Supplementary Information

COVID-19 (SARS-CoV-2) outbreak monitoring using wastewater-based epidemiology in Qatar

Jayaprakash Saththasivam [a,1], Shimaa S. El-Malah[a,1], Tricia A Gomez [a], Khadeeja Abdul Jabbar [a]; Reshma Remanan [a], Arun K K [a], Oluwaseun Ogunbiyi[a], Kashif Rasool [a], Sahel Ashhab [a], Sergey Rashkeev [a], Meryem Bensaad [b], Ayeda A Ahmed [b], Yasmin A Mohamoud [b], Joel A Malek [b], Laith J Abu Raddad [c], Andrew Jeremijenko [d], Hussein A Abu Halaweh [e], Jenny Lawler [a]* Khaled A. Mahmoud [a]*

[a] *Qatar Environment and Energy Research Institute (QEERI), Hamad Bin Khalifa University, Qatar Foundation, P. O. Box 34110, Doha, Qatar.*

[b] *Genomics Laboratory, Weill Cornell Medicine-Qatar (WCM-Q), Cornell University, Doha, Qatar*

[c] *Infectious Disease Epidemiology Group, Weill Cornell Medicine-Qatar, Cornell University, Doha, Qatar*

[d] *Hamad Medical Corporation, Doha, Qatar*

[e] *Drainage Network Operation & Maintenance Department; Public Works Authority, Doha, Qatar*

