PLOS ONE

RESEARCH ARTICLE

High throughput SARS-CoV-2 variant analysis using molecular barcodes coupled with next generation sequencing

Lyora A. Cohen-Aharonov[1], Annie Rebibo-Sabbah[1], Adar Yaacov[2,3], Roy Z. Granit[1¤], Merav Strauss[4], Raul Colodner[4], Ori Cheshin[5], Shai Rosenberg[2,3], Ronen Eavri[1]*

1 Barcode Diagnostics Ltd., Nazareth, Israel, **2** Laboratory for Computational Biology of Cancer, Sharett Institute for Oncology, Hadassah - Hebrew University Medical Center, Jerusalem, Israel, **3** The Wohl Institute for Translational Medicine, Hadassah – Hebrew University Medical Center, Jerusalem, Israel, **4** Microbiology Laboratory, Emek Medical Center, Afula, Israel, **5** Internal Medicine E, Emek Medical Center, Afula, Israel

¤ Current address: Compugen Ltd, Azrieli Center, Holon, Israel
* ronen.eav@gmail.com

Citation: Cohen-Aharonov LA, Rebibo-Sabbah A, Yaacov A, Granit RZ, Strauss M, Colodner R, et al. (2022) High throughput SARS-CoV-2 variant analysis using molecular barcodes coupled with next generation sequencing. PLoS ONE 17(6): e0253404. https://doi.org/10.1371/journal.pone.0253404

Editor: Etsuro Ito, Waseda University: Waseda Daigaku, JAPAN

Received: June 16, 2021

Accepted: April 24, 2022

Published: June 21, 2022

Data Availability Statement: Some of the data underlying the results presented in the study are personal and confidential. Upon demand, the data might be available from EMEK Medical center based on the approval of the ethics committee. Ethics Committee: Helsinky Ethics committee directed by Dr Lee Goldstein and approved by Dr Eldar Berkovits, Deputy Director of Emek Medical Center, approved study number 0196-20-EMC. Contact information for data access queries:

Abstract

The identification of SARS-CoV-2 variants across the globe and their implications on the outspread of the pandemic, infection potential and resistance to vaccination, requires modification of the current diagnostic methods to map out viral mutations rapidly and reliably. Here, we demonstrate that integrating DNA barcoding technology, sample pooling and Next Generation Sequencing (NGS) provide an applicable solution for large-population viral screening combined with specific variant analysis. Our solution allows high throughput testing by barcoding each sample, followed by pooling of test samples using a multi-step procedure. First, patient-specific barcodes are added to the primers used in a one-step RT-PCR reaction, amplifying three different viral genes and one human housekeeping gene (as internal control). Then, samples are pooled, purified and finally, the generated sequences are read using an Illumina NGS system to identify the positive samples with a sensitivity of 82.5% and a specificity of 97.3%. Using this solution, we were able to identify six known and one unknown SARS-CoV-2 variants in a screen of 960 samples out of which 258 (27%) were positive for the virus. Thus, our diagnostic solution integrates the benefits of large population and epidemiological screening together with sensitive and specific identification of positive samples including variant analysis at a single nucleotide resolution.

Introduction

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is an RNA virus that causes the coronavirus disease 2019 (COVID-19). The virus was first identified in Wuhan (China) in December 2019, spread rapidly and was declared a pandemic in March 2020 by the World Health Organization. Today, more than two years later, COVID 19 has already resulted in over 400 million confirmed infection cases and over 5.7 million deaths across the world.

Sabreen Omari, phone- 972-50-5856685, mail: sabrin_om@clalit.org.il.

Funding: This work was partly supported by a grant from the Israel Innovation Authority. The funder provided support in the form of salaries for authors [L.C.A., A.R.S., R.G., A.Y and R.E.], but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The specific roles of these authors are articulated in the 'author contributions' section.

Competing interests: L.C.A., A.R.S., R.G. and R.E. were employees of Barcode Diagnostics LTD at the time the work was conducted. This commercial affiliation does not alter our adherence to PLOS ONE policies on sharing data and materials.

COVID-19 symptoms are varied, ranging from none to severe according to the Centers for Disease Control and Prevention (CDC). Many patients develop moderate pneumonia and in the most critical cases respiratory failure and multiorgan disfunction occur. The world average mortality is 2%.

Prevention of viral transmission is essential as COVID-19 spreads through the respiratory route. Transmission prevention is achieved through social distancing, masks and hand washing in addition to early identification of potential infected individuals by different diagnostic methods.

SARS-CoV-2, a positive-sense single-stranded RNA virus is a highly pathogenic member of the coronavirus family [1]. The sequence of its whole genome was published (GenBank no. MN908947), encoding 9860 amino acids. It is composed of genes that express both structural and nonstructural proteins.

S-, E-, N- and M- genes encode for structural proteins (spike, envelope, nucleocapsid and membrane proteins respectively), ORF region encodes for non-structural proteins as RNA-dependent RNA polymerase (RdRp) or papain-like protease [2].

Several methods have been developed in order to assess the infection status of individuals. Most of them rely on the detection of viral RNA.

The gold-standard method for diagnosis of COVID-19 is real-time reverse transcription polymerase chain reaction (one step rRT-PCR) [3, 4], which detects specifically the presence of a number of viral RNA fragments. This test is performed on nasopharyngeal swabs or saliva samples [5] and is widely applied and considered very robust. Results are generally delivered within 24 hours.

