Title

High-throughput pipeline for the *de novo* viral genome assembly and the identification of minority variants from Next-Generation Sequencing of residual diagnostic samples

Running head

De novo assembly and variants from NGS clinical samples

Authors

Gallo Cassarino T *, University College London, London, UK

Sugar R, Health and Life Sciences, Intel Corporation, London, UK

Kozlakidis Z, University College London, London, UK and the Farr Institute of Health Informatics Research, London, UK

Kellam P, Imperial College London, London, UK

Pillay D, University College London, London, UK

Frampton D, University College London, London, UK

* (To whom correspondence should be addressed)

Biographical note

Tiziano Gallo Cassarino, PhD, is a Research Associate at University College London; Robert Sugar, PhD, is a software architect at Intel Health and Life Sciences; Zisis Kozlakidis, PhD, is

the ICONIC project manager and the Head of the Centre of Excellence for Infectious Diseases - BBMRI.uk; Paul Kellam, PhD, is the ICONIC Principal Investigator and Professor of Virus Genomics at Imperial College London; Deenan Pillay, PhD, Professor of Virology at University College London and Director of Africa Centre for Health and Population Studies; Dan Frampton, PhD, is a Research Fellow at University College London.

Abstract

Motivation: The underlying genomic variation of a large number of pathogenic viruses can give rise to drug resistant mutations resulting in treatment failure. Next generation sequencing (NGS) enables the identification of viral quasi-species and the quantification of minority variants in clinical samples; therefore, it can be of direct benefit by detecting drug resistant mutations and devising optimal treatment strategies for individual patients.

Results: The ICONIC (Infection response through virus genomics) project has developed an automated, portable and customisable high-throughput computational pipeline to assemble *de novo* whole viral genomes, either segmented or non-segmented, and quantify sequence variants using residual diagnostic samples. The pipeline has been benchmarked on a dedicated High-Performance Computing cluster using paired-end reads from 39 RSV, 420 HIV and 341 Influenza clinical samples. The median coverage of the generated genomes was 96% for the RSV samples, 82% for the HIV dataset and 100% for each Influenza segment. The samples were analysed in parallel, with an average duration of 3 hours per sample. The pipeline can be easily ported to a dedicated server or cluster through either an installation script or a docker image. As it enables the subtyping of viral samples and the detection of relevant drug resistance mutations within three days of sample collection, our pipeline

could operate within existing clinical reporting time frames and potentially be used as a decision support tool towards more effective personalised patient treatments.

Availability: The software and its documentation are available from https://github.com/ICONIC-UCL/pipeline

Contact: t.cassarino@ucl.ac.uk

Supplementary information: Supplementary data are available online.

Keywords

Next-Generation Sequencing, *de novo* genome assembly, genomic variants, viral genomics, bioinformatics, clinical decision support

1 Introduction

Viruses are intracellular parasites and most are characterised by a high replication rate within their host. During replication the polymerase proteins are prone to transcription errors or mutations, with RNA viruses having the highest mutation rates. New genomes containing mutations are continuously generated and selected on the basis of their fitness to infect and to replicate within the host's cells[1]. Average mutation rates of RNA viruses are about 10^{-4} – 10^{-5} errors per nucleotide copied or, based on average genomic sizes, about one mutation per genome copied[2]. Moreover, recombination of genomic parts increases the evolution dynamics and the genomic divergence within viral populations. One such example is the Human Immunodeficiency Virus (HIV), where the recombination is

considered faster than the mutation rate[3]. These and other features suggest that a viral population is actually made of an ensemble of related mutants that can be described as a quasi-species and on which the selective pressure influences all the viruses as a single unit. Such high genomic variability allows the viral populations to survive challenges mounted by the host's immune system and by antiviral agents, hindering the effective treatment of patients and making it difficult to eradicate infections[4]. Next-Generation Sequencing (NGS) coupled with bioinformatics analyses enables the high-throughput detection of genomic variants and the classification of known and novel viral species overcoming the need for expensive culturing and/or labour-intensive Sanger sequencing techniques[5].

Within the context of a clinical setting, NGS data can be applied on sequenced diagnostic samples to identify pathogens by assembling whole genomes and quantifying low-level drug resistance mutations below the 15-20% frequency sensitivity limit of the traditional Sanger sequencing technique[6].

In order to impact patient treatment pathways, it is necessary for a computational pipeline to be capable of combining a number of features at the same time. These are the ability to be pathogen agnostic, to assemble *de novo* whole genomes, to report genome variants, to be scalable for high-throughput analyses, customisable and portable to different software environments.

To this end a computational pipeline for analysing NGS data was developed that meets all of these above requirements. The input data are unprocessed paired-end reads from residual clinical diagnostic samples obtained under appropriate ethics permissions, while the output reports the consensus genome and all minority variants present in the viral quasi-species. The pipeline is part of the ICONIC project, which uses viral genomic data to provide decision

support towards the personalised treatment of patients; to guide hospital infection control responses; and to inform the surveillance and epidemiological responses to viral community outbreaks.

