Supplementary Information: Tables and Figures

NAD: Noise-augmented direct sequencing of target nucleic acids by augmenting with noise and selective sampling

Hyunjin Shim[1,,*]

Author Information

Affiliations

[1]Department of Biology, California State University, Fresno, 5241 N Maple Ave, Fresno, CA 93740, USA

*Corresponding author: Hyunjin Shim (shim@csufresno.edu)

Orcid links

Hyunjin Shim orcid=0000-0002-7052-0971

Figure S1. DNA QC of the microbial DNA standards using NanoDrop One Spectrophotometer.

lambda (aliquot for the Pseudomona samples); lambda2 (aliquot for the Salmonella samples)

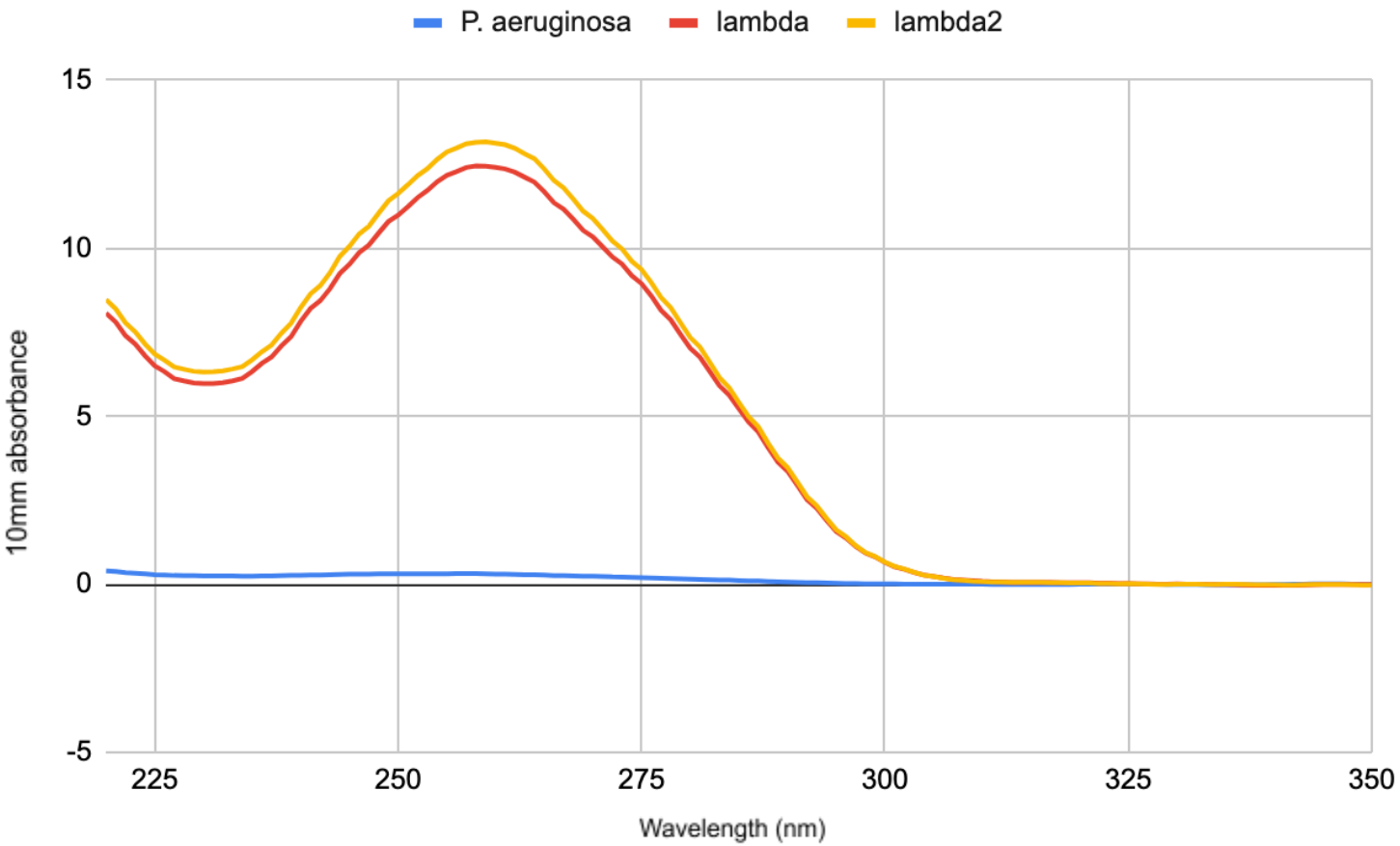
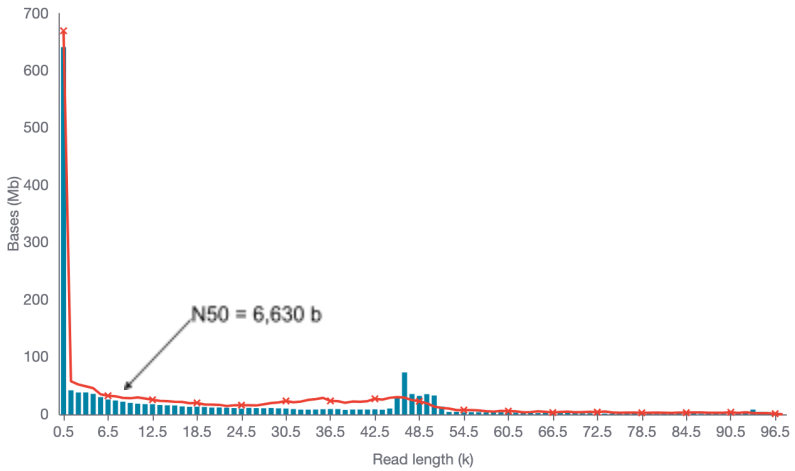
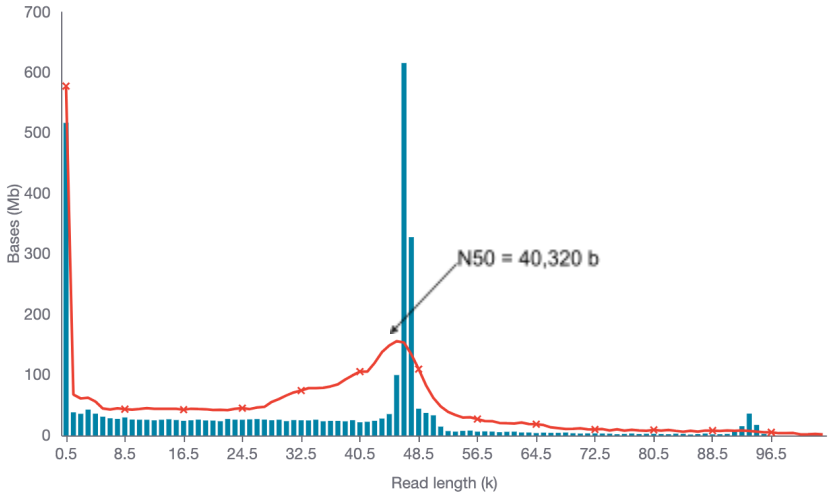


Figure S2. The read length graph from the nanopore report shows the total number of bases vs the read length in (a) P_aeruginosa_1; (b) P_aeruginosa_2; (c) Salmonella_1ng; (d) Salmonella_3ng; (e) Salmonella_5ng. The longest 1% of strands are classified as outliers and excluded in the graph to allow focus on the main body of data. The blue bars show the basecalled numbers, and the red lines show the estimated numbers. The N50 value of each experiment is displayed directly on the plot using text and arrows. The quality score graph from the nanopore shows the distribution of quality scores calculated for reads above the threshold Q-score of 10 in (f) P_aeruginosa; (g) Salmonella_1ng; (h) Salmonella_3ng; (i) Salmonella_5ng.

