4-27-2021

Highly Accurate Chip-Based Resequencing of SARS-CoV-2 Clinical Samples

Kendall Hoff

Xun Ding

Lucas Carter

John Duque

Ju-Yu Lin

See next page for additional authors

Authors

Kendall Hoff, Xun Ding, Lucas Carter, John Duque, Ju-Yu Lin, Samantha Dung, Priyanka Singh, Jiayi Sun, Filip Crnogorac, Radha Swaminathan, Emily N. Alden, Xuechen Zhu, Ryota Shimada, Marijan Posavi, Noah Hull, Darrell L. Dinwiddie, Adam M. Halasz, Glenn McGall, Wei Zhou, and Jeremy S. Edwards

LANGMUIR

pubs.acs.org/Langmuir Article

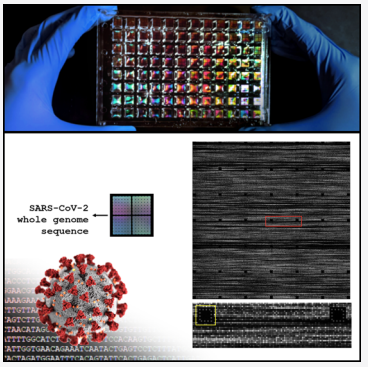
Highly Accurate Chip-Based Resequencing of SARS-CoV-2 Clinical Samples

Kendall Hoff,[#] Xun Ding,[#] Lucas Carter, John Duque, Ju-Yu Lin, Samantha Dung, Priyanka Singh, Jiayi Sun, Filip Crnogorac, Radha Swaminathan, Emily N. Alden, Xuechen Zhu, Ryota Shimada, Marijan Posavi, Noah Hull, Darrell Dinwiddie, Adam M. Halasz, Glenn McGall, Wei Zhou,* and Jeremy S. Edwards*

Cite This: *Langmuir* 2021, 37, 4763–4771 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: SARS-CoV-2 has infected over 128 million people worldwide, and until a vaccine is developed and widely disseminated, vigilant testing and contact tracing are the most effective ways to slow the spread of COVID-19. Typical clinical testing only confirms the presence or absence of the virus, but rather, a simple and rapid testing procedure that sequences the entire genome would be impactful and allow for tracing the spread of the virus and variants, as well as the appearance of new variants. However, traditional short read sequencing methods are time consuming and expensive. Herein, we describe a tiled genome array that we developed for rapid and inexpensive full viral genome resequencing, and we have applied our SARS-CoV-2-specific genome tiling array to rapidly and accurately resequence the viral genome from eight clinical samples. We have resequenced eight samples acquired from patients in Wyoming that tested positive for SARS-CoV-2. We were ultimately able to sequence over 95% of the genome of each sample with greater than 99.9% average accuracy.



INTRODUCTION

To date, there have been over 128 million confirmed cases and nearly 2.8 million deaths worldwide due to COVID-19, caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV2).[1] COVID-19 is highly contagious and rapidly spread within the human population and was defined as a global pandemic in March 2020 by the World Health Organization. Vigilant testing and tracing are essential for controlling the SARS-CoV-2 virus, so a technology to monitor the evolution of the viral genome and the emergence of virus variants and detect possible transmission chains is highly desirable.[2,3] Such a technique requires a rapid, inexpensive, and accurate tool that can detect genetic variants within the SARS-CoV-2 genome with single base resolution.

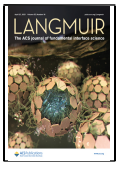
Next-generation sequencing (NGS) methods for sequencing the viral genome are established and accurate. However, it is hard to take advantage of the throughput of NGS by multiplexing a large number of samples, and major costs are associated with library preparation. Therefore, the cost of NGS sequencing of viral genomes is relatively high (cost per assembled base of a viral genome is ∼10 000 times more expensive than the human genome). To remedy this problem, a number of studies have demonstrated the capability of DNA arrays in the detection,[4] surveillance,[5,6] and screening of multiple viral strains, including coronavirus.[7,8] However, DNA arrays originally only used a limited number of oligonucleotide features, leading to a bias in genome coverage. However, improvements in DNA array technology and decreasing production costs led to the development of whole genome tiling arrays with high-density oligonucleotide features that cover each base in the genome with sense and antisense probes to the genome of interest.[9,10] These tiling arrays can be applied to resequence the genome (identify single nucleotide variants) from clinical samples at a very low cost.[11,12]

Here, we describe a full genome tiling array with more than 240 000 features that provide 2× coverage of the entire SARS-CoV-2 genome and the use of such a genome tiling array to sequence the genome from eight clinical samples from SARS-CoV-2-positive subjects. Our results indicate that we can sequence at least 95% of the viral genome with on average greater than 99.9% accuracy.

EXPERIMENTAL SECTION

Sample Preparation for Illumina Sequencing. Samples were prepared as previously described using the ARTIC sequencing methods. In brief, cDNA was prepared from total RNA extracted from clinical samples using SuperScript IV (SSIV, Thermo Scientific)

