A Modified RNA-Seq Approach for Whole Genome Sequencing of RNA Viruses from Faecal and Blood Samples

Elizabeth M. Batty[1]¶, T. H. Nicholas Wong[2,3]*¶, Amy Trebes[1]¶, Karène Argoud[1], Moustafa Attar[1], David Buck[1], Camilla L. C. Ip[4], Tanya Golubchik[4], Madeleine Cule[4], Rory Bowden[1], Charis Manganis[2], Paul Klenerman[2], Eleanor Barnes[2], A. Sarah Walker[2,3], David H. Wyllie[2,3], Daniel J. Wilson[1,2], Kate E. Dingle[3,5], Tim E. A. Peto[2,3], Derrick W. Crook[2,3]¶, Paolo Piazza[1]¶

[1] Wellcome Trust Centre for Human Genetics, Oxford, United Kingdom, [2] Nuffield Department of Medicine, University of Oxford, John Radcliffe Hospital, Oxford, United Kingdom, [3] Oxford NIHR Biomedical Research Centre, John Radcliffe Hospital, Oxford, United Kingdom, [4] Department of Statistics, University of Oxford, Oxford, United Kingdom, [5] Nuffield Department of Clinical Laboratory Sciences, Headley Way, University of Oxford, John Radcliffe Hospital, Oxford, United Kingdom

Abstract

To date, very large scale sequencing of many clinically important RNA viruses has been complicated by their high population molecular variation, which creates challenges for polymerase chain reaction and sequencing primer design. Many RNA viruses are also difficult or currently not possible to culture, severely limiting the amount and purity of available starting material. Here, we describe a simple, novel, high-throughput approach to Norovirus and Hepatitis C virus whole genome sequence determination based on RNA shotgun sequencing (also known as RNA-Seq). We demonstrate the effectiveness of this method by sequencing three Norovirus samples from faeces and two Hepatitis C virus samples from blood, on an Illumina MiSeq benchtop sequencer. More than 97% of reference genomes were recovered. Compared with Sanger sequencing, our method had no nucleotide differences in 14,019 nucleotides (nt) for Noroviruses (from a total of 2 Norovirus genomes obtained with Sanger sequencing), and 8 variants in 9,542 nt for Hepatitis C virus (1 variant per 1,193 nt). The three Norovirus samples had 2, 3, and 2 distinct positions called as heterozygous, while the two Hepatitis C virus samples had 117 and 131 positions called as heterozygous. To confirm that our sample and library preparation could be scaled to true high-throughput, we prepared and sequenced an additional 77 Norovirus samples in a single batch on an Illumina HiSeq 2000 sequencer, recovering >90% of the reference genome in all but one sample. No discrepancies were observed across 118,757 nt compared between Sanger and our custom RNA-Seq method in 16 samples. By generating viral genomic sequences that are not biased by primer-specific amplification or enrichment, this method offers the prospect of large-scale, affordable studies of RNA viruses which could be adapted to routine diagnostic laboratory workflows in the near future, with the potential to directly characterize within-host viral diversity.

Citation: Batty EM, Wong THN, Trebes A, Argoud K, Attar M, et al. (2013) A Modified RNA-Seq Approach for Whole Genome Sequencing of RNA Viruses from Faecal and Blood Samples. PLoS ONE 8(6): e66129. doi:10.1371/journal.pone.0066129

Editor: Amit Kapoor, Columbia University, United States of America

Received February 21, 2013; **Accepted** May 2, 2013; **Published** June 10, 2013

Funding: This study was supported by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre and the UKCRC Modernising Medical Microbiology Consortium, the latter funded under the UKCRC Translational Infection Research Initiative supported by Medical Research Council, Biotechnology and Biological Sciences Research Council and the NIHR on behalf of the Department of Health (grant G0800778) and the Wellcome Trust (grant 087646/Z/08/Z). The authors acknowledge the support of Wellcome Trust core funding (grant 090532/Z/09/Z) and MRC Hub (grant G0900747 91070). TEAP and DWC are NIHR Senior Investigators. THNW is an NIHR Oxford BRC Doctoral Research Fellow. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: nick.wong@ndm.ox.ac.uk

¶ These authors contributed equally to this work.

Introduction

Rapid, high-throughput and accurate whole genome sequencing of RNA viral pathogens such as Norovirus and Hepatitis C virus (HCV) has enormous potential for the investigation of local transmission and widespread dispersal [1]. Developing generic methods that yield whole genomic sequence has been challenging because the substantial genomic variation within circulating populations of RNA viruses poses difficulties for primer design [2]. Some RNA viruses are difficult or impossible to grow in tissue culture precluding obtaining highly purified viral nucleic acid in suitable concentrations for whole genome sequencing. The challenges associated with non-culture-based approaches to purifying viral RNA from samples such as faeces and blood that are heavily contaminated by RNA from other sources has severely limited the application of high-throughput sequencing [3]. Unlike whole genome sequencing of bacteria, which is now well developed [4][5][6][7][8], sequencing of viruses has therefore not benefitted from dramatic advances in sequencing capacity.

Approaches for successfully sequencing RNA viruses have previously been dependent, in some way, on target-specific primer-based amplification of viral genomes [9]. These amplicons can then be sequenced using Sanger sequencing or next-generation sequencing platforms such as Roche 454 and Illumina [10]. Recent examples of this approach include studies of Norovirus evolution [11] and HCV diversity [12]. Next-generation sequencing of amplicons and partial genome fragments from RNA viruses have been used to investigate variants within populations of Human Immunodeficiency Virus and HCV infecting individual patients [13][14][15]. An enhancement on this underlying amplification approach uses enrichment of target viral sequences incorporating primers as "bait" to capture larger genomic fragments [16][17]. All these approaches are expensive, labour-intensive, slow and inflexible, and may require *a priori* knowledge of partial or approximate virus sequence, utilising different primers for different virus strain-groups. In turn, assumptions about the sequences present can bias resulting data, altering the representation of the viral genomic sequence at a population level in a sample [18].

Here we adapt a strategy for high-throughput RNA sequencing for use on RNA viruses present in blood and faeces using RNA shotgun sequencing (RNA-Seq) [19][20]. We show that this technology can not only generate near whole genome sequences, but can also recover the sequence of multiple within-host variants of highly diverse pathogens such as Norovirus and HCV. This method may be deployed rapidly, and cheaply, using both accessible bench-top and higher capacity platforms. The approach can be used to successfully characterise un-culturable viral genomes as an alternative to a PCR-amplicon based sequencing. We anticipate that this efficient sequencing of variable RNA viruses will bring a step change to both basic and translational research.

Methods

Ethics Statement

This study was conducted in compliance with the Data Protection Act (DPA number: Z5886415), and National Health Service research governance. For Norovirus sampling, the Modernising Medical Microbiology study protocol version 1.0 was approved by the Berkshire Research Ethics Committee on the 1st October 2010 (10/H0505/83) and the UK National Information Governance Board (8-05(e)/2010). For HCV samples, ethical approval was approved by the Oxfordshire Research Ethics Committee A on the 15th April 2004, and patients gave written informed consent for these samples to be used prior to sequencing (04.OXA,010).

