

# CCQM Microbial Identity 16S rRNA Interlaboratory Study

## Supplemental Results

Nathan Olson

July 8, 2014

# Contents

<b>1</b>	<b>Biologically Conserved Positions</b>	<b>5</b>
<b>2</b>	<b>Biologically Variable Positions</b>	<b>7</b>
<b>3</b>	<b>Likely sets of variant combinations</b>	<b>10</b>
<b>4</b>	<b>Appendix</b>	<b>11</b>

## List of Figures

S1	<i>L. monocytogenes</i> base ratios . . . . .	8
S2	<i>E. coli</i> base ratios . . . . .	9

## List of Tables

S1	Biological conserved base quality statistics . . . . .	5
S2	<i>E. coli</i> variant combinations . . . . .	10
S3	<i>L. monocytogenes</i> variant combinations . . . . .	10
S4	<i>E. coli</i> positions pipeline comparison . . . . .	11
S5	<i>L. monocytogenes</i> positions pipeline pomparison . . . . .	12

# 1 Biologically Conserved Positions

None of the variants for the biologically conserved positions were called using both variant callers, indicating that the variants were potential false positives (Manuscript Table 2, Tables S4 and S5). Consensus base quality statistics for biologically conserved positions (Table S1).

Table S1: **Biologically Conserved Position Base Qualities** Characteristics of consensus based calls for conserved bases. Normalized quality values were obtained by dividing raw quality score (Raw Qual) assigned by GATK for each biologically conserved base position by the depth of coverage for that position

Org	Plat	Lab	Rep	Raw Qual	Normalized	Min	Max
Ecoli	454	LGC	1	140738.23	2.85	1.25	3.00
Ecoli	454	LGC	2	68081.73	2.85	0.63	2.98
Ecoli	454	LGC	3	128788.23	2.93	1.16	2.99
Ecoli	454	NMIA	1	11457.23	2.51	0.31	2.97
Ecoli	ION	NIMC	1	1165.23	2.78	0.59	3.14
Ecoli	ION	NIST	1	1112.23	2.48	0.51	3.16
Ecoli	Sanger	ATCC	1	34.23	17.11	9.31	31.24
Ecoli	Sanger	ISP	1	31.24	31.23	-10.00	31.24
Ecoli	Sanger	LGC	1	169.23	3.60	0.51	3.97
Ecoli	Sanger	NIST	1	115.23	3.97	-1.43	10.06
Lmono	454	LGC	1	11757.73	1.72	0.52	2.84
Lmono	454	LGC	2	115365.73	2.89	1.43	3.00
Lmono	454	LGC	3	103741.23	2.87	1.44	3.00
Lmono	454	NMIA	1	11635.23	2.41	0.79	2.92
Lmono	ION	NIMC	1	1173.23	2.81	0.33	3.14
Lmono	ION	NIST	1	1265.23	2.56	0.23	2.90
Lmono	Sanger	ATCC	1	34.23	17.11	-10.00	31.24
Lmono	Sanger	ISP	1	34.23	17.11	-10.00	31.24
Lmono	Sanger	LGC	1	169.23	3.45	1.26	3.71
Lmono	Sanger	NIST	1	242.23	3.41	2.18	3.78

A number of false positive variant calls were due to low sequencing coverage because the targeted sequencing strategy was responsible for false positive variant calls in six of the eight “454” datasets, excluding LGC *E. coli* replicate 2 and *L. monocytogenes* replicate 1. For those six datasets, a variant was called at the last position in the gap between the two sequencing regions, bases 940 and 963 relative to reference sequences for *E. coli* and *L. monocytogenes*, respectively. A 40 bp region that was not part of the targeted sequencing region had significantly lower median coverage than the targeted region (2 X vs. 30,110 X, respectively) for all “454” datasets combined (Figures 1 and 2).