2 Methods

2.1 Software Installation

The pipeline has been written in Python and developed for analysing batches of paired-end reads on a High Performance Computing (HPC) cluster or on a server running the Son of Grid Engine scheduler (which must be installed separately). The software can be built either by the provided installation script or from a Dockerfile [https://www.docker.com/]. Both of these alternatives automatically: (1) download the pipeline dependencies, (2) configure the pipeline and (3) install it on the local appliance. Finally, the pipeline can be loaded as a bundle of Environment Modules[7] which can be used immediately.

2.2 Reads Analysis

The pipeline takes as input a set of short paired-end reads, returning the consensus genome, the list of identified variants and a set of statistics and QC metrics for each analysed sample, as well as for the whole dataset. The software consists of a number of well-defined functional stages, which allow the pipeline to be run from a particular starting point or on a specific analysis step. After parsing the input arguments, the pipeline initialises and sends an array job to the HPC cluster, which runs the analysis of each sample's reads in parallel to facilitate rapid analysis. Each analysis job is composed of seven distinct stages, as shown in Figure 1.

Figure1

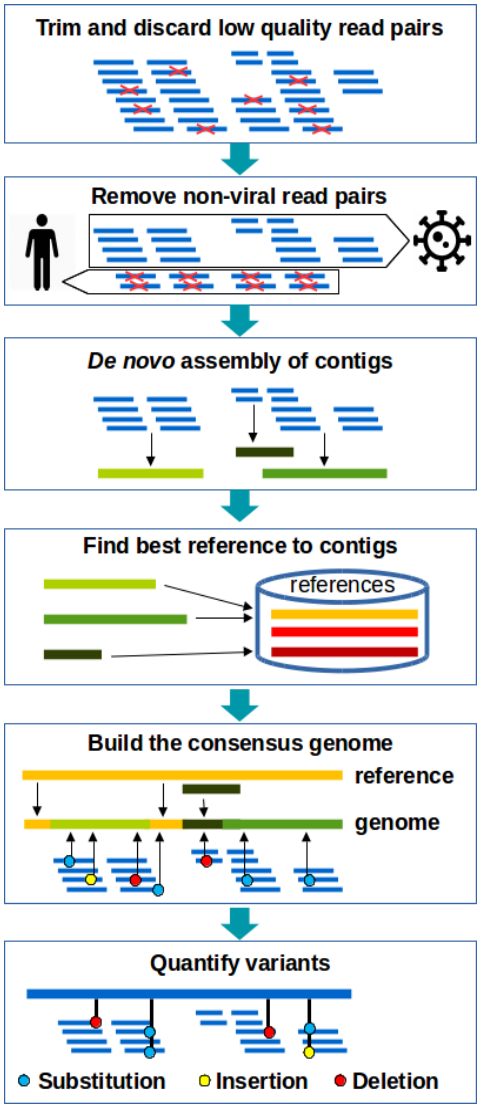

Figure 1. The workflow of the computational pipeline showing the major steps during the analysis of the sample reads. Stage 1: Trimming step to keep only reads with high quality base scores; stage 2: filtering step to discard reads unmapped or mapping to the host; stage 3: *de novo* assembly of the remaining reads into contigs; stage 4: alignment step to find the most likely reference genome for the sample; stage 5: filling gaps between contigs using the reference and creating the consensus genome by replacing contigs and reference bases with those that were most frequently found in the reads; stage 6: quantification of the genomic

variants; stage 7: generation of statistics and quality control measures for the sample consensus genome.

The first stage takes as input the short paired-end reads in either flat or compressed FASTQ format and passes them to Trimmomatic[8] (version 0.33), which removes or trims low quality read pairs (used with default settings, except for a sliding window Phred score cut-off of 30) . In the second stage, the remaining read pairs are mapped with SMALT (version 0.7.6, http://www.sanger.ac.uk/science/tools/smalt-0) against a user-defined decoy genome (which must be created by the user), containing the host genome and a set of genomes corresponding to the virus present in the sample set. Then, reads that map preferentially to the host or that are unmapped are discarded. FastQC (version 0.11.3, http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) can be optionally run before the trimming stage and after this filtering step to perform a quality control analysis and to manually spot potential any sequencing bias.

The third stage generates contigs by performing a *de novo* assembly of the filtered paired-end reads. In the current implementation, contigs are generated by the IVA *de novo* assembler[9] (version 1.0.0).

In the fourth stage, contigs are aligned with LASTZ[10] (1.07.73) against an user-made local set of genomic sequences of the sequenced virus, in order to identify the match with the highest alignment score. For segmented genomes (e.g. Influenza) this step identifies the best hit for each segment, thus allowing to utilise references with different subtypes. BLAST+[11] (version 2.2.30) is used to efficiently retrieve the reference sequence, which are written to a file.

