High sensitivity detection of coronavirus SARS-CoV-2 using multiplex PCR and a multiplex-PCR-based metagenomic method

Chenyu Li[1#], David N. Debruyne[1#], Julia Spencer[1], Vidushi Kapoor[1], Lily Y. Liu[1], Bo Zhou[2], Lucie Lee[1], Rounak Feigelman[1], Grayson Burdon[1], Jeffrey Liu[1], Alejandra Oliva[1], Adam Borcherding[3], Hongdong Tan[3,4], Alexander E. Urban[2], Guoying Liu[1], Zhitong Liu[1*]

[1] Paragon Genomics Inc., Hayward, CA 94545 USA

[2] Department of Psychiatry and Behavioral Sciences, Department of Genetics, Stanford University, CA 94305 USA

[3] MGI, BGI-Shenzhen, Shenzhen 518083 China

[4] BGI-Shenzhen, Shenzhen 518083 China

[#] These authors contributed equally to the work.

[*] Corresponding author's email: zhitong@paragongenomics.com

Abstract

Many detection methods have been used or reported for the diagnosis and/or surveillance of SARS-CoV-2. Among them, reverse transcription polymerase chain reaction (RT-PCR) is the most sensitive, claiming detection of about 5 copies of viruses. However, it has been reported that only 47-59% of the positive cases were identified by RT-PCR, probably due to loss or degradation of virus RNA in the sampling process, or even mutation of the virus genome. Therefore, developing highly sensitive methods is imperative to ensure robust detection capabilities. With the goal of improving sensitivity and accommodate various application settings, we developed a multiplex-PCR-based method comprised of 172 pairs of specific primers, and demonstrated its efficiency to detect SARS-CoV-2 at low copy numbers. The assay produced clean characteristic target peaks of defined sizes, which allowed for direct identification of positives by electrophoresis. In addition, optional sequencing can provide further confirmation as well as phylogenetic information of the identified virus(es) for specific strain discrimination, which will be of paramount importance for surveillance purposes that represent a global health imperative. Finally, we also developed in parallel a multiplex-PCR-based metagenomic method that is amenable to detect SARS-CoV-2, with the additional benefit of its potential for uncovering mutational diversity and novel pathogens at low sequencing depth.

Introduction

A severe epidemic coronavirus (SARS-CoV-2) infection, now just characterized as a pandemic by the World Health Organization (WHO), started in December of 2019 in Wuhan China and quickly spread to many countries in the world[1-3]. It has caused over 7,000 deaths as of mid-March and made daily impacts on various aspects of societal life around the globe[4]. SARS-CoV-2 is a coronavirus with positive-sense, single-stranded RNA about 30kb in length. The genome of SARS-CoV-2 is currently under careful investigation[5-9] and being extensively modeled according to the Chinese National Center for Biological information (2019 New Coronavirus Information Database, https://bigd.big.ac.cn/ncov). Of note, SARS-CoV-2 exhibits over 99% sequence similarity among many sequenced isolates, and is also highly similar to other coronaviruses[5,9].

Present methods for detecting SARS-CoV-2 have been reported and discussed[10,11], including RT-PCR, serological testing[12] and reverse transcription-loop-mediated isothermal amplification[13,14]. Currently, RT-PCR is considered the gold standard of diagnosis for SARS-CoV-2 due to its ease of use and high sensitivity. RT-PCR has been reported to detect SARS-CoV-2 in saliva[15], pharyngeal swab, blood, anal swab[16], urine, stool[17], and sputum specimens[18]. In laboratory conditions, the RT-PCR methodology has been shown to detect 4-8 copies of virus, through amplification of targets in the Orf1ab, E and N viral genes, at 95% confidence intervals[19-21]. However, only about 47-59% of the positive cases were identified by RT-PCR, and 75% of RT-PCR negative results were actually found to be positive, with repeated tests required[17,22-24]. In addition, there is evidence suggesting that heat inactivation of clinical samples causes loss of virus particles, thereby hindering the efficiency of downstream diagnosis[25].

Therefore, it is necessary to develop robust, sensitive, specific and highly quantitative methods for the delivery of reliable diagnostic assays[26,27]. The urgency to develop an effective surveillance method that can be easily used in a variety of laboratory settings is underlined by the wide and rapid spreading of SARS-CoV-2[28-30]. In addition, such method should also distinguish SARS-CoV-2 from other respiratory pathogens such as influenza virus, parainfluenza virus, adenovirus, respiratory syncytial virus, rhinovirus, human metapneumovirus, SARS coronavirus, etc., as well as mycoplasma pneumoniae, chlamydia pneumonia and bacterial pneumonia[31-34]. Furthermore, providing nucleotide sequence information through next generation sequencing (NGS) will prove to be essential for the surveillance of SARS-CoV-2's evolution[35-38]. Indeed, SARS-CoV-2 phylogenetic studies through genome sequence analysis have provided better understanding of the transmission origin, time and routes, which has guided policy-making and management procedures[8,36,37,39-41].

Here, we describe the development of a highly sensitive and robust detection assay incorporating the use of multiplex PCR technology to identify SARS-CoV-2. Theoretically, the multiplex PCR strategy, by amplifying hundreds of targets, has significantly higher sensitivity than RT-PCR and may even detect DNA molecules resulting from degraded virus genome fragments. Multiplex PCR has been shown to be an efficient and low-cost method to detect *Plasmodium falciparum* infections, with high coverage (median 99%), specificity (99.8%) and sensitivity. Moreover, this solution can be tailored to simultaneously address multiple questions of interest within various epidemiological settings[42]. Similar to a recently described metagenomic approach for SARS-CoV-2 identification[43], we also establish a user-friendly multiplex-PCR-based metagenomic method that is not only able to detect SARS-CoV-2, but could also be applied for the identification of significant sequence mutations within known viruses and to uncover novel pathogens with a limited sequencing depth of approximately 1 million reads.

Results

Mathematical model of RT-PCR

Several RT-PCR methods for detecting SARS-CoV-2 have been reported to date[15,19-21]. Among them, two groups reported the detection of 4-5 copies of the virus[19,21]. To investigate the opportunity for further improvement on the sensitivity of RT-PCR, we built a mathematical model to estimate the limit of detection (LOD) for SARS-CoV-2. The reported RT-PCR amplicon lengths are around 78-158bp, and the SARS-CoV-2 genome is 29,903bp (NC_045512.2). Thus, in this mathematical modeling, we chose 100bp amplicon length and 30kb SARS-CoV-2 genome size for the estimation. With the assumption of 99% RT-PCR efficiency[20], we found that RT-PCR assays could detect 4.8 copies of SARS-CoV-2 at 95% probability (**Fig. 1A**), which is consistent with the experimental results obtained in a previous report[19]. In this model, the probability of RT-PCR assays to detect one copy of SARS-CoV-2 is only 26% (**Supplemental Fig. 1**). This finding may explain, at least in part, the low reported 47-56% detection rates of SARS-CoV-2 in positive samples by RT-PCR[17,22]. We further found that the LOD appears to be independent of the virus genome size. For genomes of 4 to 100kb, the detection limit remains 4.8 copies at 95% probability.

One way to elevate the sensitivity is to amplify multiple targets on the same virus genome in a multiplex PCR reaction, thereby increasing the frequency of occurrence in the mathematical model. Amplifying multiple targets has the advantage of potentially detecting fragments of degraded virus genome while withstanding sequencing variations, thus allowing for the detection of upcoming mutants. The amplification efficiency of multiplex PCR is critical for LOD. We estimated that the efficiency of our multiplex PCR technology is about 26% by using Unique Molecular Identifier (UMI)-labeled primers to count the amplified products after NGS sequencing (**Supplemental Fig. 2** and **Supplemental Table 1**). However, the amplification efficiency could be lower, and not all amplicons amplified successfully if the template used is one single strand of cDNA. Thus, more amplicons are potentially required for multiplex PCR to detect limited copies of viruses.

Mathematical model of multiplex-PCR-based detection method

We thus designed a panel of 172 pairs of multiplex PCR primers for the sensitive detection of SARS-CoV-2 (**Fig. 1B**). The average amplicon length is 99bp. The amplicons run across the entire SARS-CoV-2 genome with a 76bp gap (76±10bp) between each amplicon. Since the observed efficiency of multiplex PCR is about 26% in amplifying the four DNA strands of a pair of human chromosomes, we assumed an efficiency of 6% in amplifying a single-strand of cDNA. In addition, it has already been reported that 79% of variants are recovered when directly amplifying 600 amplicons from a single cell using our technology[44]. Therefore, we assume that 80% of amplicons would be amplified successfully. Using the same mathematical model described above, we estimated that our specific SARS-CoV-2-designed panel can detect 1.15 copies of the virus at 95% probability (**Fig. 1C**). Again, the LOD is independent of virus genome size.

We also designed a second pool of 171 multiplex PCR primers. These primers overlap with those in the previous 172-primer pool. Together, these two overlapping pools of primers deliver full coverage of the entire virus genome. If using both pools in detection, the calculated detection limit is 0.29 copies at 95% probability.