Sample Collection and Preparation

Faecal samples for Norovirus sequencing. Faecal samples were collected during gastroenteritis outbreaks at the Oxford University Hospitals NHS Trust Hospitals in 2010–2011, as part of the infection service provision of the Trust. Reverse Transcriptase Polymerase Chain Reaction (RT-PCR) [21] was used for initial diagnostic confirmation of Norovirus in all faecal specimens. Viral copy numbers were further determined by quantitative PCR (qPCR) using genogroup specific Taqman probes and primers [22].

Blood samples for HCV sequencing. Blood samples were obtained from subtype-3a HCV infected, treatment naive patients with chronic infection. Plasma was obtained by centrifugation and stored within 5 hours of collection at −80°C. Patients were recruited from the Oxford University Hospitals NHS Trust Hospitals.

Total RNA Isolation

Faecal samples (Norovirus). Total RNA from faecal samples was isolated using the Fujifilm Quickgene DNA tissue kit SII under the manufacturer's RNA extraction from stool protocol for the Fujifilm Quickgene Mini-80 nucleic acid isolation system (Fujifilm Corp., Tokyo, Japan). Three hundred microlitres of supernatant from a 10% clarified emulsion of faeces was used to prepare the lysate. Fifty microlitres of RNA was eluted from the Mini-80 device. The resulting RNA samples were initially quantified by Nanodrop spectrophotometer to estimate concentrations. Samples were stored in −80°C freezers between uses.

Plasma samples (HCV). Plasma was concentrated by high speed centrifugation (23,600× g for 1 h) at 4°C. In order to test whether RNA samples available in archival sources are amenable for high throughput sequencing using a simple workflow, we also used RNA extraction methods currently implemented in many clinical and research laboratories. For HCV, viral RNA was extracted using a QIAmp viral RNA minikit (Qiagen, Hilden, Germany) following the manufacturer's protocol.

Library Preparation for Sequencing

Norovirus amplicon preparation for Sanger sequencing. Reverse transcription and first strand cDNA synthesis were performed using the Accuscript High Fidelity 1st strand cDNA synthesis kit (Agilent, Santa Clara, California, USA) following the manufacturer's published protocol. PCR amplification was then performed using specific in-house primers covering the seven overlapping amplicons of the genome of Norovirus GII.4 (see Table S1), including internal primers to increase depth of coverage using Sanger sequencing. PCR products were then confirmed by agarose gel electrophoresis. The reaction products were separated and detected with a Prism 3730 automated DNA sequencer (Applied Biosystems, Foster City, CA, USA). Sequences were assembled from the resultant chromatograms with the STADEN suite of computer programs [23]. Primer sequences at the 5′ and 3′ termini of the overlapping PCR amplicons were excluded.

Hepatitis C Sanger sequencing. One-step reverse transcription and first-round PCR in two reactions amplified a 4-kb and a 7-kb product. Second-round nested PCR reactions used >20 PCR primers in pairs generating 10 overlapping viral genomic fragments each of approximately 1 kb, that were sequenced and aligned manually to a reference sequence as previously described [24].

Illumina RNA-Seq Library Preparation. Total RNA quantity and integrity were assessed using Quant-IT RiboGreen RNA Assay Kit (Invitrogen, Carlsbad, CA, USA) and Agilent Tapestation 2200 R6K. Libraries for Illumina sequencing were constructed from 100 ng of total RNA using the NEBNext mRNA Sample Prep Kit 1 (New England Biolabs, Ipswich, MA, USA), following the manufacturers' guidelines with minor modifications: end repair in 50 µl reaction volume (40 µl DNA, 5 µl buffer and 5 µl enzyme), post fragmentation clean-up with 1:2.8× volume Agencourt Ampure RNAClean XP (Beckman Coulter, Pasadena, CA, USA); post cDNA synthesis clean-up with 1:1.25× volume Ampure XP Beads; post end repair, A-Tailing and adapter ligation clean-ups with 1:1.8× volume Ampure XP Beads and post-PCR library clean-ups with 1:1× volume Ampure XP Beads. Additionally, upon ligation of Illumina Adapters (Multiplexing Sample Preparation Oligonucleotide Kit) each library was size selected with two Ampure Bead steps (firstly, 1:0.7× volume and secondly, the supernatant from the first bind was taken for a 1:1.7× volume clean-up), selecting 200–600 bp fragments in 30 µl 10 mM Tris-Cl, pH 8.5. Pre-PCR workflow was partially performed using a

Beckman Biomek FX and post-PCR steps were performed using a Beckman Biomek NX^{P} and Biomek 3000.

The following custom primers (25 μM each) were used for the PCR enrichment step:

Multiplex PCR primer 1.0

5′-AATGATACGGCGACCACCGAGATCTA-CACTCTTTCCCTACACGACGCTCTTCCGATCT-3′

Index primer

5′-CAAGCAGAAGACGGCATACGAGAT[INDEX]CAGT-GACTGGAGTTCAGACGTGTGCTCTTCCGATCT-3′

Amplified libraries were analysed for size distribution using the Agilent Tapestation 2200 D1K. Libraries were quantified by quantitative RT-PCR using Agilent qPCR Library Quantification Kit and a Mx3005P instrument (Agilent) and relative volumes were pooled accordingly. Finally, a second quantitative RT-PCR was performed to measure the relative concentration of the pool compared to a previously sequenced mRNA library in order to determine the volume to use for sequencing.

Improved Fragmentation Library Preparation for Illumina. To increase the library insert size of the sequenced libraries and to further investigate secondary structures within the Norovirus genome, a modified fragmentation method was developed. The total RNA was fragmented using 2 μl fragmentation buffer diluted 1:4 in nuclease-free water (fragmentation buffer included in the NEBNext mRNA Library Prep Master Mix Set), incubated at 94°C for 5 minutes, placed immediately to ice and followed by adding 2 μl stop solution (not diluted). Post-fragmentation clean-up through to adapter ligation clean-up were as above. Size selection was omitted to avoid losing longer fragments, and instead, an additional bead clean-up after ligation was performed to remove any remaining adapter dimer prior to PCR amplification.

Illumina Amplicon Library Preparation. Amplicons were quantified and quality assessed using Quant-IT Qubit dsDNA High Sensitivity Assay (Invitrogen) and 1% E-gel (Invitrogen) respectively. Fifty nanograms were sheared to 400 bp using Covaris^{TM} DNA Shearing (Woburn, MA, USA) and concentrated using 1:1× volume Ampure XP Beads. Eluted DNA fragments were then processed through standard library preparation procedures (end repair through to Ampure Bead size selection using the NEBNext DNA Library Prep Master Mix Set for Illumina). Post-reaction clean-ups were 1:1.8× volume Ampure XP Beads and post-PCR libraries were cleaned with 1:1× volume Ampure XP. Ten cycles of PCR amplification were performed using custom primers as above, and the final library was pooled with Norovirus positive samples from the pilot study.

Sequencing. Multiplex libraries were prepared using bar-coded primers and a median insert size of 150 bp (increased to 200 bp with the modified method). The pooled libraries from 3 initial Norovirus and 2 HCV samples were sequenced on an Illumina MiSeq with 150 bp paired end reads following standard Illumina protocols. An average of 0.5 Gb of sequence was produced per sample. A larger pool of 77 Norovirus samples was sequenced on an Illumina HiSeq 2000 with 100 bp paired end reads following standard Illumina protocols.