The fifth stage creates the sample consensus genome. It produces a first draft sequence by aligning, with LASTZ, the contigs to the (previously found) reference genome and by filling the gaps between them using the reference bases. Since there is not a required minimum contig coverage, a draft genome can be built with less contig bases than those present in the reference.Whenever aligned contigs overlap, the contig with the highest read depth of coverage is used in the draft genome. Afterwards, an iterative approach is used to generate the final consensus genome. Firstly, filtered reads are mapped to the draft genome from which a pileup file is generated with Samtools[12] (version 1.2). The pileup file is then parsed to count the read bases (or gaps) in each alignment position and the base (or gap) with the highest count overwrites the corresponding aligned draft sequence base. The mapping and pileup steps are repeated until the genome converges to a stable sequence, i.e. when there are no more substitutions, insertions or deletions to be made. A maximum of 10 cycles are performed to avoid potentially infinite iterations.

In the sixth stage, the reads are mapped for the last time to the consensus genome and the variants – i. e., the differences (substitutions, insertions and deletions) of the read bases with respect to the consensus genome – at each genomic position are quantified using Samtools mpileup and saved to a file. To reduce the number of false positive variants, the pipeline allows user-defined cut-offs of variant read depth, variant frequency and consensus base frequency. The variants are also reported at variant frequency thresholds of 20% and 2%, in accordance with current Public Health England reporting practices and two additional bins of 10% and 5%.

In the seventh and last stage, summary statistics and plots are generated for the consensus genome; these metrics include genome length, read depth of coverage distribution, number of variants and strand biases.

When all the samples have been analysed, the pipeline collects the genome metrics of each sample and aggregates them to provide a set of statistics for the whole batch. Each stage saves to disk its results, and optionally, it is possible to keep all intermediate files created during the analysis; however, this requires considerably more disk space. The pipeline records all the steps performed during the analysis, saves the final status of each sample analysis and the cause of any premature halt (e.g. when *de novo* assembly fails due to an insufficient number of reads), thus assisting in identifying problematic samples.

The analysis stages can be either modified through the configuration file, which stores all the parameters used during the analysis, or substituted with user-made plugins. New reference datasets and decoy genomes can be added to the pipeline just by including their paths to the configuration file, thus allowing the pipeline to analyse NGS data of any virus.

Additional details regarding input options, configuration parameters, supporting data and result descriptions can be found in the software documentation .

2.3 Ethics permissions

The ICONIC study has REC approval (13/LO/1303) received on 20th August 2013, IRAS project ID 131373. The favourable opinion applies to all NHS sites taking part in the study, while additional permissions have been obtained from the NHS/HSC R&D offices of all partner sites prior to the start of the study.

3 Results

The capability of the pipeline to build *de novo* genomes was assessed using publicly available read datasets and, as an example, the results on a set of 39 Human Respiratory Syncytial Virus (RSV) samples are shown. The pipeline performances were also tested on two bigger clinical sample datasets sequenced as 300 basepair paired-end reads at the Wellcome Trust Sanger Institute on an Illumina MiSeq machine. The first set comprises 420 ICONIC samples of HIV, whereas the second set consists of 341 ICONIC samples of the outbreaks-responsible Influenza virus.

3.1 Comparison with publicly available RSV genomes

A test set of 39 RSV publicly available sequences, A and B subtypes, with deposited reads was selected from Agoti et al[13] and downloaded from the Sequence Read Archive (SRA). We created a local BLAST database of reference sequences from full RSV genomes available in GenBank (as of 2015-10-01) and a decoy genome made of the human genome and the RSV reference sequences. The average duration of a sample analysis was around 3 hours. With respect to the test set, the genomes built with the pipeline were longer in 19 samples (49%), of comparable length in 13 samples (33%), and shorter only in 7 (18%) due to the small number of reads remaining after the quality filtering stage (Details can be found in the Table_S1 of the supplementary material). The median coverage of the built genomes was 96% of the corresponding reference. Therefore, our pipeline can build consistently full-size

genomes and also contribute to the deposited public sequences by improving their genomic coverage.

3.2 Analysis of HIV clinical samples

The pipeline analysed 420 HIV ICONIC samples using a local BLAST database of complete genomes for all HIV-1 subtypes taken from the Los Alamos HIV Sequence Database (http://www.hiv.lanl.gov/content/sequence/HIV/mainpage.html, October 2015), and a decoy genome made of the human genome (version hg38) and the HIV sequences of the BLAST database. The pipeline assembled *de novo* 377 genomes out of 420 samples (~90%), three quarters of which covered at least 94% of the respective reference sequence. In the remaining 43 samples only a few hundreds reads were left after the human contaminant ones were discarded, causing the assembly to fail. The consensus genomes were aligned against the HXB2 reference sequence (Gen-Bank accession: K03455) with LASTZ to identify the regions where the pipeline was most successful in assembling the consensus sequence. Figure 2 shows that all the assembled genomes cover every gene of HXB2, confirming that the pipeline could assemble full-length, clinically relevant genomes, for every sample for which the sequencing was successful. However the genome depth of coverage decreases towards the end of the sequence with two regions with lower depth than the surrounding sequence.