While, real time PCR combined with robotic systems, can be used as a high throughput method, the number of amplicons that can be detected are limited, and fluorescent probes and special reagents are needed in order to increase the sensitivity of the method. Moreover, as the global spread of the virus continues despite the use of preventative measures and positive effects of vaccination [6], SARS-CoV-2 mutations generate different viral variants, some with the potential to become more virulent and contagious or resistant to vaccination. The formation and spread of new variants, requires the development and implementation of cost-effective diagnostic methods that can both detect infection and new single nucleotide variants (SNVs) in real time to warn authorities prior to their spread.

Next-generation diagnostic solutions, which are based on the current PCR test coupled with next generation sequencing (NGS) offer major advantages, such as significantly larger scale testing (several hundred-fold scale up versus the currently available test), reduced cost per test, reduced amount of reagents per test and potential readout of viral sequence variation.

In NGS-based methods, individual samples can be uniquely labeled with molecular barcodes and multiple fragments can be amplified in parallel and pooled together for high throughput detection and processing of thousands of specimens together. A few different protocols have been developed [7, 8] and one of them has even received an Emergency Use Authorization from the FDA [9]. As next-generation sequencers are now widely available, including inside hospitals, this platform offers a key solution for COVID-19 mass detection worldwide.

Here we propose a solution that allows high throughput testing by barcoding and pooling test samples, up to hundreds at a time, in one well (as illustrated in Fig 1). The protocol combines multiplexed PCR, patient specific molecular barcoding, pooling, NGS, bioinformatics and machine learning analysis, for identification of positive individuals by detection of three different viral genes and one human gene (as internal control) from a pool of thousands of specimens, enabling sensitive and accurate molecular-based diagnosis of COVID-19. Using this method, we have identified six known and one new viral-variants in a mass screening

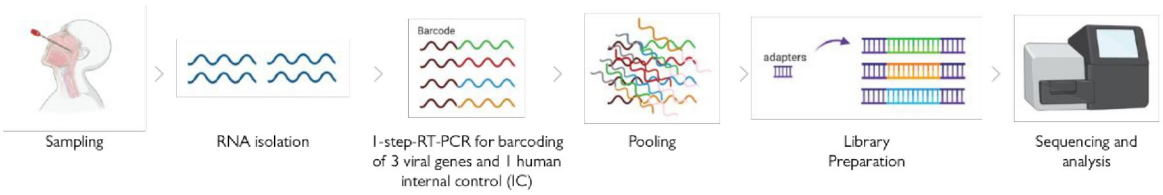

Fig 1. A schematic showing the overall workflow, including viral RNA extraction, one-step reverse transcription-PCR using barcoded-specific primers, and next-generation sequencing to quantify the amounts of barcoded amplicons.

https://doi.org/10.1371/journal.pone.0253404.g001

analysis of 960 samples. The test can prove to be vital in the identification of various viral variants as part of population and epidemiological screening.

Results

Barcodes

A bank of 2133 unique 10-bases and about 21,000 unique 12-bases barcodes was generated. For the purpose of this research, the most distinguishable 96 barcodes were selected for the experiments (see Methods).

Pilot

Positive and negative SARS-CoV-2 standards (Bio-Rad) were run as a pilot with 17 barcodes and a multiplex of two viral genes (N1 and E) and one human internal control (RNaseP). Following multiplexing, amplicons were pooled. A single library was prepared and ran on a MiSeq (Illumina) instrument.

Following demultiplexing, number of reads for each barcode was obtained. RNase P reads were used for internal control and were required to be positive in order to consider the result as valid.

The viral genes were used for the diagnosis and detection of SARS-CoV-2 specifically in each specimen. For each barcode, the associated viral reads were counted and the results are presented in Fig 2a. A clear distinction was observed between positive and negative samples.

Barcode stability

Following this experiment, the stability of 96 barcodes was tested using positive SARS-CoV-2 standard as an identical template for each barcode. Furthermore, an additional viral region (N2) was amplified and added to the multiplex. The stability of 96 barcodes is presented in Fig 2b and 2c. No specific influence of the barcode sequence was noticed.

Moreover, the N2 viral fragment added additional information regarding the viral identification and can provide an improvement in diagnosis of the specimens.

New genes

In order to improve specificity and sensitivity, additional viral regions for barcoded amplification were examined. Each new set of primers was first tested on positive and negative SARS--CoV-2 standards. The count of reads is presented in Fig 2d. Human RNase P served as the internal control gene. A threshold was set to determine the genes that should be part of the multiplexes. This threshold represents the minimal number of reads of positive samples for a specific target in order to include this specific amplified amplicon in the multiplex assay. As