**Correspondence to: J Lawler* JLawler@hbku.edu.qa

**Correspondence to: KA Mahmoud* kmahmoud@hbku.edu.qa

[1] *Contributed equally to this work.*

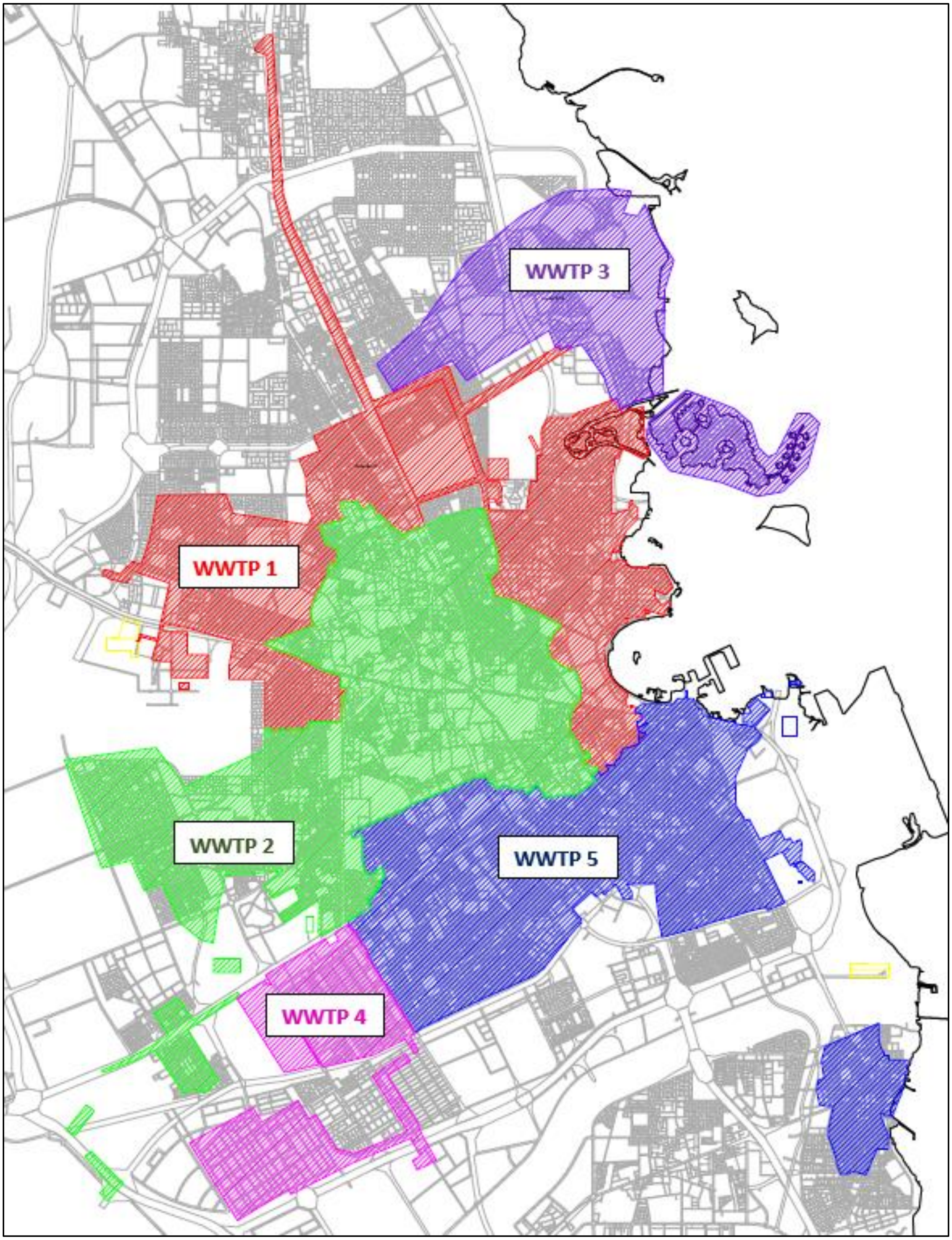

Figure S1: Locations of five major sewage treatment plants in Qatar

Table S1

Human RNAP internal control levels across different WWTPs

8/7/2020	Ct RNAP Assay	Ct N1 Assay	Ct N2 Assay	Avg (N1,N2)	Normalizing factor	N1/N2, RNAP Normalized
WWTP 1	45	30.2	30.1	30.15	1.410658307	21.373
WWTP 2	35.6	29.9	30.5	30.2	1.115987461	27.06123596
WWTP 3	35.5	29.4	29.8	29.6	1.112852665	26.59830986
WWTP 4	33.4	27.5	27.9	27.7	1.047021944	26.45598802
WWTP 5	34.2	28.2	28.5	28.35	1.072100313	26.44342105

12/7/2020	Ct RNAP Assay	Ct N1 Assay	Ct N2 Assay	Avg (N1,N2)	Normalizing factor	N1/N2, RNAP Normalized
WWTP 1	35.7	29.9	31.1	30.5	1.119122257	27.2535014
WWTP 2	37	29.8	30	29.9	1.159874608	25.77864865
WWTP 3	35.2	29.6	30.1	29.85	1.103448276	27.0515625
WWTP 4	31.9	29.6	28.5	29.05	1	29.05
WWTP 5	34	28.6	29.7	29.15	1.065830721	27.34955882

19/7/2020	Ct RNAP Assay	Ct N1 Assay	Ct N2 Assay	Avg (N1,N2)	Normalizing factor	N1/N2, RNAP Normalized
WWTP 1	36.113243	31.6	32.1	31.85	1.132076583	28.25984784
WWTP 2	34.205288	30.8	31.1	30.95	1.072266082	28.64689538
WWTP 3	33.64798	30.2	30.7	30.45	1.054795611	28.76023799
WWTP 4	32.25049	29.6	30.5	30.05	1.010987147	28.43058114
WWTP 5	34.05516	29.9	30.6	30.25	1.067559875	28.93299232

26/7/2020	Ct RNAP Assay	Ct N1 Assay	Ct N2 Assay	Avg (N1,N2)	Normalizing factor	N1/N2, RNAP Normalized
WWTP 1	45	32.4	33.3	32.85	1.410658307	23.287
WWTP 2	35.537243	30.7	31.9	31.3	1.114020157	28.09643956
WWTP 3	35.80509	33.2	33.2	33.2	1.122416614	29.57903471
WWTP 4	34.03934	31.5	31.7	31.6	1.06706395	29.61397019
WWTP 5	32.37331	30.6	30.9	30.75	1.014837304	30.3004234

