Accurate detection and quantification of SARS-CoV-2 genomic and subgenomic mRNAs by ddPCR and meta-transcriptomics analysis

Annarita Oranger
University of Bari Aldo Moro https://orcid.org/0000-0003-4402-244X
Caterina Manzari
University of Bari Aldo Moro
Matteo Chiara
University of Milan
Elisabetta Notario
University of Bari Aldo Moro
Bruno Fosso
Institute of Biomembranes
Antonio Parisi
Istituto Zooprofilattico Sperimentale della Puglia e Basilicata
Angelica Bianco
Istituto Zooprofilattico Sperimentale della Puglia e Basilicata
Michela Iacobellis
Servizio Centralizzato Aziendale di Citopatologia e Screening
Morena d'Avenia
Servizio Centralizzato Aziendale di Citopatologia e Screening
Anna Maria D'Erchia
University of Bari Aldo Moro
Graziano Pesole (graziano.pesole@uniba.it)
University of Bari Aldo Moro https://orcid.org/0000-0003-3663-0859

Article

Keywords: Structural Proteins, Discontinuous Transcription, Nucelocapsid and Spike Proteins, Molecular Diagnostic Tools, Patient Stratification

Posted Date: April 15th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-403215/v1

Version of Record: A version of this preprint was published at Communications Biology on October 22nd, 2021. See the published version at https://doi.org/10.1038/s42003-021-02748-0.

Accurate detection and quantification of SARS-CoV-2 genomic and subgenomic mRNAs by ddPCR and meta-transcriptomics analysis

Abstract

SARS-CoV-2 replication requires the synthesis of a set of structural proteins expressed through discontinuous transcription of ten subgenomic mRNAs (sgmRNAs). Here, we have fine-tuned a droplet digital PCR (ddPCR) assays to accurately detect and quantify SARS-CoV-2 genomic ORF1ab and sgmRNAs for the nucleocapsid (N) and spike (S) proteins. We analyzed 166 RNAs from anonymized COVID-19 positive subjects and we found a recurrent and characteristic pattern of sgmRNAs expression in relation to the total viral RNA content. Further, we observed that expression profiles of sgmRNAs analyzed in a subset of 110 samples subjected to meta-transcriptomics sequencing were highly correlated with those obtained by ddPCR.

Our results, providing a comprehensive and dynamic snapshot of SARS-CoV-2 sgmRNAs expression and replication, may contribute to provide a better understanding of SARS-CoV-2 transcription and expression mechanisms, and support the development of more accurate molecular diagnostic tools and for the stratification of COVID-19 patients.

Introduction

The COVID-19 (Coronavirus Disease 2019) outbreak caused by severe acute respiratory syndrome-coronavirus 2 (SARS-CoV-2) hit the world with a global pandemic. More than 115 million of confirmed infections and more than 2,800,000 deaths have been recorded worldwide since the first reported case in late December 2019, (WHO Coronavirus Disease (COVID-19) Dashboard; data last updated: 2021/4/6).

SARS-CoV-2 is an enveloped positive-sense single-stranded RNA betacoronavirus. The genome sequence (gRNA) is ~30 kb in size and shows the typical arrangement of betacoronavirus genomes. The replicase gene, which consists of two long, overlapping open reading frames encoding for polyproteins, ORF1a and ORF1b, extends over the 5' proximal two thirds of the genome, while the 3′ terminal region of the genome encodes 4 structural proteins required for the assembly of the viral capsid: spike (S), envelope (E), membrane (M) and nucleocapsid (N) and 8 other less well characterized proteins which are not universally conserved among coronaviruses.

SARS-CoV-2 infection initiates with the attachment of the virion to the surface of target cells, mediated by the binding of the S glycoprotein to the angiotensin converting enzyme 2 (ACE2) receptor[1]. Proteolytic cleavage of S protein by a cathepsin, TMPRRS2 or other proteases, followed by the fusion of the viral and cellular membranes allows the entry of the virus in the cytosol[2,3]. Subsequently, the gRNA is translated to produce the polyproteins that are post-translationally processed by viral encoded proteases to produce 16 non structural proteins (nsp) whose function is related to the synthesis of viral genomic RNA and to escape immune response [4–6].

Structural and accessory proteins are translated from a set of nested transcripts, called subgenomic mRNAs (sgmRNAs). These sgmRNAs are produced during viral genome replication through a complex template-switching discontinuous transcription mechanism which occurs during the synthesis of the negative stranded RNA, mediated by short, conserved transcriptional regulatory sequences (TRSs) that punctuate the viral genome and are found upstream of each major ORF (TRS-B) and in the 5' UTR (TRS-L)[7]. Briefly, when the replicase complex encounters a TRS-B element during the elongation of nascent minus-strand RNA, the complementarity with the TRS-L in the 5' UTR can mediate a relocation of the complex to the 5' UTR, resulting in the synthesis of a discontinuous negative sense sgmRNA. Transcription of negative sense sgmRNAs generates positive sense sgmRNAs, encoding for S, E, M, N and accessory proteins, which are 5' and 3' coterminal with the gRNA[8–10].

The reverse transcription-quantitative polymerase chain reaction (RT-qPCR) is considered the current gold standard method for the diagnosis of COVID-19. Although standard tests for COVID-19 are based on naso-pharyngeal swab samples, in principle RT-qPCR allows the detection of

SARS-CoV-2 genomic RNA from different types of specimen[11–15]. This method provides a relative quantification of the total amount of viral RNAs, expressed in the form of a threshold cycle (Ct) value, which represents the PCR cycle number at which the target product crosses the threshold for detection. Although different probes/primer sets have been reported to have different levels of sensitivity, a Ct cut-off value of positivity (usually 40) has been established for all approved commercial SARS-CoV-2 molecular diagnostic kits[16]. The major limitation of SARS-CoV-2 diagnostic RT-qPCR approaches is that they detect all viral RNA unspecifically, regardless of the nature and function (genomic, subgenomic or even degradation products). Indeed, this approach can not discriminate between sgmRNA molecules and the genomic sequence of the corresponding gene. The net result is that the Ct value derives from the sum of genomic and subgenomic RNAs. On the other hand, as sgmRNAs are transcribed in infected cells and encode for structural viral proteins, assembling in virion particles, their specific detection could provide evidence of active viral replication rather than the possible presence of residual viral RNA.

In this work, we evaluated the application of droplet digital (ddPCR) assay to perform a proof of concept study for the detection and quantification of SARS-CoV-2 sgmRNAs coding for S and N proteins, from RNA samples obtained from naso-pharyngeal swabs of COVID-19 positive subjects. By analyzing 166 RNA samples, we derived some relevant observations concerning the patterns of expression of SARS-CoV-2 sgmRNAs in infected subjects, with possible implications also in diagnostics. Firstly, we found that, while sgmRNA copy numbers are positively correlated with the total number of gRNA copies, N and S sgmRNAs have a characteristic range of expression, which remains similar across all samples, regardless of the viral RNA content. Additionally, we observed that sgmRNAs expression levels are reduced in RNA samples with a low viral RNA content, thus indicating that these samples are mainly characterized by residual genomic SARS-CoV-2 material with scarce or no active viral transcription.

Moreover, by performing bioinformatics analyses of meta-transcriptomics sequencing of a subset of 110 RNA samples, for the reconstruction of the viral genome sequence, we demonstrated that sgmRNA expression patterns recovered by NGS sequencing are highly consistent with those inferred by ddPCR, thus suggesting that this approach can provide an accurate overview of the expression patterns of SARS-CoV-2 sgmRNAs and extend our understanding of the mechanism of transcription of the genome of this novel pathogen.