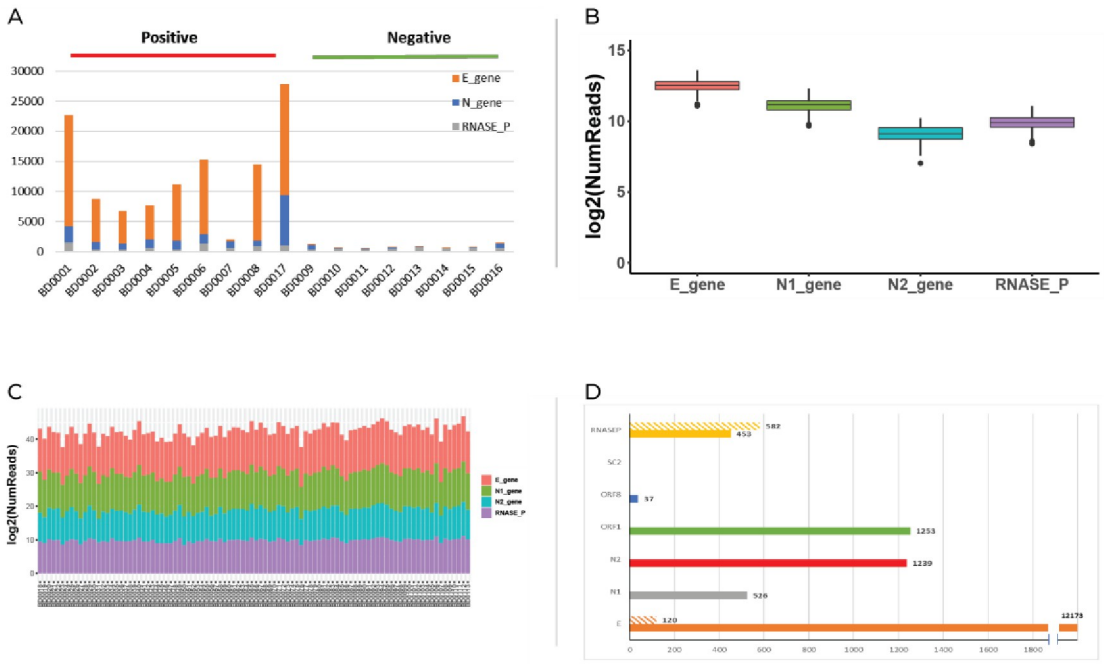

Fig 2. Identification and optimization of viral and human control sequence reads: A. 17 clinical RNA samples (9 positive and 8 negative) analyzed by molecular barcoding of 2 viral genes and one internal control human gene and Next-Generation Sequencing. B and C. Stability of each of the 96 barcodes used in the experiment. The same amount of positive standard was used with all the barcodes for comparison. Count of reads is presented in boxplots for each target (E, N1, N2 and RNASE P) (B) and in columns for each barcode (C) D. Analysis of 6 viral genes and one human internal control (plain column- positive standard, striped column- negative standard).

https://doi.org/10.1371/journal.pone.0253404.g002

the SC2 assay failed to yield any reads for both positive and negative templates, it was not comprised in the further multiplexes. Additionally, we observed a weaker signal for ORF8 in comparison to ORF1. Thus, N1, N2, E, and ORF1 were selected as the multiplex target fragments for the next step.

Multiplex optimization

In order to choose the most effective combination of genes targeted, the following multiplexes were tested on 96 clinical RNA specimens for optimization (Fig 3). Multiplex 1 included viral genes N1, N2 and E; multiplex 2 included viral genes N1, N2 and ORF1; and multiplex 3 included viral genes N1, E and ORF1. Human RnaseP was included in all three mixes as internal control.

Our results demonstrate that when genes N1 and N2 are multiplexed together, a non-specific DNA fragment is generated due to the physical proximity of the two amplicons (unshown data). This 944bp DNA fragment was formed as the elongation product of the N1 forward primer and N2 reverse primer. As this fragment is added to the library preparation and run together with all the amplicons, it can lead to a competition in the NGS run and to a lower number of reads that are obtained from the fragments of—interest.

Detection of SARS-CoV-2 in a 960 samples pool

In order to scale up our technology, 960 clinical RNA specimens were tested using multiplex 3. Amplicons were pooled (each plate separately), libraries were prepared for each plate with a