Table S2

Estimated Infected Population

Date	WWTP	Mean Flow, x 10^6 L/day	Mean C_{NH4}, mg/L	Estimated Total Population, P	N1		N2	
					C_{RNA} copy/L	Infected Population	C_{RNA} copy/L	Infected Population
21/6/2020	WWTP 1	145.0	12.5	302131	42268	10855 ± 2253	36556	9388 ± 1977
	WWTP 2	283.4	20.1	949293	542056	271997 ± 51050	245184	123030 ± 23220
	WWTP 3	41.7	33.9	235763	161877	11961 ± 2461	293075	21656 ± 4277
8/7/2020	WWTP 1	133.9	9.9	221016	53736	12746 ± 2193	85397	20256 ± 5329
	WWTP 2	256.6	19.6	838067	74835	33997 ± 2917	79202	35981 ± 6955
	WWTP 3	43.9	32	234267	91947	7152 ± 1339	120100	9342 ± 1968
	WWTP 4	62.3	21.5	223392	452188	49919 ± 14405	485050	53547 ± 18016
	WWTP 5	241.5	21.2	853300	242727	103800 ± 11533	238948	102184 ± 14160
12/7/2020	WWTP 1	132.1	14.7	323650	60333	14113 ± 4181	58933	13786 ± 1744
	WWTP 2	286.3	19.1	911423	74835	37941 ± 3195	100652	51030 ± 5424
	WWTP 3	44.0	33.4	244894	82904	6458 ± 1026	94840	7388 ± 1085
	WWTP 4	61.8	32.6	335742	125170	13696 ± 5827	206282	22572 ± 8539
	WWTP 5	232.1	24.0	928400	140660	57811 ± 19221	123779	50873 ± 3945
19/7/2020	WWTP 1	136.7	11.6	264233	22620	5474 ± 999	24052	5821 ± 1193
	WWTP 2	282.6	19.2	904397	25793	12908 ± 2841	35810	17922 ± 5968
	WWTP 3	44.1	30.6	225058	37208	2908 ± 866	43569	3405 ± 986
	WWTP 4	59.3	28.1	277591	103239	10836 ± 1817	108080	11344 ± 2300
	WWTP 5	199.8	19.8	659340	85943	30407 ± 5609	82468	29177 ± 3870
26/7/2020	WWTP 1	131.5	12.3	269649	13499	3144 ± 790	14980	3489 ± 857
	WWTP 2	287.0	19.0	908928	42268	21483 ± 4244	34445	17507 ± 3500
	WWTP 3	44.3	29.5	217862	7889	619 ± 261	15898	1247 ± 406
	WWTP 4	63.5	37.7	399287	24703	2780 ± 719	38796	4366 ± 1026
	WWTP 5	217.0	16.7	603983	45203	17370 ± 3474	62434	23991 ± 4713
12/8/2020	WWTP 1	140.7	14.3	335402	35737	8906 ± 1886	47774	11905 ± 2450
	WWTP 2	276.4	20.4	939638	42268	20685 ± 4095	57106	27946 ± 5453
	WWTP 3	44.8	25.6	190942	24703	1958 ± 554	30582	2423 ± 648
	WWTP 4	62.5	29.2	303987	48342	5347 ± 1213	52232	5777 ± 1295
	WWTP 5	217.3	19.9	720712	31246	12023 ± 2472	39967	15379 ± 3101

