A Quantitative Metagenomic Sequencing Approach for High Throughput Gene Quantification and Demonstration with Environmental Antibiotic Resistance Genes

Bo Li [1], Xu Li [2] and Tao Yan[1,] *

[1] Department of Civil and Environmental Engineering, University of Hawaii at Manoa, Honolulu, HI 96822, United States

[2] Department of Civil and Environmental Engineering, University of Nebraska-Lincoln, Lincoln, NE 68588, United States

Running Title: metagenomic sequencing for gene quantification

*Corresponding author: Tao Yan, University of Hawaii at Manoa, Department of Civil and Environmental Engineering, 2540 Dole Street, 383 Holmes Hall, Honolulu, HI 96822. Phone: 808-956-6024. Fax: 808-956-5014. E-mail: taoyan@hawaii.edu

Table S2. Information of DNA samples and sequencing data of DNA samples.

Sample	DNA Concentration (ng/μL)	Volume (μL)	Total DNA Quantity (ng)	reads	base	Length (bp)	Q20(%)	gc(%)	SRA accession numbers
E1	24	24	576	4.0E+07	6,051,450,300	150	97.88%	51.43%	SRX7970926
E2	26.5	24	636	4.1E+07	6,155,636,400	150	98.21%	51.23%	SRX7970927
E3	28.4	24	681.6	4.1E+07	6,196,367,100	150	98.29%	51.29%	SRX7970928
M1	39	23	897	4.1E+07	6,208,866,000	150	98.40%	48.88%	SRX7970929
M2	36.2	23	832.6	4.1E+07	6,215,574,000	150	98.10%	47.79%	SRX7970930
M3	36.8	23	846.4	4.1E+07	6,201,976,800	150	98.02%	47.04%	SRX7970931

Table S3. qPCR assays and thermal cycling conditions targeting 16S rRNA gene and ARGs

Gene marker (target size)	Class of antibiotics or Integron	Forward (F), reverse (R), and probe (P) sequences (5' - 3')*	Conc. (μM)	Cycling condition	Ref.
16S rRNA(142 bp)		F: CGGTGAATACGTTCYCGG	0.2	95.0 °C for 2 min and followed with 40 cycles of 95.0 °C for 15 s, 60.0 °C for 30 s. and 72.0 °C for 30 s.	(1)
		R: GGWTACCTTGTTACGACTT	0.2		
		P: FAM/CTTGTACAC/ZEN/ACCGCCCGTC/3IABkFQ	0.1		
ermB (91 bp)	macrolide	F: GGATTCTACAAGCGTACCTTGGA	0.2	95.0 °C for 2 min and followed with 40 cycles of 95.0 °C for 15 s, 60.0 °C (69.9 °C *sulI*) for 30 s. and 72.0 °C for 30 s	(2)
		R: GCTGGCAGCTTAAGCAATTGCT	0.2		
		P: FAM/CACTAGGGT/ZEN/TGCTCTTGCACACTCAAGTC/IABkFQ	0.1		
qnrS (118 bp)	quinolone	F: CGACGTGCTAACTTGCGTGA	0.2		(3)
		R: GGCATTGTTGGAAACTTGCA	0.2		
		P: FAM/AGTTCATTG/ZEN/AACAGGGTGA/IABkFQ	0.1		
tetO (171 bp)	tetracycline	F:ACGGARAGTTTATTGTATACC	0.2		(4)
		R: TGGCGTATCTATAATGTTGAC	0.2		
		P: FAM/CGTAGATGA/ZEN/AGGCACAACAAGGAC/IABkFQ	0.1		
tetM (88 bp)		F: GGTTTCTCTTGGATACTTAAATCAATCR	0.2		(5)
		R: CCAACCATAYAATCCTTGTTCRC	0.2		
		P: FAM/ATGCAGTTA/ZEN/TGGARGGGATACGCTATGGY/IABkFQ	0.1		
sulI (163 bp)	sulfonamide	F: CGCACCGGAAACATCGCTGCAC	0.2		(6)
		R: TGAAGTTCCGCCGCAAGGCTCG	0.2		
		P: FAM/TTCTTGGGC/ZEN/GCCACCGTTGGCCTT/IABkFQ	0.1		