Received: October 6, 2020
Revised: March 31, 2021
Published: April 13, 2021



https://doi.org/10.1021/acs.langmuir.0c02927
Langmuir 2021, 37, 4763–4771

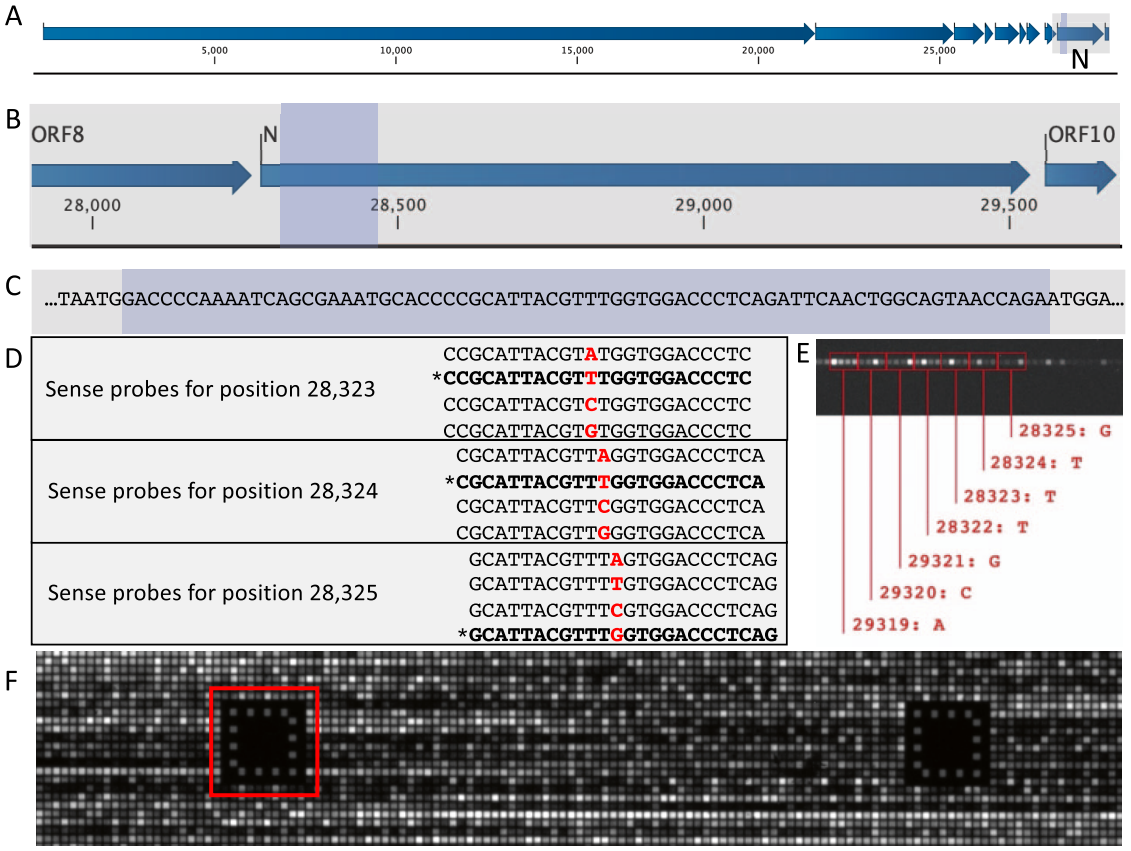


Figure 1. (A) ~30 000 base SARS-CoV-2 genome. (B) Zoomed into the N gene of the SARS-CoV-2 genome covering ~2000 bases. (C) Position 28 274-29 533 of the SARS-CoV-2 genome, which is amplified by the CDC N1 primers. (D) Three different sense probe sets; each probe set consists of four features synthesized on the genome tiling array to interrogate the middle base position (highlighted in red font). The feature whose sequence is consistent with the reference (NC_045512.2) is highlighted in bold and denoted with an asterisk. (E) Extracted regions from the tiling array for genome positions 29 319-29 321 and 28 322-28 325 illustrating how the feature with the highest intensity is used to call the base at each position in the SARS-CoV-2 genome. (F) An image illustrating the resulting confocal scan of the genome tiling array when hybridized to a SARS-CoV-2 sample. One alignment marker is highlighted, which is used for correctly extracting the intensities for each probe set.

and random hexamer priming. The resultant cDNA was amplified in two polymerase chain reaction (PCR) reactions using the ARTIC Pool1 and Pool2 SARS-CoV-2 v3 primer sets and Q5 high fidelity DNA polymerase (NEB). Following PCR, samples were purified using AMPure XP SPRI beads (Beckman Coulter). Illumina adaptors were added using the NEBNext Ultra II DNA Library Prep Kit (NEB), and SPRI bead purification was repeated.

Sample Preparation for Sequencing on Chips. To prepare samples for hybridization to the chips, 0.05 μL of the purified PCR product was amplified using the ARTIC protocol and Pool1 and Pool2 v3 primer sets for 35 cycles with 50 μM biotin-11-dUTP (Jena Biosciences) added to the reaction mixture. Pool1 and Pool2 were combined for each sample and fragmented using DNase I (D4263, Sigma-Aldrich). Two thousand Kunitz units of lyophilized DNase I was resuspended on ice using 2 mL of 1× DNase I Buffer (10 mM Tris−HCl pH 7.5, 2.5 mM $MgCl_2$, 0.1 mM $CaCl_2$). The resuspended enzyme was diluted 1000-fold using 1x DNase I Buffer, and an equal volume was added to samples prewarmed to 37 °C. Samples were incubated for 30 min at 37 °C, and the reactions were stopped by adding ethylenediaminetetraacetic acid (EDTA) to a final concentration of 12.5 mM and incubating for 20 min at 75 °C.

Hybridization. Forty-five microliters of the fragmented sample was hybridized overnight at 45 °C to the chip in a 60 μL final volume containing 5 mM EDTA, 6.25 mM HEPES pH 8.0, 312.5 mM NaCl, 1.25% Ficoll 400, 0.5 nM Cy3-AM1 (GCTGTATCGGCTGAATCG-TA). Following hybridization, chips were washed for 10 min at room temperature in Wash A (2× SSC, 0.1% TWEEN-20) and then for 10 min at 39 °C in Wash B (0.5×SSC, 0.1% TWEEN-20). Chips were stained for 15 min at room temperature using 0.02 mg/mL Cy3-Streptavidin (Thermo) in 4× SSC and washed for 5 min at room temperature using 4× SSC. Chips were scanned using a custom-built confocal scanner for 0.5, 1, 4, and 8 s in the green (Cy3) channel in 4× SSC.

Reverse Transcription Polymerase Chain Reaction (RT-PCR) Using the CDC N1 and N2 Primer Sets. RT-PCR was performed using the CDC N1 and N2 primer sets. Five hundred copies of the SARS-CoV-2 genome (Twist Biosciences) were amplified in a 25 μL reaction volume using the SuperScript IV One-Step kit (Thermo Fisher) containing 250 nM of each primer (Primer mix 1−152), 50 μM biotin-11-dUTP (Jena Biosciences), and 0.5 μL of RT enzyme mix. Cycling was performed as follows: 12 min at 45 °C, 2 min at 98 °C, 40 cycles of 10 s at 98 °C, 11 s at 61 °C, and 11 s at 72 °C, followed by a final extension of 2 min at 72 °C. The PCR product was hybridized to the array as described above with one modification: hybridization was performed for 90 min instead of overnight.

Base Calling Approach. *Base Calling for N1 and N2 Amplicon Experiments.* RT-PCR products are hybridized to chips and imaged on the custom-built confocal scanner. A synthetic alignment marker sequence, "Cy3-AM1", is added to the hybridization mixture containing the RT-PCR products and hybridization buffer. This sequence hybridizes in a square pattern at predetermined regularly spaced locations across the chip, as illustrated in Figure 1. The images are stitched together and gridded to create a composite image and using the positional information from the Cy3-AM1 sequences, and intensities for each feature on the chip are extracted from the image and stored in a. csv text file. Each base has two corresponding probe sets: one for the sense strand and one for the antisense strand. Each probe set consists of four features, one for each base, ATCG; thus,