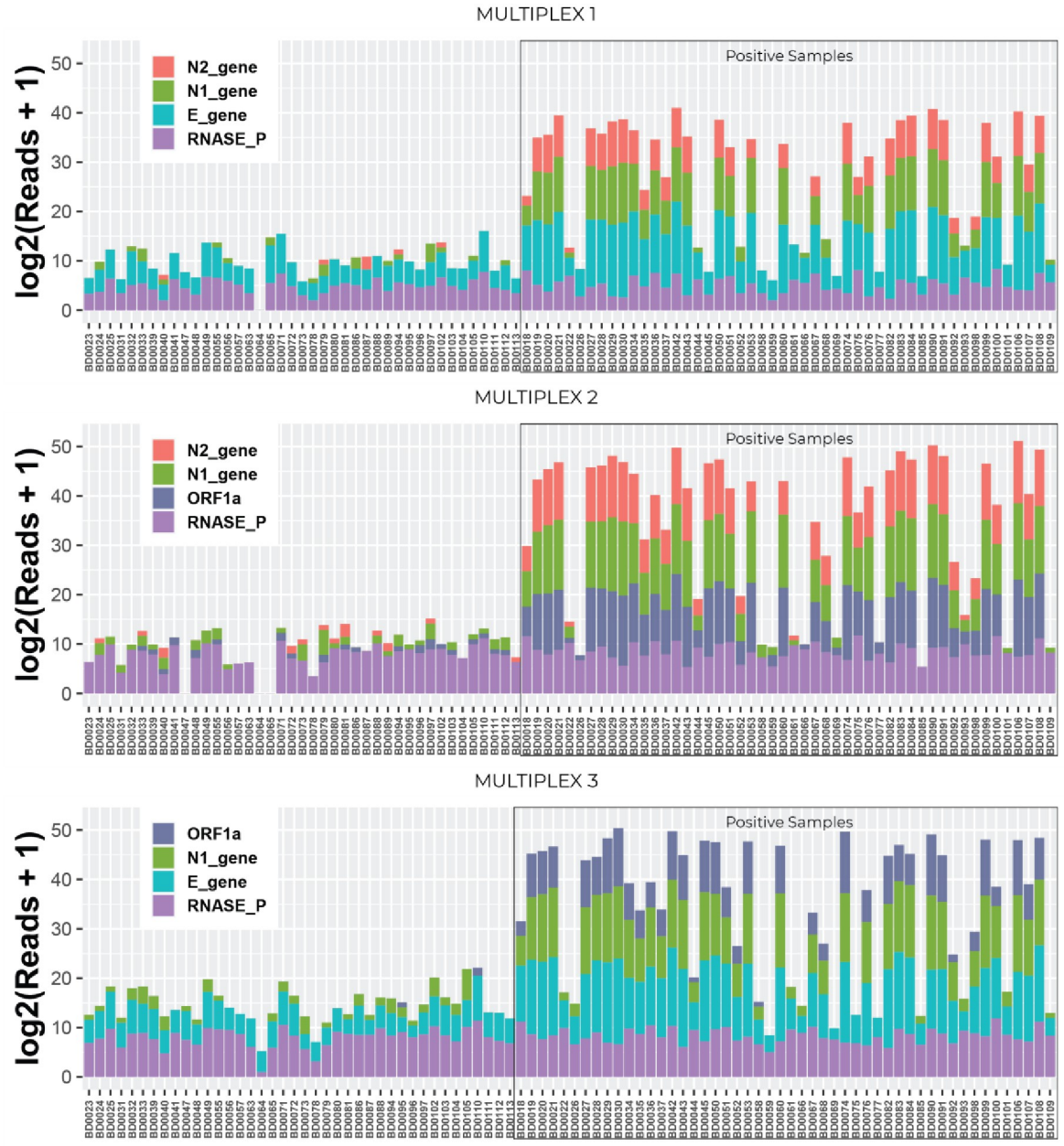

Fig 3. Multiplex optimization. Three different multiplexes including three viral genes among N1, N2, E and ORF1a genes, and including one human internal control gene RNAseP were tested separately with the same clinical specimens in order to compare between them and determine the optimal combination to be use for the test.

https://doi.org/10.1371/journal.pone.0253404.g003

different index and all ten libraries were sequenced together using the Next-Seq instrument (illumina). Following demultiplexing, obtained sequences were analyzed and positive samples were identified using the unique and specific barcodes (Fig 4a). The number of viral reads in our pool was in negative correlation with the Ct number obtained by regular real-time PCR. When the Ct was high, the number of viral reads in the pool were low (Fig 4b). Our results indicate that at a Ct higher than 30 the number of viral reads is similar between positive and negative specimens.