Data analysis. Reference sequences for each organism were obtained from GenBank. For Norovirus the sequences were GII.4, and the Norovirus Hu/GII.4/Orange/NSW001P/2008/AU was used as a reference (accession number GQ845367). This was selected as the most common match to the Sanger sequenced 288 bp RT-PCR amplicon used in the diagnostic RT-PCR stage. For HCV the reference sequence used was the subtype 3a strain (accession number AF046866).

Sequences were mapped to the organism-specific reference using Stampy v1.0.14 [25] with no BWA pre-mapping. Bases and single nucleotide variants (SNVs) were called using the SAMtools ''mpileup'' command with options '-E -M0 -Q30 -q30 -o40 -e20 -h100 -m2 -D -S' and BCFtools [26]. Sites were filtered to avoid unreliable calls using the following criteria:

a) A minimum depth of 5 reads at each position
b) A minimum average base quality of 10
c) A minimum SNV quality of 25
d) At least 75% of reads at the position support the call and the position was called as homozygous

Positions were called as heterozygous based on the SAMtools genotype calls. An evolutionary tree was created using BEAST (Bayesian evolutionary analysis sampling trees) depicting all the full genomic sequences with relatedness. Analysis was performed using BEAST v.1.7.5 combining two random number seed chains (10 million iterations each, saving 1 in 1,000 iterations, with a 1 million iteration burn-in) using: HKY substitution; estimated frequency; strict clock; and constant population size coalescent tree prior. This maximum clade credibility tree was computed using TreeAnnotator v.1.7.5 and plotted with Figtree v.1.4.0. [27].

The sequences reported in this paper have been deposited in the European Nucleotide Archive Sequence Read Archive under study accession number ERP002219. The Sanger sequences used for validation for Norovirus comparison have been submitted to EMBL under accession numbers HF952119-HF952135. For HCV, the Sanger sequences used for comparison were Genbank accession numbers GQ356201 and KC836883.

Results

1.) Proof of principle experiment using the Illumina MiSeq bench-top sequencer: Norovirus and HCV

We chose two RNA viruses for the initial proof of principle study; three faecal samples RT-PCR positive for Norovirus, and two blood samples RT-PCR positive for HCV. For the libraries produced directly from total RNA, 0.12–1.90% (4,188/3,458,332 reads – 50,138/2,671,058 reads) of the total reads mapped to the reference genomes (Table S2), indicating that a low percentage of the total RNA was from the virus of interest. However, despite the low percentage of reads which mapped, sufficient coverage of the virus genome was obtained to achieve a near complete sequence. The coverage of the sequence assembly varied, being lower at the 5′ and 3′ termini of genomes sequenced from cDNA, and at the ends of each of the 7 overlapping PCR amplicons (Figures 1 and 2). This was expected due to the recognised difficulty of recovering suitable short fragments for Illumina sequencing from the ends of DNA molecules [28][29][30]. The regions of high and low coverage in the RNA samples were consistent across different genomes from the same organism (Figure 2), which suggest that the amount of coverage obtained is due to an intrinsic property of the viral RNA [31].

Single nucleotide variants (SNVs) were called using SAMtools and BCFtools [26]. After filtering to remove unreliable base calls (see Methods), 97.4–99.1% of the genome could be called in the Norovirus samples, and 97.3–98.9% of the genome in the HCV samples. The majority of the positions which could not be called were clustered at the ends of the genome, and were removed during filtering due to the limited depth of coverage (below 5 reads) and quality at each position.

Sanger sequencing for validation. The SNVs called from our Illumina MiSeq sequencing approach were compared with

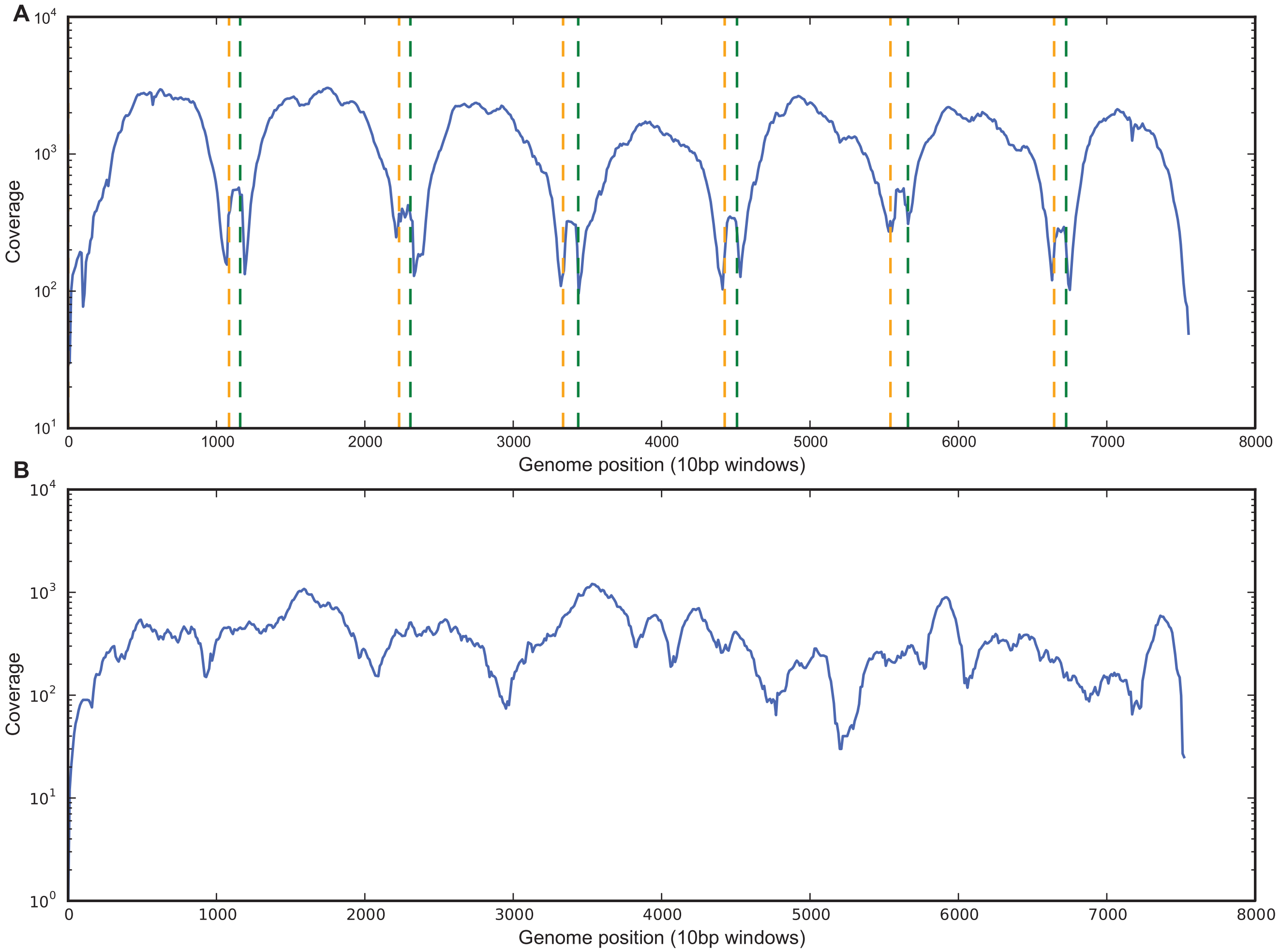

Figure 1. Coverage profiles of one Norovirus sample from amplicon and direct RNA sequencing. A – Coverage across the genome for one Norovirus sample sequenced from PCR amplicons (others similar). Green and orange dotted lined mark the locations of the PCR primers used to generate the amplicons. B – coverage across the genome for the same Norovirus sample sequenced directly from RNA.
doi:10.1371/journal.pone.0066129.g001