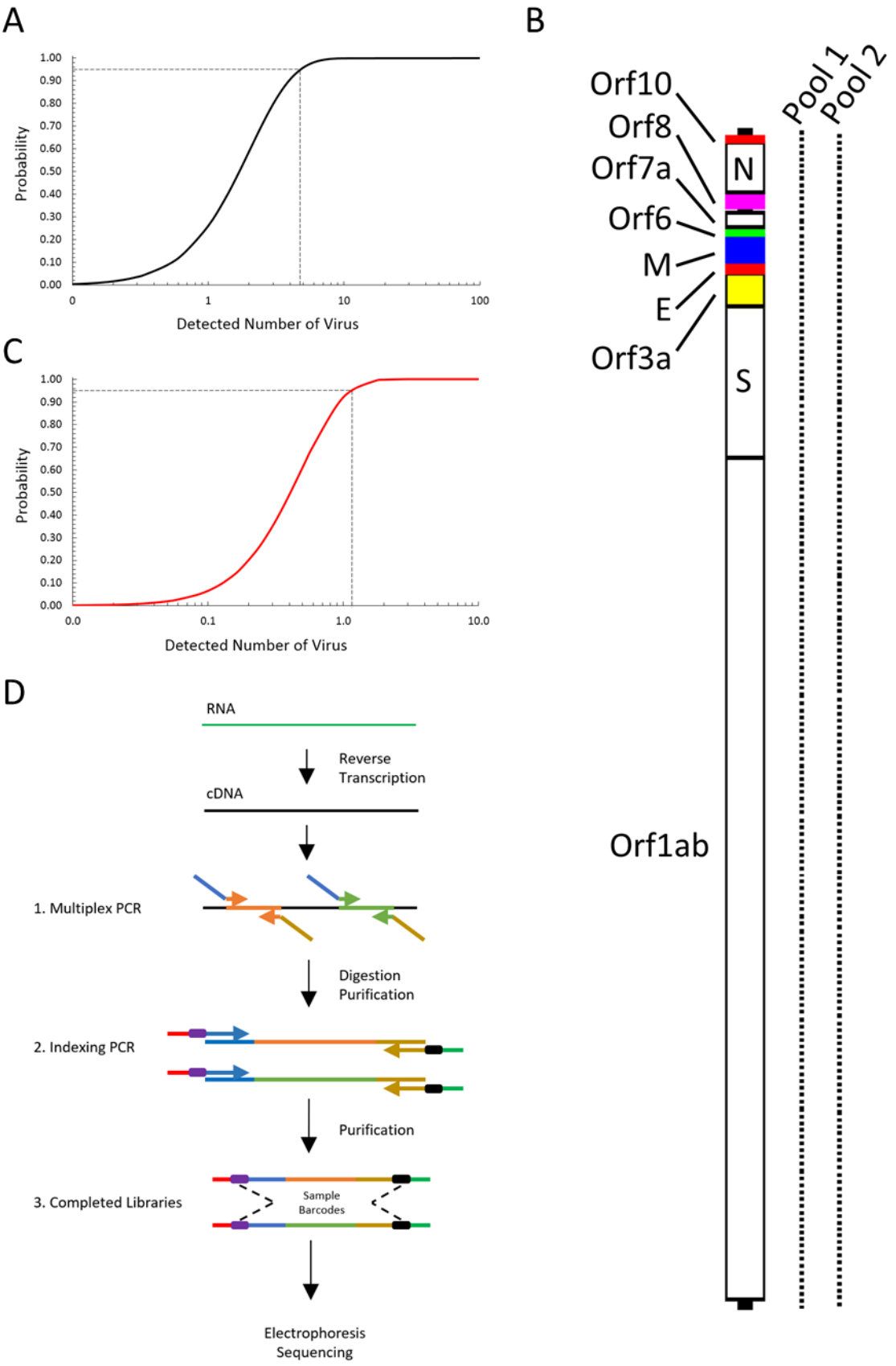


Figure 1. Mathematical model, primer design and workflow

(**A**) A mathematical model of RT-PCR based on Poisson process. The LOD is 4.8 copies of virus at 95% probability. (**B**) Two overlapping pools of multiplex PCR primers, shown on the right of the genome of SARS-CoV-2, were designed to span the entire virus genome. Pool 1, containing 172 pairs of primers, covers 56.9% of the viral genome and was used in the detection. Pool 2 contains 171 pairs of primers and covers 56.4% of the genome. Both pools are used to cover the full length of the genome. (**C**) A mathematical model of multiplex PCR with pool 1 of the primers. The LOD is 1.15 copies of virus at 95% probability. (**D**) The workflow of the multiplex PCR method. The prepared libraries can be detected using high-resolution electrophoresis, and sequenced together with other samples using high-throughput sequencing.

Detecting limited copies of SARS-CoV-2

The workflow was designed so that the multiplex PCR products are further amplified in a secondary PCR, while sample indexes and NGS sequencing primers are added (**Fig. 1D**). The PCR products were first analyzed by electrophoresis to identify potential positives. Since dozens of amplicons could be amplified from a single copy of SARS-CoV-2, electrophoresis peaks with defined peak sizes were expected. Multiplex PCR amplifies SARS-CoV-2, as well as other coronaviruses due to their high sequence similarities. In that context, electrophoresis analysis provides a fast and sensitive indication of infection from that family of viruses. In addition, the generated library allows for further investigation through NGS sequencing to provide definitive identification of the specific virus family member.

Two plasmids, containing the full sequence of S and N genes of SARS-CoV-2, respectively, were used to validate our multiplex PCR method. 28 amplicons are expected to be amplified within our 172-amplicon panel. To simulate the use of real clinical samples, these two plasmids were spiked into the cDNA made from human total RNA. The copy number of each plasmid was determined by droplet-based digital PCR in QX200 from Bio-Rad®[45]. The two plasmids were diluted from approximately 9,000 copies to below one copy, and were amplified in multiplex PCR reactions. The library peaks of expected sizes were obtained from 8,900 to 2.8 copies of plasmids (**Fig. 2A**). Quantification of peaks demonstrated a wide dynamic range from 1 to about 1,000 copies of plasmids (**Fig. 2B**). The yield of the libraries started to saturate when the copy number was 1,700. It is possible that the saturation point could be even lower when all of the 172 amplicons are amplified from positive clinical samples, and the library peak could be observed with even fewer copies of virus. In contrast, the detected quantities of a single target on N gene by RT-PCR rapidly dropped when using 2.85 copies (**Fig. 2B** and **Supplemental Fig. 3**).

Estimated from the aforementioned mathematical model, 28 amplicons have a 16% chance to detect one single copy. To test this hypothesis, we amplified about one copy of plasmid in multiplex PCR reactions. The theoretical calculation gives a 66% probability to sample 1.1 copies, and a 12% chance to detect them based on a multiplex PCR efficiency of 6%. In reality, we experimentally observed a significantly higher 56% probability to detect 1.1 copies (**Fig. 2C**). These results suggest that the efficiency of multiplex PCR is actually higher than the previously estimated 6% when single-stranded cDNA was amplified.

When the amplified products were sequenced, we found that the recovered reads were within a range of about 20-fold relative depth with about 1.4 to 2.8 plasmids, and uniformly distributed across the GC range (**Fig. 2D**). When detecting down to 1.4 copies of plasmids, only the reads of one amplicon were about 100-fold lower than the average. Approximately 96% of the amplicons were recovered with 14 copies of plasmids, 77% with 2.8 copies, and 37% with 0.6 copies (**Fig. 2E**).