Figure2

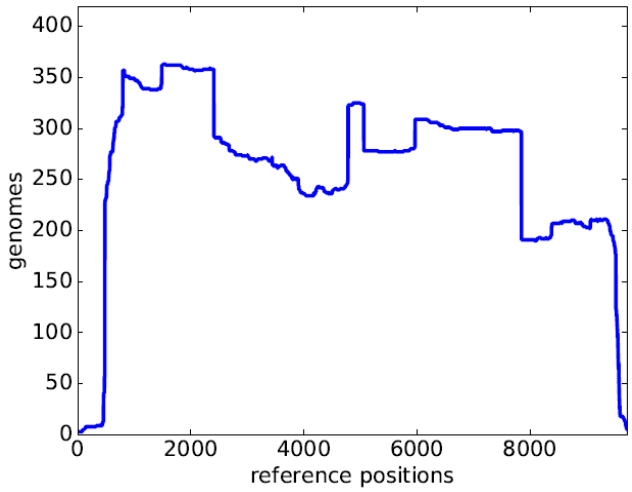

Figure 2. Genome depth of coverage from the analysis of the HIV batch. The blue line represents the number of genomes aligned at each position of the HXB2 sequence, showing that all the genomes cover each gene of HXB2. The depth decreases towards the end of the sequence, from around 350 genomes to about 200, and that it is much lower between about 2000 and 4000, and after around 8000, than the surrounding depth.

The variants identified in the last step of the pipeline above the 1% frequency cutoff, are aggregated from all the sample results and reported as a distribution of points against the genome positions, allowing to qualitatively assess the degree of variability along the viral genome through the samples in the batch. As displayed in Figure 3, mutations are spread all over the HIV genome and are more frequent in the Gag and Env regions, confirming their known high variability.

Figure3

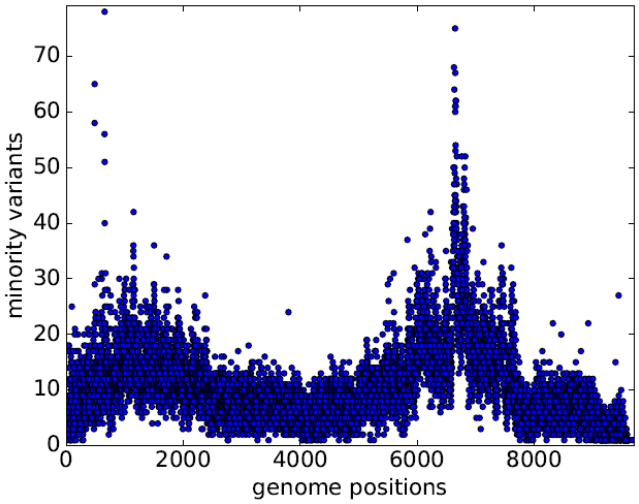

Figure 3. Variants within the HIV sample dataset. Each dot represents the absolute number of variants above the 1% frequency cutoff, at each position of the consensus genome across all the samples. The regions with the highest variation correspond to the Gag and Env genes.

For each sample, the pipeline generated plots to visualise several measures on the reads mapped to the consensus genome, for example the reads depth of coverage as shown in Figure S1 (Supplementary Material). These plots can be useful to check the possible causes for which the pipeline does not build the consensus genome even if a sample has been sequenced successfully.

3.3 Analysis of Influenza clinical samples

The database used as reference set for the analysis of the 341 ICONIC Influenza samples was created using the Human Influenza full genomes belonging to any serotype downloaded from the NCBI's Influenza Virus Resource (October 2015). As for the HIV analysis, the decoy genome was made of the human genome (version hg38) and the Influenza reference

sequences. The pipeline reported 169 genomes on a total of 341 (~50%) samples, while the rest did not contain enough viral reads to be *de novo* assembled into contigs. To overall assess the degree to which the consensus genomes covered the Influenza segments, these were aligned against a H1N1 sequence (strain: A/California/07/2009). Although the Influenza segments are relatively short (around 2000 bases for the longest one), the pipeline was able to create either partial or full-length segments (an example is shown by the genome depth in Figures S2 and S3 for segment 1 and 4 respectively). The length of each assembled segment depends on the number of filtered reads available for that particular region, which can be assessed from the distribution of the median read depth (Figure S4 in the Supplementary Material). The segment coverage across the entire dataset is plotted in Figure S5 (Supplementary Material) as the distribution of the consensus genome coverage against its reference (Ns in the consensus genome are ignored); the median coverage was 100% for all segments and at least 50% for the first three segments in half of the dataset samples. The sequence variants identified in the last step of the pipeline are aggregated from all the sample results and reported for each segment as well. In each sample, the plot of reads depth along the consensus genome qualitatively identified regions where the sequencing have been successful; Figure 4 illustrates an example in one Influenza sample, showing that the number of aligned reads is higher at the ends of the segment than in the central region. Moreover, the number of soft-clipped reads highlights regions with many mismatches, in which the alignment to the consensus genome is more difficult than in the central region.