Finally, since we observed a good correlation between N and S sgmRNAs expression levels, as derived from sequencing data, and their copy number, as determined by ddPCR (although less significant for S sgmRNA), we could envisage that this and other similar/equivalent approaches

could also be used to derive a realistic estimate of the viral load associated with SARS-CoV-2 RNA samples for which meta-transcriptomic data are available.

Results

Samples used in this study

For this study, we used RNA samples extracted from 166 nasopharyngeal swab remnants of unidentified subjects with a positive diagnosis of COVID-19. No clinical data were recorded, but only age and gender (Table 1 and Supplementary Data 1). The presence of viral RNA in all samples was confirmed by a RT-qPCR assay (see Methods section).

Accuracy and limit of detection (LoD) of ddPCR assays targeting SARS-CoV-2 genomic and subgenomic mRNAs

Two pairs of primers were designed to quantify S and N sgmRNAs by ddPCR, with a forward primer targeting the LS sequence in the 5' UTR and a reverse primer in the 5′-proximal end of the S and N gene, respectively. Both sgmRNAs primer pairs could detect specifically N and S subgenomic transcripts, excluding N and S genomic sequences. A pair of primers in the ORF1ab (nsp8 sequence) was also designed to quantify gRNA levels. Primer pairs sequences are reported in Supplementary Table 1. Ten serial dilutions of a confirmed SARS-CoV-2 positive RNA sample (RNA 121) were used to evaluate the accuracy of our ddPCR assays, for every target region (Supplementary Data 2-4). As shown in Figure 1a-c, all SARS-CoV-2 targets showed a very good linear correlation between the expected and observed copy numbers (R^2 =0.992, 0.996 and 0.990 respectively for ORF1ab and N and S sgmRNAs). The Limit of Detection (LoD) of each ddPCR assay was assessed also by analyzing the same serial dilutions used for the evaluation of the accuracy (Supplementary Data 2-4). Likewise, LoD, which is defined as the limit of an analyte/target detectable by a molecular assay with 95% of confidence and generally expressed as copies/reaction, was calculated for each target by regression probit analysis. As shown in Figure 1d-f, LoD was 3.1 (95% CI: 1.8-108), 2.6 (95% CI: 1.3-24) and 3.4 (95% CI: 1.8-14.6) copies/reaction, respectively for ORF1ab, N and S sgmRNAs.

Absolute quantification of SARS-CoV-2 ORF1ab mRNA and N and S sgmRNAs by ddPCR

ddPCR assays for the quantification of SARS-CoV-2 ORF1ab gRNA and of the N and S sgmRNAs were applied to 166 RNA samples obtained from naso-pharyngeal swabs of COVID-19 positive subjects (Supplementary Data 5). Overall, a high variability in copy numbers was observed among

the RNA samples, for all the 3 targets. ORF1ab was consistently associated with a higher number of RNA copies, compared to N and S sgmRNAs with S sgmRNA consistently associated with the lowest copy number in all samples. Based on observed gRNA and sgmRNAs copy numbers (see methods), samples were stratified in three main groups: "high" (n =21), "middle" (n = 82) and "low" (n = 63) viral RNA content. Interestingly, both N and S sgmRNAs displayed a higher number of copies in samples of the "high" group (sgmN median = 49,060 copies/ng RNA, IQR =179,187; sgmS median = 3,953 copies/ng RNA, IQR =12,093) and "middle" group (sgmN median = 387 copies/ng RNA, IQR =1,266; sgmS median = 27.5 copies/ng RNA, IQR = 80.2), while in samples belonging to the "low" group, both N and S subgenomic transcripts were barely detectable (sgmN median = 6.2 copies/ng RNA, IQR =24.2; sgmS median = 0 copies/ng RNA, IQR =1.8). In particular N sgmRNA was absent or scarcely detectable (<1 copy/ng RNA) in 12 out of 15 samples with less than 10 copies/ng RNA of ORF1ab, while S sgmRNA was absent or scarcely detectable in 36 out of 45 samples with less than 100 copies/ng RNA of ORF1ab..

As shown in Figure 2a-b, N and S sgmRNAs progressively decreased in parallel with total viral RNA content. All sgmRNAs displayed the same expression trend with statistically significant copy number differences between the three groups of samples (p-value<0.0001).

Ratios of N and S sgmRNAs to viral genome copy numbers (ORF1ab RNA target) were calculated for the three groups of samples with different viral RNA content. As shown in Figure 2c-d, N and S sgmRNAs had a significant higher expression ratio in the "high" group if compared to the "middle+low" group (p-value = 0.0089 and 0.0002 respectively for N sgmRNA and S sgmRNA), suggesting that subjects with high SARS-CoV-2 gRNA content are characterized by a very active ongoing transcription process.

Finally, based on the assumption that ~ 20 picograms of RNA are the average content of a mammalian cell[17], the expected number of N and S sgmRNAs per cell were calculated. As shown in Table 2, when the sgmRNAs average copies number/ng of input RNA were considered, samples belonging to the "high" group may contain an average of 2,478 and 184 molecules of respectively N and S sgmRNAs per cell. The number of sgmRNAs copies per cell decreased in the "middle" group (~ 25 and 2 molecules per cell respectively for N and S sgmRNAs), while in the "low" group, on average, a cell may contain any copy of N and S sgmRNAs.

Overall these data demonstrate that absolute copy number of SARS-CoV-2 gRNA and sgmRNAs are correlated and that, in very low viral load specimens, as could be inferred from the gRNA copy number, sgmRNAs are scarcely or not present, thus providing indications of a possible absence of active viral replication in these samples.

Meta-transcriptomics analysis shows high correlation with ddPCR estimates of sgmRNAs expression

A subset of the 166 RNA samples used in this study (113 samples) was subjected to meta-transcriptomics sequencing. Since meta-transcriptomic sequencing provides an accurate representation of all the RNA species present in a sample, these data were also used for deriving an estimate of the expression levels of SARS-CoV-2 sgmRNAs. A method based on the count of sgmRNA junctions spanning reads was applied (see Methods section). gRNA abundance was estimated by counting only reads covering the same genomic region in ORF1ab (nsp8 gene) used to estimate gDNA copy number in ddPCR assay. Only samples with 100 or more viral reads (110 out of 113 sequenced samples, Supplementary Data 6) were considered for the study of transcriptional profiles. As shown in Figure 3, highly significant levels of correlation were observed when expression levels, as estimated by analysis of NGS meta-transcriptomic data, were compared with copy number estimates obtained by ddPCR, for all the three target genes included in our assay. Lower, but still very significant, levels of correlation were observed for the S sgmRNA with respect to other targets (R^2 =0.78 for ORF1ab mRNA; R^2 = 0.77 for N sgmRNA; R^2 = 0.65 for S sgmRNA; $p<0.0001$ for all targets). Consistent with data obtained in ddPCR, we observed that the viral gRNA was consistently associated with an increased (more than 5 fold on average) number of reads if compared to sgmRNAs (Figure 4).

When expression patterns of all sgmRNAs were compared, we also observed that, as previously noted in other studies[18], sgmRNAs encoded by the most 3' end terminal portion of the genome, displayed higher levels of expression, with the N sgmRNA showing the highest number of reads in all samples herein considered. Conversely, sgmRNAs for the S and ORF3a proteins show ~ 3 fold decrease in the number of reads if compared with the average of other sgmRNAs. Importantly all the sgmRNAs displayed characteristic abundance/level of expression which was largely consistent across all samples (Supplementary Figure 1). Furthermore, by NGS analysis and consistently with our ddPCR data of N and S sgmRNAs relative expression, a significantly higher subgenomic to genomic RNA ratio was observed for all canonical sgmRNAs, in the "high" RNA content group if compared to "middle"+"low" groups (Supplementary Figure 2). This is consistent with higher levels of viral transcription and replication in samples with high viral load.

