METHOD ARTICLE

REVISED Methods for sequencing the pandemic: benefits of rapid or high-throughput processing [version 2; peer review: 2 approved]

Previously titled: Sequencing the pandemic: rapid and high-throughput processing and analysis of COVID-19 clinical samples for 21st century public health

Megan L. Folkerts[1], Darrin Lemmer[1], Ashlyn Pfeiffer[1], Danielle Vasquez[1], Chris French[1], Amber Jones[1], Marjorie Nguyen[1], Brendan Larsen[2], W. Tanner Porter[1], Krystal Sheridan[1], Jolene R. Bowers[1], David M. Engelthaler[1]

[1]Pathogen Genomics Division, Translational Genomics Research Institute, Flagstaff, AZ, 86005, USA
[2]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, 85721, USA

v2 **First published:** 26 Jan 2021, **10**:48
https://doi.org/10.12688/f1000research.28352.1
Latest published: 21 Feb 2022, **10**:48
https://doi.org/10.12688/f1000research.28352.2

Abstract

Genomic epidemiology has proven successful for real-time and retrospective monitoring of small and large-scale outbreaks. Here, we report two genomic sequencing and analysis strategies for rapid-turnaround or high-throughput processing of metagenomic samples. The rapid-turnaround method was designed to provide a quick phylogenetic snapshot of samples at the heart of active outbreaks, and has a total turnaround time of <48 hours from raw sample to analyzed data. The high-throughput method, first reported here for SARS-CoV2, was designed for semi-retrospective data analysis, and is both cost effective and highly scalable. Though these methods were developed and utilized for the SARS-CoV-2 pandemic response in Arizona, U.S, we envision their use for infectious disease epidemiology in the 21st Century.

Keywords

Genomic epidemiology, SARS-CoV2, targeted genomics, sequencing methods, phylogenetics,

This article is included in the Bioinformatics gateway.

Open Peer Review

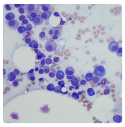
Approval Status ✓ ✓

	1	2
version 2 (revision) 21 Feb 2022	✓ view	✓ view
version 1 26 Jan 2021	? view	✗ view

1. **Bronwyn L. MacInnis**, Broad Institute of Harvard and MIT, Cambridge, USA

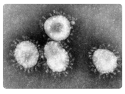
2. **Keith Crandall**, The George Washington University, Washington, USA
The George Washington University, Washington, USA

Any reports and responses or comments on the article can be found at the end of the article.



This article is included in the Cell & Molecular Biology gateway.

This article is included in the Emerging Diseases and Outbreaks gateway.



This article is included in the Coronavirus collection.

Corresponding author: Megan L. Folkerts (mfolkerts@tgen.org)

Author roles: Folkerts ML: Conceptualization, Investigation, Methodology, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Lemmer D**: Data Curation, Formal Analysis, Investigation, Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Pfeiffer A**: Data Curation, Investigation; **Vasquez D**: Investigation; **French C**: Data Curation, Formal Analysis, Software; **Jones A**: Investigation; **Nguyen M**: Investigation; **Larsen B**: Supervision; **Porter WT**: Formal Analysis, Visualization, Writing – Review & Editing; **Sheridan K**: Investigation; **Bowers JR**: Conceptualization, Data Curation, Project Administration, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Engelthaler DM**: Conceptualization, Funding Acquisition, Project Administration, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was funded in part by the NARBHA Institute and the Arizona Department of Health Services.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Folkerts ML, Lemmer D, Pfeiffer A *et al.* **Methods for sequencing the pandemic: benefits of rapid or high-throughput processing [version 2; peer review: 2 approved]** F1000Research 2022, **10**:48
https://doi.org/10.12688/f1000research.28352.2

First published: 26 Jan 2021, **10**:48 https://doi.org/10.12688/f1000research.28352.1

REVISED **Amendments from Version 1**

In this version we have made changes to the text to reflect suggestions and omissions that were pointed out by reviewers. Specific changes include the clarification of the novel contributions of this work, as well as its scope, and statistical analysis of sequencing metrics reported. Additionally, we have highlighted improvements to the protocol that have been made since the time of first submission. In this revision, we have amended the title, and ALL tables (revised versions are included at the end of the attached revised document). Figures have not been changed, but have been re-uploaded for formatting purposes. We have also updated the supplementary data for this manuscript, and a new DOI has been added to the existing reference.

Any further responses from the reviewers can be found at the end of the article

Introduction

With the advent of rapid and inexpensive next-generation sequencing, genomic epidemiology has proven to be an invaluable resource for the elucidation of disease outbreaks. Extending beyond traditional shoe-leather approaches, rapid-turnaround sequencing methods have allowed researchers to quickly gain insight into the genetic nature of pathogens at the heart of active outbreaks[1–6]. By monitoring pathogen evolution over the course of an outbreak, large-scale genomics have the potential to allow for transmission mapping for infection control and prevention[7,8], to distinguish independent cases from those part of active clusters[9], and to identify epidemiological patterns in time and space on both local and global scales[7,10–12].

The most recent example of this has been the collaborative genomic efforts mounted in response to the SARS-CoV-2 outbreak. Not long after the initial cases were identified, whole-genome sequencing quickly established the etiologic agent as a novel coronavirus[13]. Following the rapid spread of SARS-CoV-2, current next-generation technology and analysis pipelines allowed viral sequencing to take place on an unprecedented global scale, with collaborative consortia forming world-wide for the specific purpose of tracking and monitoring the pandemic[14–16].

In most instances, including with the SARS-CoV-2 outbreak, genomic epidemiology has provided a retrospective view of pathogen spread and evolution well after the information is useful in the public health response to the outbreak[1,2]. Genomic epidemiology should guide contemporaneous outbreak control measures, but can only do so if the data are generated and interpreted in real-time quickly enough to inform a response[17]. As technology has advanced, the potential exists to move beyond providing a retrospective genomic snapshot of an outbreak months after its occurrence, to providing actionable data in real-time for current outbreaks within hours after cases are identified[4]. Real-time genomic tracking has already proven valuable in a number of instances, including the recent West-African Ebola outbreak[3,17].

This is not to discredit the value of large-scale, retrospective studies. While rapid-turnaround genomics may prove essential for outbreak containment, retrospective studies will continue to be necessary to track pathogen evolution, gauge success of public health interventions, and to evaluate pathogen/host movement and behavior. With the recent SARS-CoV-2 pandemic, retrospective studies have so-far proven successful in identifying the timing and sources of outbreaks on a local[18] and global scale[5,19], in evaluating the effectiveness of early interventions[5], and in identifying super-spreader events[20]. Thus, in addition to real-time monitoring, high-throughput, cost-effective sequencing and analysis are needed to gain a better understanding of pandemics.

Here, we report two Illumina-based sequencing and analysis strategies for either real-time monitoring or large-scale, high-throughput targeted genomic sequencing of complex samples. The plexWell method is a novel means of library construction that can facilitate high-throughput sample processing. Though these strategies were developed for use with the current SARS-CoV-2 pandemic, we envision their potential use in any situation in which a genomic response is needed.

Methods

Sample information

Remnant nasopharyngeal swab specimens or extracted RNA were obtained from, or received by, the TGen North Laboratory in Flagstaff, AZ. All samples had previously tested positive for SARS-CoV-2 by RT-PCR.

RNA extraction

RNA was extracted using the MagMax Viral Pathogen II kit and a Kingfisher Flex automated liquid handler (ThermoFisher Scientific), with a DNase treatment incorporated to maximize viral RNA recovery from low viral-burden samples, defined as having an RT-PCR cycle-threshold (Ct) value above 33.0. These methods allowed for the rapid, scalable processing of a small number to hundreds of samples at once with minimal personnel, and prevented RNA extraction from becoming a bottleneck to overall throughput (https://doi.org/10.17504/protocols.io.bnkhmct6). Remnant RNA was obtained from the TGen North Laboratory, and had been extracted following their FDA-authorized protocol for the diagnosis of COVID-19.

Targeted amplification

SARS-CoV2 RNA was amplified for both of the sequencing methods described below following the nCoV-2019 sequencing protocol V.1[21] and using the ARTIC v3 primer set[22]. Adapters were added to the resulting amplicons by one of the following means described below. The full sample processing workflow, starting with raw RNA and ending with deliverable data, is illustrated in Figure 1 for reference.