https://doi.org/10.1021/acs.langmuir.0c02927
Langmuir 2021, 37, 4763−4771

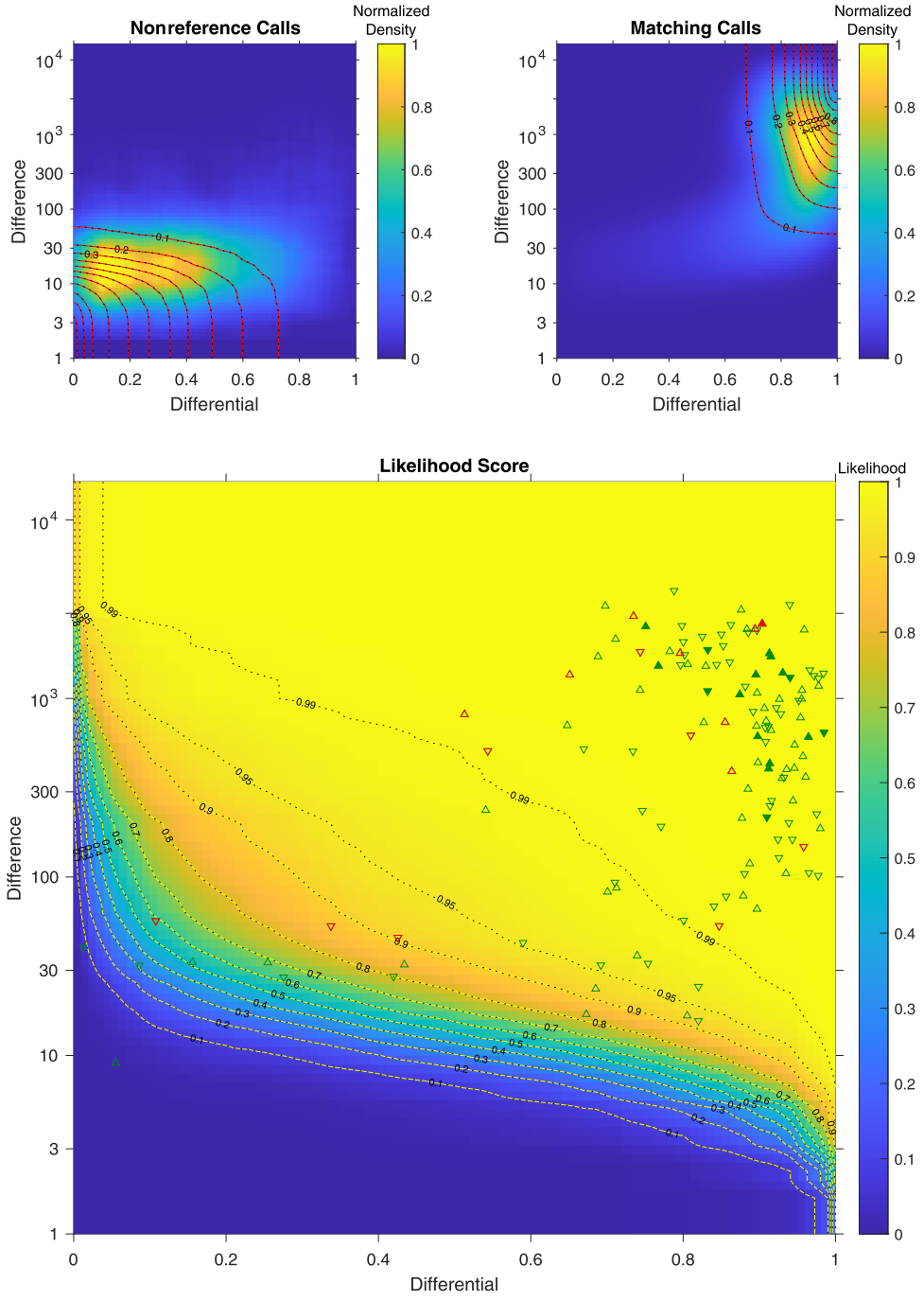


Figure 2. Development of the maximum likelihood base caller for SARS-CoV-2 genome sequencing using full genome tiling arrays. (A) Density plot derived from a two-dimensional (2D) histogram of the incorrect calls from all tiling array probe sets including sense and antisense data for a single exposure. This image was constructed by "calling" each base in the genome using all probe sets. With this approach, each base is called twice, once from the sense probe sets and once from the antisense probe sets. The difference and differential of a call are included in the histogram if the base call does not match the reference. Contours indicate a likelihood function proportional to the two-dimensional cumulative sum of the density; the sum is normalized to indicate the fraction of wrong calls whose quality parameters are higher than the given point; higher values indicate a higher likelihood that a call is "wrong". (B) Same as A, except the 2D histogram is for the correct calls from all tiling array probe sets. Contours indicate a cumulative sum of the density, normalized to indicate the likelihood that a call is correct. It can be observed that the distribution of the difference and differential of "correct" calls is very different from the "incorrect" calls. (C) Using the observation from panels A and B, we constructed a function to assign the likelihood that a probe set is calling the correct base for a given position. The dotted contours define the (combined) likelihood that a probe set is correctly calling the correct base, based on the difference and differential score for that probe set. The triangle points on the plot illustrate the different and differential values for probe sets for all variant sites that have been reported in Wyoming samples in the GISAID database as of August 2020. The green triangles indicate that the base call from this scan suggests a reference call, whereas a red triangle indicates that the call suggests a nonreference base at this position. If the triangle points up, this is from the sense probe set, whereas a downward pointed triangle indicates the data is for the antisense probe set. The triangle outline is filled in if this probe set from this scan resulted in the highest likelihood for the correct call among all scans for this position. This image is the 4 s scan of the WY64 sample. To call all of the bases, we construct a similar likelihood function for each scan, and this information was combined as described in the methods to make the final base call.