A number of false positive variants were called due to contaminants. A low level of contaminating reads (150) were present in the LGC *L. monocytogenes* rep 1 dataset. A BLAST analysis of a representative of these reads indicated that they were from *E. coli* (E value of 0.0), a well known contaminant of molecular biology reagents (Section 4). A number of *Escherichia coli* strains E value of 0.0. in LGC Lmono rep 1 454 dataset. False positive variant calls were also attributable to the sequencing strategy and the variant calling algorithm. Resulting in a number of variants called due to strand bias. Strand bias was identified as the cause of the false positive variant call because greater than 99% of the reads were covering the variant bases were in the same direction. The strand bias was a product of the amplicon-based sequencing. For the *E. coli* dataset the variant was at the 3 prime end of the region 1 amplicon (Fig. 1) and the variant in the *L. monocytogenes* was at the 5 prime end of the region 2 amplicon (Fig. 2). As a result a majority of the reads covering the variants were in a single direction as the reads in the other direction were not long enough to cover the variant. For whole genome sequencing data, read direction biases can indicate a systematic error. The UnifiedGenotyper variant caller takes into consideration strand bias resulting in the false positive variant calls and reports strand bias using the Fisher exact test statistic (see Table 2). A filtering step is

commonly performed when calling SNPs that would have identified these as false positive variants due to the low number of reads with the variant base.

## 2 Biologically Variable Positions

To determine the ratio of bases at the biologically variant positions, a novel Bayesian analysis based on binomial sampling theory was developed (Supplemental Computational Methods). According to the binomial distribution, the observed base ratios, while precise (due to high coverage), differed significantly from all potential copy ratios. Subsequently given, the observed base ratios, a Bayesian approach was used to identify the most probable copy ratio out of the possible abundant base ratios assuming *E. coli* and *L. monocytogenes* have seven and six 16S gene copies respectively (Figs. S1 and S2).

