of Intrapatient HIV-1 Evolution (A. K. Chakraborty, Ed.)', *eLife*, 4: e11282. DOI: 10.7554/eLife.11282.

Supplementary Table S1: statistics for the processed samples. In the 'Notes' column: 'QC failure' indicates those samples that failed QC checks for assembly in the IVA publication; 'Contig correction' indicates those samples specifically chosen as examples of the need for shiver's correction of structural problems in contigs; 'Lane differences' indicates the sample chosen to illustrate differences in assembly output for technical replicate samples (i.e. the contigs for the two Hiseq lanes for this sample).

Sample	Miseq (M) or Hiseq (H) data	Closest identified real reference	shiver consensus length	Extra length in shiver consensuses (c.f. real reference consensus)	Extra length in shiver consensuses (c.f. real reference consensus) at the ends	Extra length in shiver consensus (c.f. real reference consensus) internally
ERR732065	M	B.AU.87.MBC925.AF042101	8198	174	0	174
ERR732066	M	B.DK.01.CTL_035.EF514710	7354	306	193	113
ERR732067	M	02_AG.GH.97.97GH_AG1.AB049811	5767	-54	0	-54
ERR732068	M	B.US.86.5096_86.AY835749	4561	2	0	2
ERR732069	M	B.US.00.THRO_TF1.JN944930	5695	54	0	54
ERR732070	M	A1.KE.04.04KE354207V3.KT022363	8077	410	289	121
ERR732071	M	11_cpx.CM.04.1230_24.KP718938	8230	471	12	459
ERR732072	M	B.KR.92.HP_10_02SHJ8_6986.KJ140255	8072	220	79	141
ERR732073	M	C.ZA.03.03ZASK107B1.DQ056410	9018	84	70	14
ERR732074	M	C.TZ.08.707010457_CH457.w8.KC156220	9048	86	0	86
ERR732076	M	C.ZA.99.99ZALT21.EU293446	9053	81	1	80
ERR732077	M	C.ZA.03.03ZASK107B1.DQ056410	9053	243	70	173
ERR732078	M	C.ZA.99.99ZALT21.EU293446	9053	78	0	78
ERR732079	M	C.ZA.03.03ZASK107B1.DQ056410	9053	283	70	213
ERR732080	M	C.ZA.03.03ZASK107B1.DQ056410	9029	189	70	119
ERR732081	M	C.ZA.03.03ZASK107B1.DQ056410	9023	185	70	115
ERR732082	M	C.IN.00.DEMC00IN008.KP109483	9038	258	185	73
ERR732083	M	B.TW.94.TWCYS_LM49.AF086817	9002	74	0	74
ERR732085	M	B.KR.93.HP_17_02LSP11_2268.KJ140262	9000	230	120	110
ERR732086	M	BF1.BR.10.10BR_RJ075.KT427651	9002	140	59	81
ERR732087	M	BF1.BR.10.10BR_RJ075.KT427651	9002	149	59	90
ERR732088	M	C.ZM.02.02ZMBC.AB254149	9027	1	0	1
ERR732089	M	B.KR.93.HP_17_02LSP11_2268.KJ140262	9005	173	121	52
ERR732090	M	B.KR.93.HP_17_02LSP11_2268.KJ140262	9002	132	0	132
ERR732091	M	01_AE.GB.10.Donor_N094_20_Month.KP873161	9048	18	3	15
ERR732092	M	01_AE.GB.10.Donor_N094_20_Month.KP873161	9030	4	3	1
ERR732093	M	B.KR.04.04WK7_HIV_1_wk.DQ295194	7312	86	-1	87
ERR732094	M	C.ZA.03.03ZASK107B1.DQ056410	9017	88	70	18
ERR732095	M	C.ZA.03.03ZASK107B1.DQ056410	9019	92	73	19
ERR732096	M	C.ZA.03.03ZASK107B1.DQ056410	9017	80	70	10
ERR732097	M	C.ZA.03.03ZASK107B1.DQ056410	9045	108	70	38
ERR732098	M	B.KR.93.HP_17_02LSP11_2268.KJ140262	9020	311	0	311
ERR732099	M	B.KR.93.HP_17_02LSP11_2268.KJ140262	9020	303	80	223
ERR732100	M	01_AE.GB.10.Donor_N094_20_Month.KP873161	9030	6	3	3
ERR732101	M	01_AE.GB.10.Donor_N094_20_Month.KP873161	9036	0	3	-3
ERR732102	M	C.TZ.08.707010457_CH457.w8.KC156220	9014	34	0	34
ERR732103	M	B.KR.93.HP_17_02LSP11_2268.KJ140262	9000	223	0	223
ERR732104	M	C.ZA.99.99ZALT21.EU293446	9029	56	0	56
ERR732105	M	C.ZA.99.99ZALT21.EU293446	9045	74	0	74
ERR732106	M	B.US.03.933384.KT124807	8961	447	374	73
ERR732107	M	B.US.06.06US_SAJ_C166_SG.JF689864	8967	608	380	228
ERR732108	M	B.US.13.862898.KT124796	9010	385	372	13
ERR732109	M	B.US.07.HIV_US_BID_V4516_2007.JQ403096	4559	293	293	0
ERR732110	M	B.AU.95.C24.AF538304	9037	334	115	219
ERR732111	M	B.CY.05.CY124.FJ388933	9064	1262	1077	185
ERR732112	M	B.JP.x.DR1712.AB604946	8127	80	0	80
ERR732113	M	B.KR.93.HP_17_02LSP11_2268.KJ140262	7353	207	0	207
ERR732114	M	B.KR.93.HP_17_02LSP11_2268.KJ140262	9008	299	0	299
ERR732115	M	B.KR.93.HP_17_02LSP11_2268.KJ140262	7957	501	232	269
ERR732116	M	B.KR.93.HP_17_02LSP11_2268.KJ140262	8999	270	80	190
ERR732117	M	01_AE.GB.10.Donor_N094_20_Month.KP873161	5004	53	3	50
ERR732118	M	01_AE.GB.10.Donor_N094_20_Month.KP873161	9070	65	22	43
ERR732119	M	C.ZA.05.05ZASK245B1.DQ369982	7355	145	70	75
ERR732120	M	B.US.87.5113_87.AY835758	7351	131	0	131
ERR732121	M	B.US.87.5113_87.AY835758	7351	131	0	131
ERR732122	M	B.US.04.ES8_43.EF363126	7328	216	-1	217
ERR732123	M	B.UY.99.99UY_TRA0177.JN235965	7763	596	299	297

Sample	Miseq (M) or Hiseq (H) data	Closest identified real reference	shiver consensus length	Extra length in shiver consensuses (c.f. real reference consensus)	Extra length in shiver consensuses (c.f. real reference consensus) at the ends	Extra length in shiver consensus (c.f. real reference consensus) internally
ERR732124	M	B.KR.05.05YJN2.JQ316134	1911	0	0	0
ERR732126	M	B.US.07.07US_SAJ_C161_H1.JF689883	7346	379	298	81
ERR732127	M	B.US.11.ES22_27.KF384808	1910	0	0	0
ERR732128	M	B.US.08.HIV_US_BID_V4489_2008.JQ403094	7345	492	293	199
ERR732129	M	B.GB.05.MM43d368_GN1.HM586209	9023	364	381	-17
ERR732130	M	B.GB.05.MM43d368_GN1.HM586209	9042	382	382	0
ERR732131	M	B.GB.05.MM43d368_GN1.HM586209	9022	377	381	-4
ERR732132	M	B.GB.05.MM43d368_GN1.HM586209	9020	365	381	-16
17621_3_80	H	0107.CN.07.JL070032.KC990127	8993	168	78	90
17653_3_25	H	B.JP.x.JRC65B.AB565502	9007	73	0	73
17653_3_36	H	B.FR.11.DEMB11FR001.KF716496	8932	253	257	-4
17653_3_56	H	02_AG.CM.01.01CM_0002BBY.AY371122	8994	643	638	5
17653_3_62	H	09_cpx.SN.95.95SN7808.AY093604	9080	454	336	118
17653_3_64	H	22_01A1.CM.01.01CM_0001BBY.AY371159	9057	712	647	65
17653_3_72	H	B.US.11.CP7_2B.KF384805	8999	158	89	69
17653_3_74	H	B.AU.86.MBC200.AF042100	9031	19	0	19
17654_3_46	H	B.YE.02.02YE508.AY795905	9026	327	299	28
17654_3_71	H	02_AG.CM.02.02CM_4082STN.AY371141	9042	640	638	2
17654_3_72	H	B.JP.98.DR1120.AB480698	9071	140	3	137
17654_3_78	H	B.KR.04.04KMK5.JQ316126	9046	205	92	113
17795_3_40	H	B.AU.86.MBC200.AF042100	9026	1	0	1
17796_3_1	H	14_BG.ES.05.X1870.FJ670522	9076	104	48	56
17796_3_29	H	B.DE.04.963987.KT124812	8971	361	293	68
17796_3_30	H	B.DE.86.D31.U43096	8966	61	0	61
17796_3_35	H	B.US.07.07US_SAJ_C166_MS.JF689886	9009	351	304	47
18209_3_31	H	C.ZA.03.03ZASK107B1.DQ056410	9039	110	70	40
18209_3_36	H	C.ZA.04.SK133B1.AY772698	8984	496	73	423
18209_3_38	H	B.ES.09.DEMB09ES007.KC473841	9003	145	137	8
19561_3_127	H	C.ZA.03.SK041B1.AY772693	9044	218	122	96
19562_3_109	H	01_AE.VN.97.97VNAG218.FJ185255	9041	299	257	42
19562_3_2	H	B.US.07.HIV_US_BID_V3120_2007.JQ403078	9014	369	293	76
19562_3_30	H	C.ZA.07.705010162_CH162.mo6.KC156115	8971	8	0	8
19562_3_31	H	B.JP.08.NMC104_clone_01.AB731663	9018	140	3	137
19562_3_46	H	B.US.x.AC_16_0_Days_Consen_fa.DQ127537	8904	370	300	70
19562_3_50	H	B.US.85.5077_85.AY835769	9057	70	0	70
19562_3_51	H	B.US.x.CR0192W.FJ469704	9040	315	275	40
19562_3_6	H	D.KE.11.DEMD11KE003.KF716476	9046	400	150	250
19893_3_71	H	01_AE.TH.05.05TH342968.JN248342	9018	340	310	30
19960_3_11	H	B.KR.92.HP_10_02SHJ8_6986.KJ140255	9032	221	104	117
19960_3_116	H	B.GB.x.MANC.U23487	9003	30	0	30
19960_3_119	H	B.FR.83.HXB2_LAI_IIIB_BRU.K03455	9051	87	0	87
19960_3_12	H	B.US.03.CR0154X.FJ469701	8980	309	281	28
19960_3_146	H	B.US.06.502_0346_wg02.JF320097	9026	378	293	85
19960_3_15	H	12_BF.UY.99.URTR23.AF385934	8978	9	0	9
19960_3_16	H	B.US.05.05US_SAJ_NVS12.JF689852	9046	411	309	102
19960_3_17	H	B.BR.04.BREPM1066.FJ195090	9028	64	49	15
19960_3_18	H	B.CY.08.CY226.JF683775	9002	962	298	664
19960_3_22	H	02_AG.DE.09.701114.KT124792	9120	429	309	120
19960_3_28	H	17_BF.BO.02.BO02_BOL119.EU581827	8948	293	296	-3
19960_3_40	H	A1.KE.99.KSM4021.AF457075	8986	690	298	392
19960_3_44	H	B.DE.86.HAN.U43141	9002	96	46	50
19960_3_49	H	B.TH.04.04TH803686.JN248333	9005	327	288	39
19960_3_6	H	BC.CN.07.jx070017.KF250384	8962	374	293	81
19960_3_70	H	BC.BR.92.92BR023.HM100716	9001	13	0	13
19960_3_9	H	B.KR.05.05CSR3.DQ837381	9038	155	38	117
20004_3_146	H	B.US.06.502_0346_wg02.JF320097	8951	339	294	45
20004_3_155	H	A1D.KE.06.06KE894822V7.KT022417	9020	517	285	232
20004_3_56	H	B.US.00.ES1_20.EF363123	8994	41	0	41
minimum	N/A	N/A	1910	-54	-1	-54
median	N/A	N/A	9009	205	70	73
mean	N/A	N/A	8535.7	239.4	143.2	96.2
maximum	N/A	N/A	9120	1262	1077	664

Sample	Number of bases called differently with higher coverage mapping to the shiver reference than to the real reference	Number of bases called differently with higher (or equal) coverage mapping to the real reference than to the shiver reference	Number of positions where at least one of the corrected contigs agrees with the shiver consensus	Number of positions where all corrected contigs disagree with the consensus
ERR732065	2	0	8142	7
ERR732066	1	0	7332	0
ERR732067	5	2	4911	5
ERR732068	1	0	4536	3
ERR732069	4	0	5632	20
ERR732070	16	0	8016	28
ERR732071	26	0	8178	13
ERR732072	42	4	7992	40
ERR732073	51	1	8962	56
ERR732074	53	3	9019	41
ERR732076	9	0	9049	4
ERR732077	19	0	9053	0
ERR732078	10	0	9053	0
ERR732079	20	1	9046	7
ERR732080	19	0	9020	9
ERR732081	32	0	8958	7
ERR732082	14	0	9025	13
ERR732083	27	9	8987	15
ERR732085	40	2	7029	44
ERR732086	16	0	8989	13
ERR732087	7	0	8998	4
ERR732088	52	0	9003	24
ERR732089	43	1	8992	13
ERR732090	57	6	8981	21
ERR732091	23	2	9019	29
ERR732092	42	4	8996	28
ERR732093	5	0	4867	2
ERR732094	30	2	8964	55
ERR732095	38	0	8914	106
ERR732096	31	1	8974	43
ERR732097	38	1	8994	59
ERR732098	2	0	9018	2
ERR732099	13	0	9019	1
ERR732100	23	2	9016	14
ERR732101	15	18	9004	32
ERR732102	55	1	8980	34
ERR732103	37	0	8827	14
ERR732104	25	0	9019	10
ERR732105	15	0	9040	13
ERR732106	19	0	8922	54
ERR732107	4	0	8960	7
ERR732108	2	0	9004	6
ERR732109	0	1	4535	2
ERR732110	16	0	9019	18
ERR732111	8	0	9027	37
ERR732112	5	0	8093	2
ERR732113	2	0	7327	6
ERR732114	5	0	8971	37
ERR732115	3	0	7315	9
ERR732116	22	0	8974	25
ERR732117	12	2	4944	9
ERR732118	8	1	9057	13
ERR732119	11	0	7325	10
ERR732120	6	0	7335	6
ERR732121	8	0	7333	0
ERR732122	3	0	7169	5
ERR732123	6	0	7738	8