https://doi.org/10.1021/acs.langmuir.0c02927
Langmuir 2021, 37, 4763−4771

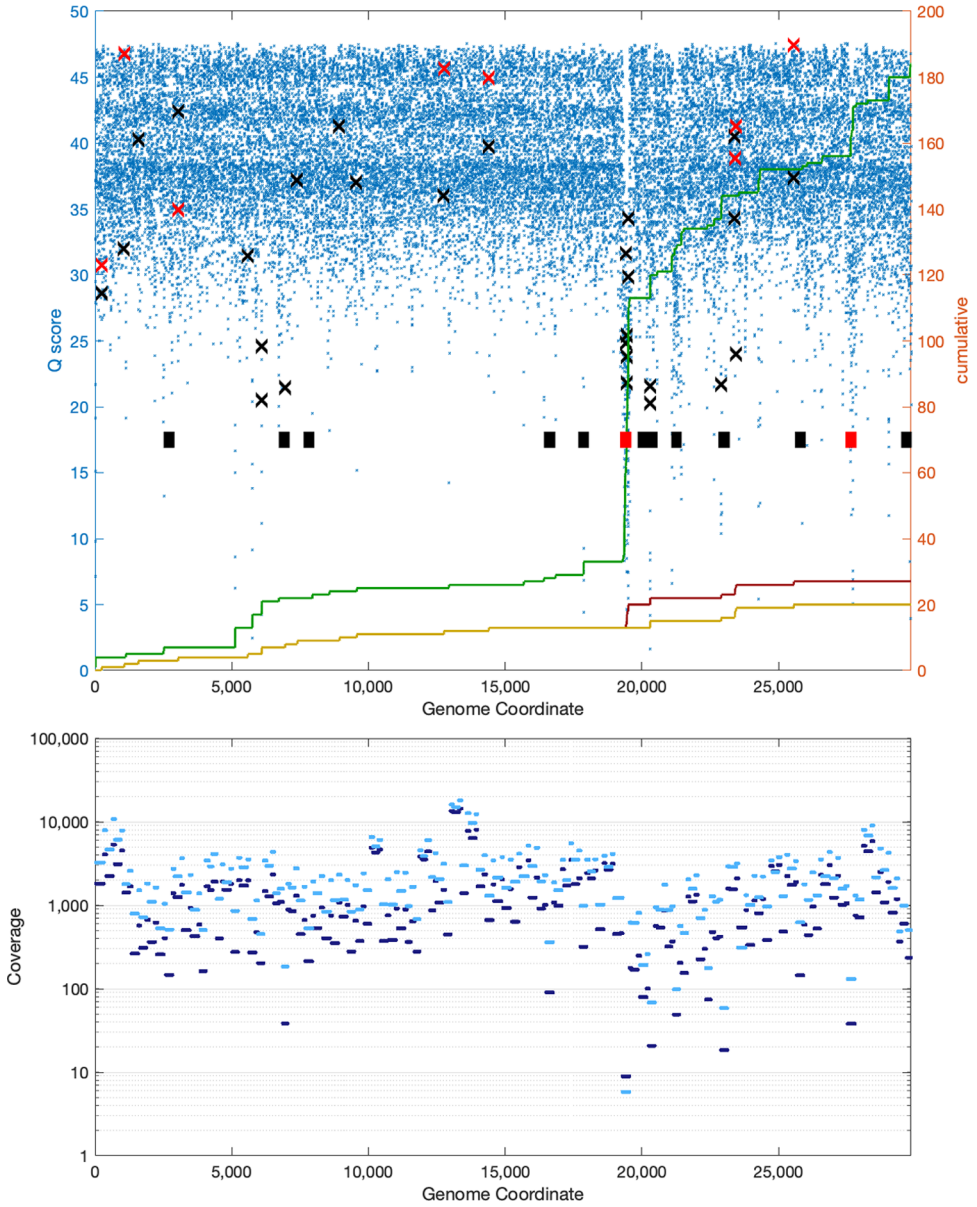


Figure 3. Sequencing accuracy across the SARS-CoV-2 genome. (A) The Phred score (left axis) for all bases in the SARS-CoV-2 genome from the tiling array full genome sequencing of WY64. The positions of all variant calls are highlighted by Black "X", and a Red X indicates this is a correct variant call (confirmed by the Illumina short read sequencing data). The cumulative sum of noncalls (blue line), variant calls with a Phred score greater than 20 (cyan line), and variant calls that have a Phred score greater than 20 and pass the low coverage filter (red line) is shown on the secondary *Y*-axis. (B) Comparison of the tiling array genome sequencing quality scores and variant calls to the amplicon coverage from short read Illumina sequencing data. The light blue (right axis) lines indicate the sequence coverage from the WY64 sample, and the dark blue lines indicate the average sequencing coverage over all Wyoming GISAID samples as of 8/2020.

there are a total of eight features per base position. The intensity for each feature is stored in the. csv file. Feature intensities within a probe set are ranked separately for the sense and antisense probe sets for each base ($I_3 \geq I_2 \geq I_1 \geq I_0$). The difference ($D = I_3 - I_0$) and differential $\left(d_{\text{rel}} = \frac{I_3 - I_2}{D}\right)$ are calculated for the sense and antisense probe sets separately.

A putative base is called for the probe set with the highest intensity for the sense set and separately for the antisense set. If the differential is smaller than a threshold (2% default) and/or if the difference is smaller than a threshold (20 for low-intensity chips or 100 for high-intensity chips), the base is called "*N*" for unknown. If a base is not called *N*, the differentials for sense or antisense set are compared, and the set with a larger differential is used to call the base. FASTA files are generated from these parameters and aligned to references.

Whole Viral Genome Base Calling Using Likelihood Maps for Multiple Exposures. Multiple successive images (0.5, 1, 4, and 8 s exposures) are taken for one chip, resulting in a set of N_E measures (typically $N_E = 3$ or 4). For each position within the genome, we obtain $2N_E$ sets (sense and antisense) of four intensities $\{\vec{I}_j\}_{j=1\cdots 2N_E}$, $\vec{I} = [I_A, I_T, I_C, I_G]$. The intensities within each set are sorted $I_{\max} \equiv I_3 \geq I_2 \geq I_1 \geq I_0 \equiv I_{\min}$, and the base corresponding to the highest one is tentatively called for the respective location. Ideally, the call from each of the sets would indicate the same base, but this is not always the case. We use the difference $D \equiv I_{\max} - I_{\min}$, differential $d_{\text{rel}} \equiv (I_{\max} - I_2)/D$, and signal magnitude $I_{\max} = I_3$ to assign a credibility score to each base call from the sense and antisense probe sets at each exposure.

We follow a Bayesian-inspired approach, where we rely on a reference genome to indicate whether a preliminary call is likely

https://doi.org/10.1021/acs.langmuir.0c02927
Langmuir 2021, 37, 4763−4771

Table 1. Sequencing Accuracy and Coverage for the Eight Clinical Samples[a]

	Q20 filter				low coverage filter				
sample	noncalls	variants	accuracy	coverage	noncalls	variants	accuracy	coverage	illumina coverage
WY24	353	39	99.89%	98.82%	1404	36	99.90%	95.30%	99.86%
WY26	493	47	99.87%	98.35%	1431	35	99.90%	95.21%	99.07%
WY32	615	48	99.86%	97.94%	1473	32	99.92%	95.07%	99.76%
WY36	354	52	99.85%	98.82%	1365	36	99.90%	95.43%	99.07%
WY41	810	68	99.79%	97.29%	1709	49	99.85%	94.28%	99.59%
WY44	266	33	99.92%	99.11%	1295	26	99.94%	95.66%	99.00%
WY59	570	57	99.83%	98.09%	1512	41	99.88%	94.94%	98.63%
WY64	183	27	99.94%	99.39%	1247	16	99.97%	95.83%	99.18%

[a]The number of "noncalls" refers to the number of base positions that were not called (or removed) because they have a Phred score lower than 20 or reside in one of the six lowest short read Illumina sequencing coverage regions. "Variants" is the number of putative variants from the sample called by the genome tiling array sequencing. "Accuracy" is the number of correctly called positions divided by the number of base calls. "Coverage" refers to the percentage of the SARS-CoV-2 genome for which we have called a base. The "illumina coverage" column refers to the sequencing coverage of the samples in the GISAID database.

correct. After identifying the base with the highest intensity for each of $2N_G$ sets for a given exposure time, we group the sets into two groups, those that match the reference call and those that do not (Figure 2). The distributions of the "quality parameter" values (D,d_{rel}) for these groups follow different geometric patterns, as illustrated in Figure 2. Each of the parameters taken separately is informative, in that higher values indicate higher confidence that a call is correct. However, the geometry of the distributions in Figure 2A,B illustrates the difficulties associated with comparing two calls, where one has a higher absolute difference and the other has a higher relative differential.