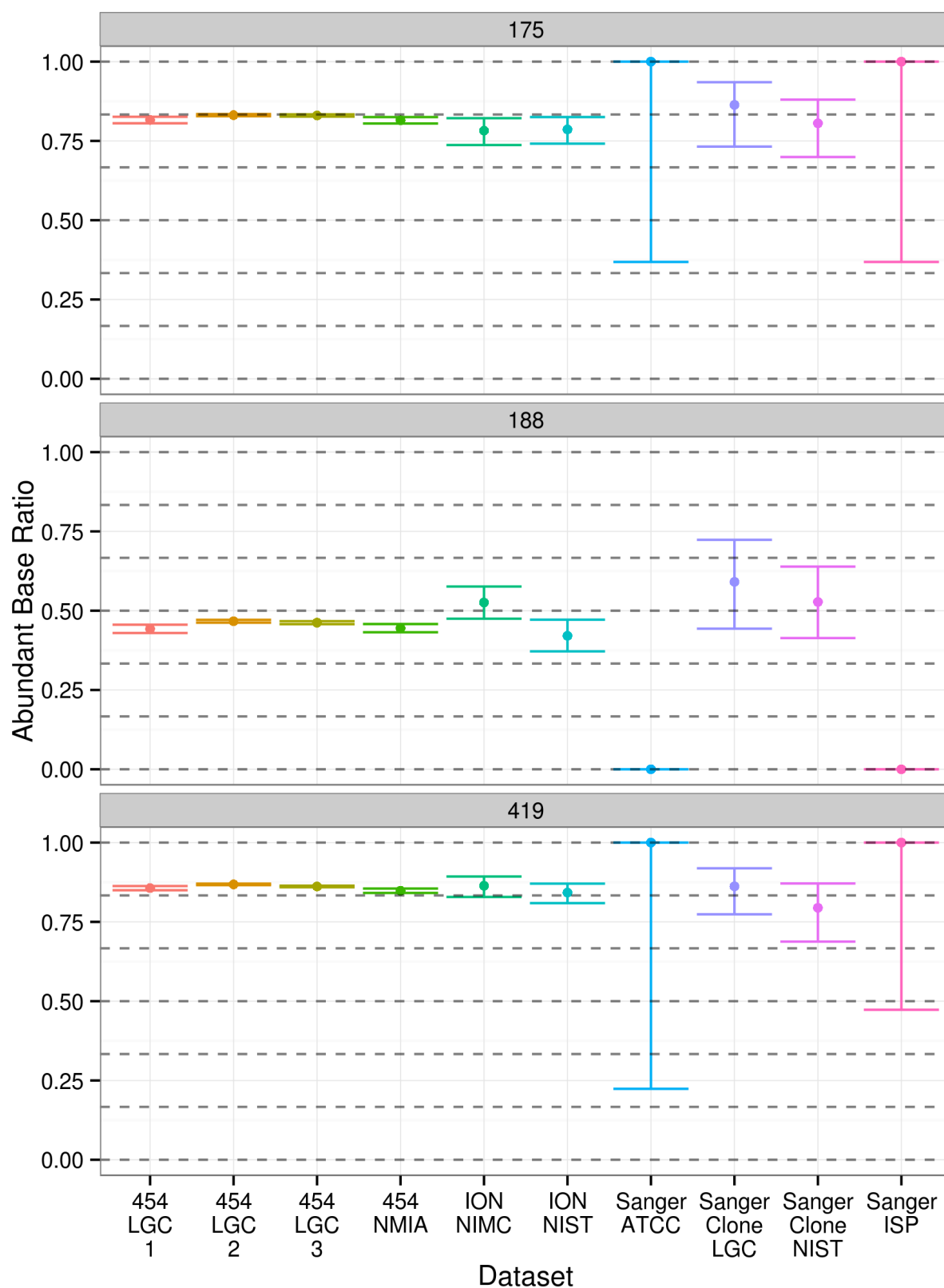


Figure S1: Base ratios at three biologically variable positions (175, 188 and 419) in *L. monocytogenes*. Variable positions shown in grey box above each graph. Error bars represent the 95 % posterior credibility interval estimated from a beta binomial distribution where  $\alpha$  is the major base count + 1 and  $\beta$  is the minor base count + 1. One sided credible intervals were calculated for prior probabilities of 0 and 1. Grey dashed lines indicate the potential base ratios assuming six gene copies (i.e. 0:6 corresponds to 0, 2:4 to 0.33, 3:3 to 0.5, 4:2 to 0.66, 5:1 to 0.83 and 6:0 to 1).