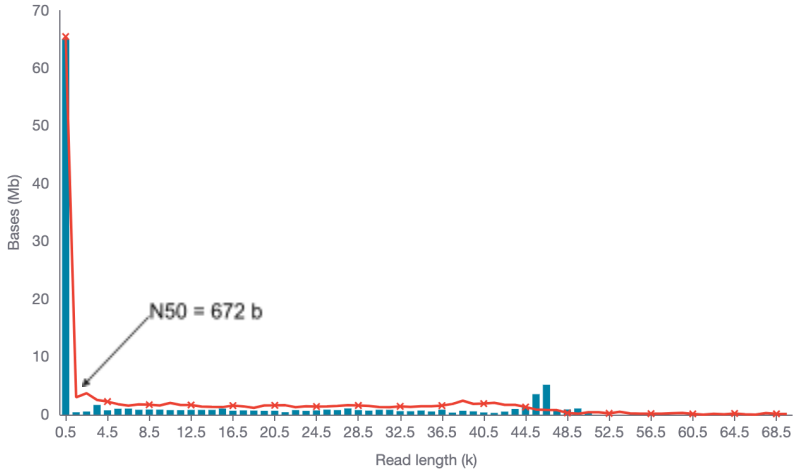
(a) P_aeruginosa_1



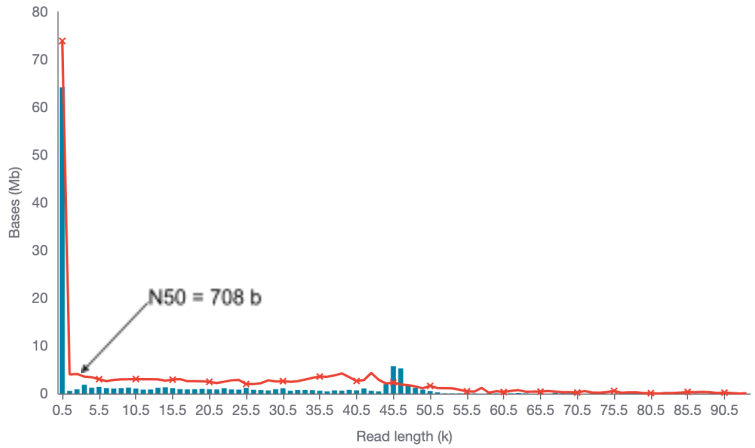
(b) P_aeruginosa_2



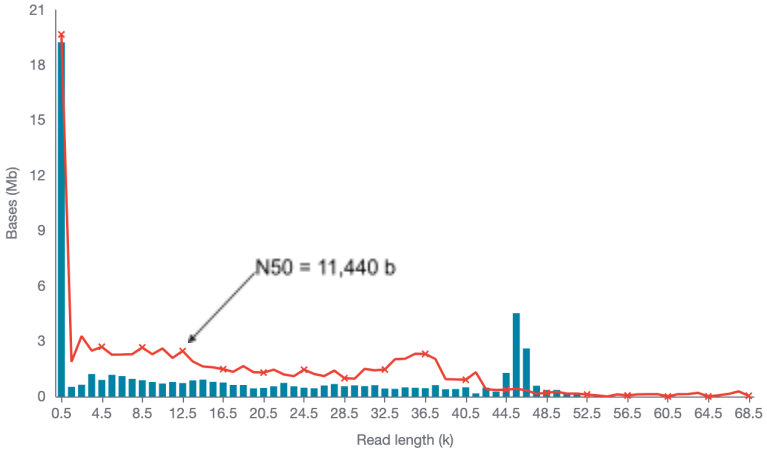
(c) Salmonella_1ng



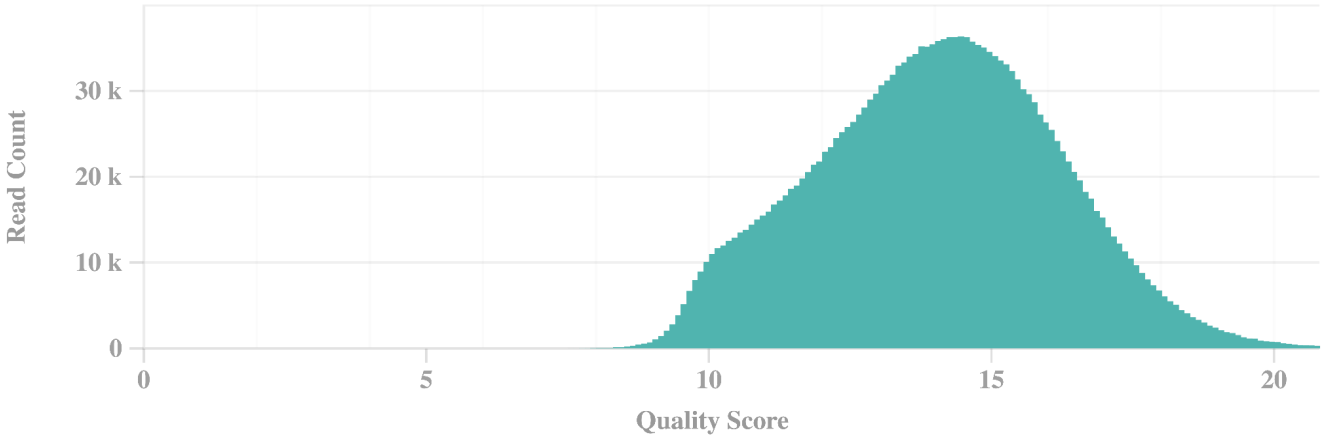
(d) Salmonella_3ng



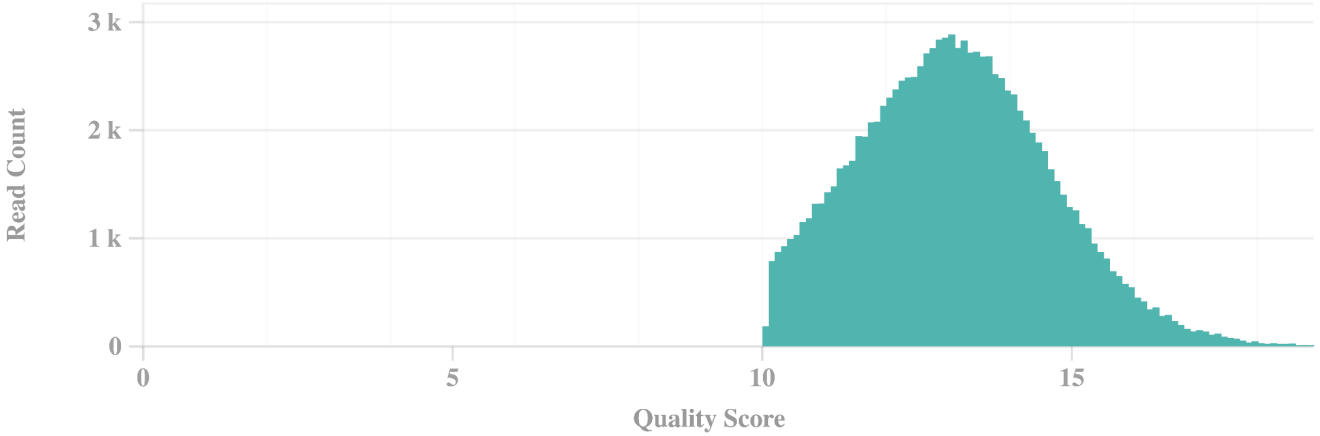
(e) Salmonella_5ng



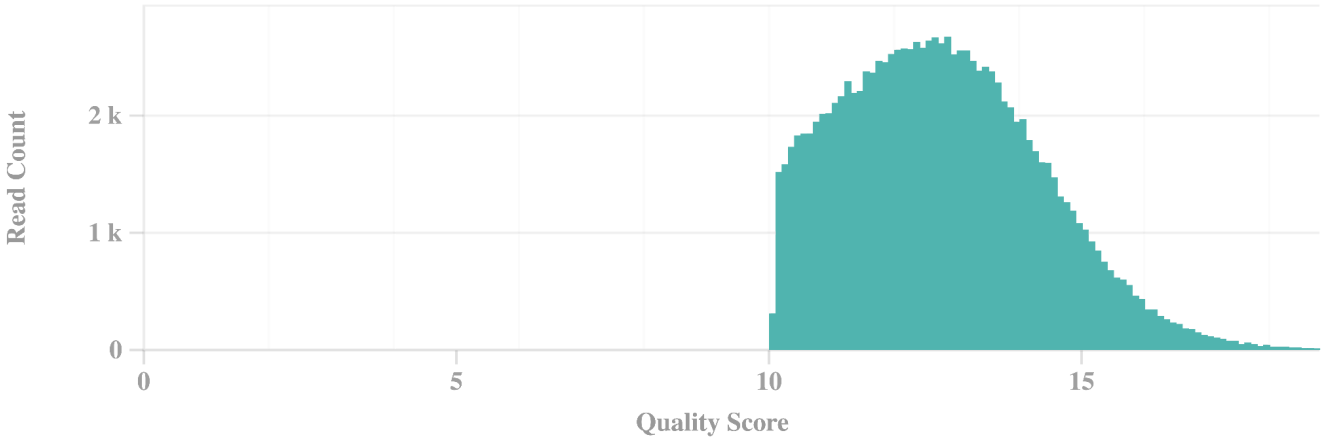
(f) P_aeruginosa



(g) Salmonella_1ng



(h) Salmonella_3ng



(i) Salmonella_5ng