To illustrate our methods, the likelihood score using one quality parameter is described. If we rely on one quality parameter x, which could be one of D,d_{rel},I_{max} to infer the likelihood that a read is correct, we can make the following argument. Without knowledge of the quality parameter x, the probability that a call is correct is approximated by the fraction of calls that are correct, $P_{correct} \approx N_{correct}/N_{total}$. Regarding the dependence on x, first note that the likelihood that a call is correct increases with x, that is, a call with parameter value $x' > x$ is more likely correct than the one with x. Second, given that a call is correct, its parameter x follows some (continuous) distribution $\rho_{correct}(x)$. This local density of correct calls is not a good estimate that a call is correct since it will decrease as x increases beyond the location of the peak density. The local probability that a call is correct should follow from the ratio of the local density of correct calls to that of all calls (correct and incorrect); however, these are too small to estimate far away from the center of the respective distributions.

We assign a likelihood that a call with parameter value $x = \lambda$ is correct by comparing the fraction of correct calls with lower or higher x values. If the likelihood of a call being correct increases with x, then all of the calls with $x < \lambda$ are less likely and those with $x > \lambda$ are more likely to be correct. A call whose x exceeds that of all correct calls is assigned the full probability $P_{correct}$. Thus, the likelihood that a call with parameter x is correct is (conservatively) estimated by

$$l_{correct}(x) = P_{correct}\int_{x}^{x_{min}} \rho_{correct}(x')dx'.$$
$$= \frac{N_{correct}}{N_{total}}\int_{x}^{x_{min}} \rho_{correct}(x')dx'$$
$$= \frac{n_{correct}(x' \le x)}{N_{total}}$$

A similar argument can be made for the likelihood that the call is wrong, except this will decrease as x increases, working out to $l_{incorrect}(x) = n_{incorrect}(x' \ge x)/N_{total}$. We obtain a normalized score by combining the two likelihoods: $S(x) = \frac{l_{correct}}{l_{correct} + l_{incorrect}}$.

To use two parameters, we construct a credibility score that maps a pair of parameters (D,d_{rel}) to a continuous value between $[0,1]$ as $S = \frac{w_{correct}}{w_{correct} + w_{incorrect} + 1}$, where

$$w_{correct}(D, d_{rel}) \equiv n_{correct}(D' \le D, d'_{rel} \le d_{rel}),\ w_{incorrect}(D, d_{rel})$$
$$= n_{incorrect}(D' \ge D, d'_{rel} \ge d_{rel})$$

Without knowledge of its quality parameters $(x,y) = (D,d_{rel})$, the probability that a given call is correct is approximated by $N_{correct}/N_{total}$. We assume that a call with parameters (x,y) is more likely to be correct than a call with lower parameters $x' \le x, y' \le x$. Let $\rho_{correct}(x,y)$ denote the normalized probability density for correct calls in two dimensions. Given that a read is correct, the probability that its quality parameters are both below (x,y) is

$$P(x' \le x, y' \le y | \text{for a correct call}) = \int_{x_{min}}^{x}\int_{y_{min}}^{y} \rho_{correct}(x', y')$$
$$dx', dy' \approx \frac{n_{correct}(x' \le x, y' \le y)}{N_{correct}}$$

Thus, $w_{correct}(D,d_{rel})/N_{total}$ approximates the probability that a call is correct and has quality parameters equal or worse than (D,d_{rel}). Effectively, this approach assumes that the likelihood that a read with D,d_{rel} is correct is proportional to the number of reference-matching reads with both $D' \le D$ and $d'_{ref} \le d_{ref}$. Similarly, $w_{incorrect}(D,d_{rel})/N_{total}$ approximates the probability that a call is wrong and has equal or better (higher) quality parameters.

For calls whose parameters are above those of every nonmatching call, $w_{incorrect} = 0$. We added the offset term in the denominator of S to avoid a "perfect" score $S = 1$ for such calls, so their rank among correct calls is still taken into account.

The score $S(D,d_{rel})$ described above is calculated efficiently for every read of every base using a 2D histogram of the distributions of matching and nonmatching calls. The raw bin counts are converted into a smoothened local density obtained as a moving average over neighboring bins, and the averages are used to compute cumulative sums in both directions. Each read is assigned a score corresponding to the 2D bin into which it falls. Reads for a given genome location are sorted by their score, and the final call is made consistent with the top-scoring read. The S score is interpreted as the likelihood that the read is correct.

FASTA files for sequencing results generated from microarray images are aligned to the reference sequences for SARS-CoV-2 and human RNase P, NC_045512.2 Wuhan seafood market pneumonia virus isolate Wuhan-Hu-1, complete genome (GISAID accession EPI_ISL_402125).

RESULTS AND DISCUSSION

SARS-CoV-2 Genome Tiling Arrays. We have constructed a genome tiling array of the SARS-CoV-2 genome (NC_045512)[13] (Figure 1A). The 3 mm × 3 mm array area is divided into 250 000 "features". Each feature contains densely packed identical single strands of DNA, each 25 bases long and