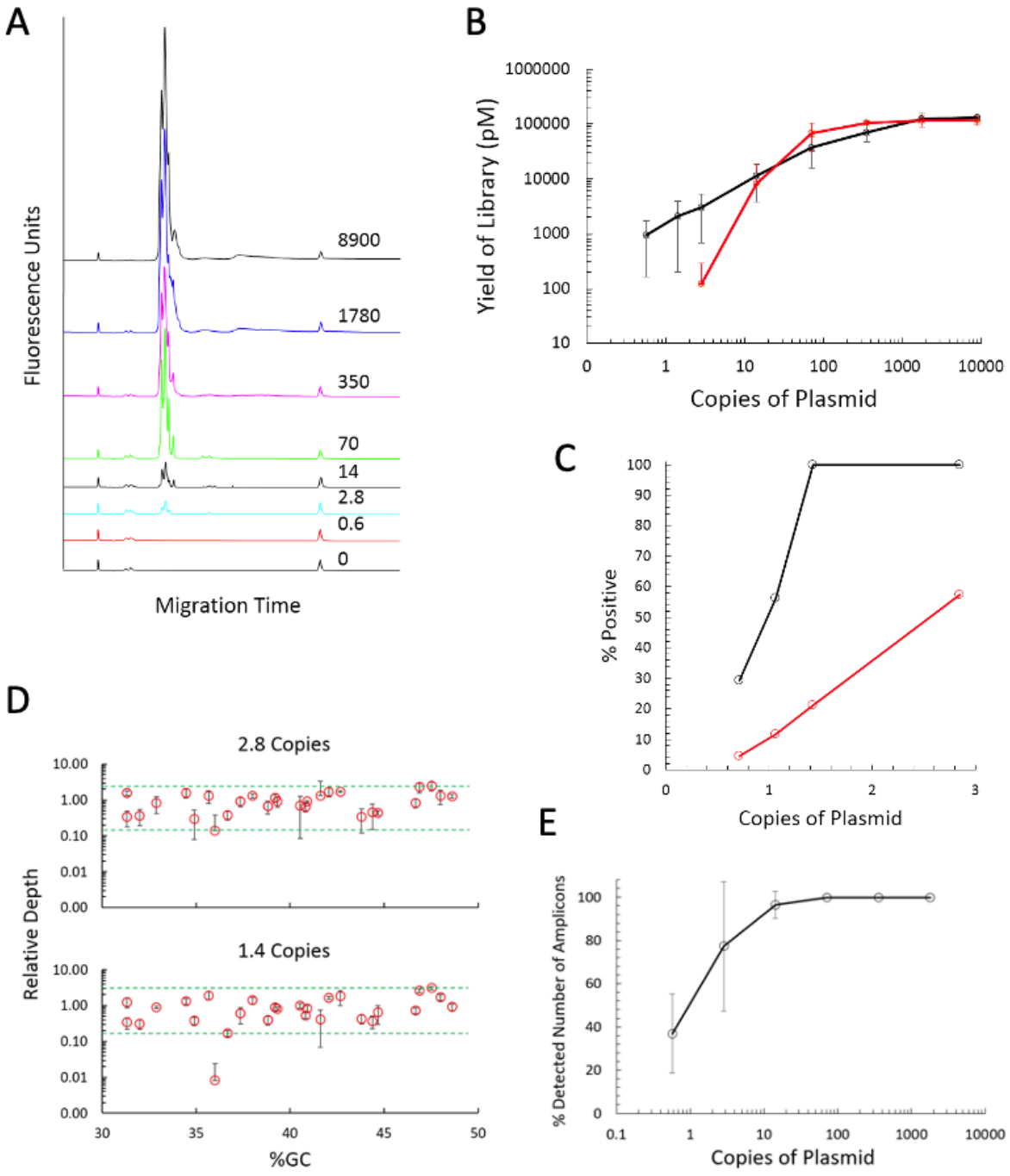


Figure 2. Detection of SARS-CoV-2 gene-containing plasmids by electrophoresis and sequencing

(**A**) Two plasmids, containing S and N genes of SARS-CoV-2, respectively, were diluted in human cDNA and amplified in multiplex PCR with pool 1 (172 pairs of primers). The number of plasmid copies per reaction, determined by ddPCR, were from 8,900 to 0.6. The resulting products obtained after multiplex PCR were resolved by electrophoresis. The specific amplification products (the library) can still be seen with 2.8 copies of plasmids. (**B**) The library yields can be detected down to 0.6 copies of plasmids (*n*=4) by multiplex PCR (black line), while only down to 2.8 copies can still be detected by RT-PCR (> 4.5-fold difference) (red

line). (**C**) Poisson process was used to estimate the chance of sampling around 1 copy of viral particles, and the mathematical model was used to estimate the chance of detecting them (red line). There is 12% of probability to detect 1.1 copies with a multiplex PCR efficiency of 6%. In reality, we observed a significantly higher 56% probability for 1.1 copies and 100% probability for 1.4 copies. (**D**) After sequencing 1.4 to 2.8 copies of plasmids, the reads of all 28 amplicons spanning both N and S genes were clustered within a 20-fold range of coverage (*n*=3). With 1.4 copies, only the reads of one amplicon were about 100-fold lower than the average (*n*=3). (**E**) About 96% of amplicons were recovered with 14 copies of plasmids, 77% with 2.8 copies, and 37% with 0.6 copies (*n*=3).

Metagenomic method design for novel pathogens

In order to discover highly mutated viruses and unknown pathogens, we subsequently developed a user-friendly multiplex-PCR-based metagenomic method. In this method, random hexamer-adapters were used to amplify DNA or cDNA targets in a multiplex PCR reaction. The large amounts of non-specific amplification products were removed by using Paragon Genomics' proprietary background removing reagent, thus resolving a library suitable for sequencing. For RNA samples, our reverse transcription reagents were additionally used to convert RNA into cDNA, resulting in significantly reduced amount of human ribosomal RNA species.

We sequenced a library made with 4,500 copies of N and S gene-containing plasmids spiked into 10 ng of human gDNA, which roughly represents 3,300 haploid genomes. Even though the molar ratios of viral targets and human haploid genomes were comparable, N and S genes which encompass about 4kb of targets, were a negligible fraction of the 3 billion base pairs of a human genome. If every region of the human genome were amplified and sequenced at 0.6 million reads per sample, only one read of viral target would be recovered. In fact, our results showed that 16% of the recovered bases, or 13% of the recovered reads, were within the viral N and S genes (**Fig. 3A** and **Supplemental Table 2**). 80% and 78% of SARS-CoV-2 and mitochondrial targets were covered, respectively (**Fig. 3B**), and the base coverage was significantly higher than human targets (**Fig. 3C**). In contrast, only 0.08% of regions in human chromosomes were amplified. Furthermore, the human exonic regions were preferentially amplified (**Fig. 3D**). This suggested that the random hexamers deselected a large portion of the human genome, while favorably amplifying regions that were more "random" in base composition. Indeed, long gaps and absence of coverage in very large repetitive regions were observed in human chromosomes (**Fig. 3E**). On the contrary, the gaps in SARS-CoV-2 and mitochondrial regions were significantly shorter (**Fig. 3F**), whereas the amplified targets overlapped and were longer than human targets (**Fig. 3G**).

The coverage was from 1000- to 10,000-fold for S and N genes, and 30- to 500-fold for the mitochondrial chromosome (**Fig. 3F**). The coverage is roughly within a 10-fold range, as is observed in human chromosomes (**Fig. 3E**). Therefore, increasing sequencing depth might not significantly improve the coverage. This 10-fold difference in coverage has been routinely observed with our multiplex PCR technology (**Supplemental Table 3** and **Supplemental Fig. 4**). We were able to detect 80% of the regions in S and N genes in libraries generated using 4,500 copies of plasmids, with an average base coverage of 5,000 for a total sequencing depth of 0.6 million reads. A few targets of 150 to 200 bp in length in S and N genes were preferentially amplified. Even when copy number went down to a few copies (3-14), these targets were still detected (**Fig.4**). They represented 14% and 11% of the target regions when 14 and 2.8 copies of plasmids were amplified, respectively.

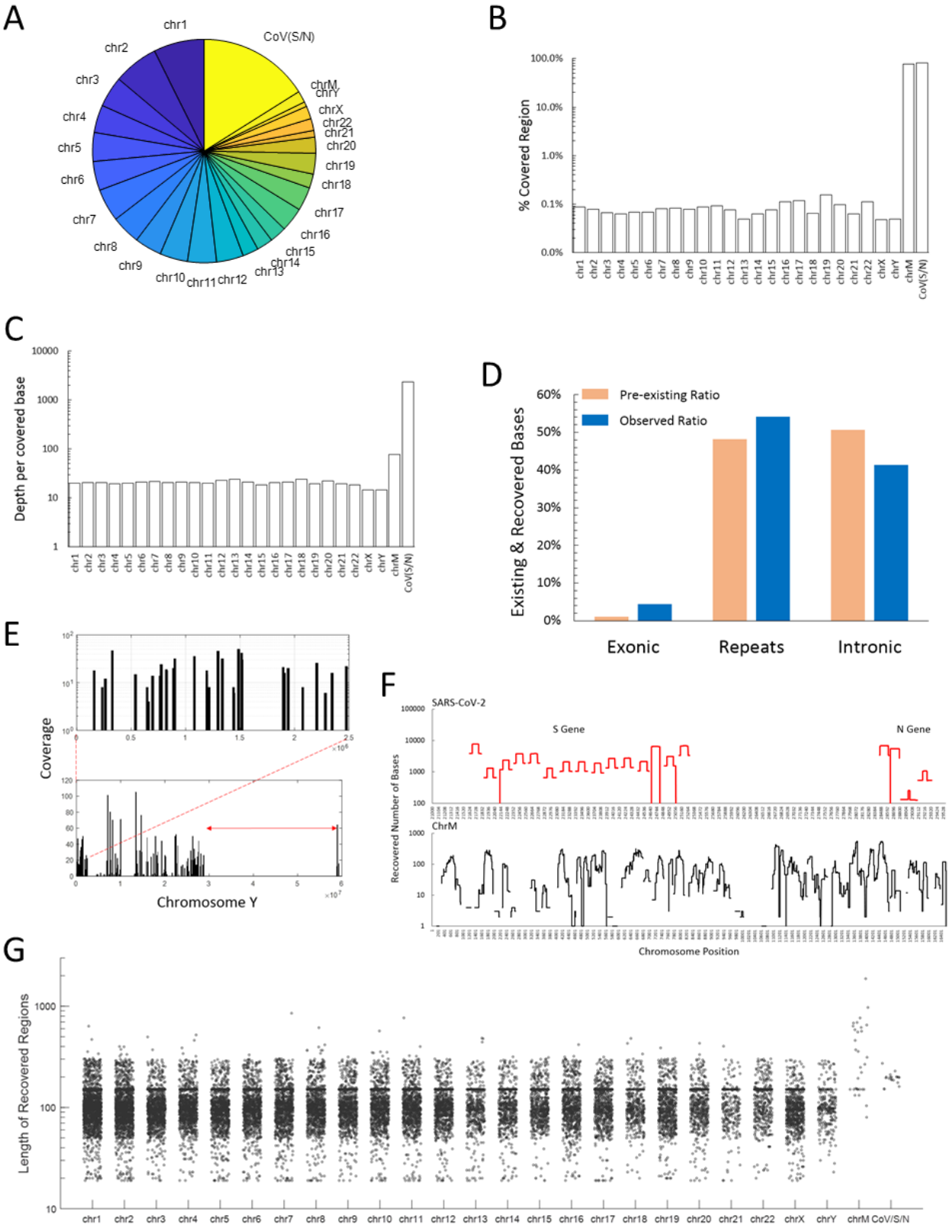


Figure 3. Multiplex PCR-based metagenomic method for the detection of SARS-CoV-2 genes