Figure4

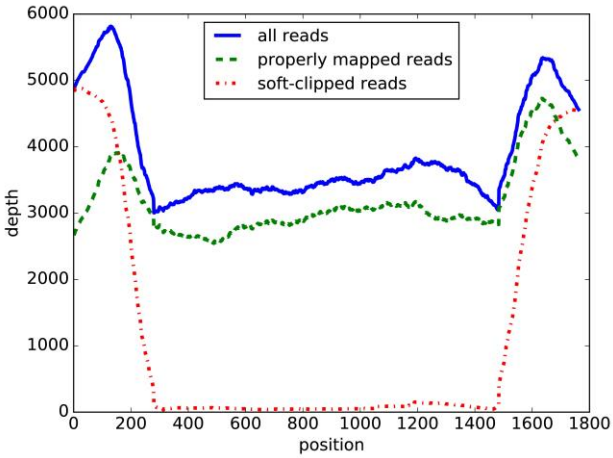

Figure 4. Read depth of coverage along the consensus sequence of segment 4 built for an Influenza sample. The blue line shows the number of aligned reads, the green dashed line shows the properly-mapped (according to the mapper) reads and the red dash-dotted line represents the number of soft-clipped reads. In regions with higher read depth, variants are called more reliably; a high number of soft-clipped reads represents mismatches and can indicate positions with high base variability.

For each consensus genome it was possible to inspect the variants in a table through: (1) their type (substitution, insertion or deletion), (2) the variant frequency and (3) the ratio between the amount of forward and reverse reads. By plotting the distribution of each single mutation against the consensus sequence, it was possible to identify genomic regions with high variability. The Influenza virus is characterized by a relatively low mutation rate; therefore, the number of variants at each position is far less with respect to high mutation rate viral genomes as HIV (as displayed for segment 1 in Figure S6 in the Supplementary Material).

3.4. Performances

The analysis was performed using University College London's HPC cluster "Legion" on 124 dedicated Dell C6220 nodes, where each node can work as a 16 core Symmetric Multi-Processing device with 64 GB of RAM. Legion runs an operating system based on Red Hat Enterprise Linux 7 with the Son of Grid Engine batch scheduler.

The pipeline processed each HIV sample in about 2 hours, while it analysed each Influenza sample in less than 5 hours, on average. The stages with the longest duration were the filtering step, in which the reads are aligned to the decoy genome and the assembly stage, where the reads are *de novo* assembled. All the software, except the scripts that build the consensus genome and manage the flow of the data through the workflow, was run in multi-threading mode using all the cores available to decrease the computational time of the analysis. The memory used through the analysis of each sample had an initial peak around 10 GB during the filtering stage (because the mapping software needs to load into memory the index of the decoy genome and the reads), whereas it reached 2 GB during the *de novo* assembly step; otherwise, the average memory load is about 500 MB.

Finally, Docker images have a negligible impact on the performances[14] and this holds true especially for our pipeline, since it is contained in a single Docker image.

A detailed analysis of the performances of the ICONIC pipeline on a dedicated server can be found in the Intel white paper “Performance Considerations of the ICONIC Next-Generation Viral Sequencing Pipeline”.

3.5 Portability

The pipeline can be installed either from a Dockerfile or by an installation script on a GNU/Linux system, or it can be ran directly as a docker image. The former installation method can be used on a dedicated server, while the latter is better suited for environments shared among multiple users, hence addressing different set up needs. Often Docker cannot be installed on academic clusters for security reasons, as the ability of running docker images is equivalent to having access to root privileges [Docker Security: https://docs.docker.com/engine/articles/security/]. In these cases Environmental Modules are used to make sure all software dependencies are loaded with the correct version. This option was necessary to address the between-institutions portability.

4 Discussion

Motivated by the increasing potential to apply NGS on clinical viral samples as a method to improve the treatment of patients affected by viral diseases, a high-throughput computational pipeline was developed as part of the ICONIC project to assemble viral consensus genomes *de novo* and to detect minority variants in viral residual clinical samples. As the pipeline accepts raw reads, it is possible to analyse sequencing data as soon as they are produced without the need of any pre-processing step. Moreover, the pipeline can analyse reads from any viral genome, segmented or not, for which at least one sequence, even partial, already exists. The consensus genome and the associated variants can be used to identify the dominant viral subtype and to quantify the variants occurring at specific positions and identifying the presence of drug resistance mutations. Furthermore, the possibility to run the pipeline in high-throughput mode is essential to facilitate analyses of

potentially large numbers of patient samples during a seasonal outbreak of conditions of viral origin. Given the variety of information technologies employed within different clinical environments, the pipeline was designed to be easily ported to any cluster or server running a GNU/Linux system.

A set of publicly available RSV sample sequences were compared to the genomes generated by the pipeline presented here, starting from the deposited reads. Full genomes were generated in most of the samples, with half giving longer alignments to reference than those that are currently publicly available. Therefore, the pipeline can successfully and efficiently assemble *de novo* viral genomes and could potentially be used to replace and update the data deposited on public archives. The subsequent analyses of the ICONIC HIV and Influenza reads on an HPC cluster, sequenced from clinical residual samples, confirmed the above results. However, the ability of the pipeline to assemble a sample genome depends on the read depth and coverage, which is affected by the efficiency of the amplification primers. In the case of HIV, multiple primers were required to fully cover the viral genome, so the portion of the DNA covered by multiple amplicons had higher genome depth of coverage compared to the surrounding sequences. Instead, viral genomes that require only one primer, like the individual segments of Influenza, can either be assembled or not depending on the degree of amplification. In particular, most of the genomes built for the HIV samples fully cover all the genes, thus enabling the quantification of variants along the whole sequence and allowing the identification of potential drug resistance mutations within the genes usually neglected during targeted sequencing of Gag-Pol. Indeed, the highest sequence variability is found in the Env gene, which can be a suitable target for new drug treatments. Fewer genomes cover the extremities than the central part of the reference, mostly caused by a higher nucleotide variability than the rest of the sequence,

which results in a more complex assembly step. The reason for some missing sample genomes was due to the very low amount of viral reads (few hundreds) that were left after the trimming and filtering stages. Such outcomes can either indicate a failed PCR amplification, or a contaminated/degraded sample, and can help to identify errors in the library preparation.