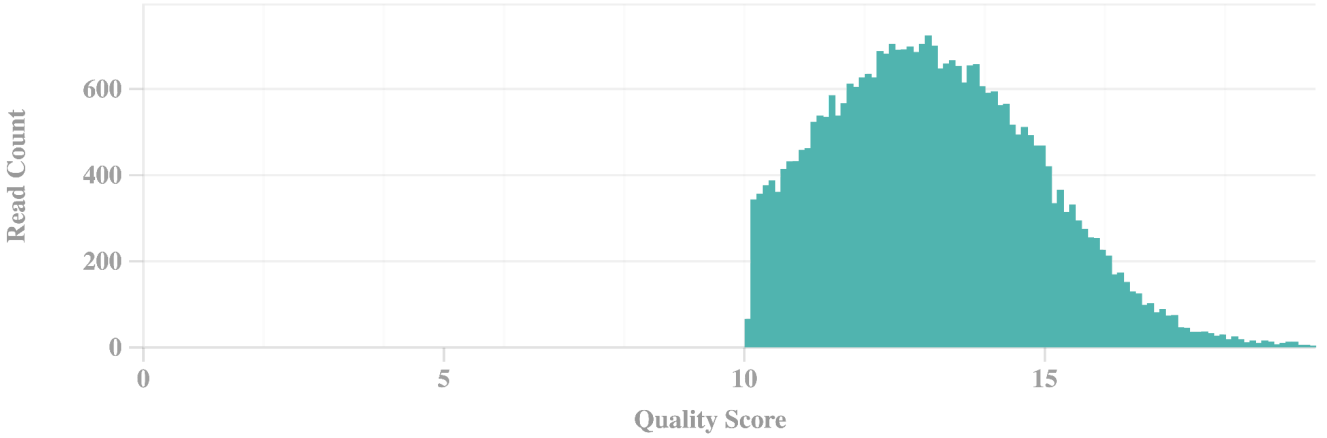
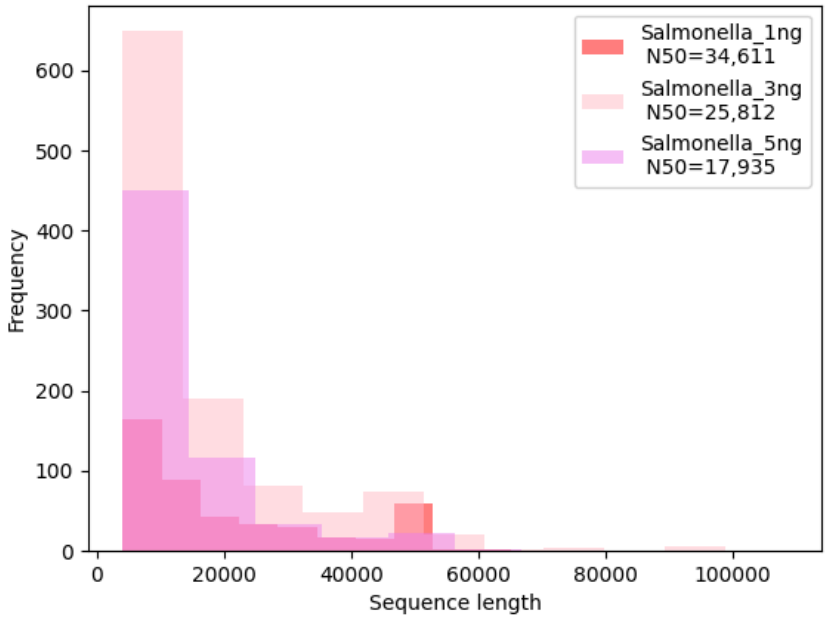


Figure S3. Comparison of the three decision types in adaptive sampling of the NAD samples. (A) Full length of WIMP classified reads accepted in the Pseudomonas sample. (B) Full length of WIMP classified reads accepted in the Salmonella samples.

(a) Salmonella samples



(b) Pseudomonas samples

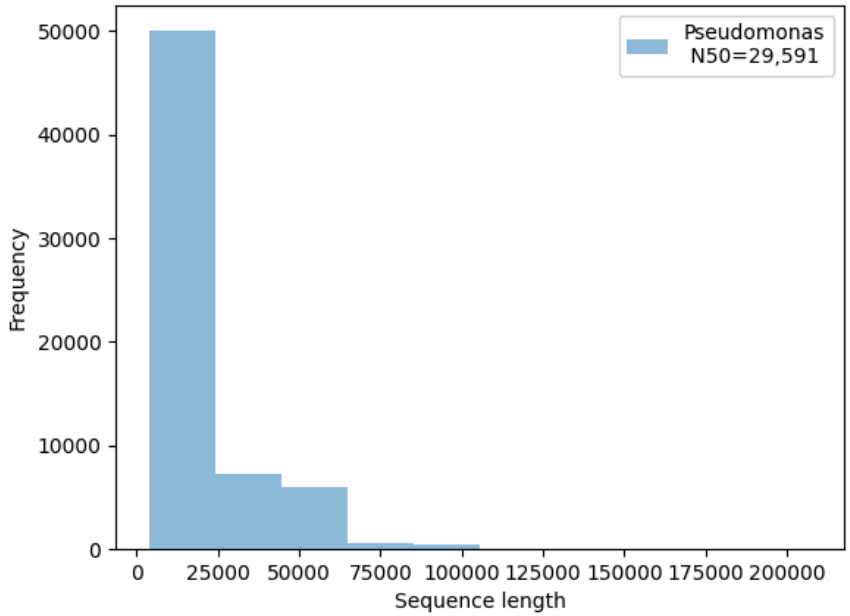
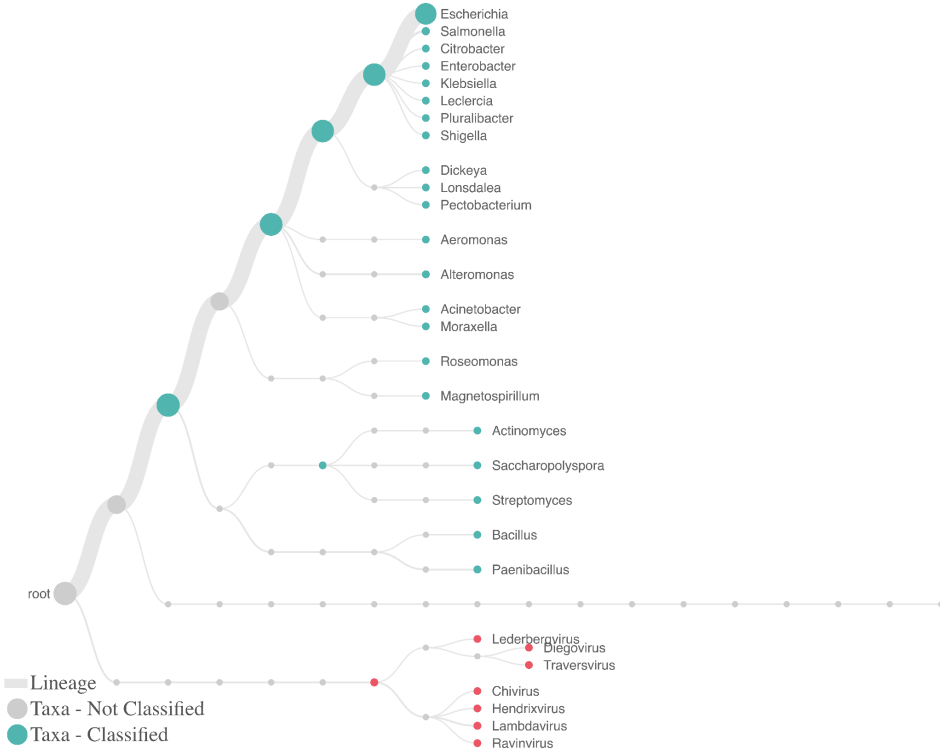