16/8/2020	WWTP 1	142.1	13.3	314926	28253	7108 ± 1547	33436	8412 ± 1793
	WWTP 2	287.8	20.3	973710	46747	23823 ± 4682	57106	29103 ± 5669
	WWTP 3	47.0	24.2	189668	49993	4163 ± 987	60605	5047 ± 1156
	WWTP 4	60.9	29	294466	35737	3855 ± 928	52232	5635 ± 1268
	WWTP 5	223.0	19.7	732183	97835	38633 ± 7451	103507	40873 ± 7870
23/8/2020	WWTP 1	155.5	15.0	388763	20196	5561 ± 1254	27153	7477 ± 1617
	WWTP 2	295.9	19.7	971594	24703	12944 ± 2645	46374	24300 ± 4771
	WWTP 3	47.0	27.2	213180	30215	2516 ± 667	55432	4616 ± 1074
	WWTP 4	58.0	31	299832	34557	3551 ± 869	34445	3540 ± 867
	WWTP 5	230.2	19.4	744313	59130	24103 ± 4734	84056	34264 ± 6634
30/8/2020	WWTP 1	136.5	16.0	364115	14199	3433 ± 846	34779	8409 ± 1793
	WWTP 2	286.2	19.8	944605	13313	6748 ± 1479	23884	12106 ± 2488
	WWTP 3	46.7	23.4	181997	55563	4591 ± 1069	66024	5456 ± 1234
	WWTP 4	62.7	32.4	338461	34581	3838 ± 925	42096	4672 ± 1085
	WWTP 5	210.3	20.1	704505	27573	10268 ± 2143	48155	17933 ± 3580

Table S3

Estimation of total infected population

Date	Corrected 22 days of Cumulative Daily Positive Cases*	Total Estimated** Infected Population (N1)
21/6/2020	308,190	542,313 ± 51,159
8/7/2020	191,990	239,646 ± 18,858
12/7/2020	167,450	129,622 ± 20,788
19/7/2020	147,600	73,401 ± 6,677
26/7/2020	127,080	51,752 ± 5,594
12/8/2020	129,810	53,731 ± 5,312
16/8/2020	131,090	84,730 ± 9,037
23/8/2020	117,690	50,871 ± 5,673
30/8/2020	114,470	31,181± 3,081

* Corrected based on 10% diagnosis ratio

** Mathematical model calculation based on N1 assay

ST 1: Modeling Approach

As mentioned in the main text the central limit theorem formula explained below gave reliable results for N and δN for most of our data sets. In other words, the Monte-Carlo-Bayesian calculation explained in the following section were not required for most of the data sets. However, because the central limit theorem formula is derived by following the steps of the Bayesian probability theory calculation, we present this calculation first, and we then present the derivation of the central limit theorem formula in the next section.

ST 1.1 Monte Carlo-Bayesian Approach

The calculation to infer the number of infected individuals from the measured RNA concentration is performed as follows:

In the main text, we discussed how Equation (1) could be used to give a good estimate for the infected population (N) from the measured RNA concentration C_{RNA}. However, there is a conceptual complication with this approach. Since there are person-to-person variations in the parameters α and β, each individual person among the N people that form the infected population has his/her individual values of α and β. As a result, it is conceptually more natural and practically easier to calculate C_{RNA} for a given value of N rather than calculate N for a given value of C_{RNA}. When calculating C_{RNA} from N, the person-to-person variations can be introduced straightforwardly and with rigorous mathematical justification. We can use random-variable generation tools to generate a large set of α and β values and then, assuming that N is known, choose N individual values of α and β to calculate C_{RNA}. By repeating this random-variable based calculation many times, we can obtain the probability distribution for C_{RNA} values. In reality, C_{RNA} is measured, and the task is to infer N, or more accurately the probability distribution for possible values of N that could produce the measured value of C_{RNA}. The Bayesian approach reconciles these two opposing situations, what can be calculated easily and what needs to be calculated in reality. In this approach we first calculate C_{RNA} probability distributions for a broad range of N values and then use Bayesian analysis to extract a probability distribution for N given a certain value of C_{RNA}, which will be set to the actually measured value. As such, the Bayesian approach can be thought of as a form of reverse engineering.