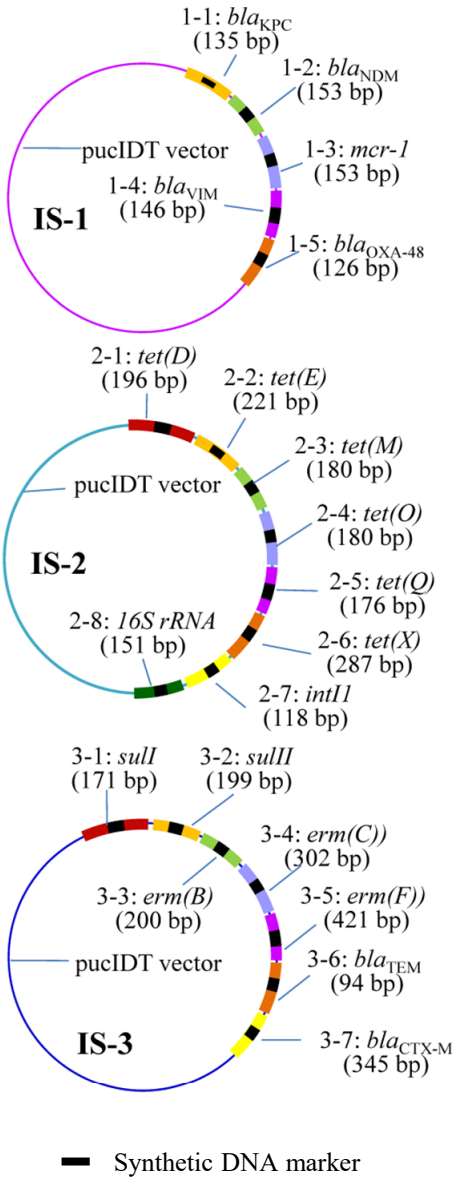


Figure S1. Three plasmids for the three synthetic DNA internal standards (IS-1, IS-2 and IS-3) which contain five, eight and seven different synthetic DNA internal standard fragments (ISFs), respectively. The pUCIDT containing bla_{TME} gene, and all blaTEM genes were excluded in sequencing results of samples to avoid false detection.

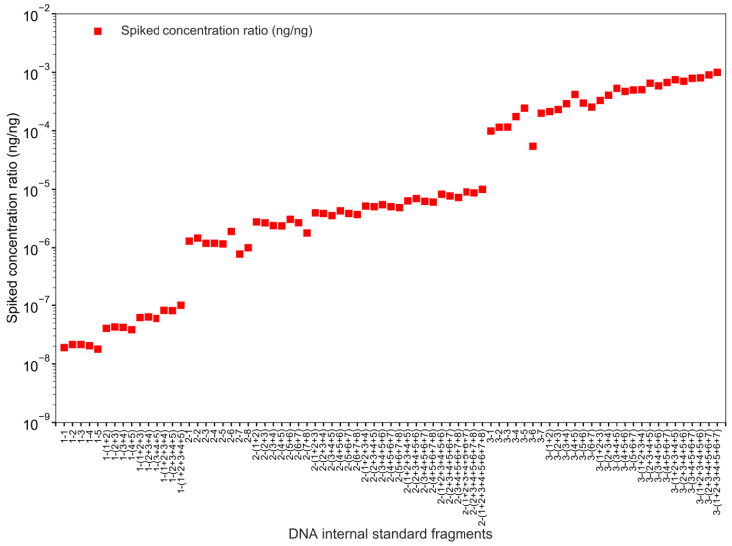


Figure S2. Spiked concentration ratio of DNA internal standard fragments(ISFs)