Rapid-turnaround adapter addition and sequencing

For samples requiring immediate attention, e.g. those from patients potentially involved in an active outbreak, adapters were added with the DNA Prep kit (Illumina) as previously

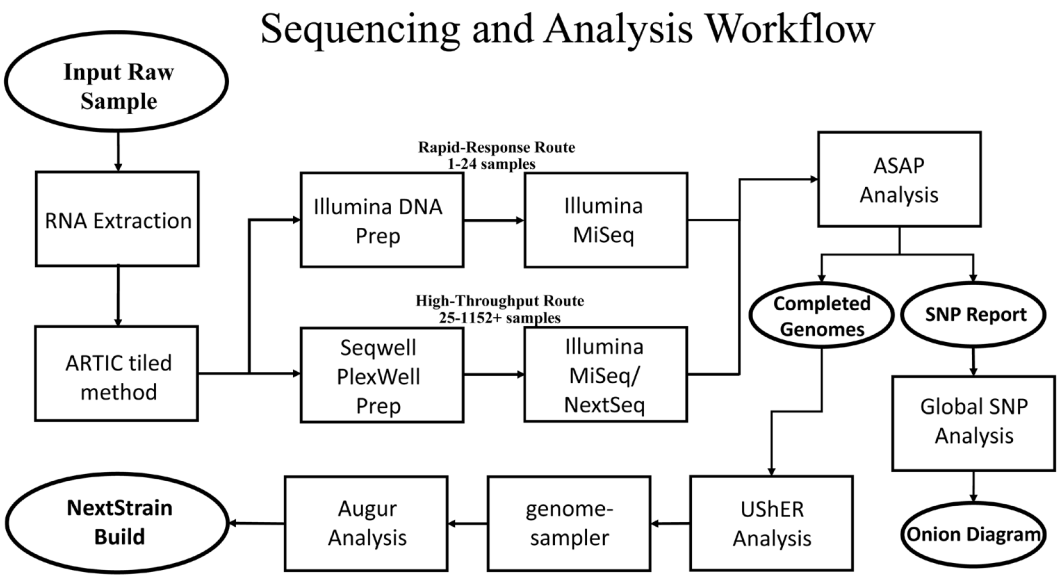


Figure 1. General workflow for the processing of complex samples. Basic workflow for processing of either high-throughput or rapid-response samples.

described[23]. Amplicons were sequenced on the MiSeq platform, using a Nano 500 cycle kit with v2 chemistry (Illumina) (https://doi.org/10.17504/protocols.io.bnnbmdan). The batch size used for this method was 2–8 samples, as that was the typical sample volume requiring rapid turnaround in our facility. However, the maximum number of clusters obtainable from a Nano 500 cycle kit is 1 million, thus, with a theoretical minimum target number of reads of ~42,000, this method would be suitable for up 24 samples per kit. For samples with low Ct values, (less than 33) this was demonstrated to be sufficient to obtain a complete genome, based on results obtained from the high-throughput sequencing method, which was optimized first. For Ct values less than 33, read counts as low as ~19,000 per sample were sufficient to provide a 99% coverage breadth (Supplementary data).

High-throughput adapter addition and sequencing

In instances where retrospective data were needed from large numbers of samples, adapters were added with the plexWell384 kit (Seqwell). Samples were multiplexed in batches of 1,152 and sequenced on a NextSeq 550 with v2 chemistry and 150 X 150 bp reads (Illumina). When batch sizes were not large enough to fill a NextSeq run, samples were sequenced on a MiSeq, with V3 chemistry (Illumina) (https://doi.org/10.17504/protocols.io.bnkimcue). At the time of submission of this manuscript, a maximum of 1,152 barcode combinations were available (this has since increased to 2,304) and thus, the targeted read depth was an order of magnitude higher than for rapid sequencing, at nearly 348,000 reads per sample. The increased number of reads was necessary to account for high Ct value (ie, low viral load) samples, which were not separated into distinct sequencing runs, and which were pooled equally by the plexWell system.

Data processing and analysis

Read data were analyzed using the statistics package "RStudio" (version 4.1.1)[24] The "stats" package was used to perform Kruskal-Wallis, and determine standard deviation[25]. Confidence intervals based on z-scores were calculated using the "BSDA" package[26]. Read data are reported in terms of mean, standard deviation (SD) and 95% confidence interval (95%CI).

Virus genome consensus sequences were built using the Amplicon Sequencing Analysis Pipeline (ASAP)[27,28]. First, reads were adapter-trimmed using bbduk[29], and mapped to the Wuhan-Hu-1 genome[30] with bwa mem[31] using local alignment with soft-clipping. Bam alignment files were then processed to generate the consensus sequence and statistics on the quality of the assembly by the following: 1) Individual basecalls with a quality score below 20 were discarded. 2) Remaining basecalls at each position were tallied. 3) If coverage ≥10X and ≥80% of the read basecalls agreed, a consensus basecall was made. 4) If either of these parameters were not met, an 'N' consensus call was made. 5) Deletions within reads, as called during the alignment, were left out of the assembly, while gaps in coverage (usually the result of a missing amplicon) were denoted by lowercase 'n's. Only consensus genomes covering at least 90% of the reference genome with an average depth of ≥30X were used in subsequent analyses. Consensus genomes generated using these methods include those from Ladner *et al.*[18] and are similar to those in Peng *et al.*[32] and Stoddard *et al.*[33]. Statistics reported for each sample included: total reads, number

of reads aligned to reference, percent of reads aligned to reference, coverage breadth, average depth, and any SNPs and INDELs found in ≥10% of the reads at that position.

Phylogenetic inference

Rapid phylogenetic inference was used when it was critical to rapidly answer two initial genomic epidemiological questions: i) are the samples in this set closely related, and ii) to what samples in the public database are these most closely related? To answer the former, SNP comparisons were made among a sample set of interest and output in text format directly from the ASAP results, foregoing the computational time of generating phylogenetic trees.

To answer the latter, a database of all SNPs in the GISAID global collection[15] was generated and periodically updated by downloading all the genomes and filtering them for quality and completeness, then aligning to the WuHan-Hu-1 reference to identify any variants, as described[34]. The list of all variants and metadata for each sample were then put into a relational database for fast querying. The format of this database was one table of SNPs, as generated by Chan *et al.*[34] with columns for GISAID ID, position, reference base, and variant base, and a second table of metadata which contained the exact data fields, as downloaded from GISAID. The two tables were linked by the GISAID ID number. The GISAID database cannot be shared via secondary publications as per the database access agreement, however the database is available for download by the public from GISAID's website[15]. The list of SNPs common to samples in a given set was then compared to this global SNP database by first querying the number of global genomes containing each of the individual SNPs, sorting them from most common to least common, then further querying for number of global genomes containing combinations of SNPs by adding in each successive SNP in turn, to determine how many GISAID genomes shared all or a subset of the SNPs from the set. A stacked Venn diagram to illustrate globally-shared SNPs was constructed using the statistical package "R" (Version 4.0.2)[25] and the "ggplot2"[35] and "ggforce"[36] R-packages. This information was then relayed to public health partners, who used these preliminary analyses to guide contact tracing efforts and track the spread of the virus in real-time.

For subsequent, more sophisticated phylogenetic analyses, phylogenetic trees were constructed in NextStrain[16], with genomes from GISAID subsampled by uploading our genomes of interest to the UCSC UShER (UShER: Ultrafast Sample placement on Existing tRee) tool[37], identifying relevant genomes (those most related to our genomes of interest), and further reducing that set of genomes when necessary with genome-sampler[18,38].

Consent

Samples used were remnant, de-identified samples from a clinical diagnostics lab. No ethical approval was required for their inclusion in this study.

Results

High-throughput workflow

For the plexWell workflow, turnaround time from raw sample to sequence data was approximately 72 hours, and to phylogenetic inference approximately 10 hours, for 1,152 samples (Table 1). Cost per sample was ~$60, which is comparable to similar sample prep methods (Table 1)[39]. A complete breakdown of run statistics by Ct value can be found in Table 2. Success rates, defined as the percentage of samples with greater than 90% genome coverage, were similar to other previously described methods[39,40]. Of a random subset of 897 samples analyzed, approximately 83% of samples with a Ct <33 yielded genomes with ≥90% breadth of coverage of the reference genome, i.e., a complete genome, with an average breadth of coverage of the reference genome of 91% (SD 20.10, 95%CI 89.79-92.81). (Figure 2, Table 2). Failure to obtain a complete genome when Ct was <33 was attributed to sample preparation error, sample degradation, or poor sequencing run metrics. Beyond a Ct of 33, success rate dropped drastically. Between Ct values of 33 and 35, complete genome success rate was ~41% (Table 2), with an average breadth of coverage of 78% (SD 23.00, 95%CI 72.50-83.20). Above a Ct of 35, ~18% of samples yielded a complete genome, and breadth of coverage dropped below 57% (SD 31.65, 95%CI 48.70-59.08) (Table 2).