Importantly, the remarkable levels of agreement observed both with ddPCR data and with previous studies for the characterization of expression patterns of the SARS-CoV-2 genomes[18], suggested that the approach developed for the quantification of sgmRNAs in this study is accurate, indicating that estimates of gRNA and sgmRNAs copy numbers from meta-transcriptomics sequencing of

SARS-CoV-2 positive samples, could be used as a proxy to provide a reasonable estimate of both genomic and subgenomic viral RNA content.

Discussion

Qualitative and quantitative measures of viral load in COVID-19 patients provide effective tools for monitoring the progression of the disease and the severity of the infection. Changes in viral load are observed during the progression of the infection, reaching the peak during the first 4-6 days from the onset of the infection with a following gradual decrease[19]. Wölfel et al. reported that 8 days after the onset of the symptoms, virus isolated from COVID-19 patients failed to infect cell cultures, demonstrating a decline in virus infectiousness while Bullard and colleagues showed that infectivity (defined by virus growth in VERO cell culture) is remarkable only when RT-qPCR Ct value are lower than 24[20,21]. While recent studies reported detectable levels of viral shedding both in asymptomatic and symptomatic COVID-19 patients even after weeks from symptom resolution[22], the correlation between detectable viral RNA and transmissibility is still unclear. A positive RT-PCR result does not necessarily indicate potential for viral transmission as this mode of testing can not discriminate between viable and inactive viral particles and furthermore readouts from different tests/commercial kits can not be easily compared[23–25].

In this context, there is a critical need to develop novel, more effective diagnostic tools capable of detecting active SARS-CoV-2 replication in the upper respiratory tract, which could provide useful indications for a more precise stratification of the patients and for the study of their potential infectivity. Current SARS-CoV-2 nucleic acid amplification diagnostic assays are qualitative or semi-quantitative, and since they can not discriminate genomic from subgenomic RNAs, do not provide any information concerning ongoing viral replication. In the light of all these considerations, subgenomic mRNAs could represent a relevant epidemiological target for a more specific diagnosis of COVID-19. Since these transcripts code for structural proteins required for assembly of novel viral particles, specific assays based on their detection may facilitate the development of novel diagnostic methods to better track the progression and to define robust guidelines for COVID-19 patients release.

In the current study, we present a method to accurately detect and quantify SARS-CoV-2 RNAs by ddPCR and meta-transcriptomics approaches. By using RNA extracted from nasopharyngeal swabs of COVID-19 positive subjects, we were able to specifically quantify the viral genome, by targeting the ORF1ab gene, and the N and S sgmRNAs.

Compared to qPCR, ddPCR allows a more precise absolute quantification of a target molecule without the need of a standard curve and, moreover, it shows a wide dynamic range with high

sensitivity, even in the detection of low copy number of the target[26–29]. Several applications of ddPCR have been recently reported for the quantification of SARS-CoV-2 RNA genome, showing higher sensitivity, specificity and reproducibility respect to RT-qPCR[30–33].

In our study, we demonstrated that a significant reduction of subgenomic RNAs species is consistently detectable in samples with a low content of gRNA. Indeed N and S smgRNAs were no longer quantifiable or barely detectable when gRNA copies/ng RNA were respectively less than 10 and 100. Samples with low viral gRNA content are thus characterized by very scarce or absent subgenomic transcripts, suggesting no or very limited SARS-CoV-2 transcription. Furthermore, regardless of the viral genome content of each sample, we demonstrated that SARS-CoV-2 N and S sgmRNAs have a specific expression pattern if compared to full-length gRNA, which remains almost constant in all analyzed samples and is correlated to the gene position respect to the proximity to the 3' terminal portion of the genome.

Bioinformatics analyses of a subset of RNA samples, subjected to meta-transcriptomics sequencing, confirmed and extended results obtained by ddPCR on sgmRNAs expression. Indeed, N and S sgmRNAs observed expression patterns were largely consistent across all the samples, demonstrating that the approach used was highly reproducible and reliable. Moreover, when expression patterns of all sgmRNAs were considered, results were largely concordant with those reported by previous studies[18], suggesting that meta-transcriptomics sequencing can provide an accurate overview of the transcriptional dynamics of the SARS-CoV-2 genome. We observed a highly significant correlation between SARS-CoV-2 meta-transcriptomics sequencing data and ddPCR results, thus suggesting that approaches based on regression models could be effectively used to derive a plausible estimate of viral content for samples subjected to meta-transcriptomics sequencing. Although we investigated by ddPCR the expression patterns of only 2 out of 9 canonical sgmRNAs, the remarkable levels of correlation between sgmRNAs expression patterns as recovered by ddPCR and NGS data, suggest that expression profiles determined by our meta-transcriptomics approach are highly reliable and can provide useful indications for the study of SARS-CoV-2 replication/transcription in vivo.

SgmRNAs can be considered markers of ongoing active viral replication and, likely, of viral infectivity, as they are synthesized exclusively inside infected cells and not packaged into virions [20,34,35]. Although our data are not related to clinical data and/or to infectiousness data on cell culture, we believe that this study provides evidence for the application of ddPCR and meta-transcriptomics for the study of the SARS-CoV-2.

In conclusion, our results might contribute to development of new effective strategies for SARS-CoV-2 RNAs detection, supporting the currently available SARS CoV-2 diagnostic tools, for a

more accurate stratification of patients, in particular of asymptomatic carriers, with relevant implications for the containment of COVID-19.

Methods

Viral RNA extraction

Remnants from 166 nasopharyngeal swabs (in UTM matrix) from COVID-19 subjects were collected at the diagnostic laboratory of Ospedale Di Venere in Bari, from May to December 2020 (Table 1 and Supplementary Data 1). 560 µL of UTM matrix were used to purify viral RNA using the QIAamp Viral RNA Mini Kit (Qiagen, Hilden, German) according to the manufacturer's instructions, without addition of poly-A RNA carrier. The eluted RNA was treated with DNase (Zymo Research Corporation, Irvine, CA, USA) and successively concentrated with RNA Clean & Concentrator Kits (Zymo Research Corporation, Irvine, CA, USA), according to the manufacturer's instructions. RNA samples were stored at -80°C until use.

RT-qPCR

Before the DNAse treatment and the concentration step, RNA samples were analyzed by an in-house reverse transcription-qPCR assay using the primers/probe for E, N, and orf1ab genes, as from the World Health Organization previously described previously [36] to confirm the presence of the viral RNA. A 25 µL reaction was set up containing 5 µL of RNA, 12.5 µL of 2× reaction buffer provided with the Superscript III One Step RT-PCR system with Platinum Taq Polymerase (Invitrogen, Carlsbad, CA, USA), 1 µL of Reverse Transcriptase/Taq mixture, 10 µM of Forward and Reverse primers and 10 µM probe. Each assay was performed in triplicate on Applied Biosystems™ 7500 Real-Time PCR Systems (Foster City, CA, USA).