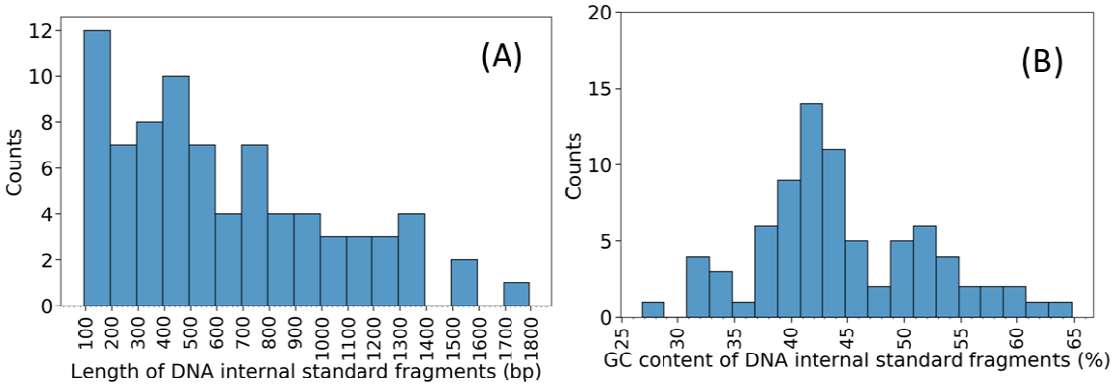


Figure S3. Length (A) and GC content (B) distribution of DNA internal standard fragments (ISFs)

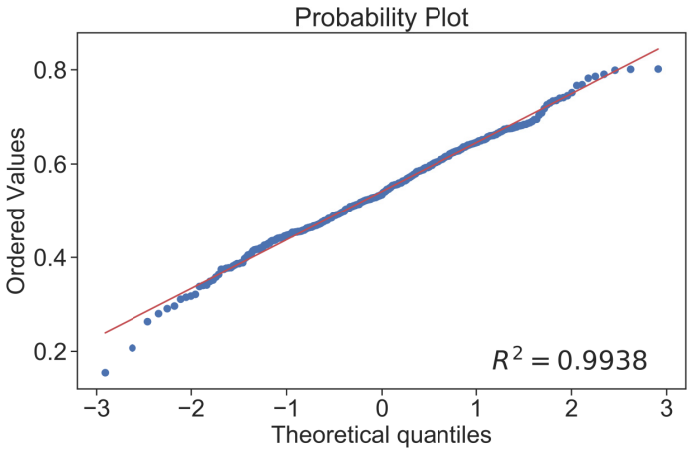


Figure S4. Normal probability Q–Q (quantile-quantile) plot for sequencing yields of ISFs over limit of quantification (LOQ)