https://doi.org/10.1021/acs.langmuir.0c02927
Langmuir 2021, 37, 4763−4771

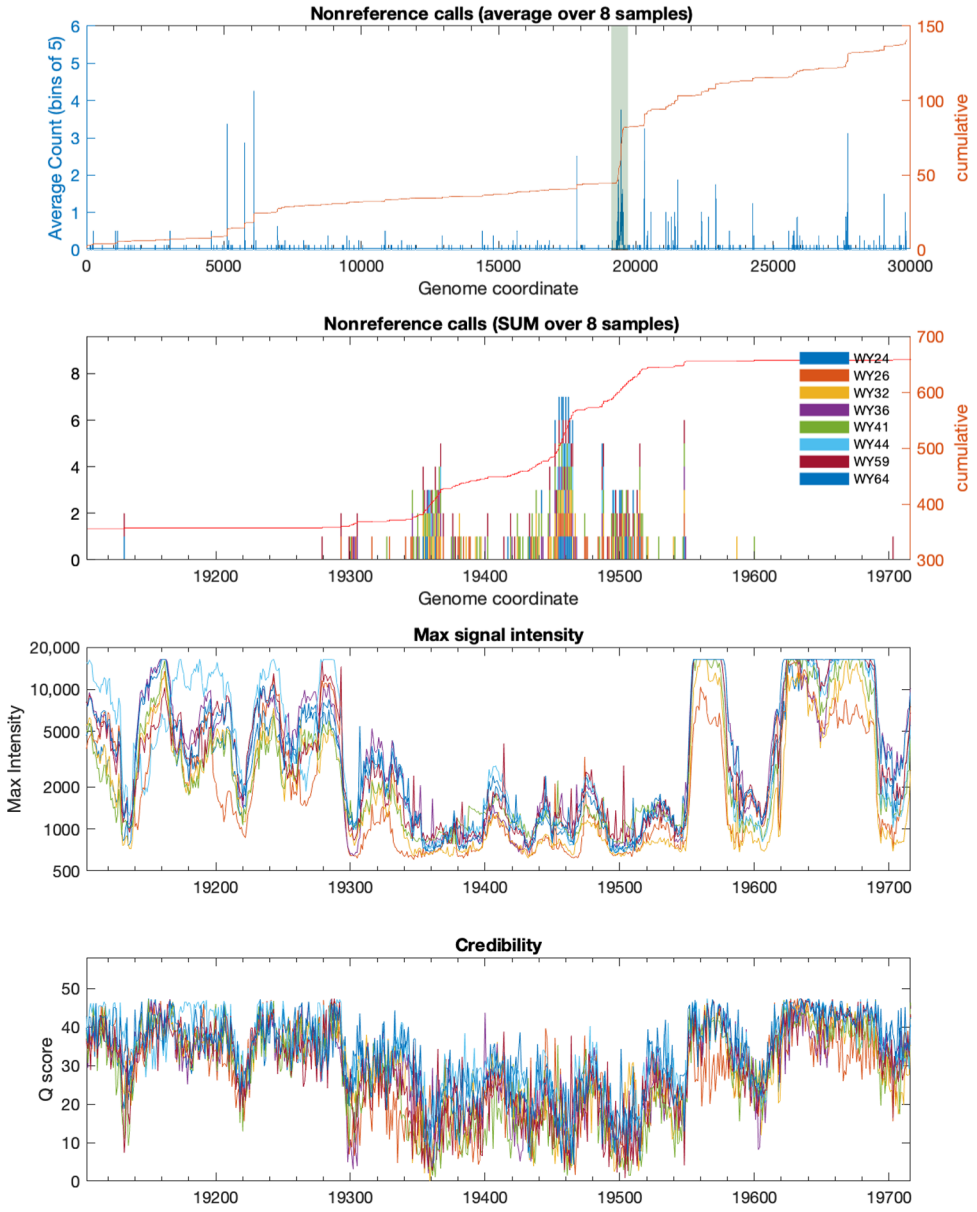


Figure 4. (A) Average number of variant base calls across all eight samples as a function of the genome coordinates. The cumulative sum of the number of variants identified is displayed on the secondary axis. The ∼300 base region between 19 300 and 19 600 is where the largest number of putative variants are called. (B) Same as panel (A), except the *x*-axis spans the region between base positions 19 100 and 19 700. Bars indicate variant calls by sample and location. The cumulative sum reflects the number of variant calls across all samples. (C) The maximum signal intensity from all exposure scans from the chip as a function of genome coordinate [same range as (B)]. From this panel, it can be seen that the maximum signal intensity (for all samples) on the genome tiling array is low in the region from bases 19 300 to 19 600. This region corresponds to the low coverage region from the short read Illumina sequencing data (see Figure 3B). (D) The Phred score of the final base calls from the tiling array as a function of the genome coordinate.

serves as a probe. Each column of the array is divided into probe sets of four features. Within each probe set, each feature consists of ssDNA with a sequence that matches the same portion of the genome except for the 13th base, which always contains an A in the first feature, T in the second, and C and G in the third and fourth, respectively (Figure 1). The probe sets are arranged in rows such that in one row the sequences in successive probe sets tile across the genome and match the genome except for the 13th base (for which only one feature within the set has the matching base) (Figure 1E). The feature within the set that matches the genome sequence will hybridize with the genome fragment. The genome fragments are each tagged with a fluorescent molecule so that in a fluorescence image (Figure 1F) the positions of the hybridized genome molecule "light up" and the base can be identified. There is the possibility of an error if there is more than one region within the genome with identical 25-mer sequences; however, the probability of this occurring is vanishingly small for a sequence of 30 000 bases.

Although the SARS-CoV-2 genome is a single-stranded RNA, during the processing of the clinical sample, the genome is reverse-transcribed and amplified generating a double-stranded DNA. Therefore, the genome tiling arrays were constructed to assay both the sense and antisense strands, and hence, there are two measurements for each base in the SARS-CoV-2 genome. Since the genome consists of ∼30 000 bases,

https://doi.org/10.1021/acs.langmuir.0c02927
Langmuir 2021, 37, 4763−4771

the tiling array possesses approximately 240 000 discrete features (~30 000 bases × 2 strands × 4 features per base). In addition to the genome tiling features, the arrays also contain alignment marks and several control features for human sequences.

Illustration of the Approach by Sequencing of the N1 Gene Sequence. As an initial test, we amplified a region of the N1 and N2 genes in the SARS-CoV-2 genome (synthesized by Twist Biosciences) using the N1 and N2 CDC primer pairs obtained from Integrated DNA Technologies (Supporting Information, Table S1). The PCR products were labeled and hybridized to the tiling array, as described in the Experimental Section. The arrays were imaged using a custom-automated fluorescent scanning confocal microscope. The fluorescence intensity and *x*,*y* coordinates of each feature were stored in a. csv file and mapped to each base in the N1 and N2 gene regions of the SARS-CoV-2 genome.

The intensities from the features containing each of the four features for each base position were ranked $\{I_0 \le I_1 \le I_2 \le I_3\}$. To call the base, we define two parameters that characterize the quality or confidence in the call: the difference ($D = I_3 - I_0$) and differential ($d_{rel} = \frac{I_3 - I_2}{D}$). Each base is called by identifying the feature with the highest intensity for both the sense and antisense features. The base call was typically consistent between the sense and antisense probe sets. However, if they are inconsistent, the probe set with the higher differential is selected as the final call. Using this approach, we were able to sequence the N1 and N2 regions of the SARS-CoV-2 genome with 100% accuracy.

Resequencing the Full SARS-COV-2 Genome. We extended the above technique to sequence all ~30 000 bases of the SARS-CoV-2 genome. We obtained eight SARS-CoV-2-positive clinical samples from the Wyoming Public Health Laboratory, which were blinded prior to resequencing (the samples were unblinding after the base calling was complete). The viral genome in these clinical samples was previously fully sequenced on an Illumina MiSeq instrument, and the results were deposited in the GISAID database. Thus, the sequences obtained from our genome tiling arrays can be directly compared with those obtained from Illumina sequencing. The samples we analyzed were USA/WY-WYPHL-00024/2020, USA/WY-WYPHL-00026/2020, USA/WY-WYPHL-00032/2020, USA/WY-WYPHL-00036/2020, USA/WY-WYPHL-00041/2020, USA/WY-WYPHL-00044/2020, USA/WY-WYPHL-00059/2020, and USA/WY-WYPHL-00064/2020. We will simply refer to these samples as WY24, WY26, WY32, WY36, WY41, WY44, WY59, and WY64, respectively. By the time we acquired these samples, there was a limited amount of material remaining because the samples had already been tested via a standard qPCR clinical test at the Wyoming Public Health Laboratory and had the complete genome sequenced at the University of New Mexico. Therefore, to resequence the genome from these samples using our genome tiling array, we started with the remaining PCR products that were amplified with the ARTIC SARS-CoV-2 v3 primer sets (the sample names were blinded) (Table S1). We further amplified the samples with labeled dUTP and then hybridized the purified PCR products to the SARS-CoV-2 genome tiling arrays, as described in the Experimental Section.