those obtained from Sanger sequencing. For sample 1, 99.1% of the genome could be called by both Sanger and Illumina sequencing, and the base calls by both methods were identical. As the Sanger sequencing for Norovirus was based on amplicon PCR amplification, only partial sequence was obtained for sample 3 since one of the PCR fragments could not be amplified. For this sample, 86.4% of the genome could be called using both methods, and the base calls were all identical. Sample 2 failed repeatedly to produce any PCR products and therefore no Sanger sequence was available for comparison. We had specifically included it in our pilot study to determine whether samples that could not be sequenced by Sanger sequencing could be successfully sequenced using the Illumina platform. Thus overall, for Norovirus we found no differences in a total of 14,019 nt that could be directly compared between Sanger and Illumina custom RNA-Seq sequencing methods. One Norovirus sample (Sample 1) was sequenced twice with Illumina technology, once using a library obtained directly from total RNA (see above) and once from 7 PCR amplicons. Sequences obtained using both methods were identical with the genome generated by Sanger sequencing.

Only one near full-length HCV Sanger sequence was available (sample 4). Due to missing bases at the ends of the genome produced by Sanger sequencing, only 8,051 bp (84% of the genome) could be called by both sequencing methods. The sequences differed at 2 positions. For sample 5, only a partial Sanger sequence (1,491 nt) of the *env* gene was available. Of the 1,491 nucleotides which could be called by both methods (15% of the genome), there were 5 SNV differences and a 1 bp insertion in the Sanger sequence relative to the Illumina sequencing. Thus overall, for HCV we found 8 variants in 9,542 nt that could be directly compared between Sanger and Illumina RNA-Seq sequencing methods, 1 variant per 1,193 nt.

Both viruses yielded heterozygous base calls by our SAMtools-based variant calling pipeline, indicating the presence of within-sample genetic variation. The three Norovirus samples showed 2, 3, and 2 distinct positions called as heterozygous, while the two HCV samples had 117 and 131 positions called as heterozygous. Although we did not investigate these positions further, they indicate the potential to directly characterize within-host viral diversity using Illumina sequencing.

2.) Validation using the HiSeq sequencer: Norovirus

To test our method on a larger scale, we sequenced 61 Norovirus patient samples on the HiSeq platform. All samples were positive for Norovirus by RT-PCR prior to sequencing. To determine the reproducibility of our approach, we split 15 larger

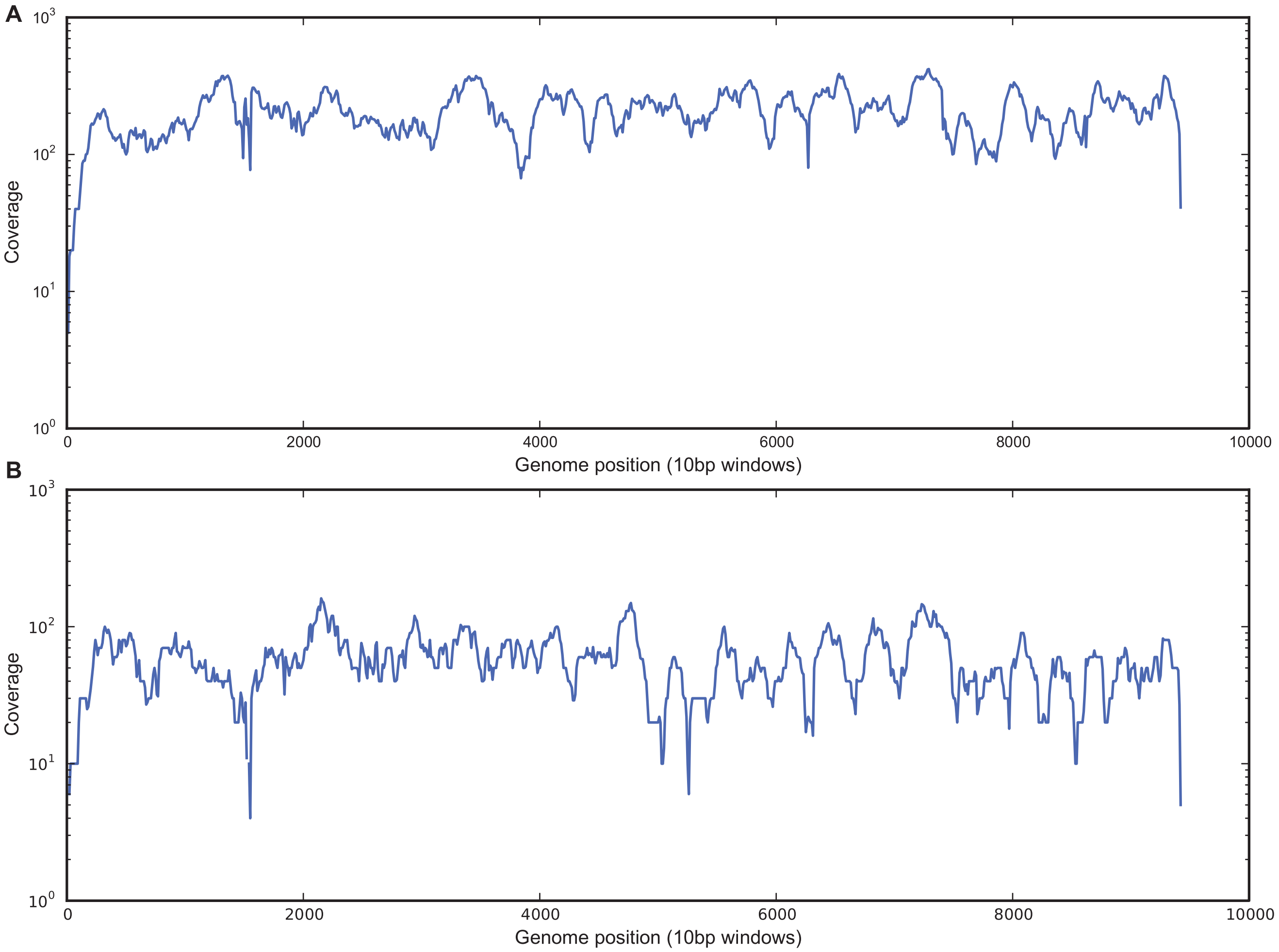

Figure 2. Coverage across the genome for two Hepatitis C samples sequenced directly from RNA.
doi:10.1371/journal.pone.0066129.g002

volume samples into duplicates. Either two libraries were prepared from the same extraction of total RNA (9 pairs), or RNA was extracted from the same faecal sample twice and two separate libraries were prepared (6 pairs). Additionally, one library previously sequenced on the MiSeq platform was also sequenced on the HiSeq (total 77 sequences generated). There was a four to seven month time difference between the extraction and re-extraction from the pairs of samples prepared from the same original source material. Although incidental, this allowed us to ascertain whether the integrity of the faecal material degraded with time. In this total set of 61 patient samples, 16 had full genomes produced by Sanger sequencing, enabling comparison and validation.