Sample	Number of bases called differently with higher coverage mapping to the shiver reference than to the real reference	Number of bases called differently with higher (or equal) coverage mapping to the real reference than to the shiver reference	Number of positions where at least one of the corrected contigs agrees with the shiver consensus	Number of positions where all corrected contigs disagree with the consensus
ERR732124	0	0	1888	0
ERR732126	6	0	7324	3
ERR732127	0	0	1857	30
ERR732128	6	0	6090	7
ERR732129	0	0	9022	1
ERR732130	0	0	9042	0
ERR732131	1	0	9005	21
ERR732132	2	4	8997	23
17621_3_80	41	0	8911	46
17653_3_25	18	0	9001	6
17653_3_36	6	0	8794	3
17653_3_56	6	0	8989	5
17653_3_62	19	0	9062	33
17653_3_64	12	0	9048	9
17653_3_72	7	0	8957	3
17653_3_74	23	0	9029	2
17654_3_46	18	0	9013	11
17654_3_71	28	1	9036	5
17654_3_72	28	0	9069	2
17654_3_78	5	0	9038	8
17795_3_40	31	24	9019	7
17796_3_1	6	0	9026	2
17796_3_29	50	3	6740	3
17796_3_30	23	23	7319	2
17796_3_35	34	1	8035	13
18209_3_31	50	11	9017	19
18209_3_36	21	0	8980	4
18209_3_38	12	0	8999	4
19561_3_127	10	0	9039	5
19562_3_109	9	0	9038	3
19562_3_2	17	0	9005	10
19562_3_30	8	0	8960	11
19562_3_31	13	0	9012	3
19562_3_46	18	1	8894	3
19562_3_50	14	0	9052	5
19562_3_51	13	0	9033	7
19562_3_6	18	0	9043	3
19893_3_71	10	0	9014	3
19960_3_11	14	0	9023	9
19960_3_116	8	0	8997	6
19960_3_119	13	0	9039	12
19960_3_12	8	1	8980	0
19960_3_146	19	0	8965	64
19960_3_15	5	0	8973	2
19960_3_16	20	0	9042	4
19960_3_17	11	0	9024	4
19960_3_18	32	0	8998	4
19960_3_22	0	1	9116	4
19960_3_28	5	0	8942	6
19960_3_40	10	0	8983	3
19960_3_44	21	0	8999	3
19960_3_49	15	0	8992	13
19960_3_6	18	0	8959	3
19960_3_70	20	0	8993	8
19960_3_9	11	0	9029	6
20004_3_146	12	0	8941	10
20004_3_155	4	0	9018	2
20004_3_56	4	0	8987	5
minimum	0	0	1857	0
median	13	0	8992	7
mean	16.8	1.2	8408.7	13.7
maximum	57	24	9116	106

Sample	Number of positions where at least one contig has a base and the shiver mapping failed to call a base	Number of positions where there is no contig coverage but the shiver consensus has a base	Notes
ERR732065	0	49	QC failure
ERR732066	0	22	QC failure
ERR732067	0	852	QC failure
ERR732068	0	22	QC failure
ERR732069	0	43	QC failure
ERR732070	0	33	QC failure
ERR732071	0	39	QC failure
ERR732072	0	41	QC failure
ERR732073	0	0	
ERR732074	0	0	
ERR732076	0	0	
ERR732077	0	0	
ERR732078	0	0	
ERR732079	0	0	
ERR732080	0	0	
ERR732081	0	58	
ERR732082	0	0	
ERR732083	0	0	
ERR732085	0	1927	
ERR732086	0	0	
ERR732087	0	0	
ERR732088	0	0	
ERR732089	0	0	
ERR732090	0	0	
ERR732091	0	0	
ERR732092	0	6	
ERR732093	0	2443	QC failure
ERR732094	0	4	
ERR732095	0	0	
ERR732096	0	0	
ERR732097	0	3	
ERR732098	0	0	
ERR732099	0	0	
ERR732100	0	0	
ERR732101	0	0	
ERR732102	0	0	
ERR732103	0	159	
ERR732104	0	0	
ERR732105	0	0	
ERR732106	0	0	
ERR732107	0	0	
ERR732108	0	0	
ERR732109	0	22	QC failure
ERR732110	0	0	
ERR732111	0	0	
ERR732112	0	32	QC failure
ERR732113	0	20	QC failure
ERR732114	0	0	
ERR732115	0	633	QC failure
ERR732116	0	0	
ERR732117	0	51	QC failure
ERR732118	0	0	
ERR732119	0	20	QC failure
ERR732120	0	10	QC failure
ERR732121	0	18	QC failure
ERR732122	0	154	QC failure
ERR732123	0	17	QC failure

Sample	Number of positions where at least one contig has a base and the shiver mapping failed to call a base	Number of positions where there is no contig coverage but the shiver consensus has a base	Notes
ERR732124	0	23	QC failure
ERR732126	0	19	QC failure
ERR732127	0	23	QC failure
ERR732128	0	1248	QC failure
ERR732129	0	0	
ERR732130	0	0	
ERR732131	0	0	
ERR732132	0	0	
17621_3_80	0	36	Contig correction
17653_3_25	0	0	
17653_3_36	0	135	
17653_3_56	0	0	
17653_3_62	0	0	Contig correction
17653_3_64	0	0	
17653_3_72	0	39	
17653_3_74	0	0	
17654_3_46	0	5	
17654_3_71	0	1	
17654_3_72	0	0	
17654_3_78	0	0	
17795_3_40	0	0	Contig correction
17796_3_1	0	48	Contig correction
17796_3_29	0	2228	Contig correction
17796_3_30	0	1645	Lane differences
17796_3_35	0	967	Contig correction
18209_3_31	0	6	Contig correction
18209_3_36	0	0	Contig correction
18209_3_38	0	0	Contig correction
19561_3_127	0	0	
19562_3_109	0	0	
19562_3_2	0	0	
19562_3_30	0	0	
19562_3_31	0	3	
19562_3_46	0	7	
19562_3_50	0	0	
19562_3_51	0	0	
19562_3_6	0	0	
19893_3_71	0	1	
19960_3_11	0	0	
19960_3_116	0	0	
19960_3_119	0	0	
19960_3_12	0	0	
19960_3_146	0	0	
19960_3_15	0	3	
19960_3_16	0	0	
19960_3_17	0	0	
19960_3_18	0	0	
19960_3_22	0	0	
19960_3_28	0	0	
19960_3_40	0	0	
19960_3_44	0	0	
19960_3_49	0	0	
19960_3_6	0	0	
19960_3_70	0	0	
19960_3_9	0	3	
20004_3_146	0	0	
20004_3_155	0	0	
20004_3_56	0	2	
minimum	0	0	
median	0	0	
mean	0.0	114.1	
maximum	0	2443	

Supplementary Information for *Easy and Accurate Reconstruction of Whole HIV Genomes from Short-Read Sequence Data with SHIVER*

SI 1 Sequencing Platform Usage Statistics for HIV

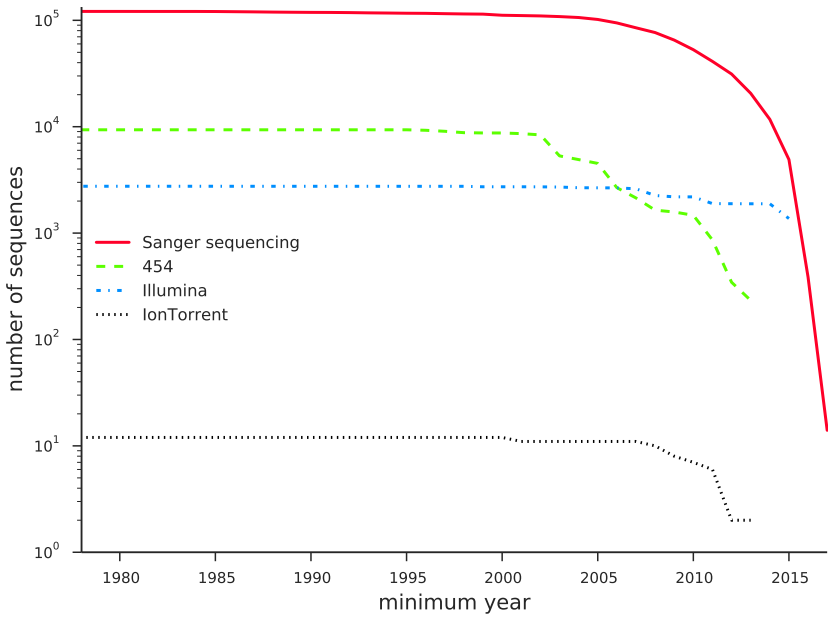

Figure S1: the number of HIV sequences available from the Los Alamos National Laboratory database on 11th Oct 2017 with sampling year and sequencing platform information available, as a function of minimum sampling year for inclusion (i.e. restricting the included sequences to increasingly recent ones).

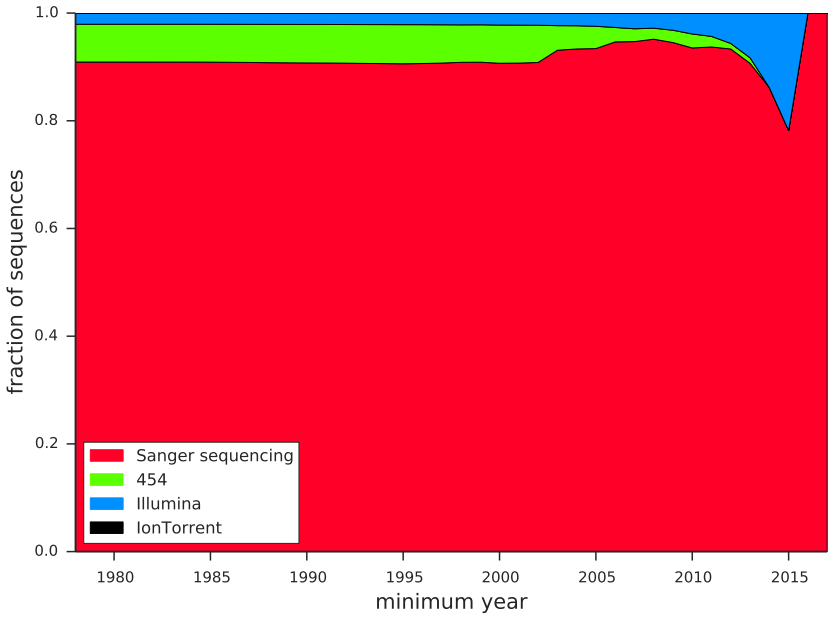

Figure S2: As Fig. S1 but showing the fraction for each platform.

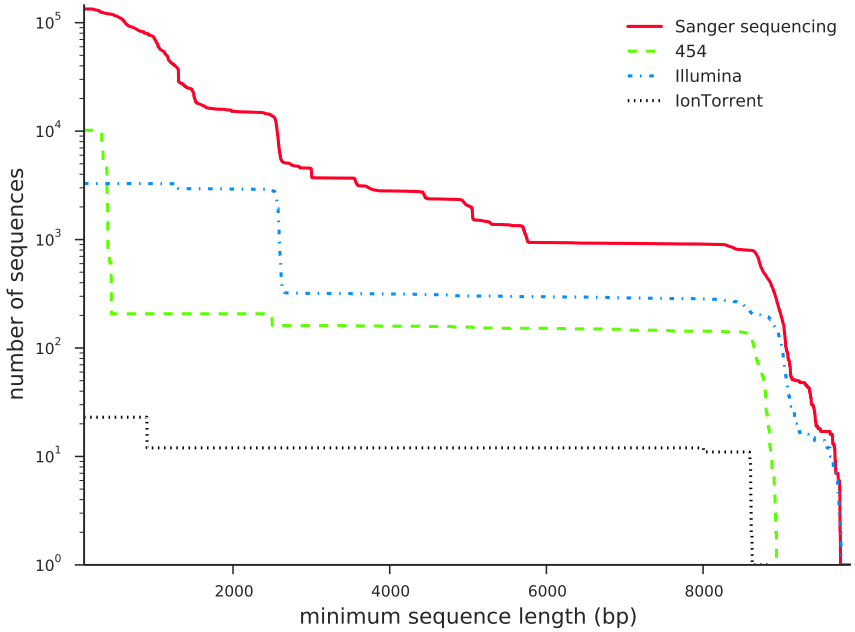

Figure S3: the number of HIV sequences available from the Los Alamos National Laboratory database on 11th Oct 2017 with sequencing platform information available, as a function of minimum sequence length for inclusion (i.e. restricting the included sequences to increasingly long ones).