SARS-CoV-2 RNAs quantification by ddPCR

A forward primer in the leader sequence (LS) and a reverse primer in the S and N coding region, respectively, were designed to specifically detect SARS-CoV-2 S and N sgmRNAs; a primers pair in the ORF1ab gene was designed to detect viral genome. Primers are listed in Supplementary Table 1. The sequence of the three amplicons was confirmed by Sanger sequencing. Before performing ddPCR assays, RNA samples were quantitatively evaluated using NanoDrop 1000 (Thermo Fisher Scientific, Waltham, MA, USA). According to availability of each sample, a variable amount (1.5-200 ng) of RNA was reverse transcribed in cDNA using the iScript™ Advanced cDNA Synthesis Kit for RT-qPCR (Bio-Rad, Hercules, CA, USA) and cDNA, diluted or as it is, was used as input in

ddPCR experiments (up to 5 µl of cDNA per 22 µl reaction). For all samples, ddPCR assays were performed using the same cDNA preparation. ddPCR experimental conditions were accurately set for each target assay. A reaction volume of 22 µl was prepared by combining cDNA (1-5 µl) with 11 µl of 2× EvagreenSupermix (Bio-Rad, Hercules, CA, USA), 220 nM ORF1ab primers or 250 nM N or S sgRNA primers and water. Emulsion was produced in the QX200 Droplet Generator (Bio-Rad, Hercules, CA, USA) according to the manufacturer's instructions. Then the droplet-partitioned samples were amplified under the following thermal cycling conditions: for ORF1ab RNA: 1 cycle at 95 °C for 5 min, 40 cycles at 95 °C for 30 s and 60°C for 1 min, 1 cycle at 4 °C for 5 min, 1 cycle at 90 °C for 5 min, final hold at 4 °C; for N sgmRNA: 1 cycle at 95 °C for 5 min, 40 cycles at 95 °C for 30 s and 58 °C for 1 min, 1 cycle at 4 °C for 5 min, 1 cycle at 90 °C for 5 min, final hold at 4 °C; for S sgmRNA: 1 cycle at 95 °C for 5 min, 40 cycles at 95 °C for 30 s and 56 °C for 1 min, 1 cycle at 4 °C for 5 min, 1 cycle at 90 °C for 5 min, final hold at 4 °C. Each RNA sample was analyzed at least in duplicate. For each experiment, a negative control (no template control) was used. Absolute quantification was performed using QuantaSoft version 7.4.1 software (Bio-Rad, Hercules, CA, USA) and the negative/positive thresholds were set manually, excluding samples with a number of droplets <10,000. QuantaSoft output results were expressed in copies/µl.

As variable RNA amounts and cDNA volumes were used, respectively, for RT and ddPCR experiments, absolute quantification of each target was obtained by calculating, firstly, the copies/µl of cDNA according to the equation (1):

$$copies/\mu l\ cDNA = \frac{dilution\ factor\ \cdot reaction\ volume\ (22\mu l) \cdot copies/\mu l}{\mu l\ cDNA\ in\ reaction} \quad (1)$$

The obtained number of copies/µl of cDNA was normalized for ng of RNA used as input in 20 µl of RT reaction volume (copies/ng RNA).

By applying a hierarchical agglomerative clustering algorithm to ddPCR absolute quantification of SARS-CoV-2 genomic and subgenomic RNAs, all samples were divided in 3 groups: the "high" , the "middle" and the "low" group, composed of 21, 82 and 63 RNA samples, respectively. Estimated number of target copies were scaled by applying base 2 logarithm. Euclidean distances were computed by means of the dist() function as implemented by the R stat package[37]. Hierarchical clustering was performed by applying the hclust() function from the same software package. Finally the cutree function was used to delineate 3 distinct clusters.

Accuracy and limit of detection of SARS-CoV-2 genomic and subgenomic ddPCR assays

Ten serial dilutions of a confirmed SARS-CoV-2 positive RNA sample (RNA 121: RT qPCR Ct = 14, corresponding to 1,536,568 copies/ng for ORF1ab RNA, 214,720 copies/ng for sgN and 14,353 copies/ng for sgS RNA in ddPCR) were prepared to analyze each target. For each dilution point, at least 9 replicates were analyzed. The accuracy was evaluated plotting expected *vs* observed target copies per reaction for each dilution point; coefficient of determination (R^2) of SARS-CoV-2 target quantification was assessed by linear regression analysis. The Limit of Detection (LoD) of ORF1ab, sgN and sgS ddPCR assays was defined by probit analysis on the same RNA dilutions used for accuracy assessment. All data are reported as copies/reaction.

Meta-transcriptomics sequencing of SARS-CoV-2 genome and data analysis

113 RNA samples were used to sequence SARS-CoV-2 genome, using the meta-transcriptomics approach. Libraries were prepared using the Truseq Stranded Total RNA with Ribo Zero plus protocol (Illumina, San Diego, CA, USA) with some changes correlated to quantity and quality of RNA extracted. Firstly, according to the availability of each RNA, a variable amount of total RNA (5-100 ng) was used as input for the library preparation. Then, as the quality of RNA was not high (RIN ranging from 2 and 5 on Agilent Bioanalyzer 2100), for all libraries, the incubation time for RNA fragmentation at 94°C was decreased to 25 seconds instead of 2 minutes, as reported in the protocol for the processing of high quality starting RNA (RIN ≥7). Finally, the number of cycles in the final enrichment PCR step was setted from 15 to 18, depending on the quantity of total RNA used as input: for RNA input of 40-100 ng, the number of cycles was setted to 15, for RNA input less of 40 ng, the number of cycles was setted from 17 to 18. Only the libraries passing the quality and quantity check were sequenced on NextSeq500 platform (Illumina, San Diego, CA, USA) to generate 2×75 bp paired-end (PE) reads. 1% of the PhiX genome library was loaded in each run. An average of 13 M PE reads were generated for sample (Supplementary Data 6). SARS-CoV-2 genome assemblies were performed by means of the "Assembly of SARS-CoV-2 from pre-processed reads" workflow as available from the COVID-19 Galaxy[38].

Detection and quantification of sgmRNA junction reads

Annotation of SARS-CoV-2 sgmRNA was obtained from http://hgdownload.soe.ucsc.edu/downloads.html in the form of a gtf file. For every subgenomic transcript, the corresponding subgenomic junction sequence was reconstructed by *in silico* juxtaposition of the LS in the 5' UTR with the first 70 residues of each gene. For the S and N sgmRNAs, the sequence of the amplicon targeted by the ddPCR assay was used. Similarly, a region

corresponding to the ORF1ab amplicon, was used to count metatranscriptomics reads associated with ORF1ab and to quantify gRNA. Metatranscriptomics reads were mapped to the ORF1ab target sequence and to the complete collection of subgenomic junction sequences with the Bowtie2[39] software, using the --sensitive preset. A custom Perl script was used to count the number of reads associated with each target region and to obtain a table of counts. Counts were log scaled, with base 2 logarithm. Graphical representation of the data was performed by using the boxplot function as available in the standard library of the R programming language. Correlation analyses were performed by means of the trendline function from the basicTrendline R package. Only samples with more than 100 reads mapping on the SARS-CoV-2 viral genome (110 out of 113 sequenced samples) were considered in the analysis.

Statistics and reproducibility

DdPCR assay accuracy was analyzed by linear correlation comparing expected *vs* obtained copies for each SARS-CoV-2 RNA target in R. LoD (SARS-CoV-2 copy number at a 95% detection rate) was calculated by probit analysis using MedCalc statistical software (version 19.6.4) on at least 9 replicates in 10 serial dilutions of a reference RNA for each target.

Kruskal-Wallis test was performed by GraphPad Prism 8.0.2 software (GraphPad Software, San Diego, CA, USA) for ddPCR SARS-CoV-2 targets analysis between different viral RNA content groups; for each RNA viral content group, we calculated N and S sgmRNAs absolute quantification as median of all RNA samples belonging to the same group with relative interquartile range (IQR).

A Mann–Whitney U test/Wilcoxon rank-sum test was performed to analyze N and S sgmRNA/ORF1ab expression ratio in the “high” viral RNA content group compared to “middle”+”low” viral RNA content groups. p-values <0.05 were considered as statistically significant.

Data Availability

The authors declare that the main data supporting the findings of this study are available within the article and its Supplementary Information file. Source data underlying figures are provided in Supplementary Data. Genomic assemblies were deposited at GISAID. A complete list of GISAID accessions is provided in Supplementary Data 6.