Uniform depth of genome coverage across samples was targeted when pooling samples for sequencing, but depth generally decreased as Ct increased. Average depth of coverage was approximately 2850X up to a Ct of 33 (SD 1712.91, 95%CI 2721.57-2978.49). Past 33, coverage dropped off sharply, with

Table 1. Workflow metrics for two separate sample processing systems.

* Listed is the number of samples that can be processed within the specified turnaround time, on a single sequencing run. This is not necessarily the upper limit of either processing system.
**A detailed breakdown of reagent costs is available in supplementary materials.
*** Time from raw sample to analyzed data.

Sequencing method	Samples able to be processed*	Cost per Sample**	Personnel Needed	Turnaround time***
Illumina DNA Prep	24	$60.30	1	48hrs
Seqwell plexWell	1152	$98.13	4–6	82hrs

Table 2. Average sequencing metrics for various RT-PCR cycle threshold values of samples sequenced using the plexWell 384 system, and either an Illumina MiSeq or an Illumina NextSeq550.

*Ct value reported is that for the nucleocapsid-2 gene.
**Uniform depth of coverage was targeted for all samples.

Average Cycle Threshold Value (n)*	Average % aligned	Average % coverage	Average depth**	% samples > 90% coverage
<20 (194)	96.64 (SD 12.27, 95%CI 94.91-98.37)	92.36 (SD 21.34 95%CI 89.35-95.39)	3371.58 (SD 2041.06 95%CI 3084.37-3658.80)	87.63
20–25 (210)	96.91 (SD 11.34 95%CI 95.38-98.44)	92.61 (SD 17.60 (95%CI 90.22-94.99)	2848.11 (SD 1523.21 95%CI 2642.09-3054.12)	85.24
25–30 (169)	95.67 (SD 12.81 95%CI 93.74-97.60)	91.11 (SD 20.12 95%CI 88.08-94.14)	2733.37 (SD 1517.23 95%CI 2504.62-2962.11)	84.62
30–33 (110)	88.58 (SD 18.99 95%CI 85.03-92.13)	87.18 (SD 21.89 95%CI 83.09-91.27)	2113.09 (SD 1389.81 95%CI 1853.37-2372.81)	74.55
33–35 (71)	74.68 (SD 25.36 95%CI 68.78-80.58)	77.85 (SD 23.00 95%CI 72.50-83.20)	1221.05 (SD 1102.35 95%CI 964.64-1477.46)	40.85
35–37 (52)	60.20 SD (32.55 95%CI 51.36-69.05)	56.91 (SD 30.39 95%CI 48.65-65.17	716.43 (SD 1140.89 95%CI 406.33-1026.42)	17.31
37+ (91)	38.28 (SD 34.42 95%CI 31.21-45.35)	52.17 (SD 32.38 95%CI 45.54-58.80)	576.09 (SD 1286.98 95%CI 311.67-840.51	18.68

an average depth of 1221X between Ct values of 33-35 (SD 1102.35, 95%CI 964.64-1477.46). Above a Ct of 35, depth dropped to 627X (SD 1233.64, 95%CI 424.93-829.32).(Table 2).

The number of reads targeted per sample for the high-throughput method was approximately 348,000, considering the limitation of the number of index combinations (i.e., samples per run) that were available at the time. In practice, reads per sample varied considerably. The average read count for samples using the plexWell method was 411,375, with 393,925 being the average number of reads that aligned to the WuHan-Hu-1 reference strain. Below a Ct of 33, average aligned reads was 475,310 (SD 260292, 95%CI 455789-494831) per sample. Between Cts of 33 and 35, the average reads aligned was 204,234 (SD 175947, 95%CI 163308-245160) per sample. Above 35, read counts dropped off sharply, with a mean of 99,388 reads aligning (SD 188499, 95%CI 68493-130283) per sample. Note that all samples were put through an equimolar pooling step by the plexWell system. This suggests this pooling is not entirely effective when Cts vary considerably in a given sample set. This was confirmed through Kruskal Wallis on the total (unaligned) read counts in each of the above-mentioned groups ($p < 0.001$).

Percent of total reads aligning to the SARS-CoV-2 reference genome decreased as the Ct value increased (Table 2, Figure 2). Up to a Ct value of 33, most (~95%) reads mapped to SARS-CoV-2 (SD 13.74, 95%CI 94.15-96.22). From Ct values of 33 to 35, this noticeably decreased, with a mean of 75% alignment (SD 25.36, 95%CI 68.78-80.58) Beyond a Ct value of 35, average percentage of reads aligning dropped to 46% (SD 35.26, 95%CI 40.47-52.03)

Rapid-Turnaround Workflow

Turnaround time for the Illumina DNA prep method was significantly faster than the high-throughput plexWell system. It took less than 48 hours to go from raw sample to deliverable, analyzed data (Table 1), and this time could potentially be reduced further by reducing cycle numbers per sequencing run. Illumina estimates a cycle time of ~2.5 minutes for the kit type described, meaning a reduction of just 50 cycles could reduce the overall run time by >2 hours[41].

Both the plexWell and DNA Prep methods use a tagmentation system for adapter addition, rather than a ligation-based approach. This results in adapters being added ~50 bp or more from the end of overlapping amplicons generated during gene-specific PCR. A second PCR step amplifies only the tagmented regions, resulting in final libraries of ~300bp. These approaches negate the need for primer trimming prior to alignment.

Generating consensus genomes and a SNP report from the sequence data, which takes approximately 15 minutes for a small (<=24 samples) dataset, quickly shows whether the samples are part of the same outbreak. To quickly find a potential origin of a cluster or sample (assuming genomes in the public domain were collected prior to the sample(s) of interest), a global SNP database can be used. Constructing or updating a global genome SNP database takes several hours, but it can be done prior to

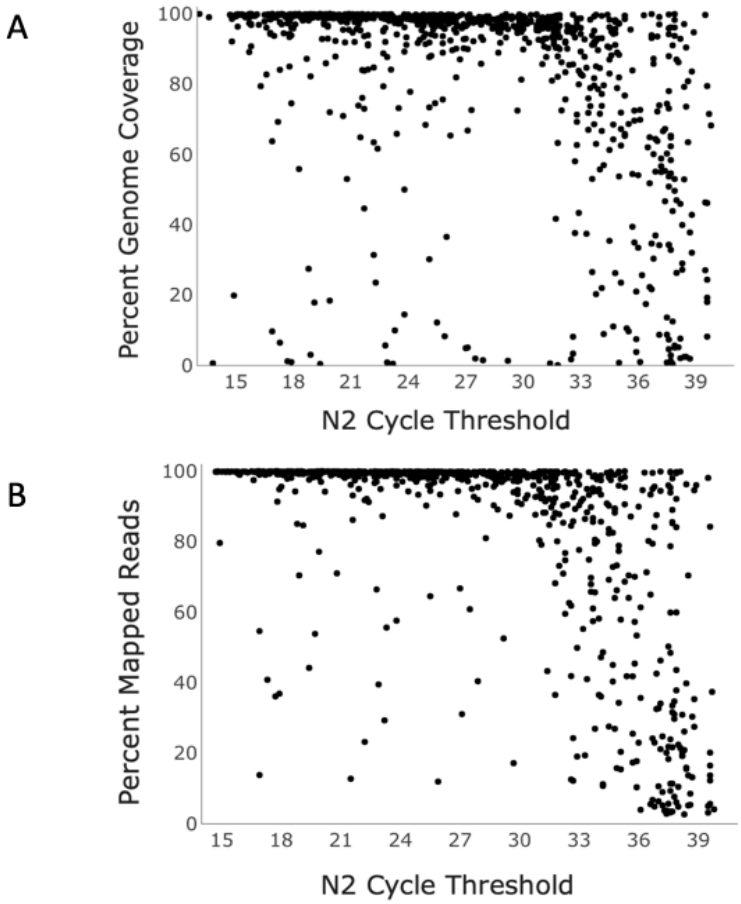


Figure 2. Sequencing outcomes of SARS-CoV2-positive samples processed using Seqwell's plexWell method. A. Nucleocapsid-2 Ct value vs percent genome coverage and **B**. percent total reads mapped to SARS-CoV-2 for 897 SARS-CoV-2-positive samples sequenced using Seqwell's plexWell 384 system.

sample sequencing (and regularly). Running the commands to query the global SNP database for particular SNPs of interest takes mere seconds.