(**A**) Random hexamer-adapters were used in multiplex PCR to amplify 4,500 copies of plasmids in the background of 3,300 haploid human gDNA molecules. The resulting libraries were sequenced at an average of 0.6 million total reads. Of the total bases recovered, 16% were on SARS-CoV-2 S and N genes. (**B**) 80% of S and N genes, and 78% of the human mitochondrial chromosome were amplified with >= 1X coverage, while only 0.08% of the human chromosomes were. (**C**) On average, S and N genes were covered at 2,346X, mitochondria at 77X, and human chromosomes at 20X. (**D**) Human exons were relatively over-amplified about 4-fold higher compared to their actual ratio within the genome. (**E**) Gaps and long regions of absence of amplification were observed for human chromosomes. An example shown here is chromosome Y. Small gaps were additionally found in the enlarged cluster of amplification. The long absent region (red double arrow) overlapped with the repetitive regions on Y chromosome. (**F**) Representation of the recovered regions in S and N genes and the human mitochondrial chromosome. The coverage was from 1,000- to 10,000-fold for S and N genes, and 30- to 500-fold for the mitochondrial chromosome. (**G**) The length of the majority of chromosomal amplification products are clustered around 100bp, while the amplified regions for the S and N genes, as well as the mitochondrial chromosome, were significantly longer.

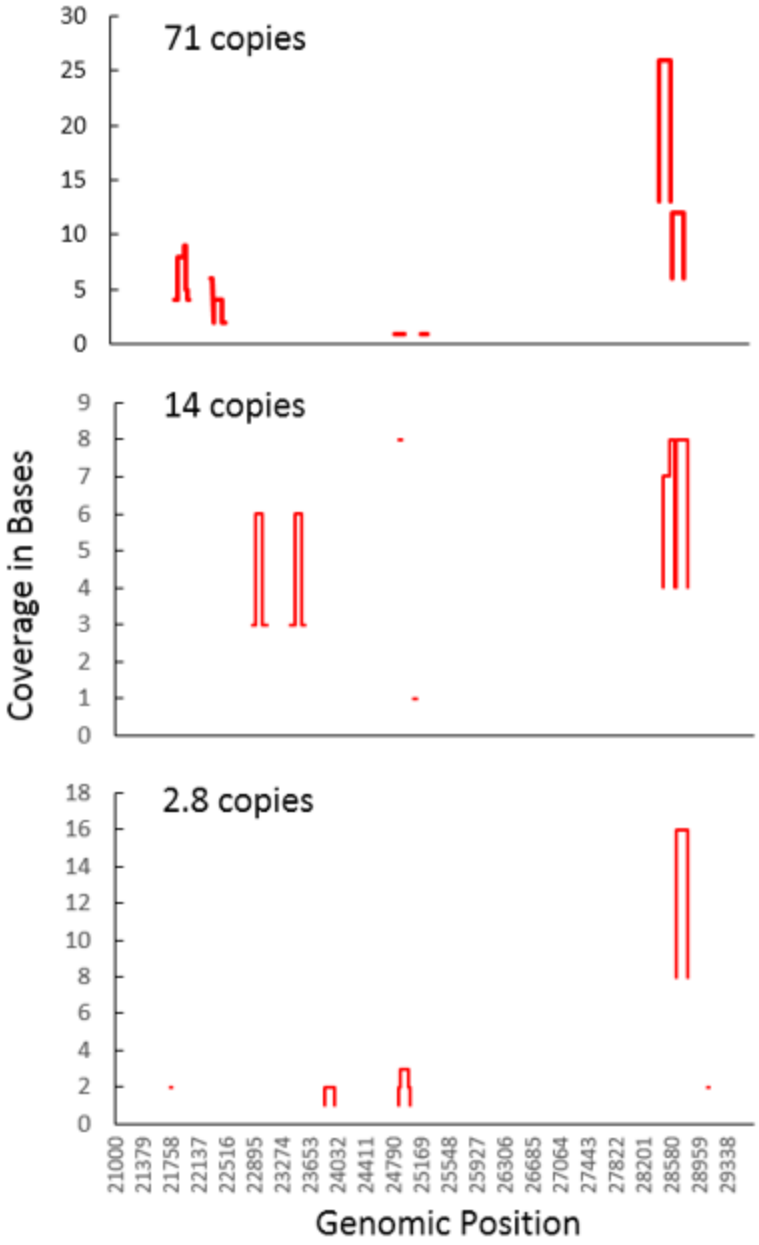


Figure 4. Metagenomic method for the detection of limited copies of SARS-CoV-2 genes

Even when low number of plasmids were used (3 to 14 copies), several targets of 150 to 200 bp in S and N genes were still detected at a total sequencing depth of about 1 million reads. These targets represented 19%, 14%, and 11% of S and N genes when 71, 14 and 2.8 copies of plasmids were used, respectively.

Discussion

Tremendous efforts have been made from around the world to provide a fast and reliable diagnostic method for SARS-CoV-2. RT-PCR is currently the preferential and most frequently used detection strategy. Yet, we found that the LOD for RT-PCR sits at around 5 copies with 95% confidence, independent of the size of the target genome. Consequently, to overcome the instability of RNA and the genomic sequence variability of virus genome due to evolution, more sensitive and robust methods are required.

Here, we report the development of a multiplex PCR assay for high-sensitivity identification of SARS-CoV-2 infection. With 172 pairs of primers, this method enables the detection of low copy numbers, potentially degraded fragments of the viral genome, or even a mutated variant, with high confidence. The 172 pairs of primers cover about 56% of the genome of SARS-CoV-2. When detecting limited copies of virus, fewer targets are successfully amplified. In the case of one copy of virus, we estimate that 20% of the viral genome, or 6kb of sequences, could be amplified. These are sufficient to produce a conspicuous peak in electrophoresis for an initial indication of positives. The use of NGS sequencing can provide nucleic acid-level information to further confirm the identification, and provide additional evidence for phylogenetic analysis. In addition, our method is robust, accurate and easily performed from different levels of expertise in various laboratory settings.

Furthermore, we also propose a metagenomic method for the potential identification of unknown pathogens. Metagenomic technology usually requires about 20 to 100 million of sequencing reads in order to detect minuscule numbers of targets embedded in the massive amounts of human background. However, our method appears to selectively amplify target sequences. This bias permits the obtention of around 16% of target bases by sequencing at a depth of about 1 million total reads. It is thus especially suitable for the detection of novel pathogens and highly genetically unstable pathogens, such as influenza viruses. The exact mechanism of deselecting human sequences is currently under investigation. The random hexamer preference, chromosomal structure, sequence composition, target length, circularity, methylation status, telomeric and centromeric regions, as well as the edge effect, may influence such outcome. Ultimately, the observed preferential amplification towards more “random” sequences could provide us with an advantageous edge for the continuous improvement of this method.

Materials and Methods

Sample preparation

The Universal Human Reference RNA was from Agilent Technologies, Inc. (Cat#74000). The plasmids containing either S or N gene of SARS-CoV-2 (pUC-S and pUC-N, respectively) were purchased from Sangon Biotech, Shanghai, China. The PCR primers used in ddPCR and RT-PCR reactions for S gene are 5'-TGTACTTGGACAATCAAAAAGAGTTGAT and 5'-AGGAGCAGTTGTGAAGTTCTTTTC; for N gene are 5'-GGGGAACTTCTCCTGCTAGAAT and 5'-CAGACATTTTGCTCTCAAGCTG, respectively. 343 pairs of multiplex PCR primers covering the entire genome of SARS-CoV-2 (the panel) were designed by Paragon Genomics, Inc. and separated into two pools. Pool 1, containing 172 pairs of primers, was used in the detection of SARS-CoV-2 by multiplex PCR.