References

1. Lan, J. *et al.* Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* **581**, 215–220 (2020).
2. Wang, Q. *et al.* Structural and Functional Basis of SARS-CoV-2 Entry by Using Human ACE2. *Cell* **181**, 894-904.e9 (2020).
3. Hoffmann, M. *et al.* SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* **181**, 271-280.e8 (2020).
4. Raj, R. Analysis of non-structural proteins, NSPs of SARS-CoV-2 as targets for computational drug designing. *Biochemistry and Biophysics Reports* **25**, 100847 (2021).
5. Jauregui, A. R., Savalia, D., Lowry, V. K., Farrell, C. M. & Wathelet, M. G. Identification of Residues of SARS-CoV nsp1 That Differentially Affect Inhibition of Gene Expression and Antiviral Signaling. *PLoS ONE* **8**, e62416 (2013).
6. Narayanan, K., Ramirez, S. I., Lokugamage, K. G. & Makino, S. Coronavirus nonstructural protein 1: Common and distinct functions in the regulation of host and viral gene expression. *Virus Research* **202**, 89–100 (2015).
7. Wang, D. *et al.* The SARS-CoV-2 subgenome landscape and its novel regulatory features. *Molecular Cell* S1097276521001660 (2021) doi:10.1016/j.molcel.2021.02.036.
8. V'kovski, P., Kratzel, A., Steiner, S., Stalder, H. & Thiel, V. Coronavirus biology and replication: implications for SARS-CoV-2. *Nat Rev Microbiol* (2020) doi:10.1038/s41579-020-00468-6.
9. Sola, I., Almazán, F., Zúñiga, S. & Enjuanes, L. Continuous and Discontinuous RNA Synthesis in Coronaviruses. *Annu Rev Virol* **2**, 265–288 (2015).
10. Romano, M., Ruggiero, A., Squeglia, F., Maga, G. & Berisio, R. A Structural View of SARS-CoV-2 RNA Replication Machinery: RNA Synthesis, Proofreading and Final Capping. *Cells* **9**, 1267 (2020).

11. Oliveira, B. A., Oliveira, L. C. de, Sabino, E. C. & Okay, T. S. SARS-CoV-2 and the COVID-19 disease: a mini review on diagnostic methods. *Rev Inst Med Trop Sao Paulo* **62**, e44 (2020).
12. Okamaoto, K. *et al.* Assessment of Real-Time RT-PCR Kits for SARS-CoV-2 Detection. *Jpn J Infect Dis* **73**, 366–368 (2020).
13. Mathuria, J. P., Yadav, R. & Rajkumar, null. Laboratory diagnosis of SARS-CoV-2 - A review of current methods. *J Infect Public Health* **13**, 901–905 (2020).
14. Afzal, A. Molecular diagnostic technologies for COVID-19: Limitations and challenges. *J Adv Res* **26**, 149–159 (2020).
15. Organization, W. H. *Diagnostic testing for SARS-CoV-2: interim guidance, 11 September 2020*. 20 p. (2020).
16. Tang, Y.-W., Schmitz, J. E., Persing, D. H. & Stratton, C. W. Laboratory Diagnosis of COVID-19: Current Issues and Challenges. *J Clin Microbiol* **58**, e00512-20, /jcm/58/6/JCM.00512-20.atom (2020).
17. Wu, J. *et al.* Ribogenomics: the Science and Knowledge of RNA. *Genomics, Proteomics & Bioinformatics* **12**, 57–63 (2014).
18. Kim, D. *et al.* The Architecture of SARS-CoV-2 Transcriptome. *Cell* **181**, 914-921.e10 (2020).
19. He, X. *et al.* Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nat Med* **26**, 672–675 (2020).
20. Wölfel, R. *et al.* Virological assessment of hospitalized patients with COVID-2019. *Nature* **581**, 465–469 (2020).
21. Bullard, J. *et al.* Predicting Infectious Severe Acute Respiratory Syndrome Coronavirus 2 From Diagnostic Samples. *Clinical Infectious Diseases* **71**, 2663–2666 (2020).
22. Walsh, K. A. *et al.* SARS-CoV-2 detection, viral load and infectivity over the course of an infection. *Journal of Infection* **81**, 357–371 (2020).
23. Sethuraman, N., Jeremiah, S. S. & Ryo, A. Interpreting Diagnostic Tests for SARS-CoV-2. *JAMA* **323**, 2249 (2020).

24. Widders, A., Broom, A. & Broom, J. SARS-CoV-2: The viral shedding vs infectivity dilemma. *Infection, Disease & Health* **25**, 210–215 (2020).
25. Criteria for releasing COVID-19 patients from isolation. https://www.who.int/news-room/commentaries/detail/criteria-for-releasing-covid-19-patients-from-isolation.
26. Hindson, C. M. *et al.* Absolute quantification by droplet digital PCR versus analog real-time PCR. *Nat Methods* **10**, 1003–1005 (2013).
27. Falzone, L. *et al.* Sensitivity assessment of droplet digital PCR for SARS-CoV-2 detection. *Int J Mol Med* **46**, 957–964 (2020).
28. Taylor, S. C., Laperriere, G. & Germain, H. Droplet Digital PCR versus qPCR for gene expression analysis with low abundant targets: from variable nonsense to publication quality data. *Sci Rep* **7**, 2409 (2017).
29. Sidstedt, M., Rådström, P. & Hedman, J. PCR inhibition in qPCR, dPCR and MPS—mechanisms and solutions. *Anal Bioanal Chem* **412**, 2009–2023 (2020).
30. Vasudevan, H. N. *et al.* Digital droplet PCR accurately quantifies SARS-CoV-2 viral load from crude lysate without nucleic acid purification. *Sci Rep* **11**, 780 (2021).
31. Alteri, C. *et al.* Detection and quantification of SARS-CoV-2 by droplet digital PCR in real-time PCR negative nasopharyngeal swabs from suspected COVID-19 patients. *PLoS ONE* **15**, e0236311 (2020).
32. de Kock, R., Baselmans, M., Scharnhorst, V. & Deiman, B. Sensitive detection and quantification of SARS-CoV-2 by multiplex droplet digital RT-PCR. *Eur J Clin Microbiol Infect Dis* (2020) doi:10.1007/s10096-020-04076-3.
33. Liu, C. *et al.* Evaluation of droplet digital PCR for quantification of SARS-CoV-2 Virus in discharged COVID-19 patients. *Aging* **12**, 20997–21003 (2020).
34. Wong, C. H. *et al. Subgenomic RNAs as molecular indicators of asymptomatic SARS-CoV-2 infection*. http://biorxiv.org/lookup/doi/10.1101/2021.02.06.430041 (2021) doi:10.1101/2021.02.06.430041.

35. Dimcheff, D. E. *et al. SARS-CoV-2 Total and Subgenomic RNA Viral Load in Hospitalized Patients*. http://medrxiv.org/lookup/doi/10.1101/2021.02.25.21252493 (2021) doi:10.1101/2021.02.25.21252493.
36. Corman, V. M. *et al.* Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. *Eurosurveillance* **25**, (2020).
37. R: The R Project for Statistical Computing. https://www.r-project.org/index.html.
38. Baker, D. *et al.* No more business as usual: Agile and effective responses to emerging pathogen threats require open data and open analytics. *PLoS Pathog* **16**, e1008643 (2020).
39. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359 (2012).

Competing interests

The authors declare no competing interests.

Ethical statement

Ethical approval was not required, as nasopharyngeal swab remnants from subjects that remained unidentified and thus not considered as human samples were used in this study.