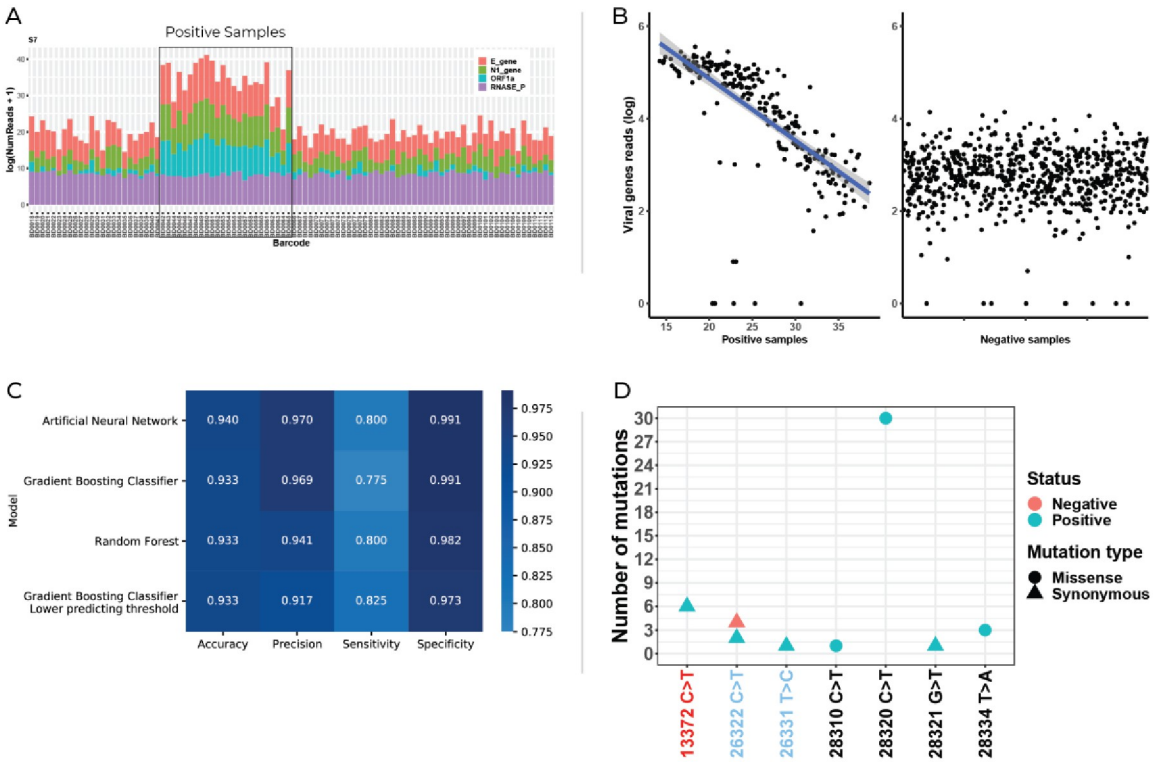

Fig 4. A. Bar charts from one representative plate of total sequencing reads from specimen tested. B. Total viral reads in positive and negative specimens. Each point represents a sample. Left panel, positive samples, X axis shows the Ct number obtained from the amplification of a fragment from N gene. Right panel, negative samples. C. Heat map showing accuracy, precision, sensitivity and specificity using four different machine learning algorithms. D. A scatter plot showing all detected genomic variants within the 960 samples tested. 6 known SNVs and one heretofore unreported SNV were found.

https://doi.org/10.1371/journal.pone.0253404.g004

Using a Gradient Boosting machine learning classification model, results were obtained with an accuracy of 93.3%, precision of 91.7%, sensitivity of 82.5% and specificity of 97.3% (Fig 4c) in a validation set. We further trained a model defining Ct<30 (N-gene) specimens as positive. The sensitivity was 100% with 98.5% specificity and a positive predictive value (PPV) of 94.7% in a validation set.

Variant detection

In addition to the detection of the virus with high specificity and sensitivity (especially for Ct <30), SARS-CoV-2 barcoded pooling provides a platform for detection of single nucleotide mutations (Fig 4d) in pre-determined gene regions. We have sequenced 3 viral fragments for detection of the virus in all 960 specimens. Following NGS analysis of the multiplexes, 7 single nucleotide variants were identified. By comparing the observed mutations to SARS_CoV-2 databases using Blastn in betacoronavirus genbank—https://blast.ncbi.nlm.nih.gov/) we have identified six known and one newly found missense mutation. Among the 258 SARS-CoV-2 PCR-positive samples, 30 carried the same N-gene missense mutation and 6 carried an additional synonymous substitution in ORF1a (Fig 4d). Eight PCR-positive and two PCR-negative samples were found with the five remaining single nucleotide variants, one variant per sample.

Discussion

SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2), a beta coronavirus, is the novel coronavirus that caused the COVID-19 outbreak originating in Wuhan, China in 2019. This virus has become a worldwide pandemic requiring immense and rapid viral detection methods throughout the world. The classic method for detection of infection involves sampling nasopharyngeal swabs from patients in order to detect viral infection. The swab is placed in transfer buffer and then lysis buffer is added to lyse the potentially infected cells. RNA is extracted from the solution then subjected to a one-step RT-PCR reaction. The RNA is first converted to cDNA by reverse transcription which is immediately followed by Multiplex quantitative Real time PCR (qPCR) with primers and a probe specific to the virus to detect viral presence.

Generally, an internal control is added in order to eliminate the possibility of false negative results and includes detection of a human gene as a proxy for host RNA content. However, some assays only amplify an exogenic internal control gene (added during the extraction step) which provides evidence for the validity of the qPCR reaction yet, is not reliable enough to exclude false negatives.

Here we describe a method we developed for combined screening of an infectious agent and simultaneously determine the presence of a genomic variant in a population, by using a DNA-barcoded sample from each subject in a population. The method includes performing one-step reverse transcription and amplification of RNA extracted from the samples with barcoded primers, pooling and sequencing the obtained amplicons, and identifying the infected individual out of all tested samples in the pool via the subject-specific barcode. The different primers target three viral genes as well as one human endogenous gene. The detection of an endogenic human internal control assay that is specific to each specimen, even though they are sequenced as a pool, is of extreme importance and represents one of the major advantages of this technique.

As the virus has spread globally, it has naturally acquired thousands of mutations. Most of these are deleterious or neutral, however, some confer selective advantage and have given rise to many viral variants that became the dominant viral strain in many countries. Some of these variants as they have been shown to increase infection rate, cause higher mortality and most importantly may evade the immune response obtained by vaccination [10]. It is therefore essential that health care authorities collect evidence on the extent and spread of such viral variants within the population. However, currently the RT-PCR method used to identify SARS--CoV-2 infection, does not enable the identification of sequence variations in specific genomic locations. Additionally, while NGS enables variant analysis and lineage tracking, it is not yet set as a standard mass-detection method which can be efficiently used for population screening.

We provide a method that can be used to identify and screen specific loci on the viral genome which are of concern at the population level. For SARS-CoV-2, one of the loci can be the entire 'Spike' sequence, while other areas can potentially be identified as more variants and lineages evolve. We suggest that once new SNVs are identified in this screen, specific samples can be identified based on their unique molecular barcode followed by full genome sequencing to provide a comprehensive sequence-analysis for lineage tracking.

Our method utilizes amplified cDNA produced from the infectious agent and at least one SNV in a sequencing read, to determine via the subject-specific barcode whether the subject is infected and simultaneously enables the identification of a genomic variant. The method presented here can potentially be modified using primers designed upstream of the variable S gene to enable the detection of the known variants (UK- B.1.1.7 lineage, South Africa- B.1.351

lineage, Brazil- P.1 lineage, NY- B.1.526 lineage) whose sequences are mutated in this specific locus (Spike gene). Our aim was to integrate sequence-analysis which can provide information at a population level by searching for SNVs on multiple short sequences from hundreds of test samples and by positively identifying SARS-CoV-2 variants at an individual level. Finally, sensitivity and specificity can be improved by applying machine learning algorithms. This integrative approach can be used to identify SARS-CoV-2 as a stand-alone method or, to be used as an additional information layer which screens the RT-PCR positive samples to identify the viral variants.