To start the calculation, we first determine the range of N values that we need to consider. For this purpose, we use an approximation based on the central limit theorem, which states that for very large values of N, variations in the different variables can be ignored and C_{RNA} will be determined by the mean values of α, β, γ and F. We can therefore obtain the initial estimate for N using the formula:

$$N_{estimate} = \frac{C_{RNA}\bar{F}}{\bar{\alpha}\bar{\beta}(1-\bar{\gamma})}$$

Where the lines above the symbols indicate that we take the mean value of the variable. As will be explained below, this estimate might or might not be accurate depending on the widths of the different probability distributions involved. However, this estimate should generally give us the overall scale of N, and hence it helps us determine the overall scale of values that we need to consider in the Monte-Carlo calculations. For example, we can set the range of N values to be from zero to $2N_{estimate}$. We could then use the values $0.1\times N_{estimate}$, $0.2\times N_{estimate}$, $0.3\times N_{estimate}$, …, $2\times N_{estimate}$ as trial values of N in the calculation described below.

Setting N to a certain value means that we are assuming a known number of infected individuals, which is an assumption that we make in this intermediate step to be used later for inferring N from C_{RNA}. For each value of N, considering that we have a population with N infected individuals, C_{RNA} is given by

$$C_{RNA} = \frac{\sum_{i=1}^{N} \alpha_i \beta_i (1-\gamma_i)}{F}$$

The index i is a counter for the number of infected individuals. Each individual has his/her own values of α, β and γ. Therefore, to generate a Monte-Carlo data point, N different values for these variables are generated randomly. The whole population produces a single value F, which also contains some randomness. Therefore, one value for F is generated randomly for one Monte-Carlo data point. If we repeat the calculation of C_{RNA} many times (M times [the number of Monte-Carlo samples]), we will obtain M different values. These values give the probability distribution of C_{RNA}, e.g. by plotting a histogram from them, for a given value of N. For example, by dividing the range of C_{RNA} values from zero to infinity into intervals of width δ, the probability of getting a value of C_{RNA} in any of the intervals can be obtained by counting the number of Monte-Carlo data points in that range and divide it by the total number of Monte-Carlo data points. We emphasize again that the calculation of many different values of C_{RNA} is needed in this first step and that in the next step of the calculation we will use only the actual, measured value of C_{RNA}.

The calculation described so far results in a set of probability distributions: for each value of N, we obtain a distribution for C_{RNA} values. Once we have all the probability distributions, we can use Bayes' rule to obtain a probability distribution for N given a certain measured value of C_{RNA}. Ignoring uncertainty in C_{RNA}, we could choose an interval range δ as described above and say that the probability $P(C_{RNA}|N)$ of obtaining the value C_{RNA} (up to the uncertainty δ) given a value N is calculated as explained above: the number of Monte-Carlo data points in the interval divided by the total number of Monte-Carlo data points for that value of N. Bayes' rule can now be applied to find the probability $P(N|C_{RNA})$ that the number of infected individuals is N given that the measured RNA concentration is C_{RNA}:

$$P(N|C_{RNA}) = \frac{P(C_{RNA}|N)P(N)}{P(C_{RNA})}$$

If we assume that we do not have any additional information apart from C_{RNA} to favor any value of N, then P(N) on the right-hand side is a constant. Since we know the measured value of C_{RNA}, then there is no uncertainty in its value and the denominator is 1. Bayes' rule then reduces to

$$P(N|C_{RNA}) = constant \times P(C_{RNA}|N)$$

The right-hand side, including the constant, can be determined as follows. For each value of N, we want to determine how many of the Monte-Carlo samples have a C_{RNA} value that matches the experimentally measured value. For this purpose, we choose a value of acceptable deviation and count how many Monte-Carlo samples are within this distance of the experimentally measured value. This way we obtain a number of counts for each value of N. This number of counts is proportional to $P(C_{RNA}|N)$, and therefore $P(N|C_{RNA})$ will also be proportional to this number of counts. To obtain a normalized probability distribution, we add up all the counts (for all values of N) and divide all the numbers of counted samples by this total number. Thus for every value of N we obtain a probability. This is the probability that the number of infected individuals is the corresponding value of N. (If the distance between N values is ΔN, the probability of the value N is actually the probability that N is between N-ΔN/2 and N+ΔN/2.)