Maximum Likelihood Method to Accurately Sequence the SARS-CoV-2 Genome. The method described above was sufficient for analyzing the N1 and N2 PCR products but was only able to resequence the full SARS-CoV-2 genome to ~98−99% accuracy. Therefore, we developed a more sophisticated maximum likelihood method for calling each base in the SARS-CoV-2 genome. We also noticed that the intensity of the probe sets varied across the entire genome, so we scanned the tiling array with three different exposure times (0.5, 1, and 4 s). The longer exposure times enabled us to accurately call bases in the "weak" intensity regions; however, the "brighter" regions were fully saturated by the longer exposures. In principle, one could combine the three different exposures of the same base position into a single, normalized intensity and call the base with the highest value. However, the resulting integrated reads would still have to be combined for the sense and antisense probe sets, and conflicting calls would require a likelihood-based criterion. This prompted us to adopt an approach that treats the $N_E = 3$ exposures as quasi-independent "reads". We assigned a credibility score to each read based on the difference, D, and differential, d_{rel}, of the intensities, as illustrated in Figure 2. Briefly, for each of the probe sets for a given base location in the genome, we sorted the intensities from the four bases (or features), identified a tentative call (the feature with the highest intensity), and computed D,d_{rel}. We grouped the base calls or reads for all sense and antisense probe sets for each base in the genome for each exposure into likely "correct" and incorrect by comparing the tentative call with the SARS-CoV-2 genome reference sequence and constructed separate density maps in the D,d_{rel} plane for the correct (reference matching) and incorrect (nonmatching) calls (Figure 2A,B). While D and d_{rel} correlate individually with a higher likelihood of correct calls, the two-dimensional densities reveal a pattern where the correct and incorrect calls are concentrated in largely nonoverlapping regions. We developed an approach that takes advantage of this feature. A read with parameters (D,d_{rel}) is assigned a score, $S = w_{correct}/(w_{correct} + w_{incorrect})$, where $w_{correct}$ is based on the number of reads with $D' < D$ and $d'_{rel} < d_{rel}$, and $w_{incorrect}$ is the number of likely incorrect reads with higher D',d'_{rel}. The score is calculated efficiently for every read of every base using a 2D histogram of the distributions of correct and incorrect calls and then computing the cumulative sums for each bin. Reads for a given location are sorted by their score, and the final call is made consistent with the top-scoring read.

Analysis of Resequencing Accuracy from USA/WY-WYPHL-00064/2020. Using the maximum likelihood method described above, we resequenced the SARS-CoV-2 genome from eight clinical samples from the Wyoming Public Health Laboratory. The probability, P, of an incorrect call was determined for each base in the genome. The "Phred scores", Q ($Q = -10 \log_{10} P$), versus genome coordinates for WY64 are shown in Figure 3 (similar figures for all samples are in the Supporting Information, Figures S25−S32). In WY64, there were 183 bases with a Phred score less than 20 (81% of these calls were correct). The short read Illumina sequencing assembly of this sample consisted of 246 Ns or uncalled bases. We filtered the 181 uncalled bases from the assembly and therefore called 29 663 bases (Table 1). Of these called bases, 27 were identified as variants (or nonreference calls). Of the 27 variant calls, we correctly detected eight variants in WY64 that were identified by Illumina short read sequencing (at positions 241, 1059, 3037, 12 756, 14 408, 23 403, 23 453, 25 563); therefore, if we assume that all other bases in this

https://doi.org/10.1021/acs.langmuir.0c02927
Langmuir 2021, 37, 4763−4771

genome are consistent with the reference genome, we made 19 sequencing errors, which is equivalent to 99.94% accuracy.

Next, we further investigated the sequencing errors and found that seven errors were made in the 100 base region beginning at position 19 417. This region is within the larger region that contained the bases that were not called from the Illumina short read sequencing data due to low sequencing coverage (Figure 3). Since we started with the same original PCR products, we filtered variants called by our tiling array within the six regions in the genome with the lowest short read coverage, with the reasoning that the ARTIC SARS-CoV-2 v3 primer sets may not be providing adequate amplification of the respective amplicons in these regions. After filtering calls in these regions, we still called all eight of the variants identified from the Illumina sequencing data, as well as nine additional variants that are presumably incorrect, which resulted in 99.97% sequencing accuracy spanning 95.8% of the SARS-CoV-2 genome. Equivalent results for all eight sequenced strains are shown in Table 1 (figures equivalent to Figure 3 for all samples are available in the Supporting Information, Figures S25−S32).

Low-Quality Assembled Regions from the Tiling Array. In the GISAID database, the WY64-assembled FASTA file contains "Ns" in the region between 19 300 and 19 547. This is consistent with the region of the genome that had the lowest short read sequencing coverage. In this region, after filtering the base calls with a Phred score below 20, our genome tiling array called 167 bases, and 160 of these bases matched the reference base. We also detected seven putative variants called by the tiling array with a Q-score higher than 20. Next, we blasted the 248 base consensus sequence (the region unassembled from the Illumina data, without filtering any bases) assembled from the tiling array against the SARS-CoV-2 reference (NC_045512) and found our sequence to be 94% similar to the reference, with 24 mismatches. This indicates that our genome tiling array can accurately identify the sequences as SARS-CoV-2 even with very little starting material.

We further explored the impact of low amplicon abundance (measured by low coverage in the Illumina short read data) by analyzing the Phred score we obtained from the genome tiling array. In Figure 4A, we plot positions within the genome, where variants are called within the SARS-CoV-2 genome. It can be seen in this figure that variants are called at a high density at a few discrete locations within the genome. Some of these locations are known to have low coverage in the Illumina short read sequencing data and are thus presumed to have poor amplification of the respective amplicons (Figure 4B). To further explore the amplification of these regions of the genome, we plotted the maximum signal intensity from each probe set for each position in the genome. Figure 4C focuses on the region discussed above (between positions 19 300 and 19 547), and it can be seen that the signal intensity is reduced in this region of the genome, which is consistent with the low short read sequencing coverage in this region.

CONCLUSIONS

As the novel SARS-CoV-2 virus continues to impact the world, it is imperative that viral genome evolution and the tracing of viral spread are closely monitored. With researchers working toward minimizing the fatalities due to COVID-19, there is a need for rapid and cost-effective monitoring of viral variants. We believe our technology meets this need. We have designed and constructed a tiled genome array for rapid and inexpensive SARS-CoV-2 genome resequencing. We have resequenced eight clinical samples to demonstrate the ability of this array to accurately sequence over 95% of the viral genome. Additionally, we have shown that the primary variable limiting our accuracy and sequencing coverage is the ARTIC multiplex PCR primer sets, which do not amplify all amplicons with sufficient efficiency. Therefore, with improved amplification of the viral genome, we anticipate that over 99% of the genome can be sequenced with an accuracy greater than 99.9% using a genome tiling array.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.langmuir.0c02927.