Using the HiSeq platform, a mean of 4.6 million reads were produced per sample. The proportion of reads that mapped to the reference genome varied across the samples, ranging from 0.01% to 97.98%; a wider range than seen in the three pilot MiSeq samples (Table S2). Although greater variability would be expected in a larger sample collection, this may also reflect variability in the proportion of the original sample that comprised the viral RNA of interest versus RNA from other sources. We observed a significant correlation between the percentage of reads which mapped to the reference genome and the viral titre estimated by qPCR (rho = 0.4, $p<0.0001$, Figure S1).

Seventy-six out of 77 sequenced preparations produced sufficient reads to allow 90–98.9% of the reference GII.4 genome bases to be called (mean 97.0%). One sample yielded a low number of reads so that only 4,449 (58.9%) of the reference bases could be successfully called (sample 6). As with the pilot MiSeq study, the bases which could not be called were clustered at the ends of the genome with specific regions consistently yielding low coverage (Figure S2). No differences in the called sequence of 15 duplicate pairs were observed. Of the 12 single samples and 4 pairs of samples where Sanger sequencing was available, the base calls were identical across all samples (total 118,757 nt, 148,421 nt including the 4 repeated samples).

We used BEAST [27] to reconstruct the evolutionary tree relating these 61 near-complete Norovirus genome sequences from patient samples (Figure 3). Viruses clustered in space and time, consistent with short-lived outbreaks, although divergent strains could be seen occurring within the same time period during ward outbreaks. Within spatio-temporal clusters, whole genome sequencing revealed low-level genetic variability (up to 13 SNVs). Between clusters, genomes diverged by up to 268 SNVs.

Improved fragmentation analysis. Although the method proved to be robust, it may suffer from the limitation that high variation in coverage across the virus genome could require very large numbers of sequence reads to achieve acceptable sequence completeness (Figure S2). Since this could have arisen from

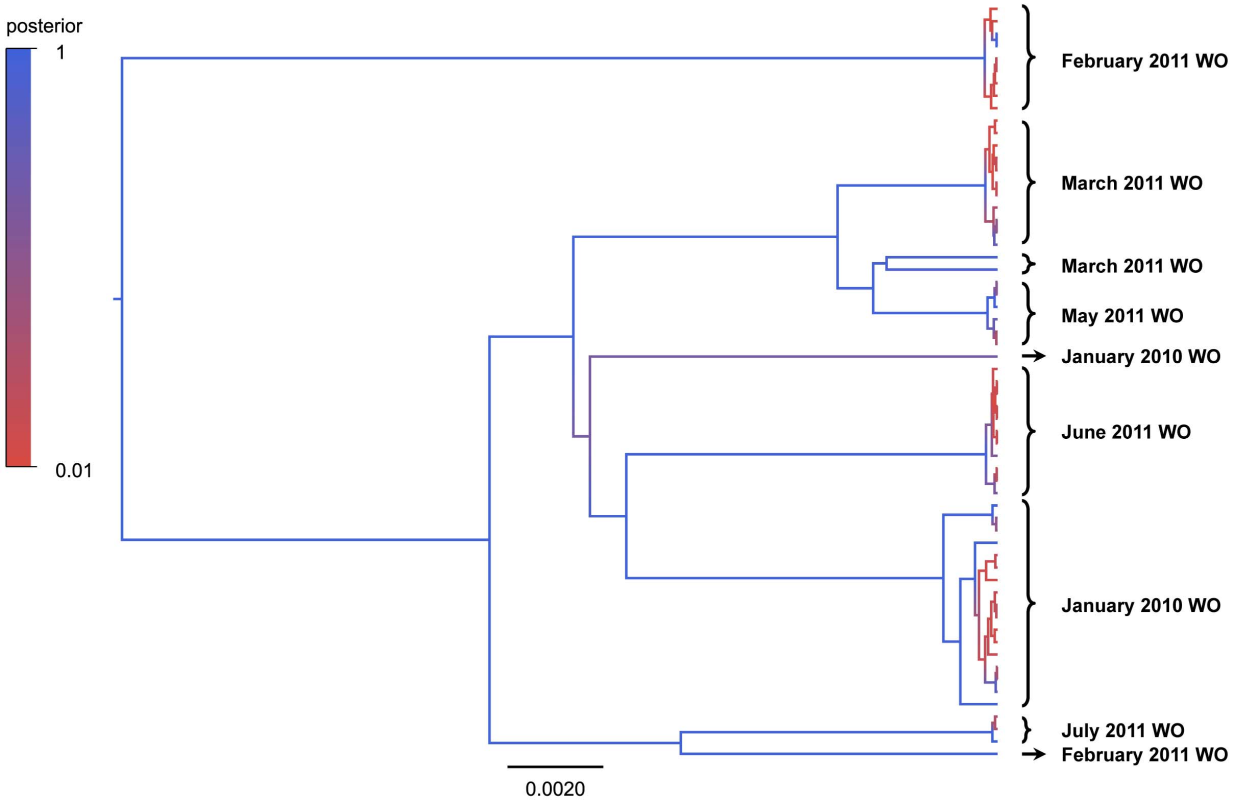

Figure 3. Evolutionary tree created by BEAST (Bayesian evolutionary analysis sampling trees) depicting all the full genomic sequences with relatedness (61 sequences, excluding repeated pairs). Clusters of genomes are visible among viruses sampled at similar points in time. Whole genome sequencing gives adequate resolution to distinguish potential divergent viral strains within the same time, as illustrated in clusters from January 2010, February 2011 and March 2011. WO = ward outbreak. Each node and branch has been coloured depicting the posterior probability supporting that clade calculated by Bayesian analysis (Dark Blue = 1 (high); Light Red = 0 (low)). Analysis was performed using BEAST v.1.7.5 combining two random number seed chains (10 million iterations each, saving 1 in 1000 iterations, with a 1 million iteration burn-in) using: HKY substitution; estimated frequency; strict clock; and constant population size coalescent tree prior. This maximum clade credibility tree was computed using TreeAnnotator v.1.7.5 and plotted with Figtree v.1.4.0.
doi:10.1371/journal.pone.0066129.g003