To rapidly visualize results of the global SNP database query, a stacked Venn diagram (aka, "onion diagram") visually describes the hierarchical nature, i.e., the parsimony, of SNPs found in the SARS-CoV-2 genomes (Figure 3), and is easily generated using the methods described above or an alternative tool.

Although fewer data are available from the Illumina DNA prep method, as it was primarily used to process five or fewer samples at a time, data suggest that it performed slightly worse than the plexWell system. Average percent coverage for samples with Ct values less than 25 was 94.53% (SD 3.74, 95%CI 91.54-97.52). At a Ct between 25 and 30, the average coverage was 87.6% (SD 5.62, 95%CI 83.71-91.49). Odds of obtaining a complete genome dropped to 0 at Ct values greater than 30, with an average percent coverage of 62.83 (SD 20.22, 95%CI 39.95-85.71) (Table 3, Supplementary Data).

A minimum of 42,000 reads were targeted for the Illumina DNA prep method, but as sample sizes were small for the outbreak in question, the average aligned reads per sample was much higher. For samples with Ct values below 25, the average number of reads aligned to the reference was 275,258 (SD 136,816, 95%CI 165,784-384732). Between Ct values of 25 and 30, average reads aligned was 200,810 (SD 86,999, 95% CI 140,523-261096). Below a Ct value of 30, average aligned reads dropped to 116,888 (SD 275257, 95%CI 53,002-180,773). Kruskal-Wallis revealed no significant difference among the average aligned reads between these groups (P= 0.08).

Discussion

The recent SARS-CoV-2 outbreak has highlighted the need for real-time sequencing and data analysis capacity in the face of active pandemics, as well as high-throughput sequencing and analysis strategies for comprehensive retrospective analysis and evaluation. We describe two different strategies that can be used in combination with Illumina platforms for the rapid or high-throughput interrogation of samples involved in disease

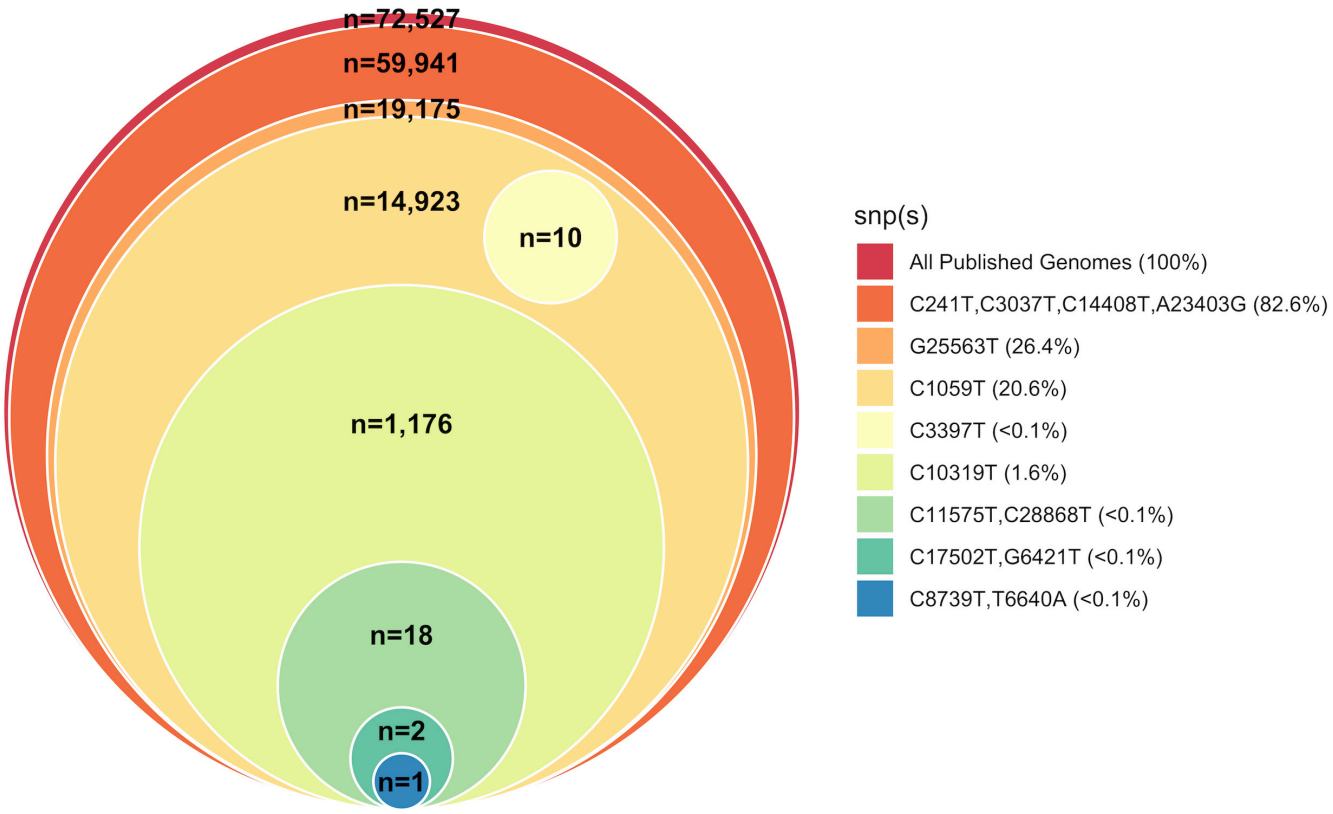


Figure 3. Phylogenetic relationship of SARS-CoV2-positive samples to Wuhan-Hu-1 reference strain. Stacked Venn "onion" diagram indicating the hierarchical nature of cluster-specific SNPs relative to the reference strain in global collection of SARS-CoV2 samples.

Table 3. Average sequencing metrics for various RT-PCR cycle threshold values of samples sequenced using the Illumina DNA Prep system and an Illumina MiSeq.

*Ct value reported is that for the nucleocapsid-2 gene.
**Uniform depth of coverage was targeted for all samples.

Average Cycle Threshold Value (n)*	Average % aligned	Average % coverage	Average depth**	% samples > 90% coverage
<25 (6)	96.56 (SD 2.38 95%CI 94.65-98.46)	94.53 (SD 3.74 95%CI 91.54-97.52)	1503.00 (SD 743.91, 95%CI 908.35-2098.83)	83.33
25–30 (8)	91.92 (SD 19.66 95%CI 68.29-95.53)	87.60 (SD 5.62 95%CI 83.71-91.49)	1115.25 (SD 470.39 95%CI 342.46-1888.04)	50%
30–37 (3)	56.46 (SD 37.84, 95%CI 13.65-99.28)	62.83 (SD 20.22, 95%CI 39.95-85.71)	851.21(SD 443.07 95%CI 349.84-1352.59)	0%

outbreaks. Though the results reported here are specific to the SARS-CoV-2 outbreak, these methods could conceivably be modified and applied to many other situations in which a genomic epidemiological response is needed.

The Illumina DNA prep-SNP-comparison analysis method is effective in providing rapid sequence data and genomic epidemiology information for small numbers of samples (<48 hours from raw sample to rough phylogenetic placement). Though throughput is limited by both the number of available indices for multiplexing and the nature of the protocol itself, this method has the advantage of being scalable to small numbers of samples, and can be performed in the course of several hours for small sample subsets.

The Seqwell plexWell system provides a scalable, cost-effective, and high-throughput method of processing thousands of samples with minimal laboratory personnel (Table 1). Here, we report the first use of plexWell for post-amplification adapter addition for whole genome sequencing of SARS-CoV2. As

the plexWell protocol calls for the pooling of samples at an initial step in the adapter tagmentation process, hundreds of samples were able to be taken through the later steps of the protocol by a single individual in an 8-hour timeframe. This, coupled with the availability of thousands of index combinations, allowed for a high-throughput, cost-effective means of processing large numbers of samples on a weekly basis with minimal laboratory personnel and infrastructure. With just two full-time staff members devoted to sequencing, and 2–4 part-time staff responsible for arraying samples prior to the ARTIC/plexWell protocol, 1152 samples were able to be successfully processed and sequenced each week. Thus, the number of genomes completed per full-time (40 hour) employee per week is 384, a novel pace for what is normally a lengthy protocol.