If there is an uncertainty δC_{RNA} in the measurement of C_{RNA}, one can use a somewhat different formula for extracting the probability $P(N|C_{RNA})$ from the Monte-Carlo C_{RNA} values. One can now take a sum over all the obtained values of C_{RNA}:

$$P(N|C_{RNA}) = constant \times \sum_{i=1}^{M} e^{-\frac{\left(C_{RNA,i}-C_{RNA,measured}\right)^2}{2\delta C_{RNA}^2}}$$

The values of C_{RNA} in the Monte-Carlo ensemble that are very close to the measured value will each contribute almost 1 to the sum. Values that are within the measurement uncertainty will also make non-negligible contributions. Values in the ensemble that are very far from the measured values make very small contributions and barely affect the sum. Once the sum is obtained for each value of N, the constant is determined by the probability normalization condition, as explained above.

ST 1.2 Central Limit Theorem

Let us go back and see if we can use the central limit theorem to avoid the long Monte-Carlo calculation described above. Specifically, let us assume that the central limit theorem gives a good approximation not only for the mean values but also for the probability distributions for C_{RNA} for any given value of N. This approximation should become good for sufficiently large values of N. In fact, according to the central limit theorem, for sufficiently large N, all the variables related to this large population should follow normal distributions, which then simplifies all subsequent calculations. With this assumption, we can proceed by saying that the total number of RNA copies M$_{RNA}$ will have a mean value

$$\bar{M}_{RNA} = N\bar{\alpha}\bar{\beta}(1-\bar{\gamma})$$

And variance

$$\delta M_{RNA}^2 = N\left((\bar{\alpha}^2 + \delta\alpha^2)\left(\bar{\beta}^2 + \delta\beta^2\right)((1-\bar{\gamma})^2 + \delta\gamma^2) - \left(\bar{\alpha}\bar{\beta}(1-\bar{\gamma})\right)^2\right)$$

The RNA concentration C$_{RNA}$ will also follow a normal distribution that is approximately centered at

$$C_{RNA} = \frac{\bar{M}_{RNA}}{\bar{F}} = \frac{N\bar{\alpha}\bar{\beta}(1-\bar{\gamma})}{\bar{F}}$$

And having a variance that is approximately given by:

$$\delta C_{RNA}^2 = (\bar{M}_{RNA}^2 + \delta M_{RNA}^2)\left(\frac{1}{\bar{F}^2} + \frac{\delta F^2}{\bar{F}^4}\right) - \frac{\bar{M}_{RNA}^2}{\bar{F}^2}$$

The above formula for δC_{RNA}^2 can be thought of as the intrinsic variance $(\delta C_{RNA}^2)_{Intrinsic}$ in C_{RNA}. Another possible source for variations in C_{RNA} is the measurement uncertainty or measurement error: if the measurement of C_{RNA} has uncertainty characterized with the standard deviation $(\delta C_{RNA}^2)_{Measurement}$, the total variance of C_{RNA} will be given by:

$$(\delta C_{RNA}^2)_{Total} = (\delta C_{RNA}^2)_{Intrinsic} + (\delta C_{RNA}^2)_{Measurement}$$

Having obtained the mean values and variances assuming a fixed value of N, we can now use these values to estimate the standard deviation in N when Bayes' rule is applied to obtain N. When we perform this inversion procedure to estimate N, we obtain a normal distribution whose standard deviation is given by

$$\delta N = \frac{\delta C_{RNA}\bar{F}}{\bar{\alpha}\bar{\beta}(1-\bar{\gamma})}$$

This is the central limit theorem formula for the standard deviation in N. Although evaluating it requires going through the few steps of calculating different mean values and standard deviations, these calculations are all straightforward algebraic calculations and do not require any random-variable sampling as in the Monte-Carlo approach described in the previous section. As a result, the central limit formula takes essentially no computational time, whereas the Monte-Carlo approach becomes increasingly time consuming as the size of the population N increases.