The analysis lasted only a few hours and required an operationally reasonable amount of memory for each sample, thus capable of processing batches of hundreds samples overnight on a typical HPC cluster. The higher execution time for the Influenza samples, compared to HIV, is due to the repetition of the commands needed to build each genomic segment. The memory required depends on the size of the decoy genome index, so it can be reduced in case of a limited amount of available memory on the user's server. Such a timeframe and reasonable resources, coupled with the easy portability, allows the pipeline to be suitable to high pressure situations, such as clinical settings, where reporting turnaround times are compressed, especially in the case of diseases of viral aetiology. For these reasons, the pipeline is utilised regularly to analyse samples from different hospitals in London (The Royal London Hospital, Guy's and St Thomas's Hospital, University College London Hospital) and from collaborations with international partners: the BaliMEI project [15], Fraunhofer Institute, University of Athens, Brussels' hospitals. It is not difficult to imagine that using current capability, a diagnostic report could be provided to a clinician within an actionable time window, as in our experience the end-to-end process from patient sampling to genome assembly and clustering reporting takes approximately five days. One of the main strengths of the pipeline is that it can be utilised on sequencing data of any known virus, to generate *de novo* full length viral genomes, in which the sample quality is very variable and the virus subtype is unknown. These features empower our software to be eventually deployed in

clinical settings as a decision support tool towards a personalised patient treatment and the improved information management of hospital infections.

Funding

This work was supported by the Health Innovation Challenge Fund T5-344 (ICONIC), a parallel funding partnership between the Department of Health and Wellcome Trust. The views expressed in this publication are those of the author(s) and not necessarily those of the Department of Health or Wellcome Trust.

Acknowledgments

We would like to thank Chiara Garattini (Intel Corporation) and Elijah Charles for the useful suggestions and discussions during the development of the software and the preparation of the manuscript; Christophe Fraser, Chris Wymant and Oliver Ratmann from the BEEHIVE consortium for their input during the pipeline's initial development and the PANGEA_HIV consortium for facilitating these discussions.

Key points

- The ICONIC high-throughput computational pipeline *de novo* assembles viral genomes and quantifies minority variants.
- It uses Illumina paired-end reads sequenced from residual diagnostic samples.

- It could operate within existing clinical reporting time frames and potentially be used as a decision support tool towards more effective personalised patient treatments.

References

1. Moya A, Holmes EC, Gonzalez-Candelas F. The population genetics and evolutionary epidemiology of RNA viruses, Nat Rev Microbiol 2004;2:279-288.
2. Sanjuan R, Nebot MR, Chirico N et al. Viral mutation rates, J Virol 2010;84:9733-9748.
3. Rhodes T, Wargo H, Hu WS. High rates of human immunodeficiency virus type 1 recombination: near-random segregation of markers one kilobase apart in one round of viral replication, J Virol 2003;77:11193-11200.
4. Domingo E, Menendez-Arias L, Quinones-Mateu ME et al. Viral quasispecies and the problem of vaccine-escape and drug-resistant mutants, Prog Drug Res 1997;48:99-128.
5. Quinones-Mateu ME, Avila S, Reyes-Teran G et al. Deep sequencing: becoming a critical tool in clinical virology, J Clin Virol 2014;61:9-19.
6. Tsiatis AC, Norris-Kirby A, Rich RG et al. Comparison of Sanger sequencing, pyrosequencing, and melting curve analysis for the detection of KRAS mutations: diagnostic and clinical implications, J Mol Diagn 2010;12:425-432.
7. Furlani JL, Osel PW. Abstract Yourself With Modules. Proceedings of the 10th USENIX conference on System administration. Chicago, IL: USENIX Association, 1996, 193-204.
8. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics 2014;30:2114-2120.
9. Hunt M, Gall A, Ong SH et al. IVA: accurate de novo assembly of RNA virus genomes, Bioinformatics 2015;31:2374-2376.
10. Harris RS. Improved pairwise alignment of genomic DNA. College of Engineering. The Pennsylvania State University, 2007.
11. Camacho C, Coulouris G, Avagyan V et al. BLAST+: architecture and applications, BMC Bioinformatics 2009;10:421.
12. Li H, Handsaker B, Wysoker A et al. The Sequence Alignment/Map format and SAMtools, Bioinformatics 2009;25:2078-2079.
13. Agoti CN, Otieno JR, Munywoki PK et al. Local evolutionary patterns of human respiratory syncytial virus derived from whole-genome sequencing, J Virol 2015;89:3444-3454.
14. Di Tommaso P, Palumbo E, Chatzou M et al. The impact of Docker containers on the performance of genomic pipelines, PeerJ 2015;3:e1273.
15. Adisasmito W, Budayanti SN, Aisyah DN et al. Phylogenetic characterisation of circulating, clinical influenza isolates from Bali, Indonesia: preliminary report from the BaliMEI project, BMC Infect Dis 2017;17:583.