Typical analyses of virus genomes include phylogenetic tree construction to understand transmission patterns. Nextstrain[16] and GISAID[15] have been crucial to the SARS-CoV-2 scientific community for global and local epidemiologic understanding, and tools for smart subsampling (e.g. genome-sampler[38]) are now necessary with the growth of the public databases. However, reconstructing phylogenies, especially paired with finding relevant subsets, takes time, and is often overkill for initial, time-sensitive public health needs. We employ a simple, rapid analysis method and visualization meant as a quick-look to determine relatedness among a sample set of interest and/or relatedness of a sample or set to the entire public database of genomes. Because of the novelty of SARS-CoV-2 and its low rate of recombination, merely comparing the low number of SNPs (at the time of first publication) across samples without applying a phylogenetic model or program is often enough to answer initial questions about COVID-19 transmission, e.g. whether samples in a given set are closely related, and/or which samples in the global database are most closely related to a given sample set. This information, then, can be used by public health officials to determine which patients might be involved in active spread of disease (requiring additional shoe-leather epidemiology to follow up) and which patients are thought to represent isolated cases without additional spread. This is not meant to replace more thorough phylogenetic analysis, but rather, to provide a quick genomic snapshot that can inform a public health ground response. Our SNP queries followed by generation of a stacked Venn diagram (onion diagram) offer a much faster alternative or antecedent to complete phylogenetic analysis, and have served as a starting point for contact tracing in active outbreaks. The data presented here using the rapid-turnaround method were obtained from coded, de-identified patient samples that were part of an active SARS-CoV2 outbreak, and the resulting phylogenetic placements were successfully used to inform public health efforts to determine the potential origin of the outbreak, and to prevent further spread[42]. Other pathogens or situations where SNP numbers are expected to be very low may also benefit from these rapid analysis methods.

For retrospective studies, time can allow for more robust phylogenetic analyses including smart subsamplers[38], such as NextStrain[16] and other commonly used phylogenetic tools; however, their employment can significantly add time to a rapid response. The UShER tool[37], which can rapidly place genomes onto an existing SARS-CoV-2 phylogenetic tree, can greatly speed-up the final analysis. Parsing the output of UShER generates a subset of public genomes that are phylogenetically close to the samples of interest. This reduces the input dataset to subsamplers such as genome-sampler[38], which significantly reduces the computation time for further subsampling based on geography and time, which in turn significantly reduces the computation time for a NextStrain analysis.

Each of the two methods have their limitations. We observed a reduced success rate of the rapid Illumina DNA Prep method over the high-throughput plexWell system, as evidenced by both decreased overall breadth of coverage and decreased success of obtaining complete genomes at Ct values above 30. It should be noted, however, that rapid prep study was conducted on a limited sample set, using samples from active outbreak clusters that were shipped from long distances through varying ambient temperatures. Samples used to evaluate the plexWell system were obtained locally and processed entirely in-house. This difference in handling, coupled with the sample size difference, may in part account for the differences in results in the two prep methods. Also, at ~$100/sample in reagent costs, the rapid Illumina DNA Prep method is less cost-effective (Table 1, Supplementary Data). And though the turnaround for the protocol is ~1.5 days, faster sequencing is achievable through other methods, such as through the use of long-read Nanopore sequencing[17,39]. Nanopore technology, however, has the disadvantage of having a higher per-base error rate when compared to short-read sequencing methods[17], thus its lack of accuracy may outweigh any potential time savings, particularly in situations where relatively few nucleotide variants can radically alter phylogenetic placement such as for SARS-CoV-2. Also, the SNP-comparison analysis is rapid and robust because of the relatively low numbers of SNPs so far documented in the SARS-CoV-2 genome (at the time of submission of this manuscript) due to the virus's novelty, the lack of recombination, and the unmatched robustness of the global SARS-CoV-2 genome database. Though this method could be applied to other types of outbreaks, rapid, precise phylogenetic placement will rely on these same factors.

The high-throughput plexWell prep system also has its drawbacks. The turnaround time of this method, when large sample numbers are processed, is not competitive[39]. The majority of processing time is lost to the ARTIC portion of the protocol, however, and not specifically to the plexWell adapter addition. And though the combinatorial index system allows for thousands of samples to be multiplexed in a single sequencing run, the SARS pandemic has demonstrated that even this may not be sufficient to meet the data challenges presented by expansive disease outbreaks.

This paper highlights the potential for rapid turnout of genomic data from epidemiological samples, giving genomics the potential to become invaluable in pandemic response. It is worth noting, however, that sequencing isn't the only potential bottleneck in the process. For data to be delivered in real-time, coordination between testing sites, sample delivery, and sequencing laboratories must be robust, so that preparation for

sequencing can begin shortly after a sample has tested positive for the pathogen of interest. Failures anywhere in this chain can lead to significant delays, which negates the utility of a real-time sequencing pipeline for investigation. Thus, in addition to improvements in the described sequencing and analysis pipeline, coordinating the efforts of testing sites, public health partners, and sequencing laboratories is critical for this technology to be most effective.

Despite overwhelming benefits of employing next generation technological advances in real-time during a public health emergency, challenges remain when using genomic epidemiology as a means of pandemic control and monitoring. It has been demonstrated that genomics are not, in and of themselves, sufficient to completely elucidate the mechanisms and transmission of all pathogen-transmitted disease, particularly when asymptomatic and mild infections are known to play a role in transmission[43] but are less likely to be identified and subsequently sequenced. Thus, the integration of genomic and traditional epidemiology is paramount to the success of this 21st century public health capability.

Data availability

Underlying data

Zenodo: Supplementary Data for "Sequencing the Pandemic: Rapid and High-Throughput Processing and Analysis of COVID-19 Clinical Samples for 21st Century Public Health" DOI: 10.5281/zenodo.5822962[44].

This project contains the following underlying data:

- Supplementary_data_final.xlsx – includes all data used in generation of figures and tables, as well as cost breakdown for two sequencing methods.

Acknowledgements

The authors would like to thank Hayley Yaglom for her helpful review of the first draft and the members of the Arizona COVID Genomics Union (ACGU), for their insight and support of this work.

References

1. Gardy JL, Naus M, Amlani A, *et al.*: **Whole-Genome Sequencing of Measles Virus Genotypes H1 and D8 During Outbreaks of Infection Following the 2010 Olympic Winter Games Reveals Viral Transmission Routes.** *J Infect Dis.* 2015; **212**(10): 1574–8.
PubMed Abstract | Publisher Full Text

2. Coll F, Harrison EM, Toleman MS, *et al.*: **Longitudinal genomic surveillance of MRSA in the UK reveals transmission patterns in hospitals and the community.** *Sci Transl Med.* 2017; **9**(413): eaak9745.
PubMed Abstract | Publisher Full Text | Free Full Text

3. Arias A, Watson SJ, Asogun D, *et al.*: **Rapid outbreak sequencing of Ebola virus in Sierra Leone identifies transmission chains linked to sporadic cases.** *Virus Evol.* 2016; **2**(1): vew016.
PubMed Abstract | Publisher Full Text | Free Full Text

4. Gardy J, Loman NJ, Rambaut A: **Real-time digital pathogen surveillance - the time is now.** *Genome Biol.* 2015; **16**(1): 155.
PubMed Abstract | Publisher Full Text | Free Full Text

5. Worobey M, Pekar J, Larsen BB, *et al.*: **The emergence of SARS-CoV-2 in Europe and North America.** *Science.* 2020; **370**(6516): 564–570.
PubMed Abstract | Publisher Full Text | Free Full Text

6. Engelthaler DM, Valentine M, Bowers J, *et al.*: **Hypervirulent *emm*59 Clone in Invasive Group A *Streptococcus* Outbreak, Southwestern United States.** *Emerg Infect Dis.* 2016; **22**(4): 734–8.
PubMed Abstract | Publisher Full Text | Free Full Text

7. Rockett RJ, Arnott A, Lam C, *et al.*: **Revealing COVID-19 transmission in Australia by SARS-CoV-2 genome sequencing and agent-based modeling.** *Nat Med.* 2020; **26**(9): 1398–1404.
PubMed Abstract | Publisher Full Text

8. Oude Munnink BB, Nieuwenhuijse DF, Stein M, *et al.*: **Rapid SARS-CoV-2 whole-genome sequencing and analysis for informed public health decision-making in the Netherlands.** *Nat Med.* 2020; **26**(9): 1405–1410.
PubMed Abstract | Publisher Full Text