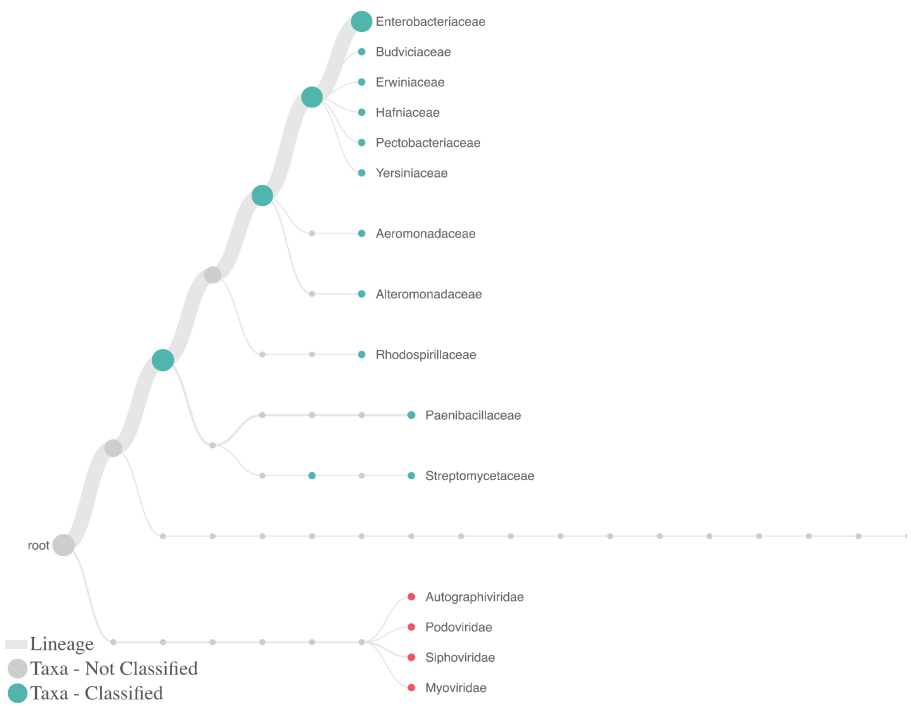
Figure S4. WIMP classification at the Family level.

(a) Salmonella_1ng

(b) Salmonella_3ng



(c) Salmonella_5ng



(d) Pseudomonas

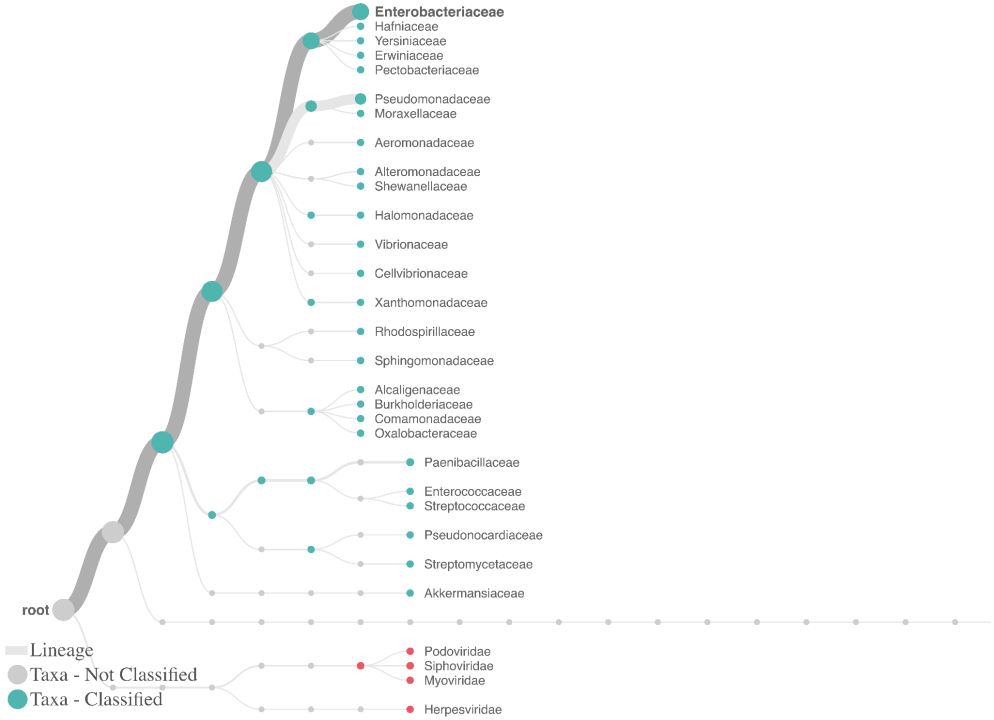
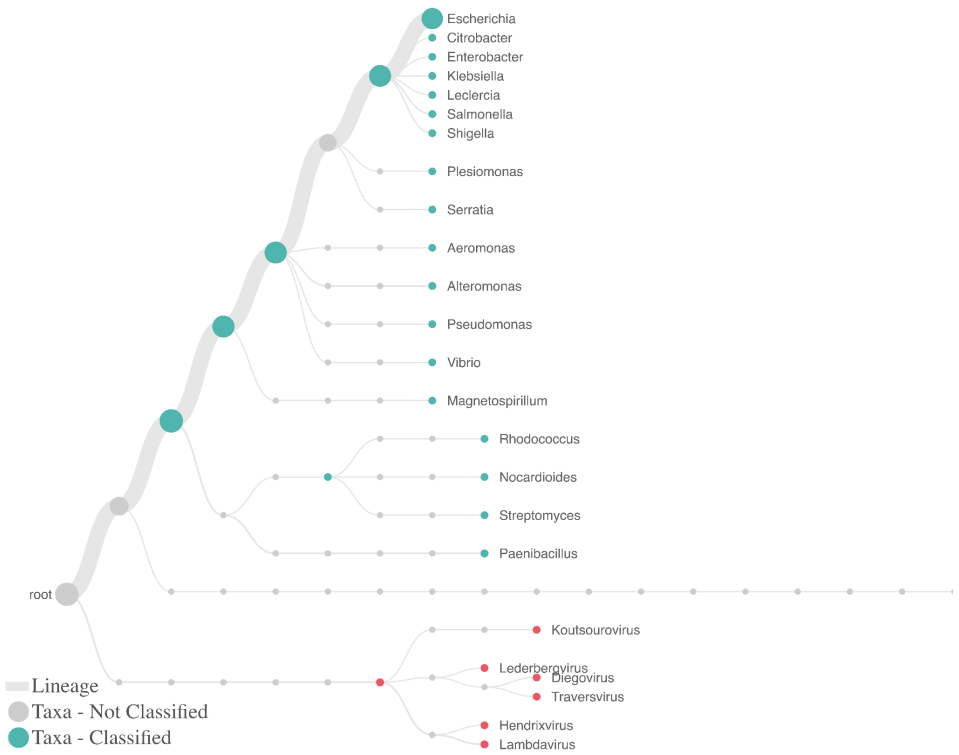
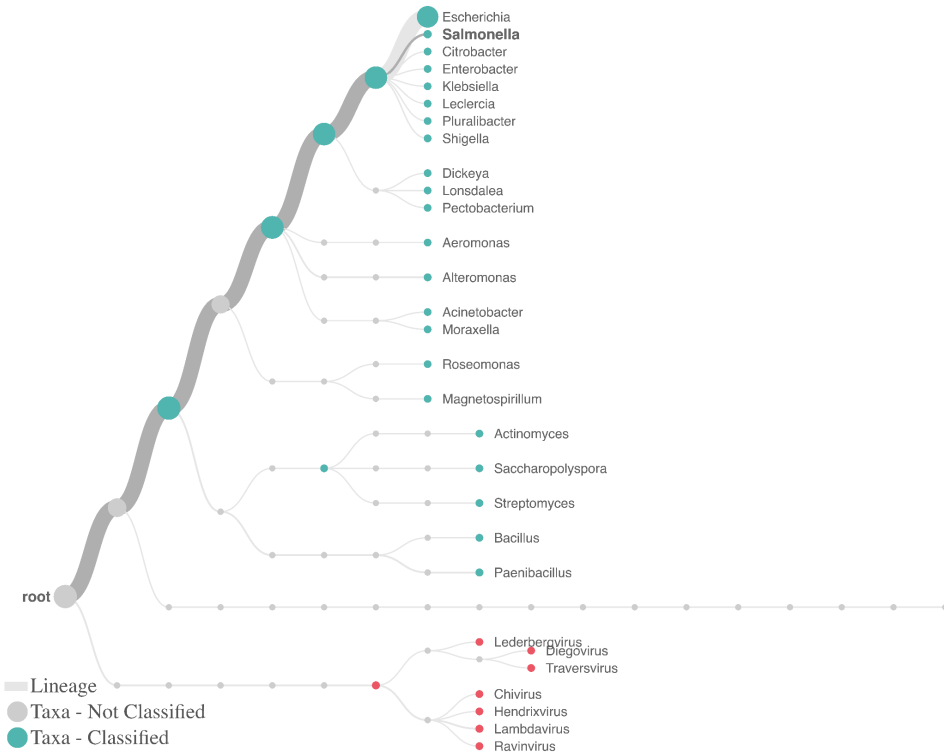