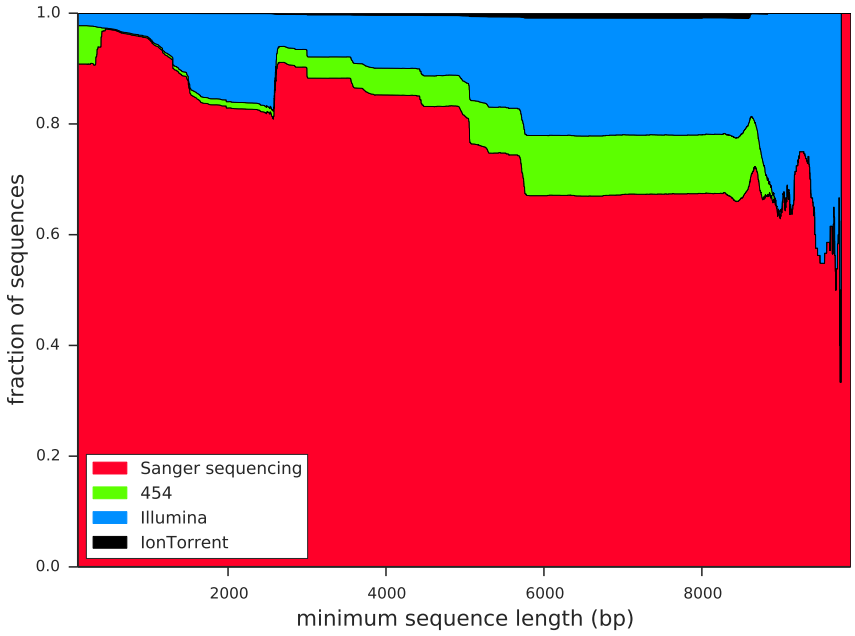

Figure S4: As Fig. S3 but showing the fraction for each platform.

SI 2 Our Method in More Detail

`shiver` is under continuing development; if at a later date description here contradicts descriptions at github.com/ChrisHIV/shiver, the latter has precedence.

2.1 Existing References

An alignment of existing reference sequences is required as input for `shiver`. Construction of a custom reference for mapping involves identifying the existing references that are closest to the sample under consideration. The greater the number and diversity of existing references given as input, the denser and broader the coverage of sequence space is, and the closer the closest reference is expected to be, with corresponding benefits for the accuracy of the results. However these existing references should be aligned to each other accurately, in order for the addition of each sample's contigs to the alignment to be meaningful; this means that producing such an input by automatically aligning a large number of diverse sequences without checking the results would be a bad idea. You will use this alignment as input for every sample in a dataset processed by `shiver`, and so we advise putting a little thought into sequence selection and manually curating the alignment if needed.

2.2 Constructing a Tailored Reference Using the Contigs

Custom reference construction begins with contig preprocessing as follows. Matches between the contigs and any existing reference from the alignment are searched for using `BLASTN` with default settings, except for the `-max_target_seqs 1` option (specifying that all reported hits are to a single reference only), and with `-word_size` set to 17 (this can be changed in `shiver`'s configuration file). Contaminant sequence is inevitable in high-throughput NGS; any contig that has no `BLASTN` hit to any of the HIV references is taken to be contamination, and is put aside for later use, leaving contigs that are putatively HIV. The `BLASTN` results are used to correct the contigs in three ways.

1. Where a single contig has multiple `BLASTN` hits (discarding any hit wholly contained inside another hit), we consider this evidence that the contig is spliced – concatenating two separated regions of the genome – due to errors *in silico* or during sequencing, as mentioned in the introduction. We correct this by cutting the contig into separate contigs at the midpoint between the hits.

2. We trim off any part of the contig that was not spanned by a `BLASTN` hit. The ends of contigs are by definition points at which the assembler has been unable to continue extending the sequence, either because of lack of reads, or because the within-sample diversity has become too great for a single, meaningful, representative sequence to be chosen. The latter possibility also means erroneous bases are more common in short stretches of sequence at the end of a contig. Trimming such sequence from the ends of contigs means the corresponding sequence from the closest existing reference will be used instead, giving a better reference for mapping. (Some assembly algorithms trim a fixed length from the ends of contigs for precisely this reason; however trimming a variable length dependent on its match to known sequence is clearly preferable.)

3. Any contig whose `BLASTN` hit is in the opposite orientation is reverse-complemented. If the assembler does not orientate the contigs, on average half of them will be in the reverse orientation. `IVA` orientates contigs such that the longest open reading frame is on the forward strand, however for very short contigs this may fail. In the process of assembling a spliced contig, an assembler may concatenate different regions in different orientations; `shiver` considers whether each separate part of a split contig requires reverse-complementation.

Contigs are then aligned to the existing reference alignment using `MAFFT`, trying both `--add` and `--addfragments` modes and using the one with the smallest maximum gap fraction (the maximum calculated over all contigs in each alignment). After alignment, a contig found to have an overly large internal deletion (by default 160bp) is split into two separate contigs at that point. This has the same role as `BLASTN`-based correction step 1 above, serving as a backup.

The alignment of contigs to the set of existing references should be visually inspected at this point. For HIV sequences, reference [1] states that "Algorithmic alignment does not necessarily retrieve the best alignment. It is important to always verify whether the sequence data are aligned unambiguously and, if necessary, manually correct the alignment." Reference [2] echoes this for any evolving pathogen: "the

'best' alignment chosen by an alignment program is not necessarily the 'true' alignment... Alignment quality should also be inspected manually in a visualisation program". The commonness of indels in HIV makes alignment more difficult, as does the fact that the contigs may be an imperfect representation of the true sample even after correction. We used `Geneious` [3] for sequence visualisation and editing where needed.

As well as revealing alignment error, inspection of the aligned contigs allows the user to check for any remaining problems with the contigs. We suggest that in general the user inspects both the alignment of the existing references with the *raw* HIV contigs (before any correction by shiver), and the alignment of the existing references with the *corrected* HIV contigs, as a check that all `shiver`'s modifications of the contigs are desired. An example of when this might not be the case is when the sample contains an indel not observed in the existing reference set, that is large enough to cause the contig to be split in two at that point, but which the user thinks might be genuine rather than an a misassembly (through previous/expert knowledge, or perhaps simply observing the same indel in multiple samples in a dataset). With sufficiently accurate mapping, reads will map here correctly whether or not the reference constructed from the contigs contains the indel, making the question moot; however with mapping inaccuracies of the kind shown in Figure 2 possible, it's best to get the reference's structure as correct as possible before mapping.

Using the alignment of contigs to existing references, the set of contigs is flattened into a single sequence as follows. At positions covered by one contig, its base (or gap character, for a deletion) is used. At positions covered by multiple contigs, we use whatever the longest contig has (be it base or gap). We used this heuristic expecting that, where sufficiently distinct haplotypes exist to result in multiple contigs covering the same place, haplotypes supported by a higher depth of reads would tend to be assembled into longer contigs. The sequence resulting from this flattening of the contigs is compared to each existing reference in the alignment in turn: we count towards similarity shared bases and gaps within contigs (known deletions), but not gaps between contigs (missing information). The existing references are ranked by their similarity to the contigs. As existing references have variable lengths (the long terminal repeat regions that flank the clinical genome are sometimes sequenced only partially or not at all), the closest reference is extended outwards using any overhanging sequence from the second closest reference, then the third longest sequence etc. terminating when both edges of the alignment are reached. This sequence – the elongated closest reference – is used to fill in any gaps between (but not inside of) the flattened contigs. This completes production of the reference tailored for this sample.

2.3 Preparing and Mapping the Reads

Before mapping to this reference, the reads are trimmed and cleaned as follows. Adapters, primers and low quality bases are trimmed using `Trimmomatic` and `Fastaq`. We then consider contaminant reads from non-HIV sources. Most of these would presumably be discarded by mapping to an HIV reference, due to lack of similarity. However there is ample opportunity for traces of human DNA to end up in a sample, and sequence of endogenous retroviruses in human DNA may resemble HIV. As a guard against this, and against any other contamination resembling HIV, we use `BLASTN` to find all read pairs that are a better match to one of the contigs previously found to be contamination, than to the tailored reference. These pairs are discarded.

The cleaned reads are mapped to the tailored reference, using `SMALT` by default (with `BWA` and `bowtie` as optional alternatives), giving a file in BAM format. Using `SAMtools` the BAM file is read into pileup format, which is parsed to give base frequencies at each position in the genome. Note that within-host diversity does not consist exclusively of point mutations: indels can be present in some reads and not others (Fig. 7 is an example), which must be dealt with in the pileup. Where some reads have a deletion relative to the reference and others do not, the deletion/gap character can simply be considered as a fifth base whose frequency can be counted like the others. Where some reads have an insertion relative to the reference and others do not, or more generally where insertions of two or more sizes are present, we find the most common insertion size and, inside that insertion, consider only those reads with an insertion of that size (thus avoiding any ambiguity in the alignment of the inserted sequences to each other). Finally, the base frequency file is parsed to call the consensus base at each position. By default the most common base is called to give the consensus, using an ambiguity code only for an exact tie in the frequency of two or more bases; optionally ambiguity codes can be used more readily, when the frequency of the most common base or bases is below a specified threshold. A consensus base is only called if the coverage equals or exceeds a minimum threshold specified by the user, to protect against the effect of residual

low-coverage contaminant reads in genomic regions lacking genuine HIV reads. By default this is 15, but this is likely to need adjusting for different datasets. (See the tool `LinkIdentityToCoverage.py` in section 3.)

2.4 Aligning Multiple Consensuses

Since we know how the consensus aligns to the reference used for mapping, and we know how that reference (constructed from the contigs) aligns to the input alignment of existing references, we can construct a global alignment of the consensuses from all samples merely by coordinate translation, negating the need for further alignment and manual curation. Two things must be excised from the consensus for this global alignment reconstruction: insertions present in the majority of reads but not in their tailored reference (which are rare, since the reference is constructed from the contigs which are constructed from the reads), and insertions present in the contigs but none of the existing references (which are rare provided the set of existing references is large and diverse). In both cases this is sequence whose alignment to the common anchor of the existing references is not known, and so coordinate translation cannot align it.

2.5 Fully Automatic `shiver`

As mentioned, `shiver` can be run from beginning to end without the break in the middle, with the single command `shiver_full_auto.sh`, for uses where visually checking the contigs is impractical. This begins with separation of contigs into HIV (those with `BLASTN` hits) and contamination as previously. Subsequent steps are as follows.

1. The need for contig correction is checked, but correction is not performed: if it is needed, processing stops. Blind trust in the accuracy of an automated alignment of contigs cut into pieces based on evidence of structural problems would be trust misplaced.

2. Each HIV contig is now certain to have a single `BLASTN` hit (discarding any smaller hits wholly contained inside others). That hit is checked to span some minimum fraction of the contig length (by default 90%) as a guard against contigs containing containing some erroneous or foreign sequence; otherwise processing stops.

3. Multiple sequence alignment is performed with these contigs and just one of the existing reference sequences, for each of the existing reference sequences separately.

4. For each such alignment, generated both with regular `mafft` and with `mafft --addfragments`, we calculate the fractional agreement between the flattened contigs and the reference, i.e. the fraction of positions spanned by the reference and at least one contig where the reference and the longest contig agree. Misalignment is penalised in this score because gaps inside contigs are taken as genuine deletions.

5. For the alignment with the highest score, the maximum gap fraction amongst the contigs in the alignment (i.e. the fraction of positions inside the contig that are gaps) is checked to be below a user-specified threshold (the default is 5%, based on analysis of thousands of such alignments that we visually checked) as a further guard against misalignment.

6. The contigs are flattened using this single existing reference to fill in any gaps between them, generating the mapping reference tailored for this sample.

Aligning contigs to the references one at a time (step 3) is simpler for the alignment algorithm than aligning to all of them at once, and means that even if misalignment occurs for what is truly the closest reference to the contigs, the alignment to the second closest can be used instead. Trimming of low-quality bases, trimming of adapter and primer sequences, removal of contaminant reads and mapping to the tailored reference all occur as described previously. For samples that cannot be processed fully automatically this way – when contig correction is required, or a contig is spanned by too small a `BLASTN` hit, or too many gaps are present after alignment – the main mode of shiver is available (for which we advise inspection of the aligned contigs).

As argued earlier, we advocate visually inspecting the aligned contigs, i.e. running the two-command implementation of `shiver` (with the check occurring between the commands). This also has the advantage of working for all samples, whereas `shiver_full_auto.sh` will not proceed if problems with the contigs

or their alignment are detected. `shiver_full_auto.sh` also does not produce a global alignment of all consensuses to each other, because the coordinate translation procedure allowing its construction is derived from each sample's alignment of contigs to all of the references at once. That alignment is produced for the two-command implementation of `shiver`, but step 3 above aligns contigs to references one at a time.

SI 3 Sample Reprocessing and Analysis

Individual steps from `shiver` can be run with stand-alone command line tools, for ease of reapplication elsewhere. For example `CorrectContigs.py` is run with a file of contigs and a file of their `BLASTN` hits to some set of references, and corrects the contigs by cutting, trimming and reverse complementing where needed. Also included in `shiver` are command-line tools for easy analysis and modification of sample output without rerunning the whole pipeline:

- Two parameters specified in the configuration file are a minimum coverage required to call a base (below this coverage, the character '`?`' is used) and a larger minimum coverage required to use upper case instead of lower, as an easy signal of increased confidence. (Note that decreasing these parameters will, in general, allow bases to be called at more positions, giving a longer consensus. However there is a trade-off: where there are fewer reads, the effect of contaminant reads on the consensus may be greater.) To regenerate a consensus with new values of these parameters, `CallConsensus.py` can be run on a sample's base frequencies file. To regenerate a coordinate-translated version of this consensus for the global alignment (of all consensuses produced by `shiver`), `TranslateSeqForGlobalAln.py` can be run on the consensus.

- Another parameter in the configuration file is the minimum read *identity* – the fraction of bases in the read which are mapped and agree with the reference – required for a read to be considered mapped, and so retained in the BAM file. If you wish to increase this after completion of `shiver`, reads with an identity below your new higher threshold can be discarded by running `RemoveDivergentReads.py` on a BAM file. Running `shiver_reprocess_bam.sh` on the resulting BAM file (or indeed any BAM file) implements just the last steps in `shiver`, namely generating pileup, calculating the base frequencies, and calling the consensus.