Reverse transcription

50ng of Universal Human Reference RNA was converted into cDNA using random primers and SuperScript™ IV Reverse Transcriptase by following the supplier recommended method (Thermo Fisher Scientific, Cat# 18090050). After reverse transcription, cDNA was purified with 2.4X volume of magnetic beads, and washed twice with 70% ethanol. Finally, the purified cDNA was dissolved in 1X TE buffer and used per multiplex PCR reaction.

Multiplex PCR panel design

Panel design is based on the SARS-CoV-2 sequence NC_045512.2 (https://www.ncbi.nlm.nih.gov/nuccore/NC_045512.2/). In total, 343 primer pairs, distributed into two pools, were selected by a proprietary panel design pipeline to cover the whole genome except for 92 bases at its ends. Primers were optimized to preferentially amplify the SARS-CoV-2 cDNA versus background human cDNA or genomic DNA. They were also optimized to uniformly amplify the covered genome.

Multiplex PCR

Plasmids pUC-S and pUC-N were combined with human cDNA and used in each multiplex PCR reaction. Paragon Genomics' CleanPlex® multiplex PCR reagents and protocol were used. Briefly, a 10µl multiplex PCR reaction was made by combining 5X mPCR mix, 10X Pool 1 of the panel, plasmid pUC-S, pUC-N and cDNA. The reaction was run in a thermal cycler (95°C for 10min, then 98°C for 15sec, 60°C for 5min for 10 cycles), then terminated by the addition of 2µl of stop buffer. The reaction was then purified by 29µl of magnetic beads, followed by a secondary PCR with a pair of primers for 25 cycles. The secondary PCR added sample indexes and sequencing adapters, allowing for sequencing of the resulting products by high throughput sequencing. A final bead purification was performed after the secondary PCR, followed by library interrogation using a Bioanalyzer 2100 instrument with Agilent High Sensitivity DNA Kit (Agilent Technologies, Inc. Part# 5067-4626).

RT-PCR

Plasmids pUC-S and pUC-N, in combination with human cDNA, were used in each reaction. Paragon Genomics' CleanPlex® secondary PCR mix was used with 100nM of each PCR primers in 10ul reactions.

The PCR thermal cycling protocol used was 95°C for 10min, then 98°C for 15sec, 60°C for 30sec for 45 cycles.

Multiplex-PCR-based metagenomic method

Paragon Genomics' CleanPlex® metagenomic reagents and protocol were used. Briefly, a 10µl multiplex PCR reaction was made by combining 5X mPCR mix, 10X random hexamer-adapters and the template DNA. The PCR thermal cycling protocol used was 95°C for 10min, then 98°C for 15sec, 25°C for 2min, 60°C for 5min for 10 cycles. The reaction was then terminated by the addition of 2µl of stop buffer, and purified by 29µl of magnetic beads. The resulting solution was treated with 2µl of CleanPlex® reagent at 37°C for 10min to remove non-specific amplification products. After a magnetic bead purification, the product was further amplified in a secondary PCR with a pair of primers for 25 cycles to produce the metagenomic library. This metagenomic library was further purified by magnetic beads before sequencing.

ddPCR

ddPCR was performed on QX200 from Bio-Rad®. Plasmids pUC-S and pUC-N at the estimated copy numbers 1 (6 repeats), 2 (3 repeats), and 100 (3 repeats) were tested. In each reaction, the ddPCR thermal cycling protocol used was 95°C for 5min, then 95°C for 30sec, 60°C for 1min with 60 cycles, 4°C for 5min and 90°C for 5min, 4°C hold. The resulting data was analyzed by following the supplier recommended method.

High throughput sequencing and data analysis

High throughput sequencing was performed using Illumina® iSeq™ 100 and MiSeq™ Sequencing Systems and MGI sequencers (DNBSEQ-G400 and its research-grade CoolMPS™ sequencing kits). The resulting data was analyzed by Paragon Genomics' Bioinformatics team with proprietary pipelines and algorithms.

References

1. Li, Q., et al., *Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia.* N Engl J Med, 2020.
2. Zhou, P., et al., *A pneumonia outbreak associated with a new coronavirus of probable bat origin.* Nature, 2020.
3. *[An update on the epidemiological characteristics of novel coronavirus pneumoniaCOVID-19].* Zhonghua Liu Xing Bing Xue Za Zhi, 2020. **41**(2): p. 139-144.
4. Ayittey, F.K., et al., *Economic Impacts of Wuhan 2019-nCoV on China and the World.* J Med Virol, 2020.
5. Wu, A., et al., *Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China.* Cell Host Microbe, 2020.
6. Zhu, N., et al., *A Novel Coronavirus from Patients with Pneumonia in China, 2019.* N Engl J Med, 2020.
7. Paraskevis, D., et al., *Full-genome evolutionary analysis of the novel corona virus (2019-nCoV) rejects the hypothesis of emergence as a result of a recent recombination event.* Infect Genet Evol, 2020. **79**: p. 104212.
8. Lu, R., et al., *Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding.* Lancet, 2020.
9. Ceraolo, C. and F.M. Giorgi, *Genomic variance of the 2019-nCoV coronavirus.* J Med Virol, 2020.
10. Yu, F., et al., *Measures for diagnosing and treating infections by a novel coronavirus responsible for a pneumonia outbreak originating in Wuhan, China.* Microbes Infect, 2020.
11. Zhang, N., et al., *Recent advances in the detection of respiratory virus infection in humans.* J Med Virol, 2020.
12. Zhang, W., et al., *Molecular and serological investigation of 2019-nCoV infected patients: implication of multiple shedding routes.* Emerg Microbes Infect, 2020. **9**(1): p. 386-389.
13. Lamb, L.E., et al., *Rapid Detection of Novel Coronavirus (COVID-19) by Reverse Transcription-Loop-Mediated Isothermal Amplification.* medRxiv, 2020: p. 2020.02.19.20025155.
14. Yu, L., et al., *Rapid colorimetric detection of COVID-19 coronavirus using a reverse tran-scriptional loop-mediated isothermal amplification (RT-LAMP) diagnostic plat-form: iLACO.* medRxiv, 2020: p. 2020.02.20.20025874.
15. To, K.K., et al., *Consistent detection of 2019 novel coronavirus in saliva.* Clin Infect Dis, 2020.
16. Chen, W., et al., *Detectable 2019-nCoV viral RNA in blood is a strong indicator for the further clinical severity.* Emerg Microbes Infect, 2020. **9**(1): p. 469-473.
17. Xie, C., et al., *Comparison of different samples for 2019 novel coronavirus detection by nucleic acid amplification tests.* Int J Infect Dis, 2020.
18. Lin, C., et al., *Comparison of throat swabs and sputum specimens for viral nucleic acid detection in 52 cases of novel coronavirus (SARS-Cov-2) infected pneumonia (COVID-19).* medRxiv, 2020: p. 2020.02.21.20026187.
19. Corman, V.M., et al., *Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR.* Euro Surveill, 2020. **25**(3).
20. Chu, D.K.W., et al., *Molecular Diagnosis of a Novel Coronavirus (2019-nCoV) Causing an Outbreak of Pneumonia.* Clin Chem, 2020.
21. Shirato, K., et al., *Development of Genetic Diagnostic Methods for Novel Coronavirus 2019 (nCoV-2019) in Japan.* Jpn J Infect Dis, 2020.
22. Ai, T., et al., *Correlation of Chest CT and RT-PCR Testing in Coronavirus Disease 2019 (COVID-19) in China: A Report of 1014 Cases.* Radiology, 2020: p. 200642.