Figure S5. WIMP classification at the Genus level.

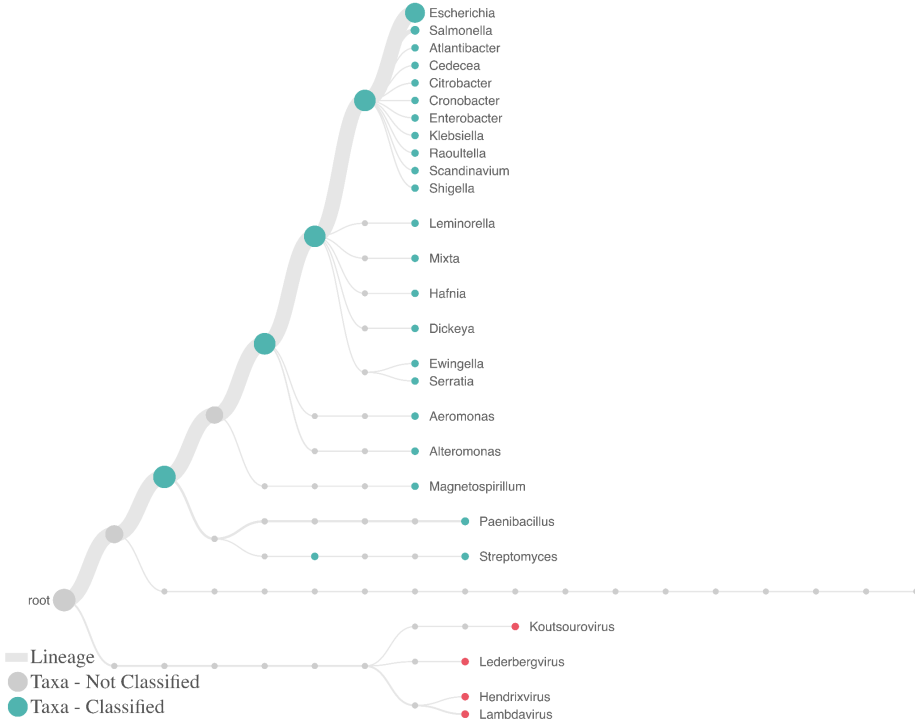
(a) Salmonella_1ng



(b) Salmonella_3ng



(c) Salmonella_5ng



(d) Pseudomonas

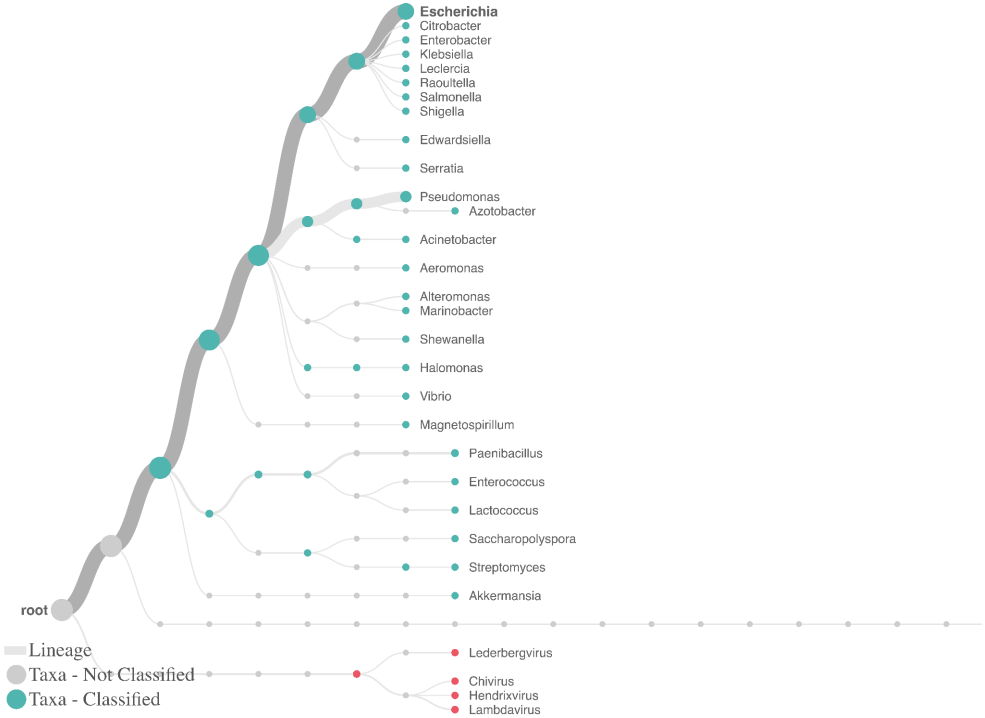