- `FindNumMappedBases.py` calculates the total number of mapped bases in a BAM file (where read length is constant this equals the number of mapped reads multiplied by read length, minus the total length of sequence clipped from reads), optionally binned by read identity. In the absence of mapped contaminant reads, and all else being equal, mapping to a reference which is closer to the true consensus should map more bases and mapped reads should have higher identities.

- `FindClippingHotSpots.py` counts, at each position in the genome, the number and percentage of reads that are clipped from that position to their left or right end. Having many such reads is a warning sign of the kind of biased loss of information shown in Figure 2B.

- `FindSubSeqsInAlignment.py` finds the location of specified sub-sequences in an alignment (allowing for gaps).

- `LinkIdentityToCoverage.py` finds, for each different coverage encountered when considering all positions in a BAM file, the mean read identity at such positions. The mean read identity tends to be lower at positions of low coverage due to a background of contaminant reads, which differ from the reference by virtue of being contamination, but which are nevertheless similar enough to be mapped. Quantifying the decline in identity at low coverage helps inform what coverage threshold is appropriate for a given data set.

- `AlignMoreSeqsToPairWithMissingCoverage.py` allows more sequences to be added to a pairwise alignment in which one sequence contains missing coverage (such as a consensus and its reference), correctly maintaining the distinction between gaps (indicating a deletion) and missing coverage.

- `AlignBaseFreqFiles.py` aligns not two sequences, but two of the csv-format base frequency files output by `shiver`. Optionally a similarity metric is calculated at each position in the alignment, between 0 (no agreement on which bases/gaps are present) and 1 (perfect agreement on which

bases/gaps are present and on their proportions). This allows comparison not just of consensus sequences between two samples but also of minority variants.

- `ConvertAlnToColourCodes.py` converts each base in a sequence alignment into a colour code indicating agreement with the consensus and indels; `AlignmentPlotting.R` takes such colour codes and visualises the alignment. These two scripts were used to produce the plots of Supplementary Information sections 4 and 5.
- `QuantifyPairwiseIndels.py` considers all possible pairs of sequences in an alignment and calculates the sizes and positions of relative indels (i.e. ignoring positions at which both have a gap). It was used to make Figure 3.
- Finally some simple tools for convenience: `FindSeqsInFasta.py` extracts named sequences from a fasta file, with options including gap stripping, returning only windows of the sequences, and inverting the search; `PrintSeqLengths.py` prints sequence lengths with or without gaps; `SplitFasta.py` splits a fasta file into one file per sequence therein.

References

[1] A. Abecasis, A. Vandamme, and P. Lemey, HIV Sequence Compendium 2006/2007 (2007).

[2] K. McElroy, T. Thomas, and F. Luciani, Microbial Informatics and Experimentation **4**, 1 (2014).

[3] Geneious version 7.1 created by Biomatters. Available from http://www.geneious.com .

Supplementary Information for *Easy and Accurate Reconstruction of Whole HIV Genomes from Short-Read Sequence Data with SHIVER*

SI 1 Sequencing Platform Usage Statistics for HIV

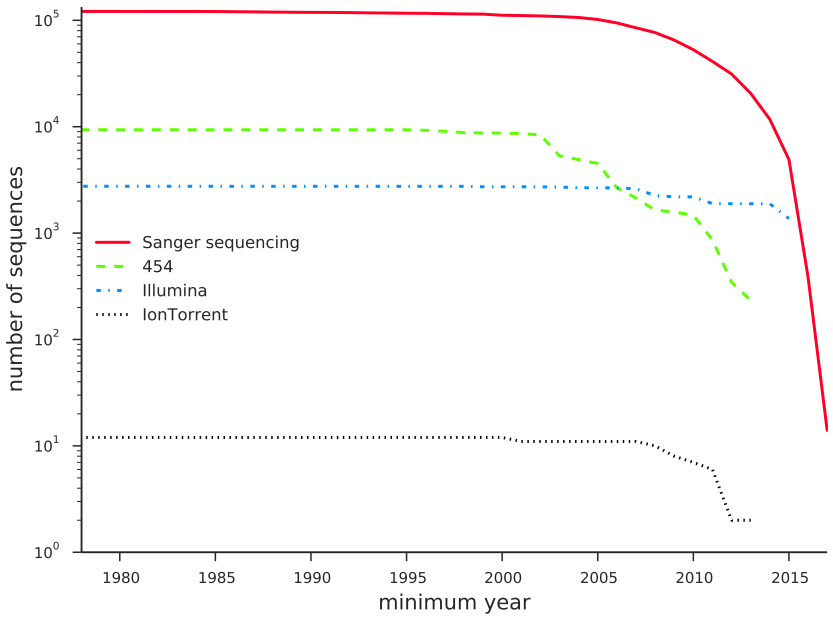

Figure S1: the number of HIV sequences available from the Los Alamos National Laboratory database on 11th Oct 2017 with sampling year and sequencing platform information available, as a function of minimum sampling year for inclusion (i.e. restricting the included sequences to increasingly recent ones).

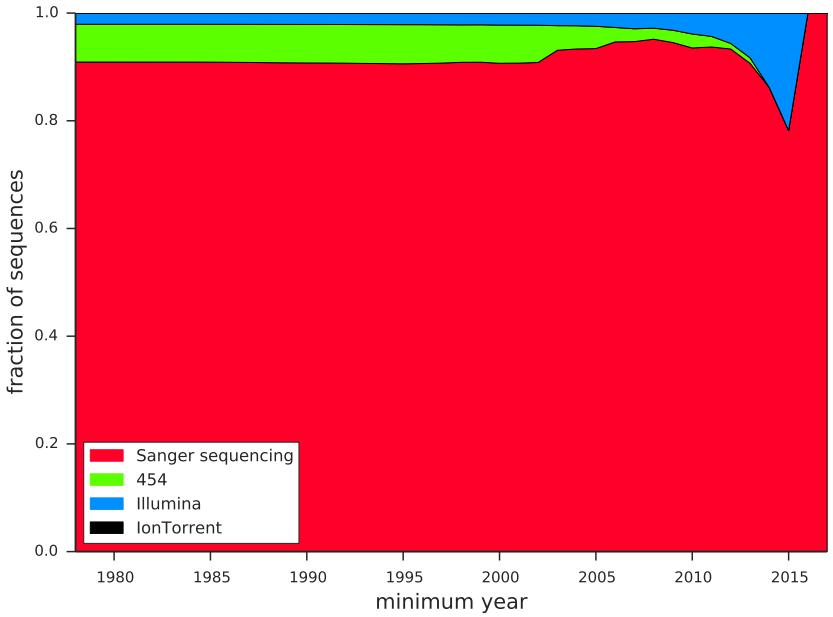

Figure S2: As Fig. S1 but showing the fraction for each platform.

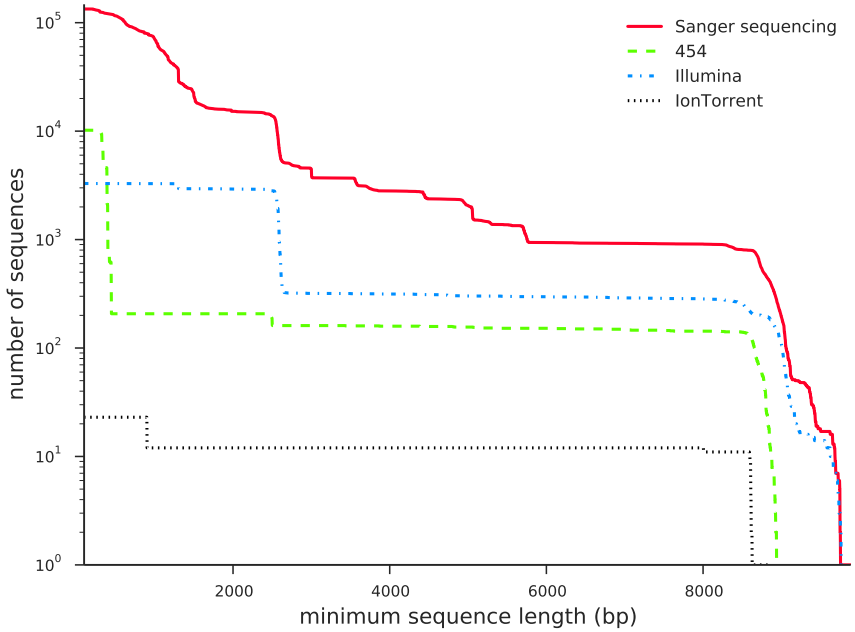

Figure S3: the number of HIV sequences available from the Los Alamos National Laboratory database on 11th Oct 2017 with sequencing platform information available, as a function of minimum sequence length for inclusion (i.e. restricting the included sequences to increasingly long ones).

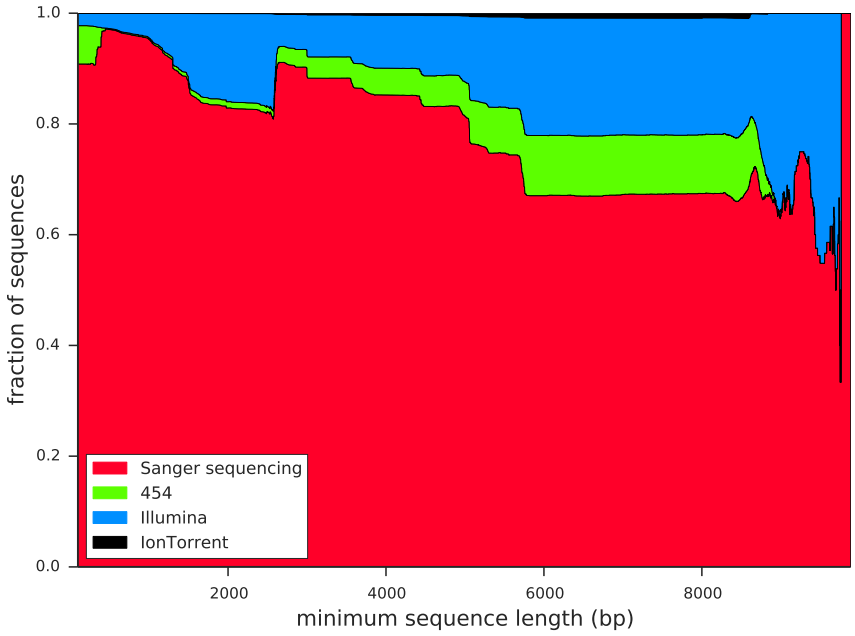

Figure S4: As Fig. S3 but showing the fraction for each platform.

SI 2 Our Method in More Detail

`shiver` is under continuing development; if at a later date description here contradicts descriptions at github.com/ChrisHIV/shiver, the latter has precedence.

2.1 Existing References

An alignment of existing reference sequences is required as input for `shiver`. Construction of a custom reference for mapping involves identifying the existing references that are closest to the sample under consideration. The greater the number and diversity of existing references given as input, the denser and broader the coverage of sequence space is, and the closer the closest reference is expected to be, with corresponding benefits for the accuracy of the results. However these existing references should be aligned to each other accurately, in order for the addition of each sample's contigs to the alignment to be meaningful; this means that producing such an input by automatically aligning a large number of diverse sequences without checking the results would be a bad idea. You will use this alignment as input for every sample in a dataset processed by `shiver`, and so we advise putting a little thought into sequence selection and manually curating the alignment if needed.

2.2 Constructing a Tailored Reference Using the Contigs

Custom reference construction begins with contig preprocessing as follows. Matches between the contigs and any existing reference from the alignment are searched for using `BLASTN` with default settings, except for the `-max_target_seqs 1` option (specifying that all reported hits are to a single reference only), and with `-word_size` set to 17 (this can be changed in `shiver`'s configuration file). Contaminant sequence is inevitable in high-throughput NGS; any contig that has no `BLASTN` hit to any of the HIV references is taken to be contamination, and is put aside for later use, leaving contigs that are putatively HIV. The `BLASTN` results are used to correct the contigs in three ways.

1. Where a single contig has multiple `BLASTN` hits (discarding any hit wholly contained inside another hit), we consider this evidence that the contig is spliced – concatenating two separated regions of the genome – due to errors *in silico* or during sequencing, as mentioned in the introduction. We correct this by cutting the contig into separate contigs at the midpoint between the hits.

2. We trim off any part of the contig that was not spanned by a `BLASTN` hit. The ends of contigs are by definition points at which the assembler has been unable to continue extending the sequence, either because of lack of reads, or because the within-sample diversity has become too great for a single, meaningful, representative sequence to be chosen. The latter possibility also means erroneous bases are more common in short stretches of sequence at the end of a contig. Trimming such sequence from the ends of contigs means the corresponding sequence from the closest existing reference will be used instead, giving a better reference for mapping. (Some assembly algorithms trim a fixed length from the ends of contigs for precisely this reason; however trimming a variable length dependent on its match to known sequence is clearly preferable.)

3. Any contig whose `BLASTN` hit is in the opposite orientation is reverse-complemented. If the assembler does not orientate the contigs, on average half of them will be in the reverse orientation. `IVA` orientates contigs such that the longest open reading frame is on the forward strand, however for very short contigs this may fail. In the process of assembling a spliced contig, an assembler may concatenate different regions in different orientations; `shiver` considers whether each separate part of a split contig requires reverse-complementation.

Contigs are then aligned to the existing reference alignment using `MAFFT`, trying both `--add` and `--addfragments` modes and using the one with the smallest maximum gap fraction (the maximum calculated over all contigs in each alignment). After alignment, a contig found to have an overly large internal deletion (by default 160bp) is split into two separate contigs at that point. This has the same role as `BLASTN`-based correction step 1 above, serving as a backup.