In order to generate a SARS-CoV-2 whole genome read with high sensitivity and specificity (98% and 97%), at least 1,000 copies of the viral genome per milliliter are required [9]. While full-genome analysis provides information on potential changes in the entire viral RNA sequence, this approach limits the potential amount of individual screening tests that can be generated in a population. Thus, the number of samples providing information from the whole viral genome limits its usage as a mass-screening tool in this current pandemic.

To overcome this limitation, additional NGS-based mass screening approaches have provided proof that such methodologies can be used to identify new variants, detect genetic epidemiology and follow viral lineages [11]. However, for low viral counts, the sensitivity and specificity of this system are not clear. The method we described here, further expands this high throughput approach through a unique individualized barcoding system, and a machine learning module to further increase sensitivity and specificity.

Based on our findings, we propose coupling individualized sample identification and pooling with machine learning approaches to optimize current NGS resources to quickly identify viral variants in the population. This approach combines the benefits of high throughput screenings by focusing on specific areas-of-interest in order to increase the number of individual screening tests and employs machine learning algorithms to increase diagnostic accuracy.

Population mass-screening diagnostic methods are required to be cost-effective. Based on a calculation of reagents and labor, we have generated a cost estimate to analyze each sample using the suggested method for SARS-CoV-2 variant detection. Given the large-capacity and pooling nature, which stand at the basis of this diagnostic method, it is estimated that an individual test would roughly cost 1–2 USD (starting from the extracted viral RNA). This estimate provides organizations which currently run large PCR-based SARS-CoV-2 screening assays, a viable cost-effective option to add an important variant-analysis layer to their data.

In conclusion, in comparison to the gold standard real time PCR method, that identifies amplified cDNA produced from the infectious agent and is only able to provide a binary answer (SARS-CoV-2 positive or negative), the method presented here, enables specific diagnosis of SARS-CoV-2 variants within a population of thousands. As this solution can either provide a stand-alone identification method for SARS-CoV-2 or be integrated with the current RT-PCR viral test, we propose adopting this technique to large population and epidemiological screening practice.

Materials and methods

Clinical specimens

Oro- and nasopharyngeal swabs were collected from patients by trained health workers with personal protective equipment and directly placed into e-swab tubes (COPAN) for conservation and transport. The study number 0196-20-EMC was granted Ethics approval by the head of the Emek Medical Center Helsinki committee Dr Lee Goldstein M.D. and signed by the Deputy Director of Emek Medical Center Dr Eldar Berkovits M.D. M.H.A.

RNA extraction

Following sampling and storage of the oro- and nasopharyngeal swabs, nucleic acids were isolated and purified from swab specimens using STARMag 96 X 4 Universal Cartridge Kit (Seegene) on Star or Starlet instruments (Hamilton) following the manufacturer's instructions.

Primer sequences

The targets analyzed were amplified using previously reported sets of known primers, while unique barcodes were added to the forward primer to enable pooling of the products before sequencing. These primers sequences are summarized in Table 1.

The complete table of primers including barcodes appears in the supplementary data of this publication.

Singleplex RT-PCR

The purified nucleic acids were reverse transcribed into cDNA and amplified using Qscript XLT one-step RT-PCR (Quantabio) and specific primers for each viral and human amplicons. Each forward primer included a 10-nucleotide barcode at the 5'-end, which was unique for each sample tested and enabled pooling of the PCR products prior to sequencing. The total volume of reaction mix (25 μl) contained the following: 12.5 μl of 2X concentrated ToughMix; 0.4 μM reverse primer, 0.4 μM barcoded forward primer; 1 μl of 25X qScript XLT and 5 μl of SARS-CoV-2 positive and negative standard (BioRad). The one step RT-PCR was conducted in a C1000 touch thermocycler (BioRad) as per the following: cDNA Synthesis (48˚C, 20 min), Initial denaturation (94˚C, 3 min), PCR cycling -25 cycles (Denaturation: 94˚C, 15s; Annealing: 60˚C, 40s; Extension: 72˚C, 30s).

Multiplex RT-PCR

The purified nucleic acids were reverse transcribed into cDNA and amplified using Qscript XLT one-step RT-PCR (Quantabio) and specific primers for 3 viral amplicons (N1, N2, E in multiplex 1; N1, N2 and ORF1 in multiplex 2; and N1, E and ORF1 in multiplex 3) and one human (RNaseP as internal control). Each forward primer included a 10-nucleotide barcode at the 5'-end, which is unique for each individual tested (the barcode generation is described in the next paragraph). The total volume of reaction mix (25 μl) contained the following: 12.5 μl of 2X concentrated ToughMix; 0.4 μM each reverse and barcoded forward primer; 1 μl of 25X qScript XLT and 8 μl of the RNA sample. The one step RT-PCR was conducted in a

Table 1. Primers sequences.