23. Xie, X., et al., *Chest CT for Typical 2019-nCoV Pneumonia: Relationship to Negative RT-PCR Testing.* Radiology, 2020: p. 200343.
24. Li, Y.Y., et al., *[Comparison of the clinical characteristics between RNA positive and negative patients clinically diagnosed with 2019 novel coronavirus pneumonia].* Zhonghua Jie He He Hu Xi Za Zhi, 2020. **43**(0): p. E023.
25. Zhang, Y., et al., 病毒核酸提取前的高温灭活过程显著降低可检出病毒核酸模板量. chinaXiv:202002.00034v1.
26. Dennis Lo, Y.M. and R.W.K. Chiu, *Racing towards the development of diagnostics for a novel coronavirus (2019-nCoV).* Clin Chem, 2020.
27. Liu, Y., et al., *Clinical and biochemical indexes from 2019-nCoV infected patients linked to viral loads and lung injury.* Sci China Life Sci, 2020.
28. Bernard Stoecklin, S., et al., *First cases of coronavirus disease 2019 (COVID-19) in France: surveillance, investigations and control measures, January 2020.* Euro Surveill, 2020. **25**(6).
29. Thompson, R.N., *Novel Coronavirus Outbreak in Wuhan, China, 2020: Intense Surveillance Is Vital for Preventing Sustained Transmission in New Locations.* J Clin Med, 2020. **9**(2).
30. Reusken, C., et al., *Laboratory readiness and response for novel coronavirus (2019-nCoV) in expert laboratories in 30 EU/EEA countries, January 2020.* Euro Surveill, 2020.
31. Lin, L. and T.S. Li, *[Interpretation of "Guidelines for the Diagnosis and Treatment of Novel Coronavirus (2019-nCoV) Infection by the National Health Commission (Trial Version 5)"].* Zhonghua Yi Xue Za Zhi, 2020. **100**(0): p. E001.
32. *[Diagnosis and clinical management of 2019 novel coronavirus infection: an operational recommendation of Peking Union Medical College Hospital (V2.0)].* Zhonghua Nei Ke Za Zhi, 2020. **59**(3): p. 186-188.
33. Cleemput, S., et al., *Genome Detective Coronavirus Typing Tool for rapid identification and characterization of novel coronavirus genomes.* Bioinformatics, 2020.
34. Shen, K., et al., *Diagnosis, treatment, and prevention of 2019 novel coronavirus infection in children: experts' consensus statement.* World J Pediatr, 2020.
35. Malik, Y.S., et al., *Emerging novel Coronavirus (2019-nCoV) - Current scenario, evolutionary perspective based on genome analysis and recent developments.* Vet Q, 2020: p. 1-12.
36. Li, X., et al., *Potential of large 'first generation' human-to-human transmission of 2019-nCoV.* J Med Virol, 2020.
37. Li, X., et al., *Transmission dynamics and evolutionary history of 2019-nCoV.* J Med Virol, 2020.
38. Ji, W., et al., *Homologous recombination within the spike glycoprotein of the newly identified coronavirus may boost cross-species transmission from snake to human.* J Med Virol, 2020.
39. Benvenuto, D., et al., *The 2019-new coronavirus epidemic: Evidence for virus evolution.* J Med Virol, 2020.
40. Benvenuto, D., et al., *The global spread of 2019-nCoV: a molecular evolutionary analysis.* Pathog Glob Health, 2020: p. 1-4.
41. Giovanetti, M., et al., *The first two cases of 2019-nCoV in Italy: Where they come from?* J Med Virol, 2020.
42. Tessema, S.K., et al., *Sensitive, highly multiplexed sequencing of microhaplotypes from the Plasmodium falciparum heterozygome.* bioRxiv, 2020: p. 2020.02.25.964536.
43. Chen, L., et al., *RNA based mNGS approach identifies a novel human coronavirus from two individual pneumonia cases in 2019 Wuhan outbreak.* Emerg Microbes Infect, 2020. **9**(1): p. 313-319.
44. Ericson, N.,et al., *Amplicon-Based Targeted Sequencing of Single Circulating Tumor Cells, Journal of Molecular Diagnostics*, AMP Abstract, 1168

45. Dong, L., et al., *Comparison of four digital PCR platforms for accurate quantification of DNA copy number of a certified plasmid DNA reference material.* Scientific Reports, 2015. **5**(1): p. 13174.

Acknowledgements

The authors would like to thank Jian Xu of MGI, BGI-Shenzhen for technical support in MGI sequencing, Dr. Jin Billy Li of Stanford University for the coordination of academic support, Dr. Alexander E. Urban of Stanford University for suggestions in preparing the manuscript, and Dr. Zihuai He of Stanford University for discussion on a potential mathematical model of the metagenomic method.

Author contributions

C.L., D.N.D., Z.L. conceived the study and drafted the manuscript. C.L., D.N.D., J.S., V.K., L.Y.L., L.L., R.F., G.B., J.L., A.O., G.L., Z.L. performed experiments, analysis and revised the manuscript. B.Z., A.E.U. performed ddPCR experiments and analysis. A.B., H.T. performed MGI sequencing and analysis.

Conflict of Interest

The authors declare no competing interests.

Additional files

Supplemental Fig 1. A mathematical model of RT-PCR.

Supplemental Fig 2. Multiplex PCR efficiency as determined by using CleanPlex® UMI technology by Paragon Genomics.

Supplemental Fig 3. Comparison of LOD between multiplex PCR and regular PCR.

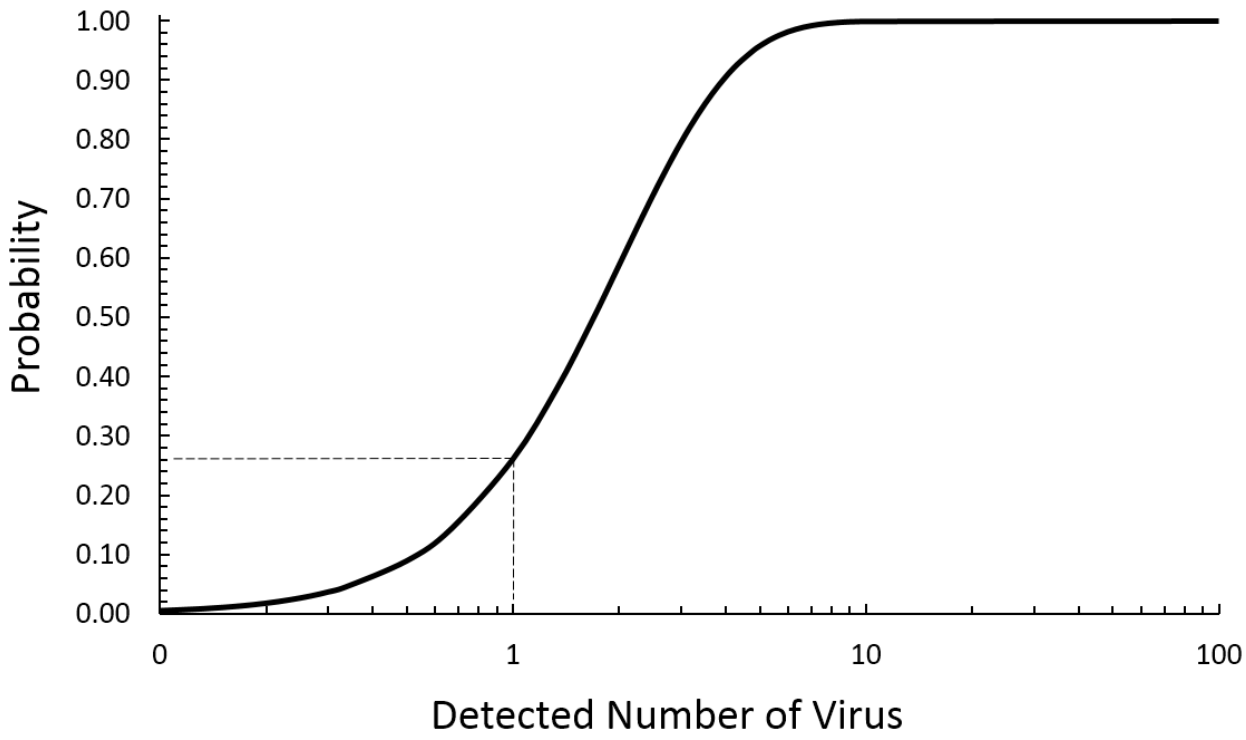
Supplemental Fig 4. Performance statistics of the amplicons retrieved from multiplex PCR method highlighting a 10-fold range read depth.

Supplemental Table 1. Multiplex PCR efficiency as determined by using CleanPlex® UMI technology by Paragon Genomics.

Supplemental Table 2. Sequencing results of the multiplex PCR-based metagenomic method using 4,500 copies of plasmids containing S and N genes of SARS-CoV-2, spiked in 10ng of human gDNA.