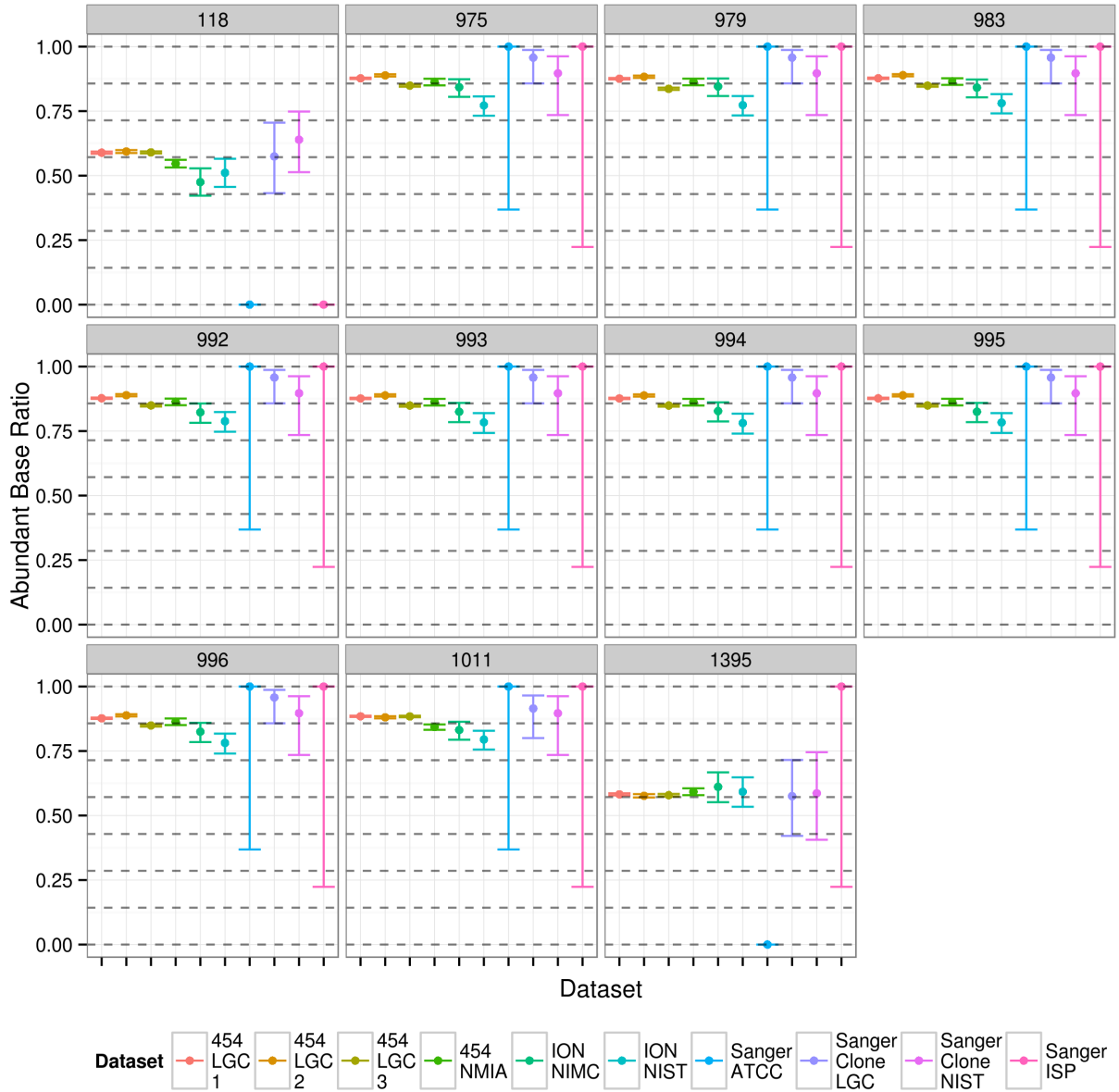


Figure S2: Base ratios at eleven biologically variable positions in *E. coli*. Variable positions shown in grey box above each graph. Error bars represent the 95 % posterior credibility interval estimated from a beta binomial distribution where  $\alpha$  is the major base count + 1 and  $\beta$  is the minor base count + 1. One sided credible intervals were calculated for prior probabilities of 0 and 1. Grey dashed lines indicate the potential base ratios assuming seven gene copies, (i.e. 0:7 to 0; 1:6 to 0.14; 2:5 to 0.26; 3:4 to 0.43, 4:3 to 0.57; 5:2 to 0.71; 6:1 to 0.86; and 7:0 to 1 ).

### 3 Likely sets of variant combinations

Most likely combination of variant strings for “454” and Sanger Clone library datasets (Table S2 and Table S3).

Table S2: **Estimated most likely set of variant combinations for *E. coli*.** See supplemental computation methods for how chimera and likelihood were calculated.

dataset	likelihood	chimera	ACCGATTGTA	ACCGATTGTG	GGTAGAATCA
Ecoli-454-LGC-1	0.04	275.55	3	3	1
Ecoli-454-LGC-2	0.03	275.27	3	3	1
Ecoli-454-LGC-3	0.04	242.13	3	3	1
Ecoli-454-NMIA-1	0.06	30.74	3	3	1
Ecoli-LGC-Sanger-Clones.csv	0.05	3.55	3	4	0
Ecoli-NIST-Sanger-Clones.csv	0.12	4.54	3	4	0
Consensus	0.04	717.62	3	3	1

Table S3: **Estimated most likely set of variant combinations for *L. monocytogenes*.** See supplemental computation methods for how chimera and likelihood were calculated.

dataset	likelihood	chimera	GCG	GTA	GTG	TCG
Lmono-454-LGC-1	0.00	47.31	2	1	2	1
Lmono-454-LGC-2	0.01	572.64	2	1	2	1
Lmono-454-LGC-3	0.01	319.50	2	1	2	1
Lmono-454-NMIA-1	0.00	55.95	2	1	2	1
Lmono-LGC-Sanger-Clones.csv	0.00	5.13	2	1	2	1
Lmono-NIST-Sanger-Clones.csv	0.01	8.38	2	1	2	1
Consensus	0.01	850.46	2	1	2	1