Name	Forward sequence (5'-3')	Reverse sequence (5'-3')	Reference
N1	GACCCCAAAATCAGCGAAAT	TCTGGTTACTGCCAGTTGAATCTG	https://www.cdc.gov/coronavirus/2019-ncov/lab/rt-pcr-panel-primer-probes.html
N2	TTACAAACATTGGCCGCAAA	GCGCGACATTCCGAAGAA	
E	ACAGGTACGTTAATAGTTAATAGCGT	ATATTGCAGCAGTACGCACACA	https://www.who.int/docs/default-source/coronaviruse/wuhan-virus-assay-v1991527e5122341d99287a1b17c111902.pdf
ORF1a	CCCTGTGGGTTTTACACTTAA	ACGATTGTGCATCAGCTGA	https://www.centerforhealthsecurity.org/resources/COVID-19/COVID-19-fact-sheets/200410-RT-PCR.pdf
ORF8	TCTAAATCACCCATTCAGTACATC	ATGAAATCTAAAACAACACGAACG	[12]
SC2	CTGCAGATTTGGATGATTTCTCC	CCTTGTGTGGTCTGCATGAGTTTAG	https://www.cdc.gov/coronavirus/2019-ncov/lab/multiplex-primer-probes.html
RNASE-P	AGATTTGGACCTGCGAGCG	TTTACATGTAAGATTTGGACCTGCGAGCG	https://www.cdc.gov/coronavirus/2019-ncov/lab/rt-pcr-panel-primer-probes.html

https://doi.org/10.1371/journal.pone.0253404.t001

C1000 touch thermocycler (BioRad) as per the following cDNA Synthesis (48˚C, 20 min), Initial denaturation (94˚C, 3 min), PCR cycling -25 for pool of 96 specimens /35 cycles for pool of 960 specimens (Denaturation: 94˚C, 15s; Annealing: 60˚C, 40s; Extension: 72˚C, 30s).

Barcode generation

Sequences of unique ten-base barcodes were created using DNABarcodes R package [13]. Barcodes were generated in such a manner that they were at least three Sequence-Levenshtein distance apart. In addition, to meet the required chemical properties, sequences that contained triplets that were self-complementary or that showed an unbalanced GC ratio were filtered out. Based on distance metrics between each pair of barcodes, most distinguishable 96 barcodes were selected for the experiments based on the distance between every possible pair among the 96 barcodes.

Library preparation and next generation sequencing

All the 96 separate reactions of one plate were pooled together. Pooled PCR products were purified using MiniElute Spin Columns (Qiagen) according to manufacturer's instructions.

Purified PCR amplicons were used as input DNA to generate Illumina-compatible sequencing libraries using NEBNext® Ultra™ II DNA Library Prep Kit for Illumina® (New England BioLabs) according to the manufacturer's instructions (fragmentation and size-selection were not required). Final sequencing libraries were sequenced on the Illumina Miseq sequencer (Illumina) using 2x150 bp paired-end reads (Micro kit).

A pool of 960 clinical samples was tested using the technology. Ten 96 plates with 96 different specimens (about 25% positive) were amplified using the barcoding-multiplex PCR. Wells from each plate were pooled together and a library was prepared using unique index. All ten libraries were sequenced together using the Next-Seq instrument (illumina).

Next generation sequence analysis

Reads processing pipeline. First, a raw FASTQ file was demultiplexed by DNA-barcodes using FASTX barcode splitter, i.e., splitting into numerous files based on DNA-barcode matching, one FASTQ per barcode (FASTX-toolkit version 0.0.14, http://hannonlab.cshl.edu/fastx_toolkit/). One mismatch [13] between the known and the observed sequence of a barcode was allowed. Then, each barcode specific FASTQ file was trimmed using FASTX trimmer, leaving 41 bases of each read, from the 11th base to the 51st base, representing a part of the designated primers. The trimmed reads were aligned to N (nucleocapsid phosphoprotein), E (envelope protein) and ORF1ab (ORF1a; ORF1b polyproteins) SARS-CoV-2 genes and human RNASE P (from NCBI Genbank) using BBMap version 38.18, which is a short-reads aligner for DNA or RNA-seq data (http://sourceforge.net/projects/bbmap/). Finally, the set of reads were quantified using Salmon version 0.14.2 [14].

Variant calling pipeline. For the purpose of variant calling, the barcodes sequences were trimmed at the 11th base, leaving the entire reads. Quality control and preprocessing of FASTQ files were performed using fastp version 0.20.1 [15]. Each FASTQ per sample was mapped to SARS-CoV-2 reference genome (GenBank no. MN908947) using bwa version 0.7.17-r1188 [16]. BAM files were sorted by SAMtools version 1.9 [17], and variants were called using bcftools version 1.9 [18], filtered by quality-measure of at least 20 and read depth of at least 50.

Machine learning models. Classification models were implemented using scikit-learn module in Python, version 0.23.2 [19]. For the initial model, the samples were randomly divided into 60% train set, 20% validation set and 20% test set. SARS-COV-2 genes counts,

normalized for total read number per sample, were selected as features, while the label of a positive/negative diagnosis was based on traditional RT-PCR. Hyperparameters were selected after an exhaustive grid search, with 5-fold cross validation. Selected hyperparameters were those which provided the maximal Area Under the Receiver Operating Characteristic (ROC) curve (AUC) scores. Then, the model was evaluated by predicting the diagnosis in the test set. Additional models, i.e., based on different gene combinations, were methodologically similar.