Figure legends

Figure 1. Accuracy and Limit of detection (LoD) of SARS-CoV-2 genomic ORF1ab RNA and N and S sub-genomic RNAs by ddPCR assays.

a)-c). Accuracy of designed SARS-CoV-2 ddPCR assays was evaluated using 10 serial dilutions, each with at least 9 replicates, using RNA 121 as reference. Accuracy was evaluated by linear regression analysis plotting expected (Y axis) *vs* ddPCR observed (X axis) copies per reaction for ORF1ab RNA (a), N sgmRNA (b) and S sgmRNA (c). Regression equation and R^2 are shown in the top corner. The gray area represents a 0.95 level of confidence interval.

d)-f). Limit of detection (LoD) of designed SARS-CoV-2 ddPCR assays was evaluated on the same serial dilutions of the reference RNA 121 used for the accuracy analysis. LoD for ORF1ab RNA (d), N sgmRNA (e) and S sgmRNA (f) was defined by probit analysis using MedCalc software.

Figure 2. SARS-CoV-2 N and S sub-genomic RNAs absolute quantification and their relative expression to genomic RNA in COVID-19 samples by ddPCR.

N sgmRNA **(a)** and S sgmRNA **(b)** copies quantification in 166 RNA samples from COVID-19 subjects. Based on ORF1ab gRNA and N and S sgmRNAs quantification, RNA samples were divided in three groups defined as "high" (n = 21, represented by the green square symbol) , "middle" (n = 82, represented by the grey triangle symbol) and "low" (n = 62, represented by the orange circle symbol) viral RNA content. Values, reported in copies/ng RNA, are expressed as the means of a duplicate assay for each sample. Samples with zero sgmRNAs copies/ng RNA were not plotted in the graphs because logarithmic axes mathematically do not contemplate 0 value. Black lines represent median with relative interquartile range (IQR); p-value < 0.0001 calculated by Kruskal-Wallis test.

N sgmRNA **(c)** and S sgmRNA **(d)** expression was evaluated as ratio respect to ORF1ab gRNA in "high" (green), "middle" (grey) and "low" (orange) viral RNA content group. Values are reported as sgmRNA/ORF1ab RNA copies/ng RNA. Significance was evaluated comparing "high" group *vs* "middle"+"low" groups by Mann–Whitney U test /Wilcoxon rank-sum. Data are presented as median with relative interquartile range (IQR). For N sgmRNA: p-value =0.008877; for S sgmRNA: p-value =0.000233.

Figure 3. Correlation between ddPCR and transcriptomics data for SARS-CoV-2 ORF1ab, N and S sgmRNA targets.

Linear correlation between ddPCR and sequencing quantification of SARS-CoV-2 ORF1ab gRNA (**a**), N sgmRNA (**b**) and S sgmRNA (**c**) calculated on 110 COVID-19 RNA samples, by means of a bivariate linear fit analysis (*p*-value <0.0001). On X axis: log2 scaled ddPCR quantification; on Y axis: log2 scaled meta-transcriptomics reads count. Regression equation and R^2 are reported in the top-left corner. The gray area represents a 0.95 level of confidence interval.

Figure 4. SARS-CoV-2 transcripts expression derived from meta-transcriptomics sequencing

Representation by boxplots of log2 scaled counts distributions of meta-transcriptomics reads assigned to each sgmRNAs and to the ORF1ab gene (gRNA). Genes are indicated on the Y axis. Log2 scaled counts on the X axis.

522 **Table 1.** RNA samples from COVID-19 confirmed anonymized subjects employed in the study.

	High Viral RNA Content*	Middle Viral RNA Content *	Low Viral RNA Content*
SARS-COV-2 positive RNAs n.	21	82	63
Patient's gender	9 F, 12 M	34 F, 48 M	32 F, 31 M
Patient's age (years)	49±20	55±22	44±20
NGS sequenced RNAs	21	60	31

523 * RNA samples were grouped applying a hierarchical agglomerative clustering algorithm to ddPCR absolute
524 quantification of SARS-CoV-2 genomic and subgenomic RNAs.

525

526

527 **Table 2.** N and S sgmRNA molecules number per cell according to viral RNA content group.

viral RNA content group	N sgmRNA (copies/ng RNA)	S sgmRNA (copies/ng RNA)	N sgmRNA copies/cell	S sgmRNA copies/cell
High	123,920	9,181	2,478	184
Middle	1,251	87	25	1.7
Low	16	1	0.3	0.02