The alignment of contigs to the set of existing references should be visually inspected at this point. For HIV sequences, reference [1] states that "Algorithmic alignment does not necessarily retrieve the best alignment. It is important to always verify whether the sequence data are aligned unambiguously and, if necessary, manually correct the alignment." Reference [2] echoes this for any evolving pathogen: "the

'best' alignment chosen by an alignment program is not necessarily the 'true' alignment... Alignment quality should also be inspected manually in a visualisation program". The commonness of indels in HIV makes alignment more difficult, as does the fact that the contigs may be an imperfect representation of the true sample even after correction. We used `Geneious` [3] for sequence visualisation and editing where needed.

As well as revealing alignment error, inspection of the aligned contigs allows the user to check for any remaining problems with the contigs. We suggest that in general the user inspects both the alignment of the existing references with the *raw* HIV contigs (before any correction by shiver), and the alignment of the existing references with the *corrected* HIV contigs, as a check that all `shiver`'s modifications of the contigs are desired. An example of when this might not be the case is when the sample contains an indel not observed in the existing reference set, that is large enough to cause the contig to be split in two at that point, but which the user thinks might be genuine rather than an a misassembly (through previous/expert knowledge, or perhaps simply observing the same indel in multiple samples in a dataset). With sufficiently accurate mapping, reads will map here correctly whether or not the reference constructed from the contigs contains the indel, making the question moot; however with mapping inaccuracies of the kind shown in Figure 2 possible, it's best to get the reference's structure as correct as possible before mapping.

Using the alignment of contigs to existing references, the set of contigs is flattened into a single sequence as follows. At positions covered by one contig, its base (or gap character, for a deletion) is used. At positions covered by multiple contigs, we use whatever the longest contig has (be it base or gap). We used this heuristic expecting that, where sufficiently distinct haplotypes exist to result in multiple contigs covering the same place, haplotypes supported by a higher depth of reads would tend to be assembled into longer contigs. The sequence resulting from this flattening of the contigs is compared to each existing reference in the alignment in turn: we count towards similarity shared bases and gaps within contigs (known deletions), but not gaps between contigs (missing information). The existing references are ranked by their similarity to the contigs. As existing references have variable lengths (the long terminal repeat regions that flank the clinical genome are sometimes sequenced only partially or not at all), the closest reference is extended outwards using any overhanging sequence from the second closest reference, then the third longest sequence etc. terminating when both edges of the alignment are reached. This sequence – the elongated closest reference – is used to fill in any gaps between (but not inside of) the flattened contigs. This completes production of the reference tailored for this sample.

2.3 Preparing and Mapping the Reads

Before mapping to this reference, the reads are trimmed and cleaned as follows. Adapters, primers and low quality bases are trimmed using `Trimmomatic` and `Fastaq`. We then consider contaminant reads from non-HIV sources. Most of these would presumably be discarded by mapping to an HIV reference, due to lack of similarity. However there is ample opportunity for traces of human DNA to end up in a sample, and sequence of endogenous retroviruses in human DNA may resemble HIV. As a guard against this, and against any other contamination resembling HIV, we use `BLASTN` to find all read pairs that are a better match to one of the contigs previously found to be contamination, than to the tailored reference. These pairs are discarded.

The cleaned reads are mapped to the tailored reference, using `SMALT` by default (with `BWA` and `bowtie` as optional alternatives), giving a file in BAM format. Using `SAMtools` the BAM file is read into pileup format, which is parsed to give base frequencies at each position in the genome. Note that within-host diversity does not consist exclusively of point mutations: indels can be present in some reads and not others (Fig. 7 is an example), which must be dealt with in the pileup. Where some reads have a deletion relative to the reference and others do not, the deletion/gap character can simply be considered as a fifth base whose frequency can be counted like the others. Where some reads have an insertion relative to the reference and others do not, or more generally where insertions of two or more sizes are present, we find the most common insertion size and, inside that insertion, consider only those reads with an insertion of that size (thus avoiding any ambiguity in the alignment of the inserted sequences to each other). Finally, the base frequency file is parsed to call the consensus base at each position. By default the most common base is called to give the consensus, using an ambiguity code only for an exact tie in the frequency of two or more bases; optionally ambiguity codes can be used more readily, when the frequency of the most common base or bases is below a specified threshold. A consensus base is only called if the coverage equals or exceeds a minimum threshold specified by the user, to protect against the effect of residual

low-coverage contaminant reads in genomic regions lacking genuine HIV reads. By default this is 15, but this is likely to need adjusting for different datasets. (See the tool `LinkIdentityToCoverage.py` in section 3.)

2.4 Aligning Multiple Consensuses

Since we know how the consensus aligns to the reference used for mapping, and we know how that reference (constructed from the contigs) aligns to the input alignment of existing references, we can construct a global alignment of the consensuses from all samples merely by coordinate translation, negating the need for further alignment and manual curation. Two things must be excised from the consensus for this global alignment reconstruction: insertions present in the majority of reads but not in their tailored reference (which are rare, since the reference is constructed from the contigs which are constructed from the reads), and insertions present in the contigs but none of the existing references (which are rare provided the set of existing references is large and diverse). In both cases this is sequence whose alignment to the common anchor of the existing references is not known, and so coordinate translation cannot align it.

2.5 Fully Automatic `shiver`

As mentioned, `shiver` can be run from beginning to end without the break in the middle, with the single command `shiver_full_auto.sh`, for uses where visually checking the contigs is impractical. This begins with separation of contigs into HIV (those with `BLASTN` hits) and contamination as previously. Subsequent steps are as follows.

1. The need for contig correction is checked, but correction is not performed: if it is needed, processing stops. Blind trust in the accuracy of an automated alignment of contigs cut into pieces based on evidence of structural problems would be trust misplaced.

2. Each HIV contig is now certain to have a single `BLASTN` hit (discarding any smaller hits wholly contained inside others). That hit is checked to span some minimum fraction of the contig length (by default 90%) as a guard against contigs containing containing some erroneous or foreign sequence; otherwise processing stops.

3. Multiple sequence alignment is performed with these contigs and just one of the existing reference sequences, for each of the existing reference sequences separately.

4. For each such alignment, generated both with regular `mafft` and with `mafft --addfragments`, we calculate the fractional agreement between the flattened contigs and the reference, i.e. the fraction of positions spanned by the reference and at least one contig where the reference and the longest contig agree. Misalignment is penalised in this score because gaps inside contigs are taken as genuine deletions.

5. For the alignment with the highest score, the maximum gap fraction amongst the contigs in the alignment (i.e. the fraction of positions inside the contig that are gaps) is checked to be below a user-specified threshold (the default is 5%, based on analysis of thousands of such alignments that we visually checked) as a further guard against misalignment.

6. The contigs are flattened using this single existing reference to fill in any gaps between them, generating the mapping reference tailored for this sample.

Aligning contigs to the references one at a time (step 3) is simpler for the alignment algorithm than aligning to all of them at once, and means that even if misalignment occurs for what is truly the closest reference to the contigs, the alignment to the second closest can be used instead. Trimming of low-quality bases, trimming of adapter and primer sequences, removal of contaminant reads and mapping to the tailored reference all occur as described previously. For samples that cannot be processed fully automatically this way – when contig correction is required, or a contig is spanned by too small a `BLASTN` hit, or too many gaps are present after alignment – the main mode of shiver is available (for which we advise inspection of the aligned contigs).

As argued earlier, we advocate visually inspecting the aligned contigs, i.e. running the two-command implementation of `shiver` (with the check occurring between the commands). This also has the advantage of working for all samples, whereas `shiver_full_auto.sh` will not proceed if problems with the contigs

or their alignment are detected. `shiver_full_auto.sh` also does not produce a global alignment of all consensuses to each other, because the coordinate translation procedure allowing its construction is derived from each sample's alignment of contigs to all of the references at once. That alignment is produced for the two-command implementation of `shiver`, but step 3 above aligns contigs to references one at a time.

SI 3 Sample Reprocessing and Analysis

Individual steps from `shiver` can be run with stand-alone command line tools, for ease of reapplication elsewhere. For example `CorrectContigs.py` is run with a file of contigs and a file of their `BLASTN` hits to some set of references, and corrects the contigs by cutting, trimming and reverse complementing where needed. Also included in `shiver` are command-line tools for easy analysis and modification of sample output without rerunning the whole pipeline:

- Two parameters specified in the configuration file are a minimum coverage required to call a base (below this coverage, the character '`?`' is used) and a larger minimum coverage required to use upper case instead of lower, as an easy signal of increased confidence. (Note that decreasing these parameters will, in general, allow bases to be called at more positions, giving a longer consensus. However there is a trade-off: where there are fewer reads, the effect of contaminant reads on the consensus may be greater.) To regenerate a consensus with new values of these parameters, `CallConsensus.py` can be run on a sample's base frequencies file. To regenerate a coordinate-translated version of this consensus for the global alignment (of all consensuses produced by `shiver`), `TranslateSeqForGlobalAln.py` can be run on the consensus.

- Another parameter in the configuration file is the minimum read *identity* – the fraction of bases in the read which are mapped and agree with the reference – required for a read to be considered mapped, and so retained in the BAM file. If you wish to increase this after completion of `shiver`, reads with an identity below your new higher threshold can be discarded by running `RemoveDivergentReads.py` on a BAM file. Running `shiver_reprocess_bam.sh` on the resulting BAM file (or indeed any BAM file) implements just the last steps in `shiver`, namely generating pileup, calculating the base frequencies, and calling the consensus.

- `FindNumMappedBases.py` calculates the total number of mapped bases in a BAM file (where read length is constant this equals the number of mapped reads multiplied by read length, minus the total length of sequence clipped from reads), optionally binned by read identity. In the absence of mapped contaminant reads, and all else being equal, mapping to a reference which is closer to the true consensus should map more bases and mapped reads should have higher identities.

- `FindClippingHotSpots.py` counts, at each position in the genome, the number and percentage of reads that are clipped from that position to their left or right end. Having many such reads is a warning sign of the kind of biased loss of information shown in Figure 2B.

- `FindSubSeqsInAlignment.py` finds the location of specified sub-sequences in an alignment (allowing for gaps).

- `LinkIdentityToCoverage.py` finds, for each different coverage encountered when considering all positions in a BAM file, the mean read identity at such positions. The mean read identity tends to be lower at positions of low coverage due to a background of contaminant reads, which differ from the reference by virtue of being contamination, but which are nevertheless similar enough to be mapped. Quantifying the decline in identity at low coverage helps inform what coverage threshold is appropriate for a given data set.

- `AlignMoreSeqsToPairWithMissingCoverage.py` allows more sequences to be added to a pairwise alignment in which one sequence contains missing coverage (such as a consensus and its reference), correctly maintaining the distinction between gaps (indicating a deletion) and missing coverage.

- `AlignBaseFreqFiles.py` aligns not two sequences, but two of the csv-format base frequency files output by `shiver`. Optionally a similarity metric is calculated at each position in the alignment, between 0 (no agreement on which bases/gaps are present) and 1 (perfect agreement on which

bases/gaps are present and on their proportions). This allows comparison not just of consensus sequences between two samples but also of minority variants.

- `ConvertAlnToColourCodes.py` converts each base in a sequence alignment into a colour code indicating agreement with the consensus and indels; `AlignmentPlotting.R` takes such colour codes and visualises the alignment. These two scripts were used to produce the plots of Supplementary Information sections 4 and 5.
- `QuantifyPairwiseIndels.py` considers all possible pairs of sequences in an alignment and calculates the sizes and positions of relative indels (i.e. ignoring positions at which both have a gap). It was used to make Figure 3.
- Finally some simple tools for convenience: `FindSeqsInFasta.py` extracts named sequences from a fasta file, with options including gap stripping, returning only windows of the sequences, and inverting the search; `PrintSeqLengths.py` prints sequence lengths with or without gaps; `SplitFasta.py` splits a fasta file into one file per sequence therein.

References

[1] A. Abecasis, A. Vandamme, and P. Lemey, HIV Sequence Compendium 2006/2007 (2007).

[2] K. McElroy, T. Thomas, and F. Luciani, Microbial Informatics and Experimentation **4**, 1 (2014).

[3] Geneious version 7.1 created by Biomatters. Available from http://www.geneious.com .

Supplementary Information for Easy and Accurate Reconstruction of Whole HIV Genomes from Short-Read Sequence Data with SHIVER

SI 6: Members of the BEEHIVE Cohorts

Swiss HIV cohort

The member of the Swiss HIV Cohort are: Aubert V, Battegay M, Bernasconi E, Böni J, Braun DL, Bucher HC, Burton-Jeangros C, Calmy A, Cavassini M, Dollenmaier G, Egger M, Elzi L, Fehr J, Fellay J, Furrer H (Chairman of the Clinical and Laboratory Committee), Fux CA, Gorgievski M, Günthard H (President of the SHCS), Haerry D (deputy of "Positive Council"), Hasse B, Hirsch HH, Hoffmann M, Hösli I, Kahlert C, Kaiser L, Keiser O, Klimkait T, Kouyos R, Kovari H, Ledergerber B, Martinetti G, Martinez de Tejada B, Marzolini C, Metzner K, Müller N, Nadal D, Nicca D, Pantaleo G, Rauch A (Chairman of the Scientific Board), Regenass S, Rudin C (Chairman of the Mother & Child Substudy), Schöni-Affolter F (Head of Data Centre), Schmid P, Speck R, Stöckle M, Tarr P, Trkola A, Vernazza P, Weber R, Yerly S.