Code availability

The R code used to generate results presented in Figs 2–4 and Python workflow used to build the machine learning models are available on https://github.com/Adarya/SARS-CoV-2_NGS_Barcodes.

Supporting information

S1 Data.
(CSV)

S2 Data.
(CSV)

S3 Data.
(CSV)

Author Contributions

Conceptualization: Annie Rebibo-Sabbah, Ronen Eavri.

Data curation: Merav Strauss, Raul Colodner, Ori Cheshin.

Formal analysis: Adar Yaacov, Roy Z. Granit, Shai Rosenberg.

Investigation: Shai Rosenberg.

Methodology: Lyora A. Cohen-Aharonov, Annie Rebibo-Sabbah.

Project administration: Lyora A. Cohen-Aharonov, Ronen Eavri.

Resources: Merav Strauss, Raul Colodner, Ori Cheshin, Ronen Eavri.

Software: Adar Yaacov, Roy Z. Granit, Shai Rosenberg.

Supervision: Ronen Eavri.

Writing – original draft: Lyora A. Cohen-Aharonov, Ronen Eavri.

Writing – review & editing: Lyora A. Cohen-Aharonov, Annie Rebibo-Sabbah, Ronen Eavri.

References

1. Machhi J. et al. The Natural History, Pathobiology, and Clinical Manifestations of SARS-CoV-2 Infections. *Journal of Neuroimmune Pharmacology* vol. 15 359–386 (2020).
2. Huang Y., Yang C., Xu X.-F., Xu W. & Liu S.-W. Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. *Acta Pharmacol. Sin.* (2020) https://doi.org/10.1038/s41401-020-0485-4 PMID: 32747721
3. Bullard J. et al. Predicting infectious severe acute respiratory syndrome coronavirus 2 from diagnostic samples. *Clin. Infect. Dis.* 71, 2663–2666 (2020). https://doi.org/10.1093/cid/ciaa638 PMID: 32442256
4. Ma L. et al. Comprehensive analyses of bioinformatics applications in the fight against COVID-19 pandemic. *Computational biology and chemistry* 95 (2021): 107599. https://doi.org/10.1016/j.compbiolchem.2021.107599 PMID: 34773807

5. Wyllie A. L. et al. Saliva or Nasopharyngeal Swab Specimens for Detection of SARS-CoV-2. *N. Engl. J. Med.* 383, 1283–1286 (2020). https://doi.org/10.1056/NEJMc2016359 PMID: 32857487

6. Hodgson S. H. et al. What defines an efficacious COVID-19 vaccine? A review of the challenges assessing the clinical efficacy of vaccines against SARS-CoV-2. *The Lancet Infectious Diseases* vol. 21 e26–e35 (2021). https://doi.org/10.1016/S1473-3099(20)30773-8 PMID: 33125914

7. Schmid-Burgk J. L. et al. LAMP-Seq: Population-scale COVID-19 diagnostics using combinatorial barcoding. *bioRxiv* 2020.04.06.025635 (2020) https://doi.org/10.1101/2020.04.06.025635

8. Yelagandula R. et al. SARSeq, a robust and highly multiplexed NGS assay for parallel detection of SARS-CoV2 and other respiratory infections. *medRxiv* 2020.10.28.20217778 (2020) https://doi.org/10.1101/2020.10.28.20217778

9. First NGS-based COVID-19 diagnostic. *Nature biotechnology* vol. 38 777 (2020). https://doi.org/10.1038/s41587-020-0608-y PMID: 32641848

10. Luchsinger L. L. & Hillyer C. D. Vaccine efficacy probable against COVID-19 variants. *Science* vol. 371 1116 (2021). https://doi.org/10.1126/science.abg9461 PMID: 33707257

11. Bhoyar R.C. et. al. High throughput detection and genetic epidemiology of SARS-CoV-2 using COVIDSeq next-generation sequencing. *PLoS One* 16(2):e0247115 (2021). https://doi.org/10.1371/journal.pone.0247115 PMID: 33596239

12. Kakhki R. K., Kakhki M. K. & Neshani A. COVID-19 target: A specific target for novel coronavirus detection. *Gene Reports* 20, (2020). https://doi.org/10.1016/j.genrep.2020.100740 PMID: 32510005

13. Buschmann T. & Bystrykh L. V. Levenshtein error-correcting barcodes for multiplexed DNA sequencing. *BMC Bioinformatics* 14, 272 (2013). https://doi.org/10.1186/1471-2105-14-272 PMID: 24021088

14. Patro R., Duggal G., Love M. I., Irizarry R. A. & Kingsford C. Salmon: fast and bias-aware quantification of transcript expression using dual-phase inference. https://doi.org/10.1038/nmeth.4197 PMID: 28263959

15. Chen S., Zhou Y., Chen Y. & Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34, i884–i890 (2018). https://doi.org/10.1093/bioinformatics/bty560 PMID: 30423086

16. Li H. & Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760 (2009). https://doi.org/10.1093/bioinformatics/btp324 PMID: 19451168

17. Li H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079 (2009). https://doi.org/10.1093/bioinformatics/btp352 PMID: 19505943

18. Narasimhan V. et al. BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics* 32, 1749–1751 (2016). https://doi.org/10.1093/bioinformatics/btw044 PMID: 26826718

19. Pedregosa F. et al. Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot. *Journal of Machine Learning Research* vol. 12 http://scikit-learn.sourceforge.net. (2011).