9. Scarpino SV, Iamarino A, Wells C, *et al.*: **Epidemiological and viral genomic sequence analysis of the 2014 ebola outbreak reveals clustered transmission.** *Clin Infect Dis.* 2015; **60**(7): 1079–82.
PubMed Abstract | Publisher Full Text | Free Full Text

10. Dudas G, Carvalho LM, Bedford T, *et al.*: **Virus genomes reveal factors that spread and sustained the Ebola epidemic.** *Nature.* 2017; **544**(7650): 309–315.
PubMed Abstract | Publisher Full Text | Free Full Text

11. Bowers JR, Kitchel B, Driebe EM, *et al.*: **Genomic Analysis of the Emergence and Rapid Global Dissemination of the Clonal Group 258 *Klebsiella pneumoniae* Pandemic.** *PLoS One.* 2015; **10**(7): e0133727.
PubMed Abstract | Publisher Full Text | Free Full Text

12. Oltean HN, Etienne KA, Roe CC, *et al.*: **Utility of Whole-Genome Sequencing to Ascertain Locally Acquired Cases of Coccidioidomycosis, Washington, USA.** *Emerg Infect Dis.* 2019; **25**(3): 501–506.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Zhou P, Yang XL, Wang XG, *et al.*: **A pneumonia outbreak associated with a new coronavirus of probable bat origin.** *Nature.* 2020; **579**(7798): 270–273.
PubMed Abstract | Publisher Full Text | Free Full Text

14. Prevention CfDCa: **SARS-CoV-2 Sequencing for Public Health Emergency Response, Epidemiology, and Surveillance (SPHERES)**. 2020.
Reference Source

15. Elbe S, Buckland-Merrett G: **Data, disease and diplomacy: GISAID's innovative contribution to global health.** *Glob Chall.* 2017; **1**(1): 33–46.
PubMed Abstract | Publisher Full Text | Free Full Text

16. Hadfield J, Megill C, Bell SM, *et al.*: **Nextstrain: real-time tracking of pathogen evolution.** *Bioinformatics.* 2018; **34**(23): 4121–4123.
PubMed Abstract | Publisher Full Text | Free Full Text

17. Quick J, Loman NJ, Duraffour S, *et al.*: **Real-time, portable genome sequencing for Ebola surveillance.** *Nature.* 2016; **530**(7589): 228–232.
PubMed Abstract | Publisher Full Text | Free Full Text

18. Ladner JT, Larsen BB, Bowers JR, *et al.*: **An Early Pandemic Analysis of SARS-CoV-2 Population Structure and Dynamics in Arizona.** *mBio.* 2020; **11**(5): e02107–20.
PubMed Abstract | Publisher Full Text | Free Full Text

19. Forster P, Forster L, Renfrew C, *et al.*: **Phylogenetic network analysis of SARS-CoV-2 genomes.** *Proc Natl Acad Sci U S A.* 2020; **117**(17): 9241–9243.
PubMed Abstract | Publisher Full Text | Free Full Text

20. Miller SL, Nazaroff WW, Jimenez JL, *et al.*: **Transmission of SARS-CoV-2 by inhalation of respiratory aerosol in the Skagit Valley Chorale superspreading event.** *Indoor Air.* 2021; **31**(2): 314–323.
PubMed Abstract | Publisher Full Text | Free Full Text

21. Quick J: **nCoV-2019 sequencing protocol V.1.** ARTICNetwork. 2020.
Publisher Full Text

22. Tyson JR, James P, Stoddart D, *et al.*: **Improvements to the ARTIC multiplex PCR method for SARS-CoV-2 genome sequencing using nanopore.** *bioRxiv.* 2020; 2020.09.04.283077.
PubMed Abstract | Publisher Full Text | Free Full Text

23. Batty EM, Kochakarn T, Wangwiwatsin A, *et al.*: **Comparing library preparation methods for SARS-CoV-2 multiplex amplicon sequencing on the Illumina**

MiSeq platform. *bioRxiv.* 2020.
Publisher Full Text

24. RStudio: **Integrated Development for R**. Version 4.1.1. PBC; 2020.
Reference Source

25. **R: A language and environment for statistical computing.** Version 4.0.2. R Foundation; 2020.
Reference Source

26. BSDA: **Basic Statistics and Data Analysis.** Version R package version 1.2.1. 2021.

27. Colman RE, Anderson J, Lemmer D, *et al.*: **Rapid Drug Susceptibility Testing of Drug-Resistant Mycobacterium tuberculosis Isolates Directly from Clinical Samples by Use of Amplicon Sequencing: a Proof-of-Concept Study.** *J Clin Microbiol.* 2016; **54**(8): 2058–67.
PubMed Abstract | Publisher Full Text | Free Full Text

28. Bowers JR, Lemmer D, Sahl JW, *et al.*: **KlebSeq, a Diagnostic Tool for Surveillance, Detection, and Monitoring of *Klebsiella pneumoniae*.** *J Clin Microbiol.* 2016; **54**(10): 2582–96.
PubMed Abstract | Publisher Full Text | Free Full Text

29. **BBMap**. 2014.
Reference Source

30. Wu F, Zhao S, Yu B, *et al.*: **A new coronavirus associated with human respiratory disease in China.** *Nature.* 2020; **579**(7798): 265–269.
PubMed Abstract | Publisher Full Text | Free Full Text

31. Li H, Durbin R: **Fast and accurate long-read alignment with Burrows-Wheeler transform.** *Bioinformatics.* 2010; **26**(5): 589–95.
PubMed Abstract | Publisher Full Text | Free Full Text

32. Peng J, Liu J, Mann SA, *et al.*: **Estimation of Secondary Household Attack Rates for Emergent Spike L452R Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Variants Detected by Genomic Surveillance at a Community-Based Testing Site in San Francisco.** *Clin Infect Dis.* 2022; **74**(1): 32–39.
PubMed Abstract | Publisher Full Text | Free Full Text

33. Stoddard G, Black A, Ayscue P, *et al.*: **Using genomic epidemiology of SARS-CoV-2 to support contact tracing and public health surveillance in rural Humboldt County, California.** *medRxiv.* 2021; 2021.09.21.21258385.
Publisher Full Text

34. Chan AP, Choi Y, Schork NJ: **Conserved Genomic Terminals of SARS-CoV-2 as Coevolving Functional Elements and Potential Therapeutic Targets.** *mSphere.* 2020; **5**(6): e00754–20.
PubMed Abstract | Publisher Full Text | Free Full Text

35. Wickham H: **ggplot2: Elegant Graphics for Data Analysis.** Springer-Verlag; 2016.
Reference Source

36. **ggforce: Accelerating 'ggplot2'.** Version R package version 0.3.2. 2020.
Reference Source

37. Turakhia Y, Thornlow B, Hinrichs AS, *et al.*: **Ultrafast Sample Placement on Existing Trees (UShER) Empowers Real-Time Phylogenetics for the SARS-CoV-2 Pandemic.** *bioRxiv.* 2020; 2020.09.26.314971.
PubMed Abstract | Publisher Full Text | Free Full Text

38. Bolyen E, Dillon MR, Bokulich NA, *et al.*: **Reproducibly sampling SARS-CoV-2 genomes across time, geography, and viral diversity [version 2; peer review: 1 approved, 1 approved with reservations].** *F1000Res.* 2020; **9**: 657.
PubMed Abstract | Publisher Full Text | Free Full Text

39. Paden CR, Tao Y, Queen K, *et al.*: **Rapid, Sensitive, Full-Genome Sequencing of Severe Acute Respiratory Syndrome Coronavirus 2.** *Emerg Infect Dis.* 2020; **26**(10): 2401–2405.
PubMed Abstract | Publisher Full Text | Free Full Text

40. Maurano MT, Ramaswami S, Zappile P, *et al.*: **Sequencing identifies multiple early introductions of SARS-CoV-2 to the New York City Region.** *medRxiv.* 2020.
PubMed Abstract | Publisher Full Text | Free Full Text

41. Illumina: **Run time estimates for each sequencing step on the Illumina sequencing platforms**.
Reference Source

42. Murray MT, Riggs MA, Engelthaler DM, *et al.*: **Mitigating a COVID-19 Outbreak Among Major League Baseball Players - United States, 2020.** *MMWR Morb Mortal Wkly Rep.* 2020; **69**(42): 1542–1546.
PubMed Abstract | Free Full Text