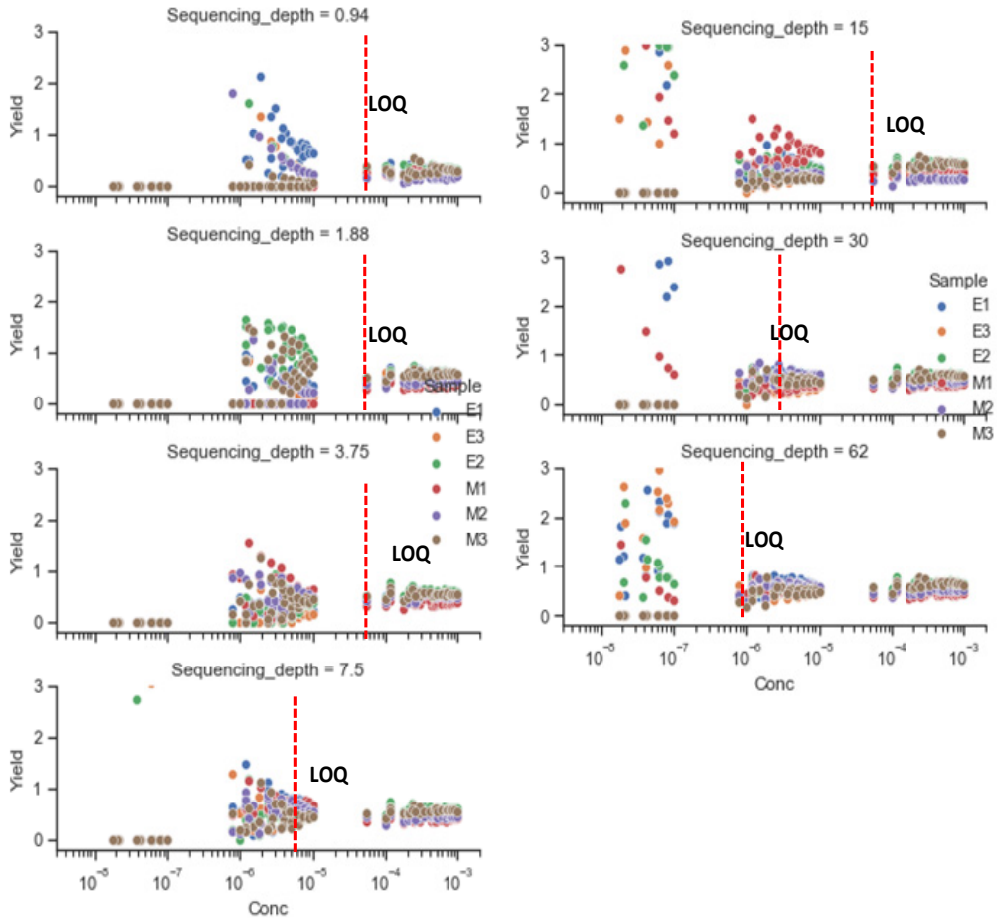


Figure S5. Sequencing yield at different spiked concentration and sequencing depth (10^8 bps)

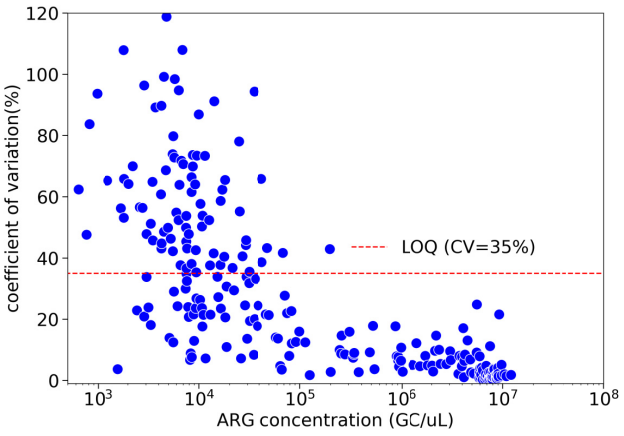


Figure S6. Coefficient of variation of individual ARGs quantification in triplicate analysis by qmNGS for the artificial *E. coli* isolate mixtures (E1-3) and cattle manure samples (M1-3) (n=246). Only CV of ARGs detected in all three replicates were summarized.

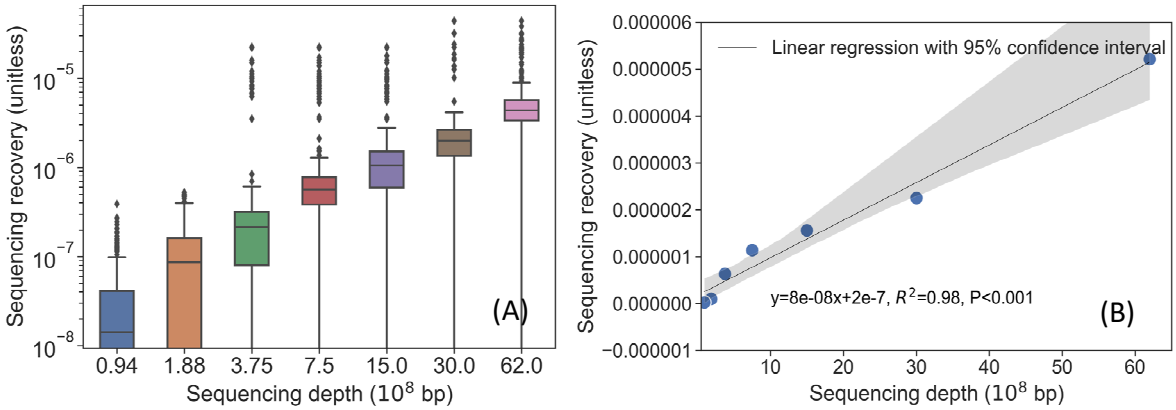


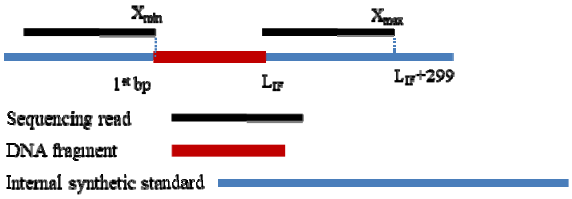
Figure S7. Recovery rate (unitless) at different sequencing depth (A); linear regression of recovery rate and sequencing depth (B)

Calculation steps of P_i

P_i is defined as: $P_i = n'_{ISF-i}/n_{ISF-i}$ (S1)