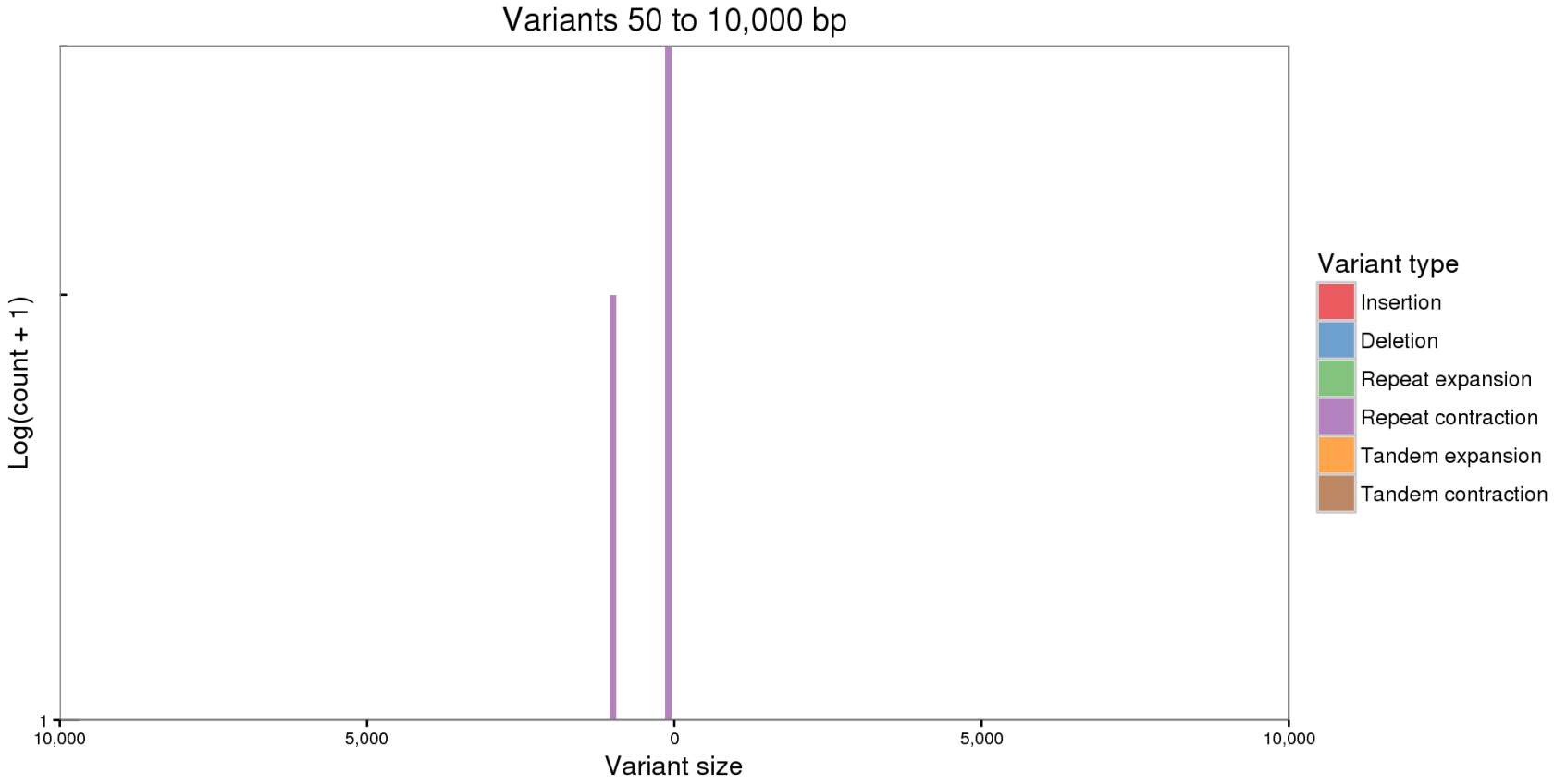
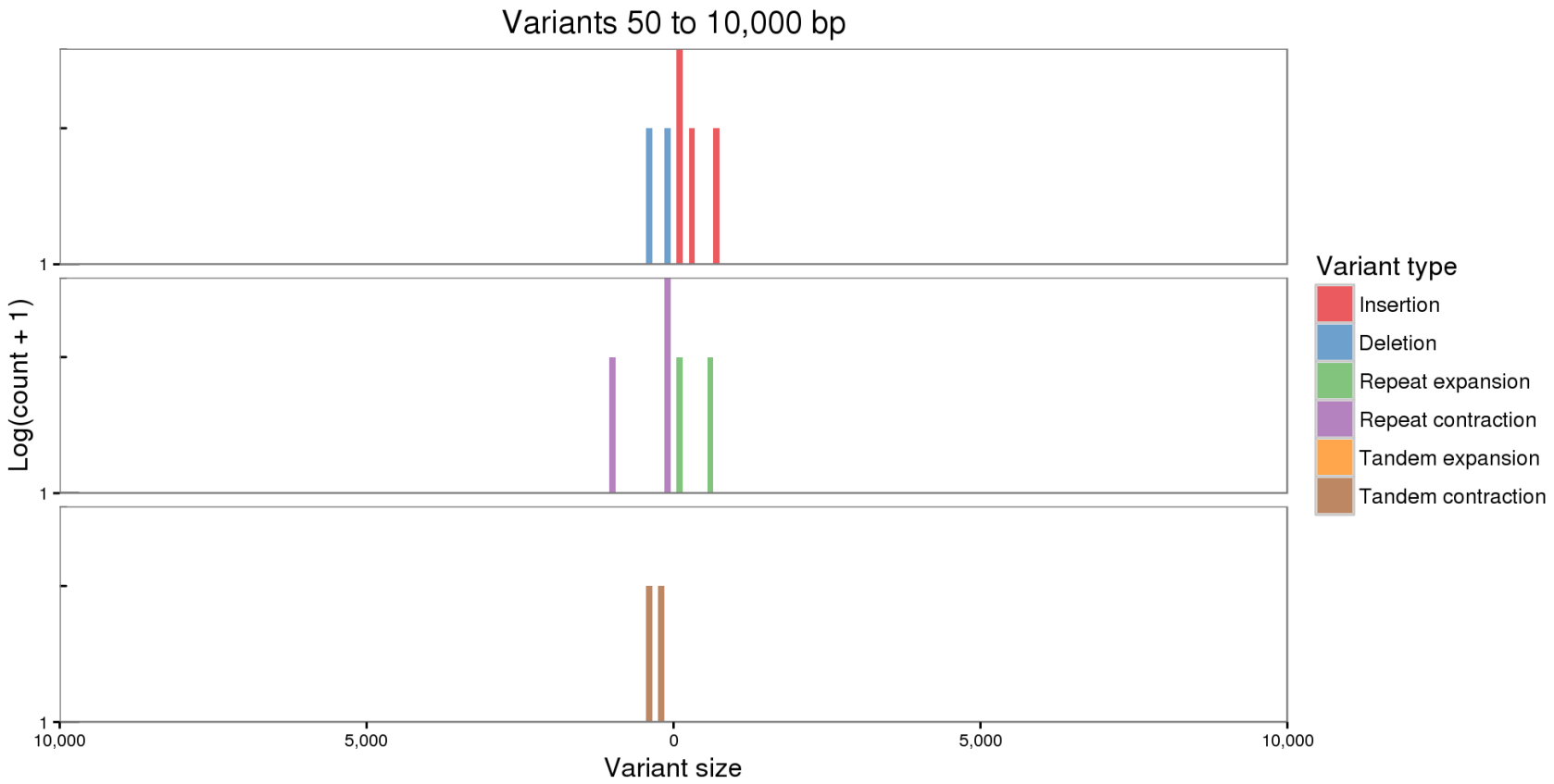


Figure S6. MUMmer size distributions of all variants between Flye-assembled genomes and reference.