43. Meredith LW, Hamilton WL, Warne B, *et al.*: **Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study.** *Lancet Infect Dis.* 2020; **20**(11): 1263–1272.
PubMed Abstract | Publisher Full Text | Free Full Text

44. Folkerts ML, Lemmer D, Pfeiffer A, *et al.*: **Supplementary Data for "Sequencing the Pandemic: Rapid and High-Throughput Processing and Analysis of COVID-19 Clinical Samples for 21st Century Public Health" [Data set].** *Zenodo.* 2021.
http://www.doi.org/10.5281/zenodo.5822962

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 2

Reviewer Report 14 March 2022

https://doi.org/10.5256/f1000research.120334.r124549

✔ **Bronwyn L. MacInnis**
Broad Institute of Harvard and MIT, Cambridge, MA, USA

The authors have addressed my comments in this version of the manuscript, congratulations on a good paper.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Virology, Genomics, Epidemiology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 07 March 2022

https://doi.org/10.5256/f1000research.120334.r124550

✔ **Keith Crandall**
[1] Computational Biology Institute, The George Washington University, Washington, DC, USA
[2] Department of Biostatistics & Bioinformatics, The George Washington University, Washington, DC, USA

The authors have responded well to my previous critiques. I think the paper now provides useful information for sequencing methods (read depth, coverage, etc.) and includes some statistics to justify some of the conclusions in the paper.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: phylogenetics, bioinformatics, computational biology, infectious disease

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 21 September 2021

https://doi.org/10.5256/f1000research.31361.r92632

Keith Crandall
[1] Computational Biology Institute, The George Washington University, Washington, DC, USA
[2] Department of Biostatistics & Bioinformatics, The George Washington University, Washington, DC, USA

This study proposes a reasonable approach to rapid assessment of viral diversity for SARS-CoV-2 taking advantage of the latest in Illumina sequencing kits and associated prep protocols as well as open source data and tools for downstream analyses. There are no statistics involved and therefore users have no idea how reliable any result might be coming out of such an approach. I would personally prefer public health decisions be based on accurate as well as rapid information. This paper speaks to rapid, but does not speak to accurate. Additionally, the paper describes some methodology that is unique to this work, but the authors do not provide any access to these crucial components. Therefore, as presented, the work seems to not be repeatable. This would need to be fixed before indexing. I have detailed these issues below.

One of the biggest issues we have when consulting with individuals or groups interested in SARS-CoV-2 sequencing is deciding on read quantity. The methods discusses 'coverage' for confidence in genome assembly, but there is no discussion at all on the number of reads targeted for sequencing. What is the target number of reads for the SNP analysis versus the whole genome analysis? Can you provide power calculations for identifying SNPs based on read coverage? Quality of assembled genomes based on read coverage? I think such analyses would be particularly helpful to individuals trying to implement this strategy in their own labs.

Is there a link to the 'global SNP database'? Is this an open access resource? If so, you should provide a link to it. If not, then the paper is not particularly useful because the work is not replicable.

Is the 'custom script' for SNP identification available? This should also be available via something

like GitHub for reproducibility.

It's too bad the authors didn't go through the effort of consenting the sample donors and including clinical data. The genomic information is highly useful (as outlined in the paper), but the genomic information COUPLED with the EHR and epidemiological data would be infinitely more useful.

There don't seem to be any statistics associated with 'transmission network' assignment. How accurate is this inference and how is a user supposed to know about the accuracy of such an assignment?

Similarly, there are no statistics associated with the phylogenetic placement of genomes. What is the confidence a user would have in such placement?

Is the rationale for developing the new method (or application) clearly explained?
Yes

Is the description of the method technically sound?
Yes

Are sufficient details provided to allow replication of the method development and its use by others?
Partly

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
No

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: phylogenetics, bioinformatics, computational biology, infectious disease

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 14 Jan 2022

Megan Folkerts, Translational Genomics Research Institute, Arizona, Flagstaff, USA

Dr. Crandall,

We want to thank you for taking the time to review our paper and provide valuable insight

and comments aimed at strengthening our manuscript. We have carefully considered the comments, and have done our best to address each of them. We hope that the revised version of this manuscript addresses the concerns you voiced, and now meets your standards. The authors welcome further constructive criticism.

Attached is a point-by-point response to both reviews. Changes made to the manuscript are clearly stated, with line numbers included for your quick reference.

Sincerely,

Megan Folkerts, MS

mfolkerts@tgen.org

Translational Genomics Research Institute, North

Response to Reviewer 2 (Dr. Keith Crandall)

Comment 1: There are no statistics involved.

Response: We agree that this was an oversight, even for a methods paper. All sequencing metrics are now reported in terms of mean, standard deviation, and 95% confidence interval.

Comment 2: There is no discussion at all on the number of reads targeted for sequencing.

Response: Thank you for this feedback. We have added read counts to all sequencing data metrics within the manuscript body. This is reported in the following paragraphs:

Lines 217-229: The number of reads targeted per sample for the high-throughput method was approximately 348,000, considering the limitation of the number of index combinations (i.e., samples per run) that were available at the time. In practice, this varied considerably. The average read count for samples using the plexWell method was 411,375, with 393,925[MOU2] being the average number of reads that aligned to the WuHan-Hu-1 reference strain. Below a Ct of 33, average aligned reads was 475,310 (SD 260292, 95%CI 455789-494831) per sample. Between Cts of 33 and 35, the average reads aligned was 204,234 (SD 175947, 95%CI 163308-245160) per sample. Above 35, read counts dropped off sharply, with a mean of 99,388 reads aligning (SD 188499, 95%CI 68493-130283) per sample. Note that all samples were put through an equimolar pooling step by the plexWell system. This suggests this pooling is not entirely effective when Cts vary considerably in a given sample set. This was confirmed through Kruskal Wallis on the total (unaligned) read counts

in each of the above-mentioned groups (p <0.001)

Lines 289-296: A minimum of 42,000 reads were targeted for the Illumina DNA prep method, but as sample sizes were small for the outbreak in question, the average aligned reads per sample was much higher. For samples with Cts below 25, the average number of reads aligned to the reference was 275,258.7 (SD 136,816.2, 95%CI 165,784.9-384732.4). Between Cts of 25 and 30, average reads aligned was 200,810 (SD 86,999.43, 95%CI 140,523.6-261096.4). Below a CT of 30, average aligned reads dropped to 116,888 (SD 275257.7, 95%CI 53,002.2-180,773.8). Kruskal-Wallis revealed no significant difference between average aligned reads between these groups (P= 0.08).

Comment 3: What is the target number of reads for the SNP analysis vs the WGS analysis?

Response: The SNP analysis was run concurrently with the WGS analysis, and the read depth targeted for the SNP analysis is the same as that of the WGS analysis. Targeted read depth, and achieved read depth, has been added to the manuscript (see response to comment 2). A previously published study reported a depth of 10 reads sufficient for SNP calling of SARS-CoV2 genomes (Ladner et al 2020). In all instances, SNPs called using the methods published here far exceed this threshold. Average read depth per sample can be found in the supplementary data.

Comment 4: Can you provide power calculations for identifying SNPs based on read coverage?

Response: We agree that providing power calculations for SNPs based on read coverage is critical for phylogenetic analyses. The purpose of this paper is simply to report sequencing and analysis methods for the purpose of generating data that can then be fed into more thorough phylogenetic analyses, as we have done with Ladner et al 2020. This is reflected in the following:

Lines 122-123: Consensus genomes generated using these methods include those from Ladner *et al*. [19] . and are similar to those in Peng et al.[43] and Stoddard et al.[44]

Comment 5: What is the quality of assembled genomes based on read coverage?

Response: Average depth of coverage, as it correlates to CT value and percent genome coverage, is summarized in Table 2 and 3. Average depth by sample across the same metrics is available in the Supplementary Methods.

Comment 6: Is there a link to the global SNP database?

Response: The global SNP database, which contains data directly downloaded from GISAID,

is not able to be shared via secondary publications, as per GISAID's data sharing agreement. The contents of the database are described, with one table generated using the Chan et al. [35] analysis on genomes downloaded from GISAID, and the other table generated directly from GISAID data which is available for public download from GISAID's website, which is linked in the reference section. This is addressed via the following:

Lines 138-143: The format of this database was simply one table of SNPs, as generated by Chan et al.[35], with columns for GISAID ID, position, reference base, and variant base, and a second table of metadata which contained the exact data fields, as downloaded from GISAID. The two tables were linked by the GISAID ID number. The GISAID database cannot be shared via secondary publications as per the database access agreement, however the database is available for download by the public from GISAID's website[16].