Symbols and the physical meanings in calculation of P_i:

L_{IF}--Length of an DNA fragment (IF) on internal DNA standard (IS), bp
$L_{IF\text{-}COV}$-- the part of the IF covered by a sequencing read, bp
$L_{IF\text{-}COV\text{-}M}$-- the part of the IF covered by a sequencing read containing gene maker, bp
S_M-- the start position of the gene marker on the IF
n_{IF}--theoretical sequence bases of an IF, bp
n'_{IF}--theoretical sequence bases of an IF containing gene marker, bp
X-- the end position of a sequencing reads on IF



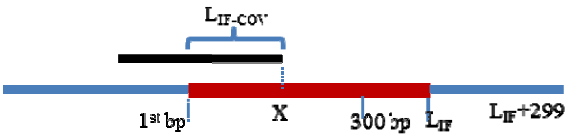
1. Calculation of n_{ISF}. In this study, the length of different IF is in the range of 94-421 bp and the sequencing read length is 300 bp. $\mathbf{n_{IF}}$ was calculated considering two length conditions of IF, including $L_{IF} \geq 300$ and $L_{IF} < 300$.

1.1 $L_{IF} \geq 300$

(1) When $1 \leq X \leq 299$ or $L_{IF}+1 \leq X \leq L_{IF}+299$

Only part base pairs of the sequencing reads are from the IF.

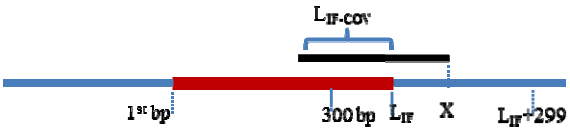
a. When $1 \leq X \leq 299$



$L_{IF-COV} = X$ bp

$\sum_1^{299} L_{IF-COV} = \frac{(1+299)\times 299}{2} = 44850$ bp

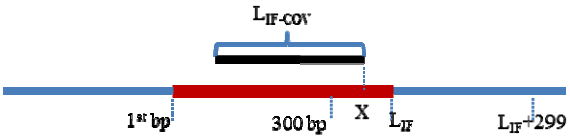
b. When $L_{IF}+1 \leq X \leq L_{IF}+299$



$L_{IF-COV} = 300 - (X - L_{IF})$ bp

$\sum_{L_{IF}+1}^{L_{IF}+299} L_{IF-COV} = \frac{[(L_{IF}+299)+(L_{IF}+1)][(L_{IF}+299)-(L_{IF}+1)+1]}{2} = \frac{(1+299)\times 299}{2} = 44850$ bp

(2) When $300 \leq X \leq L_{IF}$



Sequence bases of the entire sequencing read are from the IF.

L_{IF-COV}= 300 bp

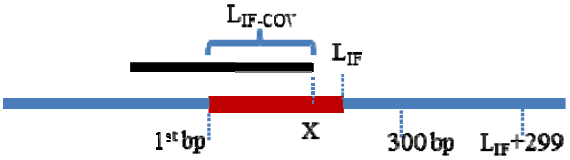
$\sum_{300}^{L_{IF}} L_{IF-COV} = 300 \times (L_{IF} - 299)$ bp

Hence,

$n_{ISF} = \sum_{1}^{L_{IF}+299} L_{IF-COV} = \sum_{1}^{299} L_{IF-COV} + \sum_{300}^{L_{IF}} L_{IF-COV} + \sum_{L_{IF}+1}^{L_{IF}+299} L_{IF-COV} =$ 300 L_{IF} bp