(a) Salmonella_1ng against *Salmonella enterica* (GCF_000006945)



(b) Salmonella_3ng against *Salmonella enterica* (GCF_000006945)



(c) Salmonella_5ng against *Salmonella enterica* (GCF_000006945)

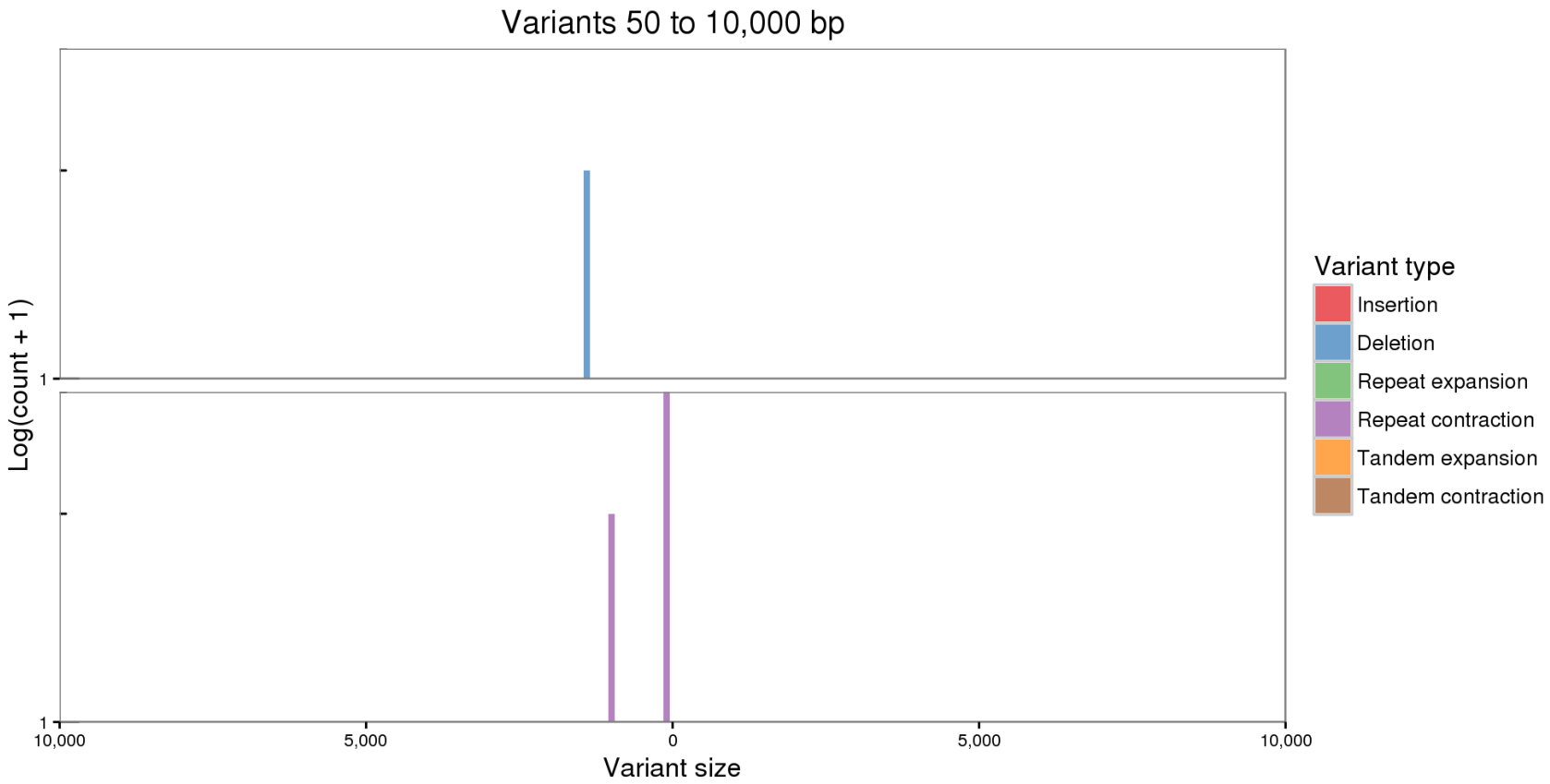


Table S1: DNA QC with the DNA Spectrophotometer (DS-11 Series from DeNovix).

Sample Name	Concentration	Units	Factor	A260	Pathlength (mm)	260/280	260/280 Alert	260/230	260/230 Alert
P. aeruginosa	15.322	ng/µL	50	0.3064	10	1.99	Below min concentration	1.21	Below min concentration
lambda (aliquot for the Pseudomona samples)	620.585	ng/µL	50	12.4117	10	1.77	Met criteria	2.08	Met criteria
lambda2 (aliquot for the Salmonella samples)	656.531	ng/µL	50	13.1306	10	1.79	Met criteria	2.08	Met criteria

Table S2: Preparation of microbial DNA standards for MinION sequencing.

Pseudomona_1	Lambda	*Pseudomonas aeruginosa*	Nuclease-free water	Total
Concentration	300 ng/µL	10 ng/µL	-	-
Volume	17 µL	30 µL	-	47 µL
Total DNA	5,100 ng	300 ng	-	5,400 ng

Pseudomona_2	Lambda	*Pseudomonas aeruginosa*	Nuclease-free water	Total
Concentration	300 ng/µL	10 ng/µL	-	-
Volume	17 µL	30 µL	-	47 µL
Total DNA	5,100 ng	300 ng	-	5,400 ng

Table S3: Preparation of microbial DNA standards for Flongle sequencing.

Salmonella_1ng	Lambda	*Salmonella enterica*	Nuclease-free water	Total
Concentration	300 ng/µL	10 ng/µL	-	-
Volume	2 µL	0.1 µL	21.4 µL	23.5 µL
Total DNA	600 ng	1 ng	-	601 ng

Salmonella_3ng	Lambda	*Salmonella enterica*	Nuclease-free water	Total
Concentration	300 ng/µL	10 ng/µL	-	-
Volume	2 µL	0.3 µL	21.2 µL	23.5 µL
Total DNA	600 ng	3 ng	-	603 ng

Salmonella_5ng	Lambda	*Salmonella enterica*	Nuclease-free water	Total
Concentration	300 ng/µL	10 ng/µL	-	-
Volume	2 µL	0.5 µL	21 µL	23.5 µL
Total DNA	600 ng	5 ng	-	605 ng

Table S4: Adaptive Sampling statistics, during which each read passing the nanopore was mapped against the reference genome of the Lambda phage (NC_001416.1).