Supplementary material

sample	Iconic-pipeline genomes (length)	Public genomes (length)
ERR303259	14995	9899
ERR303260	14941	14720
ERR303261	14995	14715
ERR303262	14941	4685
ERR303264	14958	14719
ERR303265	14995	14934
ERR303266	14994	14934
ERR303267	14998	14720
ERR303269	15003	14953
ERR303303	12495	14953
ERR303311	14994	14953
ERR303312	14212	14934
ERR303313	14998	11417
ERR303316	12640	14731
ERR303322	14914	14953
ERR323212	14420	6817
ERR323213	4789	9211
ERR323214	10173	14953
ERR331021	15254	14953
ERR376407	6703	14231
ERR376408	10578	14934
ERR376409	8016	8759
ERR376413	8023	5409
ERR376414	7991	14953
ERR376415	10601	12028
ERR376416	12369	5409
ERR376417	9838	14934
ERR376442	14941	14953
ERR381723	14642	14933
ERR381725	14642	14735
ERR381726	14641	12143
ERR438864	14674	14719
ERR438865	14645	14952
ERR438867	429	9778
ERR438868	14642	9899
ERR438904	14642	7091
ERR438905	14642	14953
ERR438910	14665	14735
ERR438932	14642	9783

Table S1. Sample identifiers from the Sequence Read Archive with the length of the associated deposited RSV genomes and of the consensus sequences built by the ICONIC pipeline.

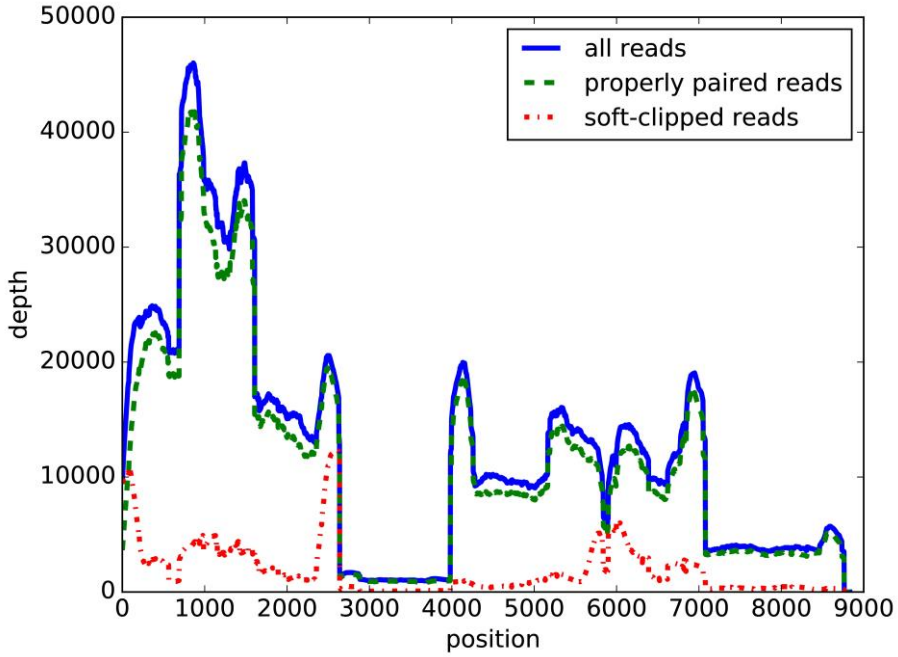

Figure S1. Read depth of coverage along the consensus sequence of a HIV sample. The blue line shows the number of aligned reads, the green dashed line shows the properly-mapped (according to the mapper) reads and the red dash-dotted line represents the number of soft-clipped reads. The higher the read depth, the higher the reliability of the base at that position, while a high number of soft-clipped reads represent mismatches and can indicate positions with high base variability.

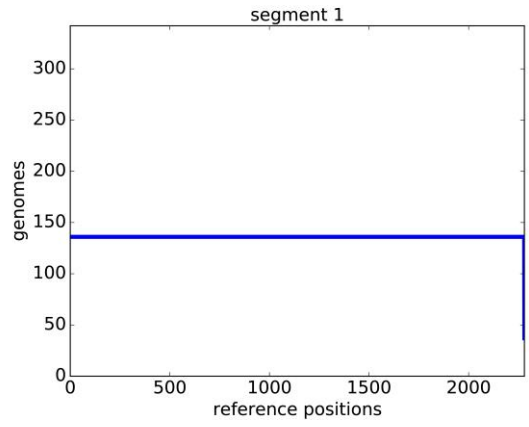

Figure S2. Genome depth of coverage of the segment 1 within the Influenza batch. The blue line shows the number of genomes aligned at each position of the reference sequence.