1.2 L_{IF}<300

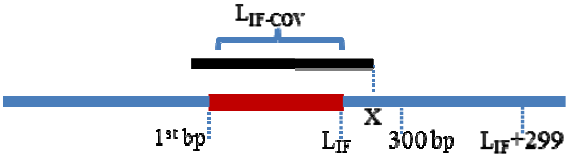
(1) When 1≤X≤L_{IF}-1



L_{IF-COV}= X

$\sum_{1}^{L_{IF}-1} L_{IF-COV} = \frac{L_{IF} \times (L_{IF}-1)}{2}$

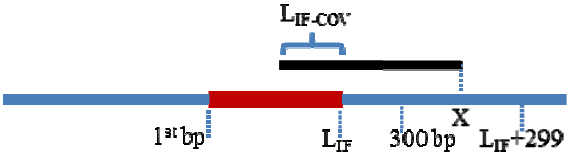
(2) When L_{IF}≤X≤300



L_{IF-COV}= L_{IF}

$\sum_{L_{IF}}^{300} L_{IF-COV} = L_{IF} \times (301 - L_{IF})$

(3) When 301≤X≤ L_{IF}+299



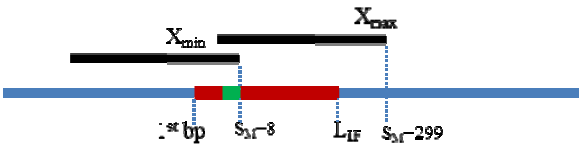
L_{IF-COV}= L_{IF}-(X-300)

$n_{ISF} = \sum_{301}^{L_{IF}+299} L_{IF-COV} = \frac{L_{IF} \times (L_{IF}-1)}{2}$

Hence, in total

$n_{ISF} = \sum_{1}^{L_{IF}+299} L_{IF-COV} = \sum_{1}^{299} L_{IF-COV} + \sum_{300}^{L_{IF}} L_{IF-COV} + \sum_{L_{IF}+1}^{L_{IF}+299} L_{IF-COV} = 300\ L_{IF}$ bp

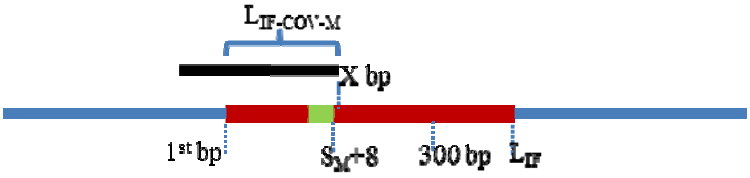
2. Calculation of n'_{ISF}. To cover the entire gene marker, the end position X of sequencing reads should be in the range of [S_M+8, S_M+299] (where S_M is the start position of a gene marker in the IF).



Gene marker

2.1 L_{IF}≥300

(1) When $S_M+8 \le X \le 299$

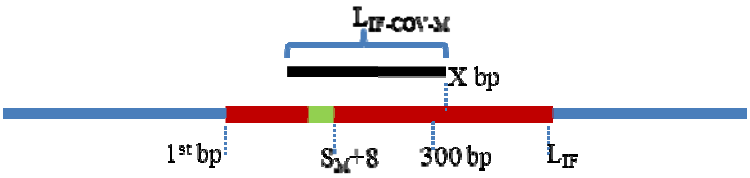


$L_{IF\text{-}COV\text{-}M}=X$

$$\sum_{S_M+8}^{299} L_{IF-COV-M} = \frac{(S_M+8+299)[299-(S_M-8)+1]}{2} = \frac{(S_M+307)(292-S_M)}{2}$$