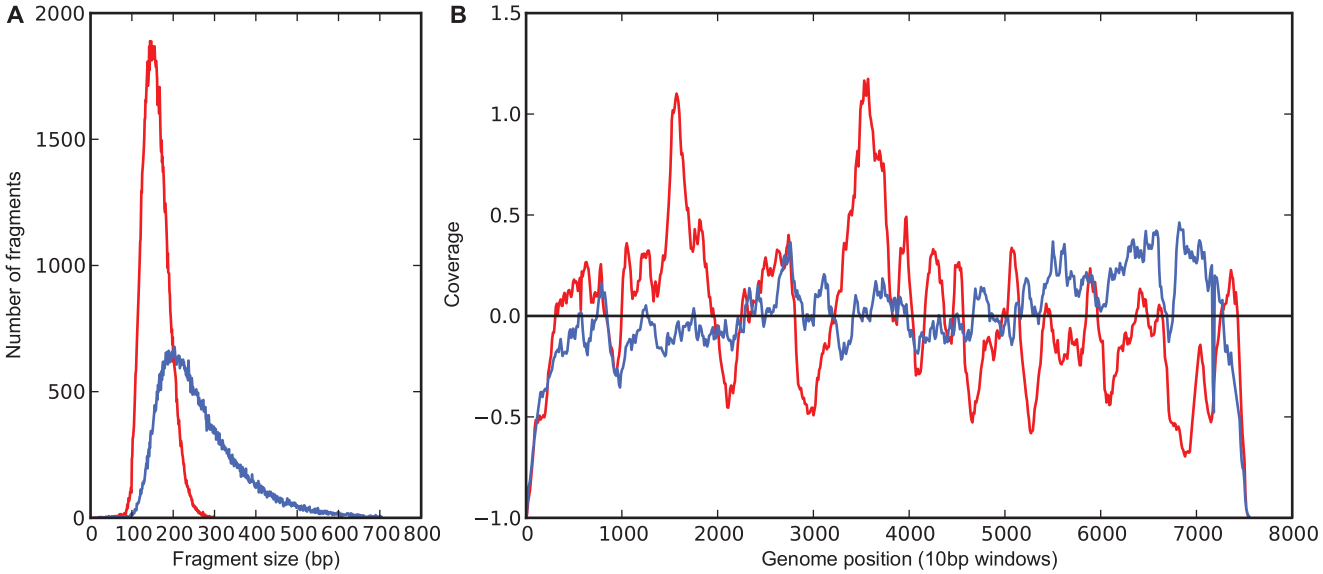

Figure 4. Comparison of the different fragmentation methods. A) fragment size distribution of a library prepared using the standard fragmentation method (red) and a library prepared using the new fragmentation method (blue). B) the coverage across the genome for the standard fragmentation sample (red) and the new fragmentation sample (blue). Data has been scaled as the difference from the median coverage for both samples.
doi:10.1371/journal.pone.0066129.g004

Table 1. A comparison of workflows and consumable costs for various viral sequencing approaches.

Method	Requires sequence specific primers/baits?	Failure Rate in Study (%)	Target Enrichment	mRNA Isolation	Material Fragmented	Size Selection	PCR cycles	Time per sample (days)	Sequencing Throughput (libraries/lane)	Consumable Cost per Sample
RNA-Seq	No	1/82 (1%)	N/A	PolyA Bead Isolation	mRNA to 150 bp	Yes	12	3.5 (same time up to 96 samples)	MiSeq = 6 HiSeq = 96	£39
Custom	No	NA (test sample only)	N/A	N/A	Total RNA to 200 bp	N/A	12	3 (same time up to 96 samples)	MiSeq = 6 HiSeq = 96	£35
Sanger	Yes	**39/55 (71%)**	Amplicon Generation	N/A	N/A	N/A	35	2	6–7	£47
Amplicon-Seq	Yes	71%*	Amplicon Generation	N/A	Amplicons to 400 bp	Yes	45	2	MiSeq = 96	£52
SureSelect Hybridisation	Yes	N/A	Probe Hybridisation	PolyA Bead Isolation	mRNA to 100 bp	N/A	10–14 +12–16	5	MiSeq = 96	£270 probes only**
Nugen Ovation	No	N/A	N/A	N/A	N/A	N/A	12 ***	2.5	HiSeq = 96	£103 Ovation

The 'Custom' protocol refers to the modified RNA-Seq method used in this study to create larger insert fragments. Alternative methods include Amplicon-seq and hybridisation capture (SureSelect Target Enrichment for Illumina Paired-End mRNA-Seq Library Prep; version 1.1). Failure rates are determined by failure to sequence at least one amplicon (<86% of the genome). The failure rate for SureSelect is not given as it was not performed in our study. Consumable costs are list price per sample and exclude sequencing.

*failure rate based on that observed with Sanger Sequencing.

**estimated cost for probes only. Extra cost incurred for Agilent library preparation kit plus additional reagents recommended by Agilent.

***linear amplification during SPIA reverse transcription has not been accounted for.

doi:10.1371/journal.pone.0066129.t001

interference of RNA secondary structures with efficient fragmentation, a modified fragmentation method was adopted that produced libraries with a wider distribution of insert sizes. This modification smoothed out coverage appreciably, improving data in those regions that previously were poorly represented (Figure 4). This improvement offers the prospect of increasing the throughput of samples from a single sequencing run.

Discussion

Here we describe a reproducible next-generation short-read sequencing method based on a modified streamlined RNA-Seq approach for producing near-complete genome sequences of Norovirus and Hepatitis C virus. This method is scalable to batch processing of large numbers of samples with a relatively modest turn-around time of 4 weeks. The method is equally suited to rapid bench-top sequencing with a turn-around of under a week (Figure S3). The current cost of £60 per sample (consumables and sequencing) is similar to that achieved for sequencing a bacterial genome in our facility (Table 1). We have shown that in two distinct clinical sample types there is sufficient virus RNA present (as low as 0.01% or ~4,200 reads mapping to that respective reference) to successfully undertake virus whole-genome sequencing. This is despite the presence of sequences from other sources such as human RNA and, in the case of faeces, food and microbial RNA. This improved method offers the prospect of large-scale affordable studies of RNA viruses and could be adapted to routine diagnostic laboratory workflows. An additional advantage of this method is that no step requiring sequence-specific primers for PCR amplification or bait-based enrichment is needed. This feature limits the possibility of primer-based bias in the processing of samples and provides the opportunity of bioinformatically detecting any other known RNA virus at the same time.

Figure 5 shows a schematic representation of the various methods available for sequencing RNA viruses. We have shown that, for Norovirus, even for samples with a low titer (~300 copies per microliter) it is possible to reconstruct near full genome sequence data (Table S2). However, in cases where the total available RNA is very limited and the viral load known to be ultra-low, other strategies may be implemented. As recently described the NuGEN Ovation RNA-Seq system provides a valid solution to amplify RNA viral genomes while removing rRNA [32]. It should be noted that this strategy adds to the cost of the preparation (Table 1). Moreover, the Single Primer Isothermal Amplification (SPIA) protocol relies on polyA degraded RNA species and non polyA containing species are expected to be lost or to only provide partial sequencing data.

One advantage of systems that can produce long cDNAs is the possibility of using the Illumina's Nextera kit which may greatly reduce the time for library construction. However, this is a speculative thought which has not been experimentally investigated. As such we prefer not to make hypothetical suggestions in the setting of this report.