List of primer sequences used (Table S1); display the full set of density plots used in the maximum likelihood method (Figures S1−S24); and plot the Q-score versus genome coordinate for all eight clinical samples (Figures S25−S32) (PDF)

AUTHOR INFORMATION

Corresponding Authors

Wei Zhou − *Centrillion Technologies, Palo Alto, California 94303, United States*; Email: wzhou@centrilliontech.com

Jeremy S. Edwards − *Department of Chemistry and Chemical Biology, University of New Mexico, Albuquerque, New Mexico 87131, United States*; orcid.org/0000-0003-3694-3716; Email: jsedward@unm.edu

Authors

Kendall Hoff − *Centrillion Technologies, Palo Alto, California 94303, United States*

Xun Ding − *Centrillion Technologies, Palo Alto, California 94303, United States*

Lucas Carter − *Centrillion Technologies, Palo Alto, California 94303, United States*

John Duque − *Centrillion Technologies, Palo Alto, California 94303, United States*

Ju-Yu Lin − *Centrillion Technologies, Palo Alto, California 94303, United States*

Samantha Dung − *Centrillion Technologies, Palo Alto, California 94303, United States*

Priyanka Singh − *Centrillion Technologies, Palo Alto, California 94303, United States*

Jiayi Sun − *Centrillion Technologies, Palo Alto, California 94303, United States*

Filip Crnogorac − *Centrillion Technologies, Palo Alto, California 94303, United States*

Radha Swaminathan − *Department of Chemistry and Chemical Biology, University of New Mexico, Albuquerque, New Mexico 87131, United States*

Emily N. Alden − *Department of Chemistry and Chemical Biology, University of New Mexico, Albuquerque, New Mexico 87131, United States*

Xuechen Zhu − *Department of Chemistry and Chemical Biology, University of New Mexico, Albuquerque, New Mexico 87131, United States*; orcid.org/0000-0001-7458-465X

https://doi.org/10.1021/acs.langmuir.0c02927
Langmuir 2021, 37, 4763−4771

Ryota Shimada − *Department of Chemistry and Chemical Biology, University of New Mexico, Albuquerque, New Mexico 87131, United States*
Marijan Posavi − *Department of Chemistry and Chemical Biology, University of New Mexico, Albuquerque, New Mexico 87131, United States*
Noah Hull − *Wyoming Public Health Laboratory, Wyoming Department of Health, Cheyenne, Wyoming 82007, United States*
Darrell Dinwiddie − *Department of Pediatrics, University of New Mexico Health Sciences Center, Albuquerque, New Mexico 87131, United States*
Adam M. Halasz − *Department of Mathematics, West Virginia University, Morgantown, West Virginia 26506, United States*
Glenn McGall − *Centrillion Technologies, Palo Alto, California 94303, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.langmuir.0c02927

Author Contributions

[#]K.H. and X.D. contributed equally to this work.

Author Contributions

K.H. and X.D. conducted the experiments. L.C., J.D., J.-Y.L., S.D., P.S., F.C., and J.S. designed and constructed the genome arrays. R.S., E.N.A., X.Z., R.S., M.P., and A.M.H. developed the base calling methods. N.H. and D.D. provided access to the clinical samples. J.S.E. and W.Z. conceived of and directed the study. All authors drafted the manuscript.

Notes

The authors declare the following competing financial interest(s): The Centrillion affiliated authors are employees and the company may commercialize the work described herein.

■ ACKNOWLEDGMENTS

The authors thank Dr. Ori Sargsyan for assistance with bioinformatics. This research was partially supported by the UNM Comprehensive Cancer Center Support Grant NCI (P30CA118100) and an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health (P20GM103451).

■ REFERENCES

(1) WHO. Coronavirus Disease (COVID-19) Dashboard. World Health Organization: Geneva, 2020. https://covid19.who.int.
(2) He, X.; Lau, E. H. Y.; Wu, P.; Deng, X.; Wang, J.; Hao, X.; Lau, Y. C.; Wong, J. Y.; Guan, Y.; Tan, X.; Mo, X.; Chen, Y.; Liao, B.; Chen, W.; Hu, F.; Zhang, Q.; Zhong, M.; Wu, Y.; Zhao, L.; Zhang, F.; Cowling, B. J.; Li, F.; Leung, G. M. Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nat. Med.* **2020**, *26*, 672−675.
(3) Gardy, J. L.; Loman, N. J. Towards a genomics-informed, real-time, global pathogen surveillance system. *Nat. Rev. Genet.* **2018**, *19*, 9−20.
(4) Drmanac, R.; Drmanac, S.; Baier, J.; Chui, G.; Coleman, D.; Diaz, R.; Gietzen, D.; Hou, A.; Jin, H.; Ukrainczyk, T.; Xu, C. DNA Sequencing by Hybridization with Arrays of Samples or Probes. In *DNA Arrays Methods and Protocols*; Humana Press: Clifton, N.J, 2001; Vol. *170*.
(5) Shendure, J.; Mitra, R. D.; Varma, C.; Church, G. M. Advanced sequencing technologies: methods and goals. *Nat. Rev. Genet.* **2004**, *5*, 335−344.
(6) Miller, M. B.; Tang, Y. W. Basic concepts of microarrays and potential applications in clinical microbiology. *Clin. Microbiol. Rev.* **2009**, *22*, 611−633.
(7) Wilson, C. H.; Tsykin, A.; Wilkinson, C. R.; Abbott, C. A. Experimental Design and Analysis of Microarray Data. In *Applied Mycology and Biotechnology*; Elsevier, 2006; Vol. *6*, pp 1−36.
(8) Schena, M.; Renu Heller, D. S.; Chai, A.; Brown, P. O.; Davis, R. W. Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 10614−10619.
(9) Kumar, D.; Sonia Sheoran, V. C.; Singh, R.; Sharma, P.; Jaiswal, S.; Iquebal, M. A.; Jaiswar, A.; Jaisri, J.; Angadi, U. B.; Rai, A.; Singh, G. P.; Kumar, D.; Tiwari, R. Characterization of genetic diversity and population structure in wheat using array based SNP markers. *Mol. Biol. Rep.* **2020**, *47*, 293−306.
(10) Mockler, T. C.; Chan, S.; Sundaresan, A.; Chen, H.; Jacobsen, S. E.; Ecker, J. R. Applications of DNA tiling arrays for whole-genome analysis. *Genomics* **2005**, *85*, 1−15.
(11) Pihlak, A.; Bauren, G.; Hersoug, E.; Lonnerberg, P.; Metsis, A.; Linnarsson, S. Rapid genome sequencing with short universal tiling probes. *Nat. Biotechnol.* **2008**, *26*, 676−684.
(12) Jaing, C. J.; McLoughlin, K. S.; Thissen, J. B.; Zemla, A.; Gardner, S. N.; Vergez, L. M.; Bourguet, F.; Mabery, S.; Fofanov, V. Y.; Koshinsky, H.; Jackson, P. J. Identification of Genome-Wide Mutations in Ciprofloxacin-Resistant F. tularensis LVS Using Whole Genome Tiling Arrays and Next Generation Sequencing. *PLoS One* **2016**, *11*, No. e0163458.
(13) Wu, F.; Zhao, S.; Yu, B.; Chen, Y. M.; Wang, W.; Song, Z. G.; Hu, Y.; Tao, Z. W.; Tian, J. H.; Pei, Y. Y.; Yuan, M. L.; Zhang, Y. L.; Dai, F. H.; Liu, Y.; Wang, Q. M.; Zheng, J. J.; Xu, L.; Holmes, E. C.; Zhang, Y. Z. A new coronavirus associated with human respiratory disease in China. *Nature* **2020**, *579*, 265−269.