ATHENA cohort (The Netherlands)

CLINICAL CENTRES:
* denotes site coordinating physician
Academic Medical Centre of the University of Amsterdam (AMC-UvA): *HIV treating physicians:* M. van der Valk*, S.E. Geerlings, M.H. Godfried, A. Goorhuis, J.W. Hovius, J.T.M. van der Meer, T.W. Kuijpers, F.J.B. Nellen, DT. van der Poll, J.M. Prins, P. Reiss, H.J. M. van Vugt, W.J. Wiersinga, F.W.M.N. Wit. *HIV nurse consultants:* M. van Duinen, J. van Eden, A.M.H. van Hes, M. Mutschelknauss, H.E. Nobel, F.J.J. Pijnappel, A.M. Weijsenfeld. *HIV clinical virologists/chemists:* S. Jurriaans, N.K.T. Back, H.L. Zaaijer, B. Berkhout, M.T.E. Cornelissen, C.J. Schinkel, K.C. Wolthers. **Admiraal De Ruyter Ziekenhuis, Goes:** *HIV treating physicians:* M. van den Berge, A. Stegeman. *HIV nurse consultants:* S. Baas, L. Hage de Looff. *HIV clinical virologists/chemists:* B Wintermans, J Veenemans. **Catharina Ziekenhuis, Eindhoven:** *HIV treating physicians:* M.J.H. Pronk*, H.S.M. Ammerlaan. *HIV nurse consultants:* E.S. de Munnik, H.A.M. van Beek. *HIV clinical virologists/chemists:* A.R. Jansz, J. Tjhie, M.C.A. Wegdam, B. Deiman, V. Scharnhorst. **Elisabeth-TweeSteden Ziekenhuis, Tilburg:** *HIV treating physicians:* M.E.E. van Kasteren*, A.E. Brouwer. *HIV nurse consultants:* R. van Erve, B.A.F.M. de Kruijf-van de Wiel, S.Keelan-Pfaf, B. van der Ven. *Data collection:* B.A.F.M. de Kruijf-van de Wiel, B. van der Ven. *HIV clinical virologists/chemists:* A.G.M. Buiting, P.J. Kabel, D.Versteeg. **Erasmus MC, Rotterdam:** *HIV treating physicians:* M.E. van der Ende*, H.I. Bax, E.C.M. van Gorp, J.L. Nouwen, B.J.A. Rijnders, C.A.M. Schurink, A. Verbon, T.E.M.S. de Vries-Sluijs, N.C. de Jong-Peltenburg. *HIV nurse consultants:* N. Bassant, J.E.A. van Beek, M. Vriesde, L.M. van Zonneveld. *Data collection:* H.J. van den Berg-Cameron, J. de Groot, M. de Zeeuw-de Man. *HIV clinical virologists/chemists:* C.A.B. Boucher, M.P.G Koopmans, J.J.A van Kampen, S.D. Pas. **Flevoziekenhuis, Almere:** *HIV treating physicians:* J. Branger*, A. Rijkeboer-Mes. *HIV nurse consultant:* C.J.H.M. Duijf-van de Ven. **HagaZiekenhuis, Den Haag:** *HIV treating physicians:* E.F. Schippers*, C. van Nieuwkoop. *HIV nurse consultants:* J.M. van IJperen, J. Geilings. *Data collection:* G. van der Hut. *HIV clinical virologist/chemist:* N.D. van Burgel. **Hiv Focus Centrum (DC Klinieken):** *HIV treating physicians:* A. van Eeden*. *HIV nurse consultants:* W. Brokking, M. Groot, L.J.M. Elsenburg. *HIV clinical virologists/chemists:* M. Damen, I.S. Kwa. **HMC (Haaglanden Medisch Centrum), Den Haag:** *HIV treating physicians:* E.M.S. Leyten*, L.B.S. Gelinck. *HIV nurse consultants:* A.Y. van Hartingsveld, C. Meerkerk, G.S. Wildenbeest. *HIV clinical virologists/chemists:* E. Heikens. **Isala, Zwolle:** *HIV treating physicians:* P.H.P. Groeneveld*, J.W.

Bouwhuis, A.J.J. Lammers. *HIV nurse consultants:* S. Kraan, A.G.W. van Hulzen. *Data collection:* G.L. van der Bliek, P.C.J. Bor. *HIV clinical virologists/chemists:* P. Bloembergen, M.J.H.M. Wolfhagen, G.J.H.M. Ruijs. **Leids Universitair Medisch Centrum, Leiden:** *HIV treating physicians:* F.P. Kroon*, M.G.J. de Boer, H. Scheper, H. Jolink, A.M. Vollaard. *HIV nurse consultants:* W. Dorama, N. van Holten. *HIV clinical virologists/chemists:* E.C.J. Claas, E. Wessels. **Maasstad Ziekenhuis, Rotterdam:** *HIV treating physicians:* J.G. den Hollander*, K. Pogany, A. Roukens. *HIV nurse consultants:* M. Kastelijns, J.V. Smit, E. Smit, D. Struik-Kalkman, C. Tearno. *Data collection:* T. van Niekerk. *HIV clinical virologists/chemists:* O. Pontesilli. **Maastricht UMC+, Maastricht:** *HIV treating physicians:* S.H. Lowe*, A.M.L. Oude Lashof, D. Posthouwer. *HIV nurse consultants:* R.P. Ackens, K. Burgers, J. Schippers. *Data collection:* B. Weijenberg-Maes. *HIV clinical virologists/chemists:* I.H.M. van Loo, T.R.A. Havenith. **MC Slotervaart, Amsterdam:** *HIV treating physicians:* J.W. Mulder*, S.M.E. Vrouenraets, F.N. Lauw. *HIV nurse consultants:* M.C. van Broekhuizen, D.J. Vlasblom. *HIV clinical virologists/chemists:* P.H.M. Smits. **MC Zuiderzee, Lelystad:** *HIV treating physicians:* S. Weijer*, R. El Moussaoui. *HIV nurse consultant:* A.S. Bosma. **Medisch Centrum Leeuwarden, Leeuwarden:** *HIV treating physicians:* M.G.A.van Vonderen*, D.P.F. van Houte, L.M. Kampschreur. *HIV nurse consultants:* K. Dijkstra, S. Faber. *HIV clinical virologists/chemists:* J Weel. **Medisch Spectrum Twente, Enschede:** *HIV treating physicians:* G.J. Kootstra*, C.E. Delsing. *HIV nurse consultants:* M. van der Burg-van de Plas, H. Heins. *Data collection:* E. Lucas. **Noordwest Ziekenhuisgroep, Alkmaar:** *HIV treating physicians:* W. Kortmann*, G. van Twillert*, R. Renckens. *HIV nurse consultant and data collection:* D. Ruiter-Pronk, F.A. van Truijen-Oud. *HIV clinical virologists/chemists:* J.W.T. Cohen Stuart, E.P. IJzerman, R. Jansen, W. Rozemeijer W. A. van der Reijden. **OLVG, Amsterdam:** *HIV treating physicians:* K. Brinkman*, G.E.L. van den Berk, W.L. Blok, P.H.J. Frissen, K.D. Lettinga W.E.M. Schouten, J. Veenstra. *HIV nurse consultants:* C.J. Brouwer, G.F. Geerders, K. Hoeksema, M.J. Kleene, I.B. van der Meché, M. Spelbrink, A.J.M. Toonen, S. Wijnands. *HIV clinical virologists:* D. Kwa. *Data collection:* R. Regez (coordinator). **Radboudumc, Nijmegen:** *HIV treating physicians:* R. van Crevel*, M. Keuter, A.J.A.M. van der Ven, H.J.M. ter Hofstede, A.S.M. Dofferhoff, J. Hoogerwerf. *HIV nurse consultants:* K.J.T. Grintjes-Huisman, M. de Haan, M. Marneef, A. Hairwassers. *HIV clinical virologists/chemists:* J. Rahamat-Langendoen, F.F. Stelma. *HIV clinical pharmacology consultant:* D. Burger. **Rijnstate, Arnhem:** *HIV treating physicians:* E.H. Gisolf*, R.J. Hassing, M. Claassen. *HIV nurse consultants:* G. ter Beest, P.H.M. van Bentum, N. Langebeek. *HIV clinical virologists/chemists:* R. Tiemessen, C.M.A. Swanink. **Spaarne Gasthuis, Haarlem:** *HIV treating physicians:* S.F.L. van Lelyveld*, R. Soetekouw. *HIV nurse consultants:* L.M.M. van der Prijt, J. van der Swaluw. *Data collection:* N. Bermon. *HIV clinical virologists/chemists:* W.A. van der Reijden, R. Jansen, B.L. Herpers, D.Veenendaal. **Medisch Centrum Jan van Goyen, Amsterdam:** *HIV treating physicians:* D.W.M. Verhagen. *HIV nurse consultants:* M. van Wijk. **Universitair Medisch Centrum Groningen, Groningen:** *HIV treating physicians:* W.F.W. Bierman*, M. Bakker, J. Kleinnijenhuis, E. Kloeze, Y. Stienstra, K.R. Wilting, M. Wouthuyzen-Bakker. *HIV nurse consultants:* A. Boonstra, P.A. van der Meulen, D.A. de Weerd. *HIV clinical virologists/chemists:* H.G.M. Niesters, C.C. van Leer-Buter, M. Knoester. **Universitair Medisch Centrum Utrecht, Utrecht:** *HIV treating physicians:* A.I.M. Hoepelman*, J.E. Arends, R.E. Barth, A.H.W. Bruns, P.M. Ellerbroek, T. Mudrikova, J.J. Oosterheert, E.M. Schadd, M.W.M. Wassenberg, M.A.D. van Zoelen. *HIV nurse consultants:* K. Aarsman, D.H.M. van Elst-Laurijssen, I. de Kroon, C.S.A.M. van Rooijen. *Data collection:* M. van Berkel, C.S.A.M. van Rooijen. *HIV clinical virologists/chemists:* R. Schuurman, F. Verduyn-Lunel, A.M.J. Wensing. **VUmc, Amsterdam:** *HIV treating physicians:* E.J.G. Peters*, M.A. van Agtmael, M. Bomers. *HIV nurse consultants:* M. Heitmuller, L.M. Laan. *HIV clinical virologists/chemists:* C.W. Ang, R. van Houdt, A.M. Pettersson, C.M.J.E. Vandenbroucke-Grauls.

COORDINATING CENTRE:

Director: P. Reiss. *Data analysis:* D.O. Bezemer, A.I. van Sighem, C. Smit, F.W.M.N. Wit, T.S. Boender. *Data management and quality control:* S. Zaheri, M. Hillebregt, A. de Jong. *Data monitoring:* D. Bergsma, S. Grivell, A. Jansen, M. Raethke, R. Meijering, T. Rutkens. *Data collection:* L. de Groot, M. van den Akker, Y. Bakker, M. Bezemer, E. Claessen, A. El Berkaoui, J. Geerlinks, J. Koops, E. Kruijne, C. Lodewijk, R. van der Meer, L. Munjishvili, F. Paling, B. Peeck,

C. Ree, R. Regtop, Y. Ruijs, M. Schoorl, A. Timmerman, E. Tuijn, L. Veenenberg, S. van der Vliet, A. Wisse, E.C. de Witte, T. Woudstra. *Patient registration:* B. Tuk.

Antwerp cohort (Belgium)

Data extraction for the Antwerp Cohort is done by Maartje Van Frankenhuijsen, MD.

PRIMO cohort (France):

Région Sud–Est:

- Thierry ALLEGRE, Centre hospitalier général d'Aix en Provence, Service d'Hématologie
- Djamila MAKHLOUFI, Jean-Michel LIVROZET, Pierre CHIARELLO, Mathieu GODINOT, Florence BRUNEL-DALMAS, Sylvie GIBERT, Hôpital Edouard Herriot de Lyon, Immunologie Clinique
- Christian TREPO, Dominique PEYRAMOND, Patrick MIAILHES, Joseph KOFFI, Valérie THOIRAIN, Corinne BROCHIER, Thomas BAUDRY, Sylvie PAILHES, Lyon La Croix Rousse, Services d'Hépato-Gastroentérologie et des Maladies Infectieuses
- Alain LAFEUILLADE, Gisèle PHILIP, Gilles HITTINGER, Assi ASSI, Véronique LAMBRY, Hôpital Font-Pré de Toulon, Médecine Interne, Hémato-Infectiologie
- Eric ROSENTHAL, Alissa NAQVI, Brigitte DUNAIS, Eric CUA, Christian PRADIER, Jacques DURANT, Aline JOULIE, Hôpital L'Archet, Nice, Service de Médecine Interne, Maladies Infectieuses et Tropicales
- Denis QUINSAT, Serge TEMPESTA, Centre Hospitalier d'Antibes, Service de Médecine Interne
- Isabelle RAVAUX, Hôpital de la Conception de Marseille, Service des Maladies Infectieuses
- Isabelle POIZOT MARTIN, Olivia FAUCHER, Nicolas CLOAREC, Hôpital Sainte Marguerite de Marseille, Unité d'Hématologie
- Hélène CHAMPAGNE, Centre Hospitalier de Valence, Maladies Infectieuses et Tropicales
- Gilles PICHANCOURT, Centre Hospitalier Henri Duffaut d'Avignon, Service Hématologie Maladies Infectieuses

Région Sud-Ouest:

- Philippe MORLAT, Thierry PISTONE, Fabrice BONNET, Patrick MERCIE, Isabelle FAURE, Mojgan HESSAMFAR, Denis MALVY, Denis LACOSTE, Marie-Carmen PERTUSA, Marie-Anne VANDENHENDE, Noëlle BERNARD, François PACCALIN, Cédric MARTELL, Julien ROGER-SCHMELZ, Marie-Catherine RECEVEUR, Pierre DUFFAU, Denis DONDIA, Emmanuel RIBEIRO, Sabrina CALTADO, Hôpital Saint André de Bordeaux, Médecine Interne
- Didier NEAU, Michel DUPONT; Hervé DUTRONC, Frédéric DAUCHY, Charles CAZANAVE, Thierry PISTONE, Marc-Olivier VAREIL, Thierry PISTONE, Gaétane WIRTH, Séverine LE PUIL, Hôpital Pellegrin de Bordeaux, Maladies Infectieuses.
- Jean-Luc PELLEGRIN, Isabelle RAYMOND, Jean-François VIALLARD, Severin CHAIGNE DE LALANDE, Hôpital Haut Lévèque de Bordeaux, Médecine Interne et Maladies Infectieuses
- Daniel GARIPUY, Hôpital Joseph Ducuing de Toulouse, Médecine Interne
- Pierre DELOBEL, Martine OBADIA, Lise CUZIN, Muriel ALVAREZ, Noemie BIEZUNSKI, Lydie PORTE, Patrice MASSIP, Alexa DEBARD, Florence BALSARIN, Myriam LAGARRIGUE, Hôpital Purpan de Toulouse, SMIT-CISIH
- François PREVOTEAU DU CLARY, Christian AQUILINA, Cité de la santé Toulouse
- Jacques REYNES, Vincent BAILLAT, Corinne MERLE, Vincent LEMOING, Nadine ATOUI, Alain MAKINSON, Jean Marc JACQUET, Christina PSOMAS, Christine TRAMONI, Hôpital Gui de Chauliac de Montpellier, Service des Maladies Infectieuses et Tropicales
- Hugues AUMAITRE, Mathieu SAADA, Marie MEDUS, Martine MALET, Aurélia EDEN, Ségolène NEUVILLE, Milagros FERREYRA, Martine MALET, Hôpital Saint Jean de Perpignan, Service des Maladies Infectieuses

- Albert SOTTO, Claudine BARBUAT, Isabelle ROUANET, Didier LEUREILLARD, Jean-Marc MAUBOUSSIN, Catherine LECHICHE, Régine DONSESCO, CHU de Nîmes-Caremeau, Service des Maladies Infectieuses et Tropicales.