	Samonella_1ng	Samonella_3ng	Samonella_5ng	P_aeruginosa_1	P_aeruginosa_2
Unblock	125212	139630	38522	1034891	998427
No_decision	32178	62732	35137	904914	910358
Stop_receiving	1331	3272	1983	52348	60647
Total	158721	205634	75642	1992153	1969432

Table S5: Adaptive Sampling statistics by WIMP classification

	Samonella_1ng	Samonella_3ng	Samonella_5ng	Pseudomonas
Input target DNA (ng)	1	3	5	300
Input noise DNA (ng)	600	600	600	5100
Input target/noise ratio	0.001666666667	0.005	0.008333333333	0.05882352941
Output target DNA (number of reads: Family)	60,589	62,735	60,165	256,693
Output noise DNA (number of reads: Family)	953	868	752	13,815
Output target/noise ratio	63.57712487	72.27534562	80.00664894	18.58074557
Output target DNA (number of reads: Genus)	433	2,013	5,804	256,622
Output noise DNA (number of reads: Genus)	161	131	175	13,629
Output target/noise ratio	2.689440994	15.36641221	33.16571429	18.82911439
Input target DNA length (estimated by Manufacturer)	6000	6000	6000	6000
Input noise DNA length (estimated by Manufacturer)	48502	48502	48502	48502
Input target/noise ratio	0.1237062389	0.1237062389	0.1237062389	0.1237062389
Output target DNA length (mean)	1807.946882	1961.964729	1999.506375	1,823.25
Output noise DNA length (mean)	443.826087	410.450382	454.525714	532.88
Output target/noise ratio	4.073548029	4.780029	4.399105075	3.42

Output target DNA length (s.d.)	1166.482912	1192.734209	1187.387932	1,196.21
Output noise DNA length (s.d.)	49.923763	64.114467	66.662283	554.10
Output target/noise ratio	23.36528422	18.60319932	17.81199021	2.16
Target DNA: no_decision	380	1862	5394	975621
Target DNA: stop_receiving	30	123	391	55543
Target DNA: unblock	23	28	19	1766
Target DNA stop_receiving/unblock ratio	1.304347826	4.392857143	20.57894737	31.45130238
Noise DNA: no_decision	0	0	0	790
Noise DNA: stop_receiving	0	0	0	75
Noise DNA: unblock	161	131	175	11574
Noise DNA stop_receiving/unblock ratio	0	0	0	0.006480041472

Table S6: Flye-assembled genomes

	Samonella_1ng	Samonella_3ng	Samonella_5ng	P_aeruginosa
Basecalling	Guppy_SUP	Guppy_SUP	Guppy_SUP	Guppy_HAC
Total length	182527	1557907	1055836	7439804
Fragment	5	44	35	98
Fragments N50	49739	54023	39216	1142881
Largest fragment	50996	158196	152933	1877226
Scaffolds	0	0	0	0
Mean coverage	165	23	32	244

Table S7: Analysis of the assembled target genomes against the target reference genome (GCF_000006945 or GCF_000006765).

"1-to-1" refers to an alignment type in MUMmer software where each position in one sequence is aligned to a single, corresponding position in the other sequence, meaning there is a direct, one-to-one correspondence between the two sequences being compared, with no duplications or major rearrangements involved.

	Salmonella enterica (GCF_000006945)	Samonella_1ng	Salmonella enterica (GCF_000006945)	Samonella_3ng	Salmonella enterica (GCF_000006945)	Samonella_5ng	Pseudomonas Aeruginosa (GCF_000006765)	P_aeruginosa
Total Bases	4857450	182527	4857450	1557907	4857450	1055836	6264404	7439804
Aligned Bases	17573(0.36%)	16793(9.20%)	1357232(27.94%)	1279936(82.16%)	899445(18.52%)	846000(80.13%)	6258121(99.90%)	6297543(84.65%)
Unaligned Bases	4839877(99.64%)	165734(90.80%)	3500218(72.06%)	277971(17.84%)	3958005(81.48%)	209836(19.87%)	6283(0.10%)	1142261(15.35%)
Total Sequences	1	5	1	44	1	35	1	98
Aligned Sequences	1(100.00%)	4(80.00%)	1(100.00%)	42(95.45%)	1(100.00%)	33(94.29%)	1(100.00%)	20(20.41%)
Unaligned Sequences	0(0.00%)	1(20.00%)	0(0.00%)	2(4.55%)	0(0.00%)	2(5.71%)	0(0.00%)	78(79.59%)
1-to-1	10	10	59	59	48	48	14	14
Total Length	16798	16793	1280926	1285001	845297	845479	6263146	6261391
Average Length	1679.8	1679.3	21710.61	21779.68	17610.35	17614.15	447367.57	447242.21
Average Identity	82.53	82.53	97.58	97.58	97.77	97.77	99.97	99.97
M-to-M	12	12	129	129	117	117	32	32
Total Length	17573	17567	1362705	1366968	922144	922744	6317822	6316059
Average Length	1464.42	1463.92	10563.6	10596.65	7881.57	7886.7	197431.94	197376.84
Average Identity	82.5	82.5	97.45	97.45	97.58	97.58	99.96	99.96
Breakpoints	24	24	258	245	234	216	62	22
Relocations	0	1	2	4	1	3	1	1
Translocations	5	0	44	0	37	0	12	0

Inversions	0	1	0	1	2	5	0	2
Insertions	15	14	182	85	160	62	5	20
Insertion Sum	4840652	162499	3581127	182225	4033162	148657	13259	53345
Insertion Average	322710.13	11607.07	19676.52	2143.82	25207.26	2397.69	2651.8	2667.25
Tandem Insertion	0	0	2	0	0	0	0	0
Tandem Insertion Sum	0	0	657	0	0	0	0	0
Tandem Insertion Average	0	0	328.5	0	0	0	0	0
Total SNPs	2834	2834	14372	14372	8440	8440	49	49
Total GSNPs	7	7	1971	1971	1199	1199	39	39
Total Indels	11	11	15771	15771	9039	9039	1835	1835
Total GIndels	0	0	2749	2749	1723	1723	1511	1511

Table S8: Analysis of the assembled noise genomes against the noise reference genome (NC_001416.1).

	Lambda phage (NC_001416.1)	Samonella_1ng	Lambda phage (NC_001416.1)	Samonella_3ng	Lambda phage (NC_001416.1)	Samonella_5ng	Lambda phage (NC_001416.1)	P_aeruginosa
Total Bases	48502	182527	48502	1557907	48502	1055836	48502	7439804
Aligned Bases	48502(100.00 %)	53662(29.40%)	48502(100.00 %)	49290(3.16%)	48422(99.84%)	49530(4.69%)	48502(100.00 %)	750859(10.09%)
Unaligned Bases	0(0.00%)	128865(70.60 %)	0(0.00%)	1508617(96.84 %)	80(0.16%)	1006306(95.31 %)	0(0.00%)	6688945(89.91 %)
Total Sequences	1	5	1	44	1	35	1	98
Aligned Sequences	1(100.00%)	2(40.00%)	1(100.00%)	3(6.82%)	1(100.00%)	3(8.57%)	1(100.00%)	50(51.02%)
Unaligned Sequences	0(0.00%)	3(60.00%)	0(0.00%)	41(93.18%)	0(0.00%)	32(91.43%)	0(0.00%)	48(48.98%)
1-to-1	1	1	2	2	1	1	2	2
Total Length	48502	48458	48501	48456	48422	48353	77602	77555
Average Length	48502	48458	24250.5	24228	48422	48353	38801	38777.5
Average Identity	99.89	99.89	99.89	99.89	99.84	99.84	99.93	99.93
M-to-M	2	2	4	4	3	3	72	72
Total Length	53709	53662	49336	49290	49599	49530	750039	750866
Average Length	26854.5	26831	12334	12322.5	16533	16510	10417.21	10428.69
Average Identity	99.89	99.89	99.81	99.81	99.82	99.82	98.91	98.91
Breakpoints	2	3	7	4	6	4	112	91
Relocations	0	0	1	1	0	0	0	0
Translocations	0	0	0	0	0	0	1	0
Inversions	0	0	0	0	0	0	0	0
Insertions	0	4	1	5	2	6	0	122
Insertion Sum	0	30099	1	57314	80	82210	0	869353
Insertion Average	0	7524.75	1	11462.8	40	13701.67	0	7125.84
Tandem Insertion	0	0	0	0	0	0	0	0

Tandem Insertion Sum	0	0	0	0	0	0	0	0
Tandem Insertion Average	0	0	0	0	0	0	0	0
Total SNPs	5	5	6	6	6	6	4	4
Total GSNPs	4	4	5	5	4	4	4	4
Total Indels	46	46	47	47	71	71	11	11
Total GIndels	42	42	46	46	62	62	11	11