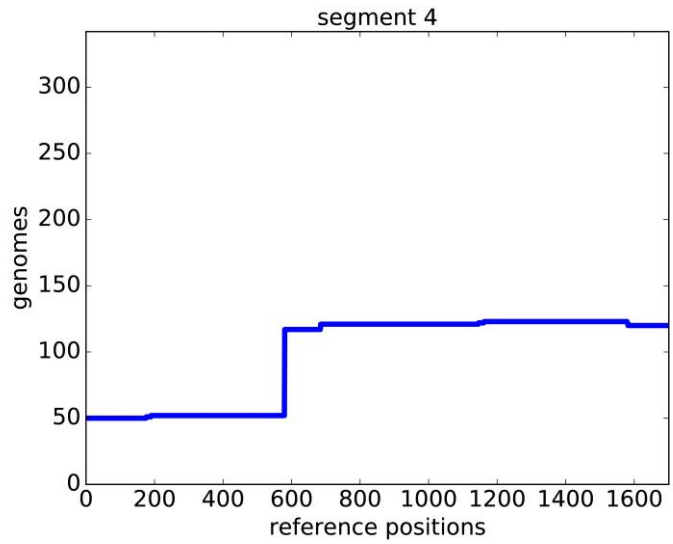

Figure S3. Genome depth of coverage of the segment 4 within the Influenza batch. The blue line shows the number of genomes aligned at each position of the reference sequence.

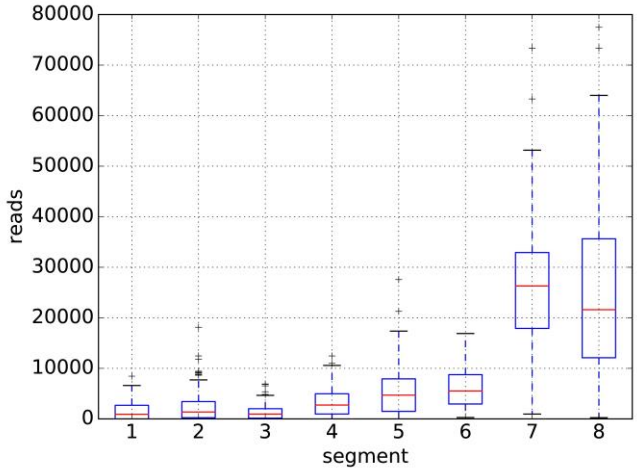

Figure S4. Median read depth of coverage for each segment across the Influenza batch. Median is shown as red line, 25 and 75 QRT as box, 95 QRT as whiskers and outliers as plus signs.

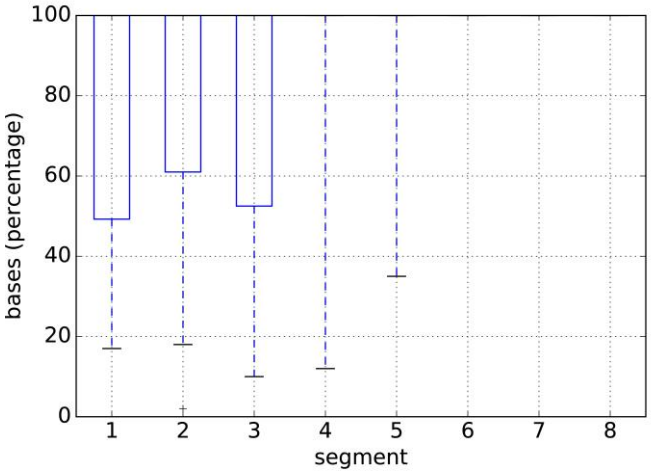

Figure S5. Segment coverage from all the sample in the Influenza batch. All segments have median equal to 100%, segments 4 and 5 have only the 95 QRT below 100%, while segments 7 and 8 have all coverages at 100%. Median is shown as red line, 25 and 75 QRT as box, 95 QRT as whiskers and outliers as plus signs.

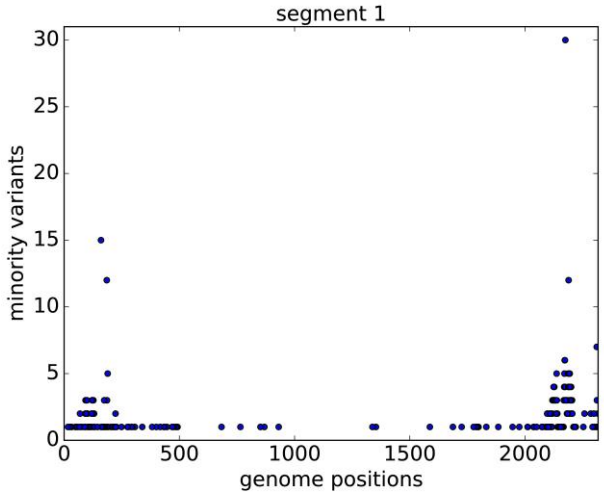

Figure S6. Number of variants within the Influenza sample dataset. Each dot represents the absolute number of variants, above 1% frequency cutoff, at each position of the consensus genome across all the samples.