Antilles:

-André CABIE, Sylvie ABEL, Sandrine PIERRE-FRANCOIS, Anne-Sophie BATALA, Christophe CERLAND, Camille RANGOM, Nadine THERESINE, CHU Fort de France, Hôpital de Jour

-Bruno HOEN, Isabelle LAMAURY, Isabelle FABRE, Kinda SCHEPERS, Elodie CURLIER, Rachida OUISSA, CHU de Pointe à Pitre/ABYMES, Service de Dermatologie / Maladies Infectieuses

-Catherine GAUD, Carole RICAUD, Roland RODET, Guillaume WARTEL, Carmele SAUTRON, CHU de la Reunion, site Felix Guyon, Service d'Immunologie

Région Est:

- Geneviève BECK-WIRTH, Catherine MICHEL, Charles BECK, Jean-Michel HALNA, Jakub KOWALCZYK, Meryem BENOMAR, Hôpital Emile Muller de Mulhouse, Hématologie Clinique
- Christine DROBACHEFF-THIEBAUT, Catherine CHIROUZE, Jean-François FAUCHER, François PARCELIER, Adeline FOLTZER, Cécile HAFFNER-MAUVAIS, Mathieu HUSTACHE MATHIEU, Aurélie PROUST - Hôpital St Jacques de Besançon, Service des Maladies Infectieuses et de Dermatologie
- Lionel PIROTH, Pascal CHAVANET, Michel DUONG, Marielle BUISSON, Anne WALDNER, Sophie MAHY, Sandrine GOHIER, Delphine CROISIER, Hôpital du Bocage de Dijon, Service des Maladies Infectieuses
- Thierry MAY, Mikael DELESTAN, Marie ANDRE, CHU de Vandoeuvre-lès-Nancy, Hôpital de Brabois, Service des Maladies Infectieuses et Tropicales
- Mahsa MOHSENI ZADEH, Martin MARTINOT, Béatrice ROSOLEN, Anne PACHART, Hôpital Louis PASTEUR de Colmar, Service d'Immunologie Clinique
- Benoît MARTHA, Noëlle JEUNET, Centre Hospitalier William Morey de Chalon Sur Saône, Service de Médecine Interne
- David REY, Christine CHENEAU, Maria PARTISANI, Michèle PRIESTER, Claudine BERNARD-HENRY, Maria PARTISANI, Marie-Laure BATARD, Patricia FISCHER, Service le Trait d'Union, Hôpitaux Universitaires de Strasbourg
- Jean-Luc BERGER, Isabelle KMIEC, Hôpital Robert Debré, Service des Maladies Infectieuses, Reims.

Région Nord:

-Olivier ROBINEAU, Thomas HULEUX, Faïza AJANA, Isabelle ALCARAZ, Christophe ALLIENNE, Véronique BACLET, Agnès MEYBECK, Michel VALETTE, Nathalie VIGET, Christophe ALLIENNE, Emmanuelle AISSI, Raphael BIEKRE, Pauline CORNAVIN, Centre Hospitalier DRON de Tourcoing, Service de Maladies Infectieuses

-Dominique MERRIEN, Jean-Christophe SEGHEZZI, Moise MACHADO, Centre Hospitalier de Compiègne, Service de Médecine Interne

-Georges DIAB, C H de la Haute Vallée de l'Oise de Noyon, Service de Médecine

Région Ouest:

- François RAFFI, Bénédicte BONNET, Clotilde ALLAVENA, Olivier GROSSI, Véronique RELIQUET, Eric BILLAUD, Cecile BRUNET, Sabelline BOUCHEZ, Pascale MORINEAU-LE HOUSSINE, Fabienne SAUSER, David BOUTOILLE, Michel BESNIER, Hervé HUE, Nolwenn Hall, Delphine BROSSEAU, Hôtel-Dieu de Nantes, CISIH Médecine Interne
- Faouzi SOUALA, Christian MICHELET, Pierre TATTEVIN, Cédric ARVIEUX, Matthieu REVEST, Helene LEROY, Jean-Marc CHAPPLAIN, Matthieu DUPONT, Fabien FILY, SOLÈNE PATRA-DELO, CÉLINE LEFEUVRE, CHRU Pontchaillou de Rennes, Clinique des Maladies Infectieuses
- Louis BERNARD, Frédéric BASTIDES, Pascale NAU, Hôpital Bretonneau de Tours, Service des maladies Infectieuses
- Renaud VERDON, Arnaud DE LA BLANCHARDIERE, Anne MARTIN, Philippe FERET, CH régional Côte de Nacre de Caen, Service de Maladies Infectieuses
- Loïk GEFFRAY, Hôpital Robert Bisson de Lisieux, Service de Médecine Interne
- Corinne DANIEL, Jennifer ROHAN, Centre Hospitalier La Beauchée de Saint-Brieuc, Médecine Interne et Maladies Infectieuses

- Pascale FIALAIRE, Jean Marie CHENNEBAULT, Valérie RABIER, Pierre ABGUEGUEN, Sami REHAIEM, Centre Hospitalier Régional d'Angers, Service des Maladies Infectieuses
- Odile LUYCX, Mathilde NIAULT, Philippe MOREAU, Centre Hospitalier Bretagne Sud de Lorient, Service d'Hématologie
- Yves POINSIGNON, Marie GOUSSEF, Virginie MOUTON- RIOUX, Centre Hospitalier Bretagne Atlantique de Vannes, Service de Medecine Interne et Maladies Infectieuses
- Dominique HOULBERT, Sandrine ALVAREZ-HUVE, Frédérique BARBE, Sophie HARET, Centre Hospitalier d'Alençon, Médecine 2
- Philippe PERRE,Sophie LEANTEZ-NAINVILLE, Jean-Luc ESNAULT, Thomas GUIMARD, Isabelle SUAUD, Centre Hospitalier Départemental de La Roche sur Yon, Service de Médecine
- Jean-Jacques GIRARD, Véronique SIMONET, Hôpital de Lôches, Service de Médecine Interne
- Yasmine DEBAB, CHU Charles Nicolle de Rouen, Maladies Infectieuses et Tropicales
- Jean-Luc SCHMIT, CHU d'Amiens, Service des Maladies Infectieuses.

Région Centre:

-Christine JACOMET, Hôpital Gabriel-Montpied de Clermont Ferrand, Service des Maladies Infectieuses et Tropicales

-Pierre WEINBERCK, Claire GENET, Pauline PINET, Sophie DUCROIX, Hélène DUROX, Éric DENES, Hôpital DUPUYTREN de Limoges, Maladies Infectieuses et Tropicales

-Bruno ABRAHAM, Centre Hospitalier de Brive, Departement de maladies Infectieuses

-Florence GOURDON, Centre Hospitalier de Vichy, Service de Médecine Interne

-Odile ANTONIOTTI, Centre Hospitalier de Montluçon, Dermatologie

Paris:

- Jean-Michel MOLINA, Samuel FERRET, Caroline LASCOUX-COMBE, Matthieu LAFAURIE, Nathalie COLIN DE VERDIERE, Diane PONSCARME, Nathalie DE CASTRO, Alexandre ASLAN, Willy ROZENBAUM, Claire PINTADO, François CLAVEL, Olivier TAULERA, Caroline GATEY, Anne-Lise MUNIER, Sandrine GAZAIGNE, Pauline PENOT, Guillaume CONORT, Nathalie LEROLLE, Anne LEPLATOIS, Stéphanie BALAUSINE, Jeannine DELGADO, Hôpital Saint Louis de Paris, Service des Maladies Infectieuses et Tropicales
- Julie TIMSIT, Magda TABET, Hôpital Saint Louis de Paris, Clinique MST
- Laurence GERARD, Hôpital Saint Louis de Paris, Service d'Immunologie Clinique
- Pierre-Marie GIRARD, Odile PICARD, Jürgen TREDUP, Diane BOLLENS, Nadia VALIN, Pauline CAMPA, Julie BOTTERO, Benedicte LEFEBVRE, Muriel TOURNEUR, Laurent FONQUERNIE, Charlotte WEMMERT, Jean-Luc LAGNEAU Hôpital Saint Antoine de Paris , Service des Maladies Infectieuses et Tropicales
- Yazdan YAZDANPANAH, Bao PHUNG, Adriana PINTO, Dorothée VALLOIS, Ornella CABRAS, Françoise LOUNI, G. Hospitalier Bichat-Claude Bernard de Paris, Service de Maladies Infectieuses et Tropicales
- Gilles PIALOUX, Thomas LYAVANC, Valérie BERREBI, Julie CHAS, Sophie LENAGAT, Hopital Tenon de Paris, Service des Maladies Infectieuses
- Agathe RAMI, Myriam DIEMER, Maguy PARRINELLO, Audrey DEPOND, Hôpital Lariboisière de Paris, Service de Médecine Interne A
- Dominique SALMON, Loïc GUILLEVIN, Tassadit TAHI, Linda BELARBI, Pierre LOULERGUE, Olivier ZAK DIT ZBAR, Odile LAUNAY, Benjamin SILBERMANN, Catherine LEPORT, Laura ALAGNA, Marie-Pierre PIETRI, G. H. Cochin de Paris, Département de Médecine Interne
- Anne SIMON, Manuela BONMARCHAND, Naouel AMIRAT, François PICHON, Myriam KIRSTETTER, G. H. Pitié-Salpétrière de Paris, Service de Médecine Interne
- Christine KATLAMA, Marc Antoine VALANTIN, Roland TUBIANA, Fabienne CABY, Luminita SCHNEIDER, Nadine KTORZA, Ruxandra CALIN, Audrey MERLET, Saadia BEN ABDALLAH, G. H. Pitié-Salpétrière de Paris, Service des Maladies Infectieuses
- Laurence WEISS, Martin BUISSON, Dominique BATISSE, Marina KARMOCHINE, Juliette PAVIE, Catherine MINOZZI, Didier JAYLE, Philippe CASTEL, Jean DEROUINEAU, Pascale KOUSIGNAN, Murielle ELIAZEVITCH, Isabelle PIERRE, Lio COLLIAS, Hôpital Européen Georges Pompidou de Paris, Service d'Immunologie Clinique

- Jean-Paul VIARD, Jacques GILQUIN, Alain SOBEL, Laurence SLAMA, Jade GHOSN, Blanka HADACEK, Nugyen THU-HUYN, Audrey MERLET, Lella NAIT-IGHIL, Agnes CROS, Aline MAIGNAN, Hôtel Dieu de Paris, Centre de Diagnostic et Thérapeutique
- Claudine DUVIVIER, Paul Henri CONSIGNY, Fanny LANTERNIER, Michka SHOAI-TEHRANI, Fatima TOUAM, Saadia JERBI, Centre Médical de l'Institut Pasteur de Paris, Service des Maladies Infectieuses
- Loïc BODARD, Corinne JUNG, Institut Mutualiste Montsouris de Paris, Département de Médecine Interne

Région Parisienne:

- Cécile GOUJARD, Yann QUERTAINMONT, Martin DURACINSKY, Olivier SEGERAL, Arnaud BLANC, Delphine PERETTI, Antoine CHERET, Christelle CHANTALAT, Marie Josée DULUCQ, Hôpital de Bicêtre, Médecine Interne
- Yves LEVY, Jean Daniel LELIEVRE, Anne Sophie LASCAUX, Cécile DUMONT, Hôpital Henri Mondor de Créteil, Immunologie Clinique
- François BOUE, Véronique CHAMBRIN, Sophie ABGRALL, Imad KANSAU, Mariem RAHO-MOUSSA, Hôpital Antoine Béclère de Clamart, Médecine Interne et Immunologie Clinique
- Pierre DE TRUCHIS, Aurélien DINH, Benjamin DAVIDO, Dhiba MARIGOT, Huguette BERTHE, Hôpital Raymond Poincaré de Garches, Service des Maladies Infectieuses et Tropicales
- Alain DEVIDAS, Pierre CHEVOJON, Amélie CHABROL, Nouara AGHER, Hôpital de Corbeil-Essonnes, Service Hématologie
- Yvon LEMERCIER, Fabrice CHAIX, Isabelle TURPAULT, Centre Hospitalier Général de Longjumeau, Service de Médecine Interne
- Olivier BOUCHAUD, Patricia HONORE, Hôpital Avicenne de Bobigny, Maladies Infectieuses et Tropicales
- Elisabeth ROUVEIX, Evelyne REIMANN, Hôpital Ambroise Paré de Boulogne, Médecine Interne
- Alix GREDER BELAN, Claire GODIN COLLET, Safia SOUAK, Hôpital du Chesnay, CH Andre Mignot du Chesnay, Maladies Infectieuses et Tropicales
- Emmanuel MORTIER, Martine BLOCH, Anne-Marie SIMONPOLI, Véronique MANCERON, Isabelle CAHITTE, Emmanuel HIRAUX, Erik LAFON, François CORDONNIER ? Ai-feng ZENG, Hôpital Louis Mourier de Colombes, Médecine Interne
- David ZUCMAN, Catherine MAJERHOLC, Dominique BORNAREL, Hôpital Foch de Suresnes , Médecine Interne
- Agnès ULUDAG, Justine GELLEN-DAUTREMER, Agnès LEFORT, Christine BAZIN, Hôpital Beaujon de Clichy, Médecine Interne
- Vincent DANELUZZI, Juliette GERBE, Centre Hospitalier de Nanterre, Service de Médecine Interne
- Vincent JEANTILS, Mélissa COUPARD, Hôpital Jean Verdier de Bondy, Service de Médecine Interne, Unité de Maladies Infectieuses
- Olivier PATEY, Jonas BANTSIMBA, Sophie DELLLION, Pauline CARAUX PAZ, Benoit CAZENAVE, Laurent RICHIER, Centre Hospitalier Intercommunal de Villeneuve St Georges, Médecine Interne
- Valérie GARRAIT, Isabelle DELACROIX, Brigitte ELHARRAR, Laurent RICHIER, Centre Hospitalier Intercommunal de Créteil, Médecine Interne, Hépato-Gastroentérologie
- Daniel VITTECOQ, Claudine BOLLIOT, Hôpital de Bicêtre, Service de Maladies Infectieuses et Tropicales
- Annie LEPRETRE, Hôpital Simone Veil d'Eaubonne, Médecine 2, Consultation ESCALE
- Philippe GENET, Virginie MASSE, Juliette GERBE, Consultation d'Immuno/Hématologie d'Argenteuil
- Véronique PERRONE, Centre Hospitalier François Quesnay de Mantes La Jolie, Service des Maladies Infectieuses
- Jean-Luc BOUSSARD, Patricia CHARDON, Centre Hospitalier Marc Jacquet de Melun, Service de Médecine
- Eric FROGUEL, Philippe SIMON, Sylvie TASSI, Hôpital de Lagny, Service de Médecine Interne.

Scientific Committee:
Véronique AVETTAND FENOEL (Virologie, Necker, Paris), Francis BARIN (Virologie, Tours), Christine BOURGEOIS (INSERM U1184 IMVA, Bicêtre), Fanny CARDON (ANRS), Marie-Laure CHAIX (Virologie, Saint Louis, Paris), Antoine CHERET (Médecine Interne, Paris), Jean François DELFRAISSY (Médecine Interne, Paris), Asma ESSAT (INSERM U1018, Bicêtre), Hugues FISCHER (TRT5), Cécile GOUJARD (Médecine Interne, Bicêtre), Caroline LASCOUX-COMBE (Médecine, Saint Louis, Paris), Camille LECUROUX (INSERM U1184 IMVA, Bicêtre), Laurence MEYER (Santé Publique, INSERM U1018, Bicêtre), Ventzislava PETROV-SANCHEZ (ANRS), Christine ROUZIOUX (Virologie, Necker), Asier SAEZ-CIRION (Institut Pasteur, Paris), Rémonie SENG (Santé Publique, INSERM U1018, Paris).

UK Register of HIV seroconverters:

We would like to thank all the UK Register participants for allowing their routine clinical data to be included. We gratefully acknowledge the work of the members of the Steering Committee and colleagues at the clinical centres. Special thanks go to the following colleagues: Kristin Kuldanek, Scott Mullaney (St Mary's Hospital, London), Carmel Young (Mortimer Market Centre, London), Antonella Zucchetti, Margaret-Ann Bevan (St Thomas' Hospital, London), Sinead McKernan (Royal Victoria Hospital, Belfast), Emily Wandolo (King's College Hospital, London), Celia Richardson, Elaney Youssef (Brighton and Sussex University Hospital), Pippa Green (Withington Hospital, Manchester), Sue Faulkner (Gloucester Royal Hospital), Rebecca Faville (Whittall Street Clinic, Birmingham), Sandra Herman, Christine Care (Royal Hallamshire Hospital, Sheffield), Helen Blackman (St Mary's Hospital, Portsmouth), and Katharine Bellenger, Keith Fairbrother (Medical Research Council Clinical Trials Unit at UCL, London).
Members of the UK Register Steering Committee: Andrew Phillips (Chair), University College London (UCL), London; Abdel Babiker, UCL, London; Valerie Delpech, Public Health England, London; Sarah Fidler, St. Mary's Hospital, London; Mindy Clarke, Brighton & Sussex University Hospitals NHS Trust, Brighton; Julie Fox, Guys and St Thomas' NHS Trust/Kings College, London; Richard Gilson, West London Centre for Sexual Health, London; David Goldberg, Health Protection Scotland, Glasgow; David Hawkins, Chelsea & Westminster NHS Trust, London; Anne Johnson, UCL, London; Margaret Johnson, UCL and Royal Free NHS Trust, London; Ken McLean, West London Centre for Sexual Health, London; Eleni Nastouli, UCL, London; Frank Post, King's College, London.
The members of the UK Register of HIV seroconverters are: N Kennedy, Monklands Hospital, Airdrie; J Pritchard, Ashford Hospital, Ashford; U Andrady, Ysbyty Gwynedd, Bangor; N Rajda, North Hampshire Hospital, Basingstoke; C Donnelly, S McKernan, Royal Victoria Hospital, Belfast; S Drake, G Gilleran, D White, Birmingham Heartlands Hospital, Birmingham; J Ross, J Harding, R Faville, Whittall Street Clinic, Birmingham; J Sweeney, P Flegg, S Toomer, Blackpool Victoria Hospital, Blackpool; H Wilding, R Woodward, Royal Bournemouth Hospital, Bournemouth; G Dean, C Richardson, N Perry, Royal Sussex County Hospital, Brighton; M Gompels, L Jennings, Southmead Hospital, Bristol; D Bansaal, Queen's Hospital, Burton-Upon-Trent; M Browing, L Connolly, Cardiff Royal Infirmary, Cardiff; B Stanley, North Cumbria Acute Hospitals NHS Trust, Carlisle; S Estreich, A Magdy, St. Helier Hospital, Carshalton; CO'Mahony, Countess of Chester Hospital, Chester; P Fraser, Chesterfield & North Derbyshire Royal Hospital, Chesterfield; SPR Jebakumar, Essex County Hospital, Colchester; L David, Coventry & Warwickshire Hospital, Coventry; R Mette, Mayday University Hospital, Croydon; H Summerfield, Weymouth Community Hospital, Dorset; M Evans, Ninewells Hospital, Dundee; C White, University Hospital of North Durham, Durham; R Robertson, Muirhouse Medical Group, Edinburgh; C Lean, S Morris, Western General Hospital, Edinburgh; A Winter, Gartnavel General Hospital & Glasgow Royal Infirmary, Glasgow; S Faulkner, Gloucestershire Royal Hospital, Gloucester; B Goorney, Salford Hope Hospital, Greater Manchester; L Howard, Farnham Road Hospital, Guildford; I Fairley, C Stemp, Harrogate Hospital, Harrogate; L Short, Huddersfield Royal Infirmary, Huddersfield; M gomez, F

young, St Mary's Hospital Isle of Wight; M Roberts, S Green, Kidderminster General Hospital, Kidderminster; K Sivakumar, the Queen Elizabeth Hospital, King's Lynn; J Minton, A Siminoni, Leeds General Infirmary, Leeds; J Calderwood, D Greenhough, J Minton, St. James' Hospital, Leeds; C DeSouza, Lisa Muthern, C Orkin, Barts & the London NHS Trust, London; S Murphy, M Truvedi, Central Middlesex Hospital, London; K McLean, Charing Cross Hospital, London; D Hawkins, C Higgs, A Moyes, Chelsea & Westminster Hospital, London; S Antonucci, S McCormack, Dean Street Clinic, London; W Lynn, Ealing Hospital, London; M Bevan, J Fox, A Teague, Guy's & St. Thomas NHS Trust, London; J Anderson, S Mguni, Homerton Hospital, London; F Post, L Campbell, E Wandolo King's College Hospital, London; C Mazhude, H Russell, Lewisham University Hospital, London; R Gilson, G Carrick, C Young Mortimer Market Centre, London; J Ainsworth, A Waters, North Middlesex Hospital, London; P Byrne, M Johnson, Royal Free Hospital, London; London; S Fidler, K Kuldanek, S Mullaney, St. Mary's Hospital, London; V Lawlor, R Melville, Whipps Cross Hospital, London; A Sukthankar, S Thorpe, Manchester Royal Infirmary, Manchester; C Murphy, E Wilkins, North Manchester General Hospital, Manchester; S Ahmad, P Green, Withington Hospital, Manchester; S Tayal, James Cook Hospital, Middlesbrough; E Ong, Newcastle General Hospital, Newcastle; J Meaden, Norfolk & Norwich University Hospital, Norwich; L Riddell, City Hospital, Nottingham; D Loay, K Peacock, George Eliot Hospital, Nuneaton; H Blackman, V Harindra, St. Mary's Hospital, Portsmouth; AM Saeed, Royal Preston Hospital, Preston; S Allen, U Natarajan, East Surrey Hospital, Redhill; O Williams, Glan Clwyd District General, Rhyl; H Lacey, Baillie Street Health Centre, Rochdale; C Care, C Bowman, S Herman, Royal Hallamshire Hospital, Sheffield; SV Devendra, J Wither, Royal Shrewsbury Hospital, Shrewsbury; A Bridgwood, G Singh, North Staffordshire Hospital, Stoke-on-Trent; S Bushby, Sunderland Royal Hospital, Sunderland; D Kellock, S Young, King's Mill Centre, Sutton-in-Ashfield; G Rooney, B Snart, the Great Western Hospital, Swindon; J Currie, M. Fitzgerald, Taunton & Somerset Hospital, Taunton; J Arumainayyagam, S Chandramani, Manor Hospital, Walsall; S Rajamanoharan, T Robinson, Watford General Hospital, Watford; M Roberts, Worcester Royal Infirmary, Worcester; O Williams, Maelor Hospital, Wrexham; B Taylor, Wycombe General Hospital, Wycombe; C Brewer, I Fairley, Monkgate Health Centre, York Hospital NHS Trust, York.

HIV-1 Seroconverter Study (Germany):

We would like to thank all members of the German HIV-1 Seroconverter Study Group who participated in this study: **Berlin**: Dres. Mayr, Schmidt, Speidel and Strohbach (Medizinisches Versorgungszentrum, Ärzteforum Seestraße), PD Dr. Arastéh (Auguste-Viktoria-Krankenhaus/ Vivantes), Dr. Cordes, Dres. Stündel and Claus, Dres. Baumgarten, Carganico, Ingiliz and Dupke, Dres. Freiwald and Rausch, Dres. Moll and Schleehauf, Dres. Hintsche and Klausen, Dres. Jessen and Jessen, Dres. Köppe and Kreckel, Dres. Schranz and Fischer, Dres. Schulbin and Speer, Dres. Glaunsinger and Wicke, Dres. Bieniek and Hillenbrand, Dres. Schlote, Lauenroth-Mai and Schuler, Dres. Schürmann and Wesselman (Charité Berlin); **Bochum**: Prof. Dr. Brockmeyer (St. Joseph-Hospital); **Dortmund**: Prof. Dr. Gehring and Dr. Schmalöer and Dr. Hower (Klinikum Dortmund); **Dresden**: Dr. Spornraft-Ragaller (Universitätsklinikum Dresden); **Düsseldorf**: Prof. Dr. Häussinger and PD Dr. Reuter (Universitätsklinik Düsseldorf); **Essen**: Dr. Esser (Universitätsklinikum Essen); **Frankfurt/Oder**: Dr. Markus; **Halle/Saale**: Dr. Kreft (Universitätsklinik Martin-Luther-Universität); **Hamburg**: Dres. Berzow, Christl and Meyer, Prof. Dr. Plettenberg, Dr. Stoehr, Dr. Graefe and Dr. Lorenzen (Institut für Infektionsmedizin, ifi, Allgemeines Krankenhaus St. Georg); Dres. Adam, Schewe and Weitner, Dr. Fenske, Dr.Hansen, Prof. Dr. Stellbrink (Infektionsmedizinisches Zentrum Hamburg, ICH); Dr. Wiemer (Bundeswehrkrankenhaus Hamburg); Dr. Hertling (Universitätsklinikum Hamburg Eppendorf); **Hannover**: Prof. Dr. Schmidt (Medizinische Hochschule Hannover); **Krefeld**: Dr. Arbter; **Ludwigshafen**: Dr. Claus (Klinikum Ludwigshafen); **Mainz**: Prof. Dr. Galle (Klinikum der Joh.-Gutenberg-Universität); **München**: Dres. Jäger and Jägel-Guedes, Dr. Postel, Prof. Dr. Fröschl and Dr. Spinner (Technische Universität München); Prof. Dr. Bogner (Klinikum der Ludwig-Maximilians-Universität); **Regensburg**: Prof. Dr.

Salzberger, Prof. Dr. Schölmerich and Dr. Audebert (Universitätsklinik Regensburg); **Salzgitter**: Dr. Marquardt (Klinikum Salzgitter); **Stuttgart**: Dres. Schaffert, Schnaitmann and Trein, Dres. Frietsch, Müller and Ulmer; **Trier**: Dr. Detering-Hübner (Gesundheitsamt Trier); **Ulm**: Prof. Dr. Kern and Prof. Dr. Dr. Kreidler (Universitätsklinik Ulm); **Weil/Rhein**: Dres. Schubert, Dehn and Schreiber; **Wiesbaden**: Dr. Güler. **Robert Koch Institute Berlin**: Dr. Barbara Gunsenheimer-Bartmeyer, MSc. Daniel Schmidt, Dr. Karolin Meixenberger, Prof. Dr. Norbert Bannert.