528

Figures

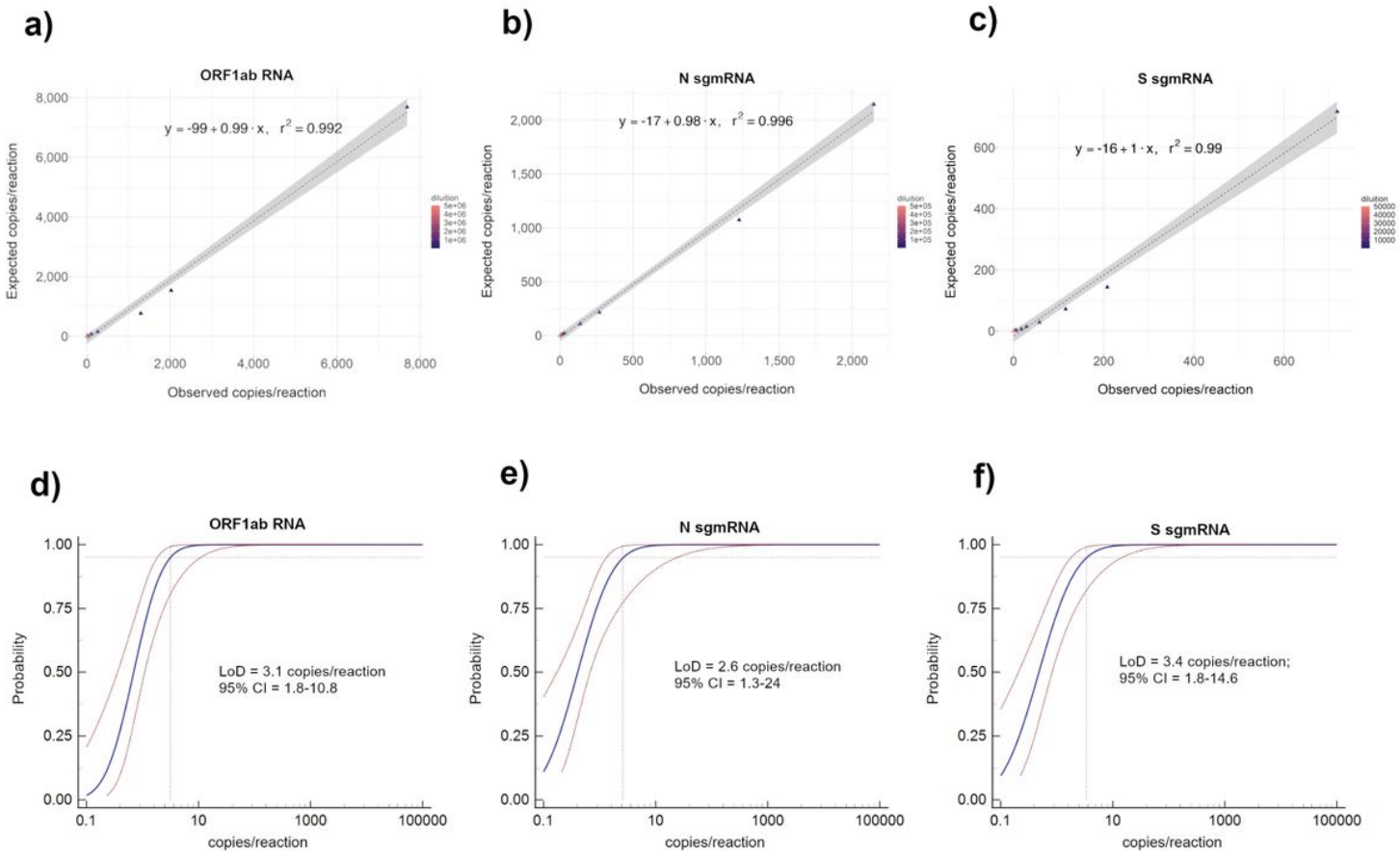

Figure 1

Accuracy and Limit of detection (LoD) of SARS-CoV-2 genomic ORF1ab RNA and N and S sub-genomic RNAs by ddPCR assays. a)-c). Accuracy of designed SARS-CoV-2 ddPCR assays was evaluated using 10 serial dilutions, each with at least 9 replicates, using RNA 121 as reference. Accuracy was evaluated by linear regression analysis plotting expected (Y axis) vs ddPCR observed (X axis) copies per reaction for ORF1ab RNA (a), N sgmRNA (b) and S sgmRNA (c). Regression equation and R2 are shown in the top corner. The gray area represents a 0.95 level of confidence interval. d)-f). Limit of detection (LoD) of designed SARS-CoV-2 ddPCR assays was evaluated on the same serial dilutions of the reference RNA 121 used for the accuracy analysis. LoD for ORF1ab RNA (d), N sgmRNA (e) and S sgmRNA (f) was defined by probit analysis using MedCalc software.

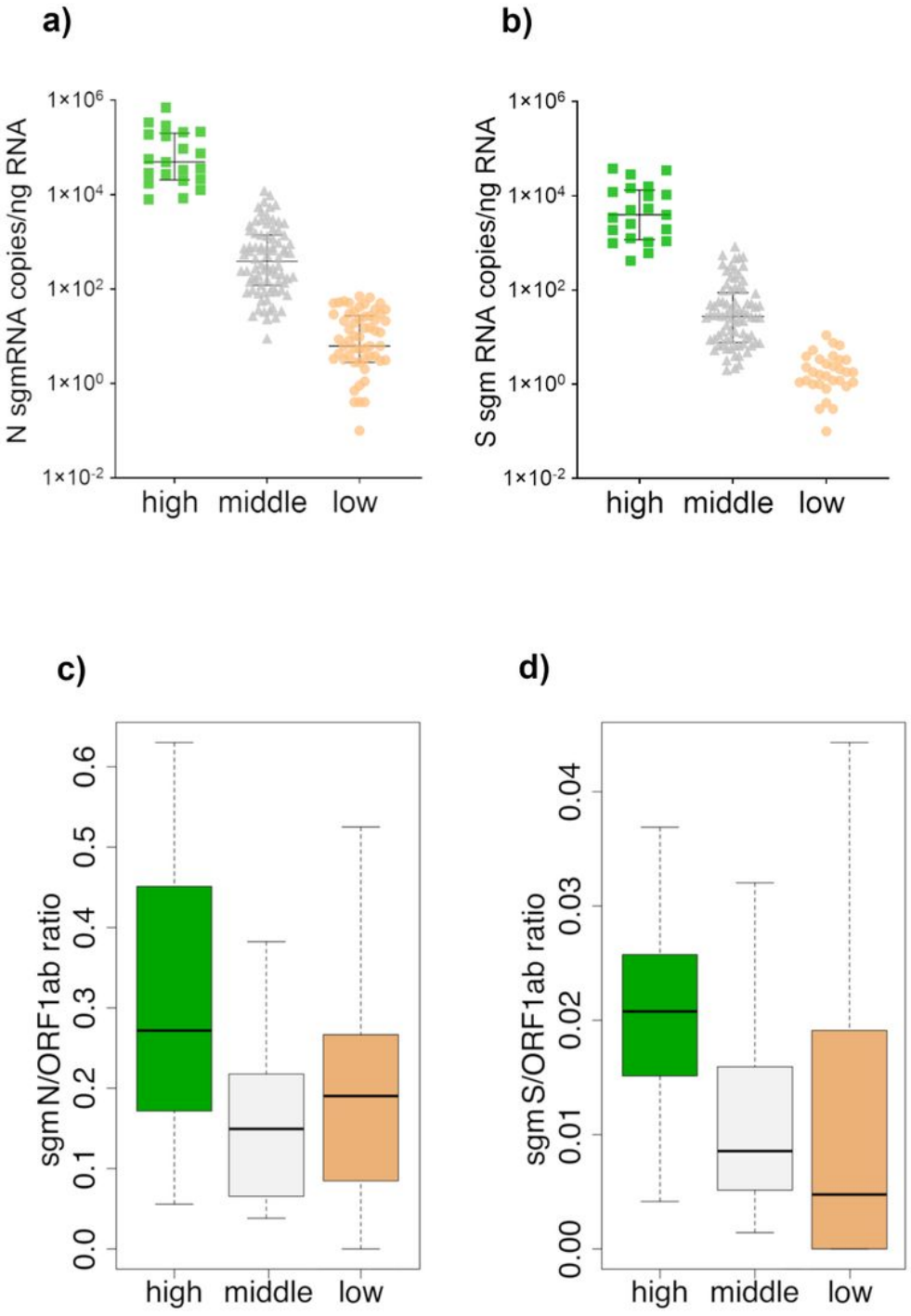

Figure 2

SARS-CoV-2 N and S sub-genomic RNAs absolute quantification and their relative expression to genomic RNA in COVID-19 samples by ddPCR. N sgmRNA (a) and S sgmRNA (b) copies quantification in 166 RNA samples from COVID-19 subjects. Based on ORF1ab gRNA and N and S sgmRNAs quantification, RNA samples were divided in three groups defined as “high” (n = 21, represented by the green square symbol) , “middle” (n = 82, represented by the grey triangle symbol) and “low” (n = 62, represented by the orange

circle symbol) viral RNA content. Values, reported in copies/ng RNA, are expressed as the means of a duplicate assay for each sample. Samples with zero sgmRNAs copies/ng RNA were not plotted in the graphs because logarithmic axes mathematically do not contemplate 0 value. Black lines represent median with relative interquartile range (IQR); p-value < 0.0001 calculated by Kruskal-Wallis test. N sgmRNA (c) and S sgmRNA (d) expression was evaluated as ratio respect to ORF1ab gRNA in “high” (green), “middle” (grey) and “low” (orange) viral RNA content group. Values are reported as sgmRNA/ORF1ab RNA copies/ng RNA. Significance was evaluated comparing “high” group vs “middle”+“low” groups by Mann–Whitney U test /Wilcoxon rank-sum. Data are presented as median with relative interquartile range (IQR). For N sgmRNA: p-value =0.008877; for S sgmRNA: p-value =0.000233.