(2) If $S_M+299 \le L_{IF}$

When $300 \le X \le S_M+299$

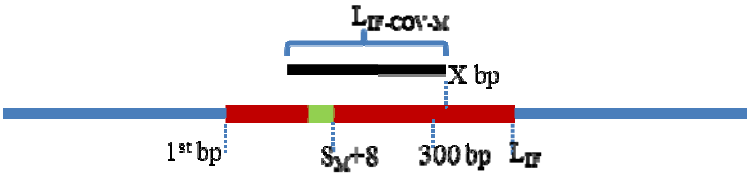


$L_{IF\text{-}COV\text{-}M}$=300 bp

$$\sum_{300}^{S_M+299} L_{IF-COV-M} = 300 \bullet S_M \text{ bp}$$

(3) If $S_M+299 > L_{IF}$

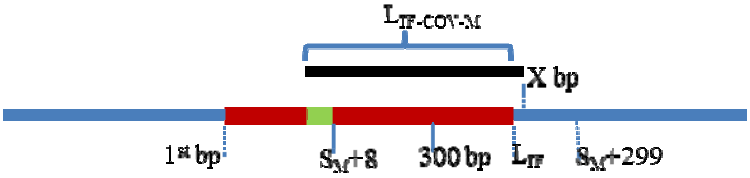
a). When $300 \le X \le L_{IF}$



$L_{IF\text{-}COV\text{-}M}$=300 bp

$$\sum_{300}^{L_{IF}} L_{IF-COV-M} = 300 \bullet (L_{IF} - 299)$$

b). When $L_{IF}+1 \le X \le S_M+299$

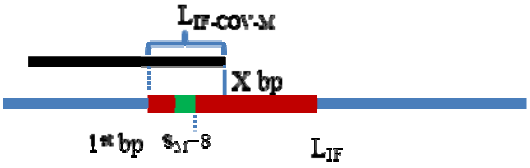


$L_{IF\text{-}COV\text{-}M}$= L_{IF}-(X-300) bp

$$\sum_{L_{IF}+1}^{S_M+299} L_{IF-COV-M} = \frac{(L_{IF}-S_M+300)(S_M-L_{IF}+299)}{2} \text{ bp}$$

2.2 L_{IF}<300

(1) When $1 \le X \le L_{IF}-1$

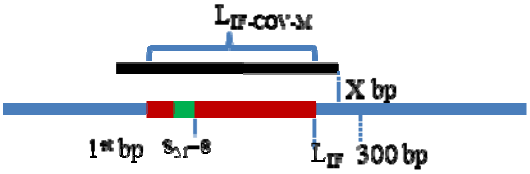


$L_{IF-COV-M}$ =X

$$\sum_{S_M+8}^{L_{IF}-1} L_{IF-COV-M} = \frac{(S_M+8+L_{IF}-1)[(L_{IF}-1)-(S_M+8)+1]}{2}$$

$$= \frac{(S_M+L_{IF}+7)(L_{IF}-S_M-8)}{2}$$

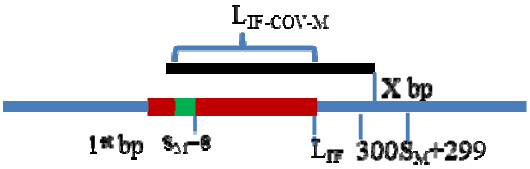
(2) When $L_{IF} \le X \le 300$



$L_{IF-COV-M}$= L_{IF}

$$\sum_{L_{IF}}^{300} L_{IF-COV-M} = L_{IF} \bullet (301 - L_{IF})$$

(3) When $301 \le X \le S_M+299$



$L_{IF-COV-M}$= L_{IF}-(X-300)

$$\sum_{301}^{S_M+299} L_{IF-COV-M} = \frac{[(L_{IF}-1)+(L_{IF}-S_M+1)](S_M-1)}{2} = \frac{(2L_{IF}-S_M)(S_M-1)}{2}$$

$$n'_{ISF} = \sum_{S_M+8}^{S_M+299} L_{IF-COV-M}$$

1. Suzuki MT, Taylor LT, DeLong EF. 2000. Quantitative analysis of small-subunit rRNA genes in mixed microbial populations via 5′-nuclease assays. Applied and environmental microbiology 66:4605-4614.

2. Böckelmann U, Dörries H-H, Ayuso-Gabella MN, de Marçay MS, Tandoi V, Levantesi C, Masciopinto C, Van Houtte E, Szewzyk U, Wintgens T. 2009. Quantitative PCR monitoring of antibiotic resistance genes and bacterial pathogens in three European artificial groundwater recharge systems. Applied and environmental microbiology 75:154-163.
3. Colomer-Lluch M, Jofre J, Muniesa M. 2014. Quinolone resistance genes (qnrA and qnrS) in bacteriophage particles from wastewater samples and the effect of inducing agents on packaged antibiotic resistance genes. Journal of Antimicrobial Chemotherapy 69:1265-1274.
4. Aminov R, Garrigues-Jeanjean N, Mackie RI. 2001. Molecular ecology of tetracycline resistance: development and validation of primers for detection of tetracycline resistance genes encoding ribosomal protection proteins. Appl Environ Microbiol 67:22-32.
5. Peak N, Knapp CW, Yang RK, Hanfelt MM, Smith MS, Aga DS, Graham DW. 2007. Abundance of six tetracycline resistance genes in wastewater lagoons at cattle feedlots with different antibiotic use strategies. Environmental Microbiology 9:143-151.
6. Pei R, Kim S-C, Carlson KH, Pruden A. 2006. Effect of river landscape on the sediment concentrations of antibiotics and corresponding antibiotic resistance genes (ARG). Water research 40:2427-2435.