Comment 7: Is the custom script for SNP identification available?

Response: The custom script consisted of a simple set of commands, described in detail in the following so that they can be reproduced:

Lines 143-147: The list of SNPs common to samples in a given set was then compared to this global SNP database by first querying the number of global genomes containing each of the individual SNPs, sorting them from most common to least common, then further querying for number of global genomes containing combinations of SNPs by adding in each successive SNP in turn.

Comment 8: Genomic information coupled with the epidemiological data would be infinitely more useful

Response: We completely support this comment. However, this was not done here as it exceeded the purpose of this paper, which was to provide methods for rapid and high-throughput sample processing and initial phylogenetic placement of genomes. We agree that obtaining patient metadata can provide for a more robust analysis and meaningful interpretation for a more complete epidemiological picture, and in our publications where that is the goal, we make every effort to arrange this.

Comment 9: There don't seem to be any statistics associated with the "transmission network"

Response: Thank you for this feedback. This language was changed to reflect the scope of the paper. We were not conducting full transmission analyses (as we overstated in the first draft), but rather, were using the obtained genomic data to guide field epidemiology. Current terminology reflects this, as per:

Lines 260-262: Generating consensus genomes and a SNP report from the sequence data, which takes approximately 15 minutes for a small (<=24 samples) dataset, quickly shows

whether the samples are part of the same outbreak.

Comment 10: There are no statistics associated with the phylogenetic placement of genomes.

Response: This is correct. We have cited our previous phylogenetic analyses and other published works that have used and described similar methods. The purpose of this manuscript is primarily to report methods of generating data that can be fed into more complex phylogenetic analyses. We agree that statistics associated with placement are critical, and we have detailed complete phylogenetic analysis of SARS-CoV-2 genomes in Ladner et al. 2020, which follows the published methods of Bolyen et al 2020. Detailed phylogenetic analyses methods and results are not included here, as we felt that this was not part of this study. We have indicated this in the following:

Lines 336-340: However, reconstructing phylogenies, especially paired with finding relevant subsets, takes time, and is often overkill for initial, time-sensitive public health needs. We employ a simple, rapid analysis method and visualization meant as a quick-look to determine relatedness among a sample set of interest and/or relatedness of a sample or set to the entire public database of genomes.

Lines 349-351: This is not meant to replace more thorough phylogenetic analysis, but rather, to provide a quick genomic snapshot that can inform a public health ground response

Competing Interests: No competing interests were disclosed.

Reviewer Report 16 March 2021

https://doi.org/10.5256/f1000research.31361.r78623

?

Bronwyn L. MacInnis
Broad Institute of Harvard and MIT, Cambridge, MA, USA

This is a methodologically sound study describing complementary approaches for generating and analyzing SARS-CoV-2 genomic data. The workflows address two common use cases: rapid turnaround of a relatively small number of samples, i.e. for cluster investigation, and larger scale but less urgently time sensitive sequencing.

The approach described here is sound and there is sufficient detail to follow the methodology.

My main issue with the manuscript is that it is not clear what the novel contribution is, since each of the approaches are previously described, using commercial and published methods. However it is nonetheless helpful to have the workflows mapped out as they are here, and to benchmark the differences between the approaches. This could be particularly helpful to the many labs now attempting viral/SARS-CoV-2 sequencing for the first time.

I would also advise that the focus on retrospective sequencing limits the potential value of the larger scale approach. It could equally be relevant for larger scale surveillance sequencing, i.e. for variants of concern, real time (or near real time).

I would like to see the costs described in more detail, and in relative terms, since costs of consumables and effort can vary widely across contexts. Please clarify what is included in costs in more detail, whether costs include effort, what the impact of effort is.

Personally I find the title to be overstated for the content of the manuscript. I would suggest a title that more clearly conveys that this is a methods paper. In particular I would delete the prefix "Sequencing the pandemic", as this suggests something being done at a global scale, and the "21st century public health" as there is no direct application to public health.

Finally, although this study clearly focuses on lab methods for sequencing, given the framing I would like to see it acknowledge that currently many of the delays in SARS-CoV-2 sequencing are not in the sample preparation and sequencing steps, rather in processes up and downstream of sequencing.

Is the rationale for developing the new method (or application) clearly explained?

Yes

Is the description of the method technically sound?

Yes

Are sufficient details provided to allow replication of the method development and its use by others?

Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Virology, Genomics, Epidemiology

I confirm that I have read this submission and believe that I have an appropriate level of

expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 14 Jan 2022

Megan Folkerts, Translational Genomics Research Institute, Arizona, Flagstaff, USA

Dr. MacInnis,

We want to thank you for taking the time to review our paper and provide valuable insight and comments aimed at strengthening our manuscript. We have carefully considered the comments, and have done our best to address each of them. We hope that the revised version of this manuscript addresses the concerns you voiced, and now meets your standards. The authors welcome further constructive criticism.

Attached is a point-by-point response to both reviews. Changes made to the manuscript are clearly stated, with line numbers included for your quick reference.

Sincerely,

Megan Folkerts, MS

mfolkerts@tgen.org

Translational Genomics Research Institute, North

Response to Reviewer 1 (Dr. Bronwyn MacInnis)

Comment 1: It is not clear what the novel contribution is

Response: We appreciate this valuable feedback and hope we have provided some clarity. This manuscript seeks to provide guidance to labs seeking to undertake rapid and or high throughput sequencing, and it reports the first use of the SeqWell plexWell system for use in conjunction with the commonly used ARTIC primer system. This novel method required few laboratory personnel and minimal equipment, and allowed for the rapid sequencing of thousands of samples per week, a novel pace for such a complex procedure. The manuscript has been reworded to highlight these developments. Changes mentioned here are reflected in the following lines of the manuscript:

Abstract: The high-throughput method, first reported here for SARS-CoV2,
Lines 45-46: The plexWell method is a novel means of sequencing adapter addition to facilitate high-throughput sample processing.
Lines 315-316: Here, we report the first use of plexWell for post-amplification adapter addition for whole genome sequencing of SARS-CoV2.

Lines: 321-325: With just two full-time staff members devoted to sequencing, and 2-4 part-time staff responsible for arraying samples prior to the ARTIC/plexWell protocol, 1152 samples were able to be successfully processed and sequenced each week. Thus, the number of genomes completed per full-time (40 hour) employee per week is 384, a novel pace for what is normally a lengthy protocol

Comment 2: Focus on retrospective sequencing limits the potential value of the larger scale approach.

Response: We agree with this comment; however, sequencing was not merely done retrospectively. All data shown for the rapid-turnaround method were done in real-time, and data were delivered fast enough to inform a public health response. The manuscript has been edited to make this clear via the following (which includes a citation of the relevant MMWR report for this study).

Lines 352-355: The data presented here using the rapid-turnaround method were obtained from patients who were part of an active SARS-CoV2 outbreak, and the resulting phylogenetic placements were successfully used to inform public health efforts to determine the potential origin of the outbreak, and to prevent further spread[45].

Comment 3: I would like to see the costs described in more detail.

Response: We have now provided a breakdown of sequencing costs in the supplementary methods. This includes the manufacturer's reagent pricing, and represents the relative cost to the user minus any institution-specific discounts or advantages.

Comment 4: I find the title to be overstated

Response: We now recognize that the title overstated the scope of this methods paper. We have changed the title to more closely represent the content of the publication.

Comment 5: I would like to see it acknowledged that currently many of the delays in SARS-CoV2 sequencing are not in the sample preparation and sequencing steps, rather in the processes up and downstream of sequencing.

Response: We agree that delays aren't just in sequencing, and have acknowledged and described the other areas of potential delay in the manuscript, as per:

Lines 398-407: This paper highlights the potential for rapid turnout of genomic data from epidemiological samples, giving genomics the potential to become invaluable in pandemic response. It is worth noting, however, that sequencing isn't the only potential bottleneck in the process. For data to be delivered in real-time, coordination between testing sites, sample delivery, and sequencing laboratories must be robust, so that preparation for sequencing can begin shortly after a sample has tested positive for the pathogen of interest. Failures anywhere in this chain can lead to significant delays, which negates the

utility of a real-time sequencing pipeline for investigation. Thus, in addition to improvements in the described sequencing and analysis pipeline, coordinating the efforts of testing sites, public health partners, and sequencing laboratories is critical for this technology to be most effective.

Competing Interests: No competing interests were disclosed.