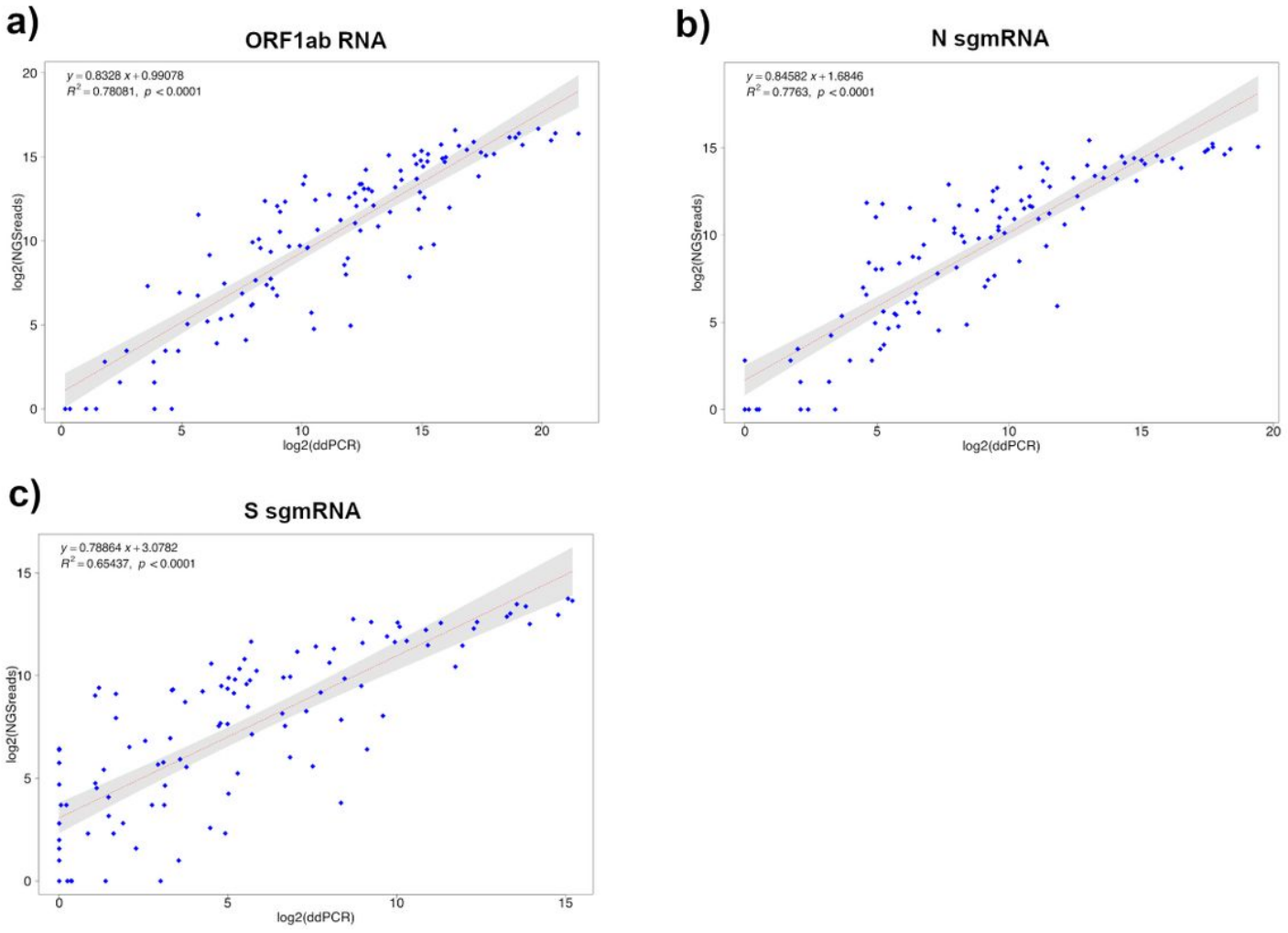

Figure 3

Correlation between ddPCR and transcriptomics data for SARS-CoV-2 ORF1ab, N and S sgmRNA targets. Linear correlation between ddPCR and sequencing quantification of SARS-CoV-2 ORF1ab gRNA (a), N sgmRNA (b) and S sgmRNA (c) calculated on 110 COVID-19 RNA samples, by means of a bivariate linear fit analysis (p-value <0.0001). On X axis: log2 scaled ddPCR quantification; on Y axis: log2 scaled meta-transcriptomics reads count. Regression equation and R2 are reported in the top-left corner. The gray area represents a 0.95 level of confidence interval.

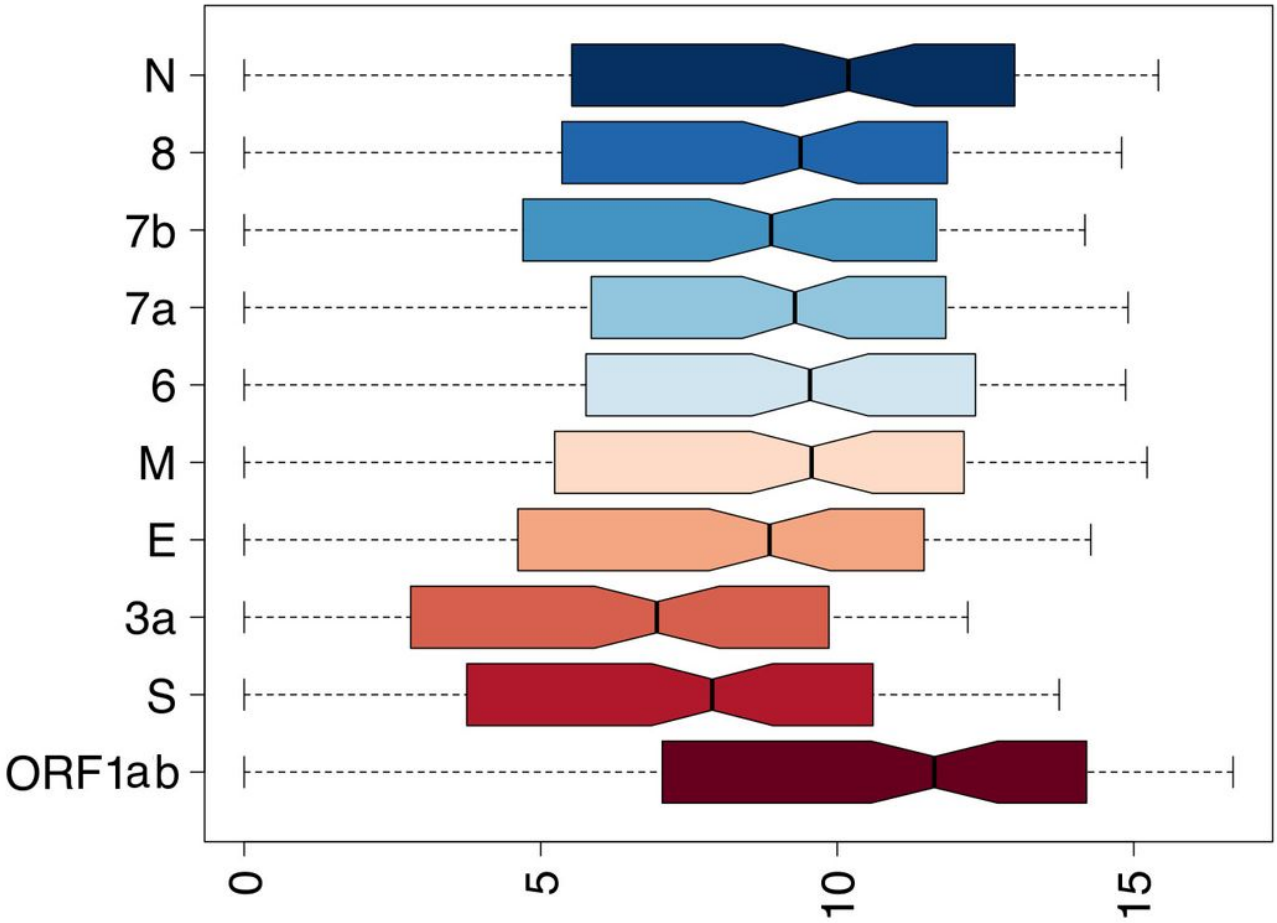

Figure 4

SARS-CoV-2 transcripts expression derived from meta-transcriptomics sequencing Representation by boxplots of log2 scaled counts distributions of meta-transcriptomics reads assigned to each sgmRNAs and to the ORF1ab gene (gRNA). Genes are indicated on the Y axis. Log2 scaled counts on the X axis.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- DescriptionofAdditionalSupplementaryFiles.pdf
- SupplementaryInformation.pdf
- SupplementaryData.xlsx