Our RNA-Seq based method requires only 12 cycles of PCR amplification in total, while all amplicon or hybridisation based methods will include over 30 cycles of PCR (Table 1), increasing the likelihood of biases being introduced through amplification. We encountered a high failure rate with amplicon based methods, while our RNA-Seq based method was less susceptible to failures in reverse transcription, implying that the method may be more robust. The speed and throughput of the Illumina workflow make it a desirable approach for large numbers of samples (Figure S3), being able to prepare up to 96 samples in 3–3.5 days and sequencing them either in batches of 6 samples over 27 hours

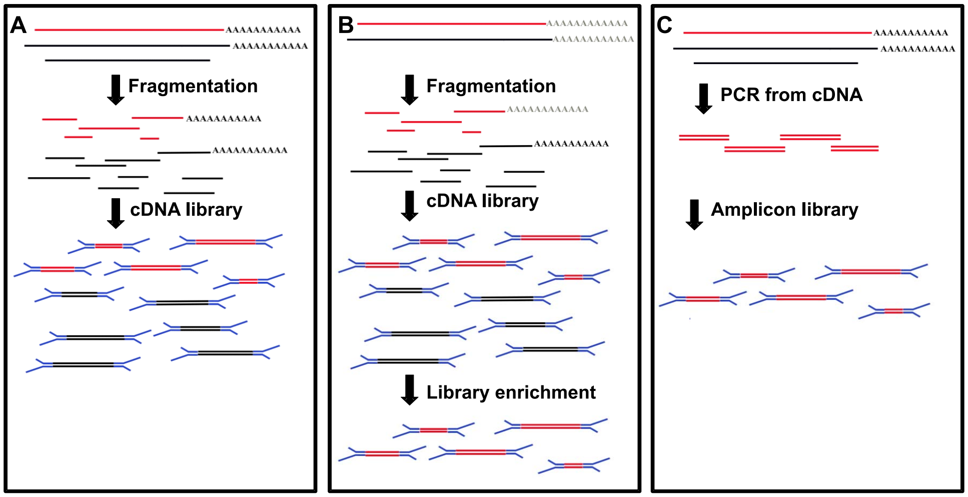

Figure 5. Schematic representation of different strategies for viral genome resequencing. A) Total RNA library: all the RNA species present in the sample are sequenced, no assumption on which genome is present, B) Hybridisation capture of a mRNA library: a good reference genome is needed to design the probes for capture, C) PCR enrichment: the desired genome is amplified from cDNA, a reference genome is needed to design specific oligos. Red lines, genomes of interest; Blue segments, Illumina adapters; Black lines, other RNA species.
doi:10.1371/journal.pone.0066129.g005

using the MiSeq or as a single batch of 96 over 14 days on a HiSeq.

The reproducibility of our sequencing method was demonstrated using multiple replicates of the same clinical samples (15 pairs) by the Illumina sequencing platform characterized by traditional Sanger sequencing. In the case of Norovirus, the sequences from strains associated with multiple temporally linked cases suggests that whole genomic sequence analysis is likely to provide a robust approach to tracking outbreaks. Furthermore, such high resolution data will also give rapid turnaround data on viral genotypes (e.g. HCV or HIV) and recognition of sudden shifts in the predominant capsid variant of Norovirus in general circulation [33][34].

The depth of genome coverage enabled identification of within-sample sequence variation that probably reflects the presence of multiple, closely-related viral variants within the patient, sometimes known as “quasi-species” [35]. Analysis of our data revealed a small number of positions variable within individual Norovirus patients (up to 3 sites in one sample), which is the first evidence of low-level within-patient diversity in this virus. In contrast HCV exhibited a larger number of sites variable within individual patients (up to 131), consistent with previous reports [36]. Although mapping or assembly of short-read sequences obtained by next-generation sequencing has limited scope for reconstructing full-length virus haplotypes, useful insights may nevertheless be gained from the marginal frequency distribution of variants.

The success of our sequencing approach depends on bioinformatics methodology for recognising the viral sequences of interest as a minor component of a heterogeneous population of sequences. The mapping-based approach used here can be regarded as a prototype for characterizing viral genomes, and other strategies using assemblers of viral sequences may be more powerful. These include using a *de novo* population consensus assembly [37][38] which may prove useful particularly when the virus population contains variation in genetic organisation such as large insertions and deletions. Current *de novo* assemblers have difficulty in robustly assembling virus genomes accurately in the presence of within-sample variability, contaminants, and variable sequence coverage [39]. However, new tools are becoming available which may combat these issues [40].

The improvements we describe in generating near whole viral genome sequences over sequences produced by Sanger sequencing [33] or Roche 454 sequencing [41] are likely to be superseded by future improvements in sequencing platforms, such as longer read and simplified sample preparation which requires lower input material [42]. It is expected that complete genome sequences may ultimately be obtained directly from clinical samples using these enhanced sequencing platforms and improved bioinformatics analysis in clinically relevant time-frames (e.g. within hours of receipt in a laboratory). These future methodologies may facilitate the discovery of new viruses once processing of all clinical samples in this way becomes routine practice. Such changes would revolutionise the diagnosis of viral infections and would also promote new avenues of research into virus evolution, antiviral resistance and personalized medicine approaches to treating specific viral genotypes.

Supporting Information

Table S1 Details of primers used for producing the seven overlapping amplicon fragments. Internal primers have been included with the abbreviation INT followed by number. All primers listed were designed in-house. *F = forward; R = reverse.*
(DOCX)

Table S2 Data generated by both Illumina platforms. Samples with * in brackets denote samples that were originally sequenced using the MiSeq. Samples with alphabetical suffices (a and b) were pairs of samples re-sequenced from the same extraction of total RNA on the Illumina HiSeq to confirm reproducibility. Samples with Roman suffices (i and ii) denote pairs of samples where RNA extracted from the same faecal sample twice and two separate libraries prepared.
(XLS)

Figure S1 A comparison of the viral titre of each sample assessed by qPCR, and the percentage of reads obtained from each sample which mapped to the Norovirus reference genome, showing the correlation between the two measures (rho = 0.4, $p<0.0001$).
(TIFF)

Figure S2 Coverage and percentage plots of 76 Norovirus HiSeq sequences. A – percentage of the 76 Norovirus HiSeq samples which could be called at each position in the genome, averaged across 5 bp windows. B – the median, tenth percentile, and ninetieth percentile of the coverage across the genome for the 76 HiSeq samples. Note: 10th percentile is 0 where not shown.
(EPS)

Figure S3 Timelines associated with platforms per optimum batch size per lane. Batch sizes are calculated by a combination of required coverage and the maximum multiplexing capacity available at the time. Sanger sequencing timeline does not take into account the high failure rate (71% in our study). PE = paired ends
(EPS)

Acknowledgments

We thank Alexander Kanapin at the Wellcome Trust Centre for Human Genetics for assistance with data analysis. We acknowledge Bethany Dearlove and David Eyre at Nuffield Department of Medicine, University of Oxford for assistance with evolutionary tree construction and thank Anthony Brown with sequence queries and submissions. We would also like to acknowledge the Oxford University Hospitals NHS Trust Microbiology Laboratory and the Infection Control Department, in particular Lily O'Connor, Lydia Rylance-Knight, Graham Pike, Lisa Butcher, Ruth Moroney, Clare Ward-Jackson, and Dayle Kinch for collecting epidemiological data during ward outbreaks.

Author Contributions

Conceived and designed the experiments: EMB THNW AT DB TEAP DWC PP. Performed the experiments: THNW AT KA MA CM. Analyzed the data: EMB THNW ASW DHW TEAP DWC. Contributed reagents/materials/analysis tools: EMB CLCI TG MC DJW. Wrote the paper: EMB THNW AT DWC PP. Care and collection of clinical data: THNW EB Provided critical review of the manuscript: RB PK EB ASW DHW DJW KED.