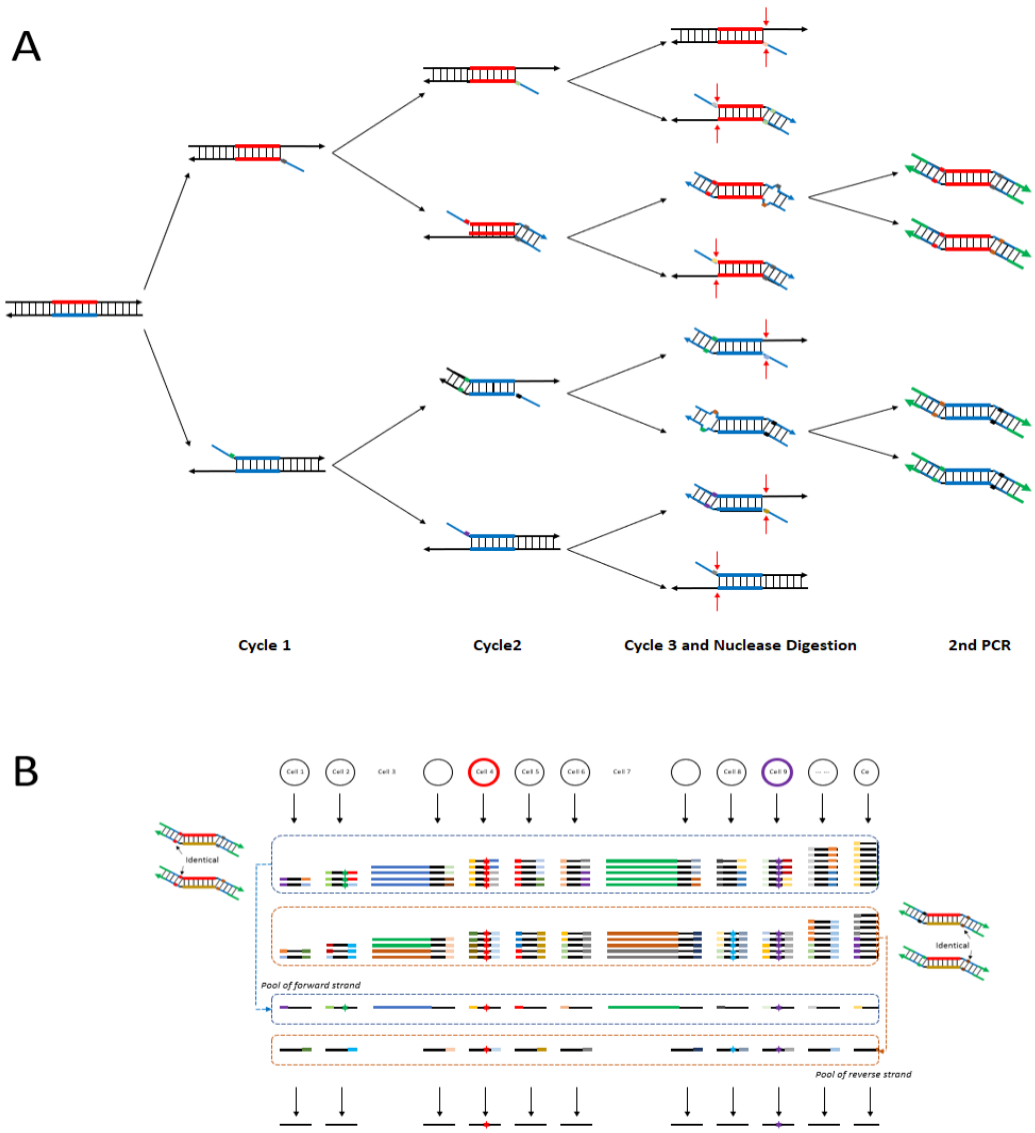
Supplemental Table 3. Performance statistics of the amplicons retrieved from our multiplex PCR method highlighting a 10-fold range read depth.

Supplementary Information



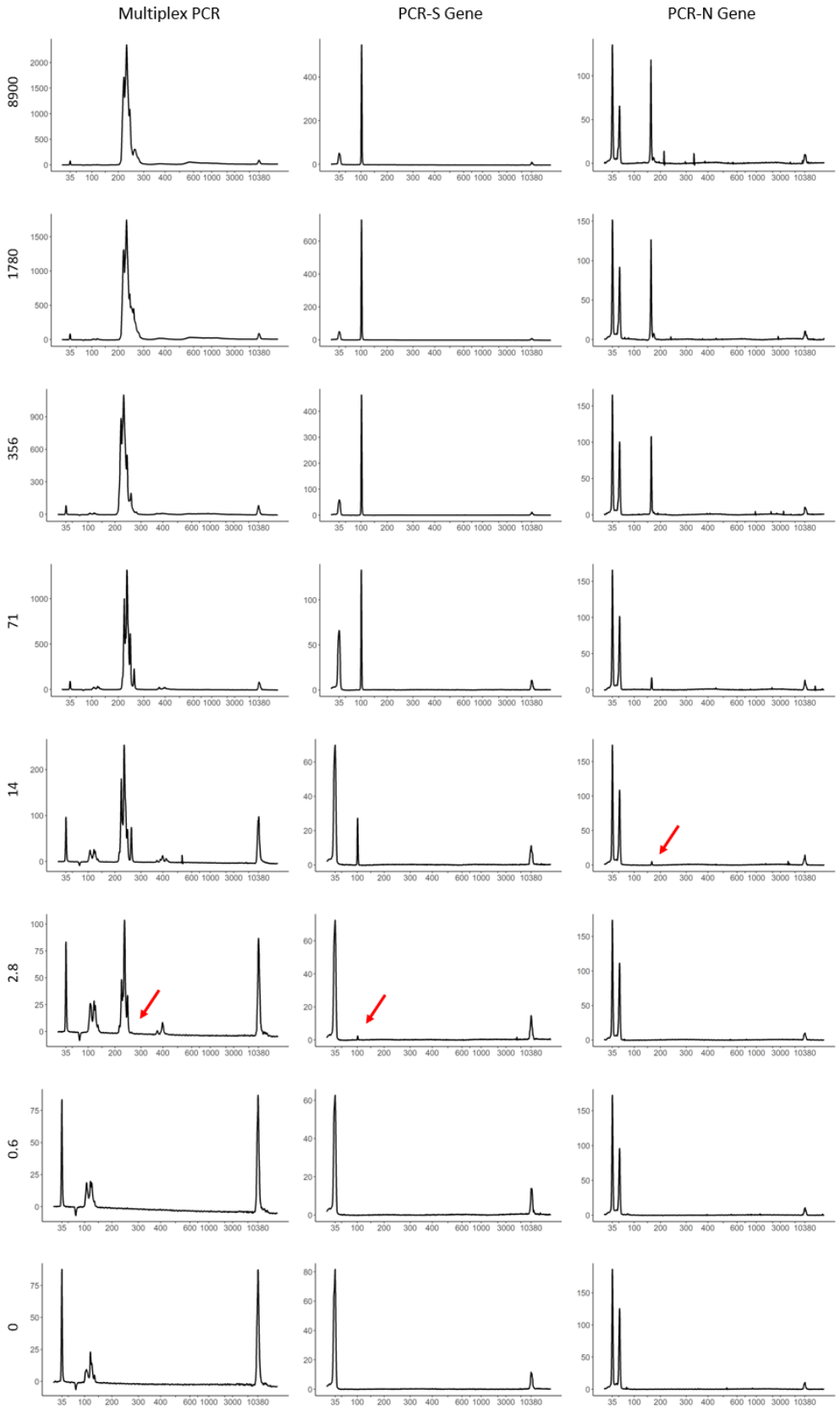
Supplemental Fig 1. A mathematical model of RT-PCR.

The same model was used to estimate the LOD of both RT-PCR and multiplex PCR, through changing the amplicon length and number, the virus genome size, as well as the intended detected copies and PCR efficiency. We found that the probability of detecting 1 copy of SARS-CoV-2 is 26% by using RT-PCR, and the LOD is independent of the length of virus genome.



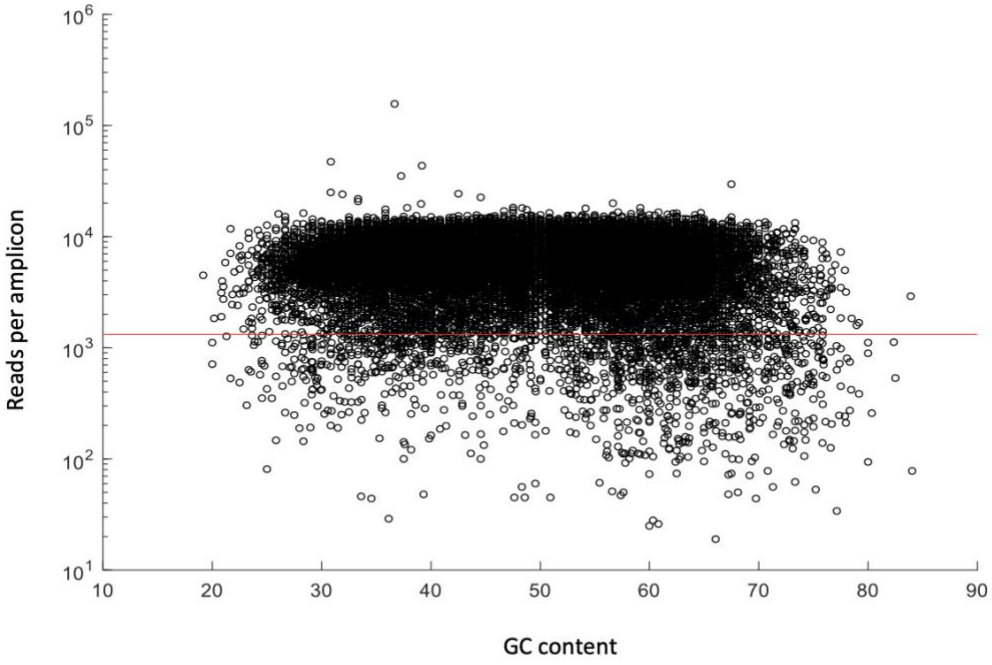
Supplemental Fig 2. Multiplex PCR efficiency as determined by using CleanPlex® UMI technology by Paragon Genomics.

A. The underlying mechanism of CleanPlex® UMI technology by Paragon Genomics, which uses three cycles of multiplex PCR to label targets with UMI. The redundant UMIs generated in the third cycle of PCR are destroyed by removing the single-stranded regions by nuclease digestion. The resulting products are further amplified by using a pair of universal primers, while sample indexes and sequencing adapters are introduced. **B.** The UMIs are initially sorted based on the UMI itself, and further on the occurrence of identical UMIs on either the 5′ or 3′ end of amplicons after sequencing, thus allowing the identification of the original template. The position of identical UMIs on either the 5′ or 3′ end of amplicons can further indicate whether the final amplification products are from the pool of the sense or antisense strand of the original templates.