References

1. Pybus O, Rambaut A (2009) Evolutionary analysis of the dynamics of viral infectious disease. Nat Rev Genet 10: 540–550.
2. Yu Q, Ryan EM, Allen TM, Birren BW, Henn MR, et al. (2011) PriSM: a primer selection and matching tool for amplification and sequencing of viral genomes. Bioinformatics 27: 266–267.
3. Patel M, Widdowson M, Glass R, Akazawa K, Vinjé J, et al. (2008) Systematic literature review of role of noroviruses in sporadic gastroenteritis. Emerging Inf Dis 14: 1224–1231.
4. Eyre DW, Golubchik T, Gordon NC, Bowden R, Piazza P, et al. (2012) A pilot study of rapid benchtop sequencing of Staphylococcus aureus and Clostridium difficile for outbreak detection and surveillance. BMJ Open: e001124
5. Walker T, Ip C, Harrell R, Evans J, Kapatai G, et al. (2013) Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. Lancet Infect Dis 13: 137–146.
6. Wilson D (2012) Insights from genomics into bacterial pathogen populations. PLoS Pathogens 8: e1002874.
7. Chin C, Sorenson J, Harris JB, Robins WP, Charles RC, et al. (2011) The origin of the Haitian cholera outbreak strain. N Engl J Med 364: 33–42.
8. Chan JZ, Pallen MJ, Oppenheim B, Constantinidou C (2012) Genome sequencing in clinical microbiology. Nat Biotechnol 30: 1068–1071.
9. Daly GM, Bexfield N, Heaney J, Stubbs S, Mayer AP, et al. (2011) A Viral Discovery Methodology for Clinical Biopsy Samples Utilising Massively Parallel Next Generation Sequencing. PLoS ONE 6: e28879.
10. Nakamura S, Yang C, Sakon N, Ueda M, Tougan T, et al. (2009) Direct metagenomic detection of viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach. PLoS ONE 4: e4219.
11. Bull RA, White PA (2011) Mechanisms of GII.4 norovirus evolution. Trends Microbiol 19: 233–240.
12. Lauck M, Alvarado-Mora M, Becker E, Bhattacharya D, Striker R, et al. (2012) Analysis of hepatitis C virus intrahost diversity across the coding region by ultradeep pyrosequencing. J Virol 86: 3952–3960.
13. Zagordi O, Klein R, Däumer M, Beerenwinkel N (2010) Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. Nucleic Acids Res 38: 7400–7409.
14. Astrovskaya I, Tork B, Mangul S, Westbrooks K, Măndoiu I, et al. (2011) Inferring viral quasispecies spectra from 454 pyrosequencing reads. BMC Bioinformatics 12 Suppl 6: 1–10.
15. Prosperi MC, Prosperi L, Bruselles A, Abbate I, Rozera G, et al. (2011) Combinatorial analysis and algorithms for quasispecies reconstruction using next-generation sequencing. BMC Bioinformatics 12: 5.
16. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, et al. (2010) Target-enrichment strategies for next-generation sequencing. Nat Methods 7: 111–118.
17. Depledge DP, Palser AL, Watson SJ, Lai IY, Gray ER, et al. (2011) Specific Capture and Whole-Genome Sequencing of Viruses from Clinical Samples. PLoS ONE 6: e27805.
18. Willerth SM, Pedro HA, Pachter L, Humeau LM, Arkin AP, et al. (2010) Development of a Low Bias Method for Characterizing Viral Populations Using Next Generation Sequencing Technology. PLoS ONE 5: e13564.
19. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10: 57–63.
20. Nagalakshmi U, Waern K, Snyder M (2010) RNA-Seq: a method for comprehensive transcriptome analysis. Curr Protoc Mol Biol Chapter 4: Unit 4.11.1-13.
21. Vennema H, de Bruin E, Koopmans M (2002) Rational optimization of generic primers used for Norwalk-like virus detection by reverse transcriptase polymerase chain reaction. J Clin Virol 25: 233–235.
22. Kageyama T, Kojima S, Shinohara M, Uchida K, Fukushi S, et al. (2003) Broadly Reactive and Highly Sensitive Assay for Norwalk-Like Viruses Based on Real-Time Quantitative Reverse Transcription-PCR. J Clin Microbiol 41: 1548–1557.
23. Staden R (1996) The Staden sequence analysis package. Mol Biotechnol 5: 233–241.
24. Humphreys I, Fleming von, Fabris P, Parker J, Schulenberg B, et al. (2009) Full-Length Characterization of Hepatitis C Virus Subtype 3a Reveals Novel Hypervariable Regions under Positive Selection during Acute Infection. J Virol 83: 11456–11466.
25. Lunter G, Goodson M (2011) Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res 21: 936–939.
26. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079.
27. Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol 29(8): 1969–1973..
28. Hansen KD, Brenner SE, Dudoit S (2010) Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucl Acids Res 38: e131.
29. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, et al. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320: 1344–1349.
30. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5: 621–628.
31. Simmonds P, Karakasiliotis I, Bailey D, Chaudhry Y, Evans DJ, et al. (2008) Bioinformatic and functional analysis of RNA secondary structure elements among different genera of human and animal caliciviruses. Nucleic Acids Res 36: 2530–2546.
32. Malboeuf CM, Yang X, Charlebois P, Henn MR, Levin JZ, et al. (2013) Complete viral RNA genome sequencing of ultra-low copy samples by sequence-independent amplification. Nucleic Acids Res 41(1): e13
33. Dingle KE (2004) Mutation in a Lordsdale Norovirus Epidemic Strain as a Potential Indicator of Transmission Routes. J Clin Microbiol 42: 3950–3957
34. van Beek J, Ambert-Balay K, Botteldoorn N, Eden J, Fonager J, et al. (2013) Indications for worldwide increased norovirus activity associated with emergence of a new variant of genotype II.4, late 2012. Euro Surveill 18: 1–2.
35. Holmes EC, Moya A (2002) Is the quasispecies concept relevant to RNA viruses? J Virol 76: 460–465.
36. Bailey JR, Laskey S, Wasilewski LN, Munshaw S, Fanning LJ, et al. (2012) Constraints on viral evolution during chronic hepatitis C virus infection arising from a common-source exposure. J Virol 86: 12582–12590.
37. Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18: 821–829.
38. Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G (2012) De novo assembly and genotyping of variants using colored de Bruijn graphs. Nat Gen 44: 226–232.
39. Henn MR, Boutwell CL, Charlebois P, Lennon NJ, Power KA, et al. (2012) Whole genome deep sequencing of HIV-1 reveals the impact of early minor

variants upon immune recognition during acute infection. PLoS Path 8: e1002529.

40. Yang X, Charlebois P, Gnerre S, Coole MG, Lennon NJ, et al. (2012) De novo assembly of highly diverse viral populations. BMC Genomics 13: 475.
41. Croville G, Soubies SM, Barbieri J, Klopp C, Mariette J, et al. (2012) Field monitoring of avian influenza viruses: whole-genome sequencing and tracking of neuraminidase evolution using 454 pyrosequencing. J Clin Microbiol 50: 2881–2887.
42. Eisenstein M (2012) Oxford Nanopore announcement sets sequencing sector abuzz. Nat Biotechnol 30: 295–296