Supplemental Fig 3. Comparison of LOD between multiplex PCR and regular PCR.

A total of 35 cycles was used in multiplex PCR, while 45 cycles was used for regular PCR. The resulting amplification products from multiplex PCR were processed as described in the Materials and Methods. The PCR products were directly resolved using high sensitivity DNA chips on a Bioanalyzer 2100 instrument. X-axis indicates fragment size (bp) and y-axis indicates fluorescence units. The arrows point to the expected specific amplification products. The number of copies is indicated on the left.



Supplemental Fig 4. Performance statistics of the amplicons retrieved from multiplex PCR method highlighting a 10-fold range read depth.

The number of sequencing reads for a majority of the recovered amplicons (Supplemental Table 2) were within a 10-fold range, representing a uniformity of 92.62 ± 1.96% at 0.2X mean (red line).

Supplemental Table 1. Multiplex PCR efficiency as determined by using CleanPlex® UMI technology by Paragon Genomics.

A 53-amplicon panel was amplified with 20 to 80 ng of gDNA (NA12878) by using CleanPlex® UMI technology by Paragon Genomics, the mechanism of which is depicted in Supplemental Fig 2. The efficiency of multiplex PCR was found to be 26%, calculated from the recovered numbers of UMI clusters that contained >= 3 members (UMI (>=3)).

DNA ng	Uniformity 0.2X Mean (%)	Total UMI	UMI (=1)	UMI (>=3)	UMI (>=5)	Raw Reads	Mapped Reads	On-target Reads	Total UMI Reads	UMI (>=3) Reads	Haploid Genomes	UMI (>=3) eff.	%UMI (>=3)
20	98.1%	530874	165687	351904	322511	5508532	4321171	4014880	3551662	3359409	1660	25.2%	66.3%
20	100.0%	448374	96774	327229	252561	3430557	2727335	2557727	2351544	2206028	1544	23.4%	73.0%
20	98.1%	613434	236067	364231	353500	8939089	7048505	6572606	5903206	5640867	1718	26.0%	59.4%
30	96.2%	982983	401164	560890	527993	9037026	7765886	7485868	6561378	6118356	2646	26.7%	57.1%
30	98.1%	812441	236456	553325	496326	7485107	6254270	5877149	5189318	4907542	2610	26.4%	68.1%
30	98.1%	722872	128591	580862	556866	9854077	7727106	7714642	7487927	7332504	2740	27.7%	80.4%
30	98.1%	709433	105870	587582	549015	8472590	6610966	6599497	6411865	6274035	2772	28.0%	82.8%
30	98.1%	693711	141968	507126	371323	4700397	3975741	3768653	3458895	3227693	2392	24.2%	73.1%
30	98.1%	929401	333715	575663	557579	12185504	10230011	9645802	8648217	8274456	2715	27.4%	61.9%
50	96.2%	1521124	504587	970308	841045	12240066	10145239	9795805	8610343	8013298	4577	27.7%	63.8%
80	96.2%	2055833	623322	1344679	1078860	14037982	12323533	11897105	10482966	9683980	6343	24.0%	65.4%
Avg.												26.1%	68.3%
STDEV												1.7%	8.2%

Supplemental Table 2. Sequencing results of the multiplex PCR-based metagenomic method using 4,500 copies of plasmids containing S and N genes of SARS-CoV-2, spiked in 10ng of human gDNA.

	Covered Region (Bases)	Recovered Bases	Existing Bases	%Base Covered	Depth per Base	%Base Recovered	# Continuous Regions	Max Length
chr1	217690	4374253	249250621	0.087%	20	7.36%	1871	635
chr2	189593	3878281	243199373	0.078%	20	6.52%	1677	400
chr3	130594	2644455	198022430	0.066%	20	4.45%	1195	498
chr4	121131	2380497	191154276	0.063%	20	4.00%	1084	521
chr5	124355	2477809	180915260	0.069%	20	4.17%	1144	298
chr6	117578	2507731	171115067	0.069%	21	4.22%	1039	331
chr7	128702	2811905	159138663	0.081%	22	4.73%	1114	853
chr8	120983	2501918	146364022	0.083%	21	4.21%	1046	615
chr9	110739	2331049	141213431	0.078%	21	3.92%	959	316
chr10	117000	2391950	135534747	0.086%	20	4.02%	1007	570
chr11	123298	2491047	135006516	0.091%	20	4.19%	1055	769
chr12	101440	2318400	133851895	0.076%	23	3.90%	899	302
chr13	56253	1362272	115169878	0.049%	24	2.29%	504	484
chr14	67012	1393000	107349540	0.062%	21	2.34%	595	317
chr15	78397	1437683	102531392	0.076%	18	2.42%	668	312
chr16	99946	2074762	90354753	0.111%	21	3.49%	829	418
chr17	95737	2028045	81195210	0.118%	21	3.41%	793	315
chr18	50052	1211867	78077248	0.064%	24	2.04%	431	481
chr19	91863	1783580	59128983	0.155%	19	3.00%	774	388
chr20	61046	1371117	63025520	0.097%	22	2.31%	525	431
chr21	29838	577980	48129895	0.062%	19	0.97%	266	403
chr22	56658	1032340	51304566	0.110%	18	1.74%	447	303
chrX	75130	1097325	155270560	0.048%	15	1.85%	736	294
chrY	28987	417843	59373566	0.049%	14	0.70%	259	289
chrM	12851	992192	16571	77.55%	77	1.67%	32	1874
CoV(S/N)	4075	9558337	5082	80.18%	2346	16.08%	32	273

Supplemental Table 3. Performance statistics of the amplicons retrieved from our multiplex PCR method highlighting a 10-fold range of read depth.

To simulate multiplex PCR with random hexamers as primers, we used a panel of 27,296 pairs of primers to perform multiplex PCR. These primers were divided into 2 overlapping primer pools, and amplification was initially performed in two separate reactions. The number of sequencing reads for a majority of the recovered amplicons were within a 10-fold range, representing a uniformity of 92.62 ± 1.96% at 0.2X mean.

Sample Number	Uniformity 0.2X Mean (%)	Mapping Rate (%)	On-Target Rate (%)	Total Reads	Mapped Reads	On-Target Reads	Primer Dimer Reads	Primer Dimer Rate (%)	Average On-Target Reads per Amplicon
1	93.6%	94.0%	96.2%	269435076	253341610	243602321	3326193	1.31%	8924
2	92.7%	97.3%	95.9%	272733262	265270282	254306010	1920151	0.72%	9316
3	92.8%	96.6%	95.8%	188776124	182357219	174778001	1492049	0.82%	6403
4	93.1%	97.2%	95.7%	265645114	258182429	247093224	1879080	0.73%	9052
5	92.6%	97.1%	95.5%	232875020	226046731	215875589	1535441	0.68%	7908
6	93.5%	97.1%	96.2%	192403072	186821637	179683779	1266338	0.68%	6582
7	91.3%	96.5%	96.5%	181465458	175141658	168975648	1423714	0.81%	6190
8	88.1%	96.7%	96.6%	191995346	185607975	179350573	1390209	0.75%	6570
9	95.0%	97.2%	95.3%	197356014	191797447	182683756	1056748	0.55%	6692
10	95.0%	96.5%	95.7%	201474266	194391114	185997107	1317629	0.68%	6814
11	95.0%	96.7%	96.5%	193125370	186751406	180154727	1036439	0.55%	6600
12	88.2%	95.5%	97.0%	199616350	190581763	184843808	1908098	1.00%	6771
13	92.0%	93.3%	96.4%	272873692	254679030	245480306	3824568	1.50%	8993
14	94.0%	93.6%	95.9%	297541418	278622642	267164257	4259698	1.53%	9787
15	91.2%	91.7%	96.3%	323595774	296801010	285698174	5184019	1.75%	10466
16	91.4%	93.9%	96.3%	248888684	233638153	224968924	5337352	2.28%	8241
17	91.5%	97.3%	96.8%	206702726	201183261	194776826	2063586	1.03%	7135
18	91.5%	98.2%	96.9%	178948198	175675252	170297061	1083447	0.62%	6238
19	90.2%	99.0%	97.1%	206684484	204582941	198563161	708813	0.35%	7274
20	91.1%	87.3%	95.0%	397262166	346621107	329334881	17048174	4.92%	12065
21	94.5%	95.5%	96.1%	215357334	205717503	197695646	2174832	1.06%	7242
22	94.1%	99.1%	95.9%	166942258	165384936	158665846	425846	0.26%	5812
23	94.2%	98.7%	96.0%	198067370	195451983	187701462	631609	0.32%	6876
24	94.2%	98.6%	96.1%	169536088	167192888	160657429	514940	0.31%	5885
25	94.8%	91.4%	94.9%	356335780	325681114	308973305	8034601	2.47%	11319
Avg.	92.6%	95.8%	96.1%					1.11%	7806
STDEV	1.96%	2.77%	0.58%					0.98%	1739
CV	2.12%	2.89%	0.60%					88.6%	22.3%