## 4 Appendix

### Full List of False Positive Variants

All variants called by the 8 pipelines used during the pipeline validation along with the suspected cause of the variant. The following abbreviations were used in Tables S4 and S5: Org - Organism, Plat - sequencing platform, Rep - replicate, Map - read mapping algorithm, Var - variant calling algorithm, POS - base position relative to the reference, DP - coverage, QUAL - confidence in variant call assigned by variant calling algorithm, MQ - mapping quality score assigned by mapping algorithm, FS - fisher strain bias test statistic, Cause - hypothesized cause of false positive variant call. See supplemental manuscript methods section for mapping algorithm and variant calling algorithm descriptions. Note that for the NIST Ion Torrent *L. monocytogenes* dataset at position 792 a variant was called by the UnifiedGenotyper Variant Calling Algorithm when the reads were mapped using both bwa and tmap, but the FS score was only above 60 when the reads were mapped with tmap. Upon manual inspection of the results we attributed the false positive to a strand bias.

Table S4: ***E. coli* Pipeline Comparison** Characteristics of variant calls for different bioinformatic pipelines.

Org	Plat	Lab	Rep	Map	Var	POS	DP	QUAL	MQ	FS	Cause
Ecoli	454	LGC	1	bwa	gatk	324	250	443.77	60.00	47.88	End of read
Ecoli	454	LGC	1	TMAP	gatk	324	250	432.77	88.54	60.26	End of read
Ecoli	454	LGC	1	bwa	gatk	325	250	308.77	60.00	53.48	End of read
Ecoli	454	LGC	1	TMAP	gatk	325	250	309.77	88.54	50.67	End of read
Ecoli	454	LGC	1	bwa	sam	396	2551	81.00	60.00		End of read
Ecoli	454	LGC	1	TMAP	sam	396	3013	37.00	56.00		End of read
Ecoli	454	LGC	1	bwa	gatk	940	19	215.77	60.00	28.54	Non-target region
Ecoli	454	LGC	1	TMAP	gatk	940	21	179.77	80.15	28.54	Non-target region
Ecoli	454	LGC	1	bwa	gatk	959	250	1222.77	60.00	9.12	End of read
Ecoli	454	LGC	2	bwa	gatk	106	250	235.77	60.00	0.00	End of read
Ecoli	454	LGC	2	TMAP	gatk	106	250	34.77	68.90	0.00	End of read
Ecoli	454	LGC	2	bwa	gatk	959	250	795.77	59.98	28.04	End of read
Ecoli	454	LGC	3	bwa	gatk	324	250	231.77	59.83	40.63	End of read
Ecoli	454	LGC	3	TMAP	gatk	324	250	739.77	88.54	63.25	End of read
Ecoli	454	LGC	3	bwa	gatk	325	250	556.77	59.83	60.23	End of read
Ecoli	454	LGC	3	TMAP	gatk	325	250	498.77	88.54	36.85	End of read
Ecoli	454	LGC	3	bwa	gatk	348	250	741.77	59.92	11.62	End of read
Ecoli	454	LGC	3	bwa	sam	417	1032	22.00	60.00		Homopolymer
Ecoli	454	LGC	3	TMAP	sam	417	1020	32.00	58.00		Homopolymer
Ecoli	454	LGC	3	bwa	gatk	940	9	91.05	60.00	0.00	Non-target region
Ecoli	454	LGC	3	TMAP	gatk	940	14	194.29	82.41	0.00	Non-target region
Ecoli	454	NMIA	1	TMAP	gatk	313	250	5630.77	80.26	453.68	Strand bias
Ecoli	454	NMIA	1	TMAP	gatk	508	250	1160.77	83.71	0.00	End of read
Ecoli	454	NMIA	1	TMAP	gatk	509	250	1208.77	83.71	0.00	End of read
Ecoli	454	NMIA	1	TMAP	gatk	510	250	1275.77	83.71	0.00	End of read
Ecoli	454	NMIA	1	TMAP	gatk	514	250	1185.77	83.71	0.00	End of read
Ecoli	454	NMIA	1	TMAP	sam	514	6337	5.46	59.00		End of read
Ecoli	454	NMIA	1	TMAP	gatk	901	208	8061.77	84.22	0.00	Non-target region
Ecoli	454	NMIA	1	TMAP	gatk	904	208	8023.77	84.22	0.00	Non-target region
Ecoli	454	NMIA	1	TMAP	gatk	934	250	8711.77	71.76	0.00	Non-target region
Ecoli	454	NMIA	1	TMAP	gatk	935	250	8708.77	71.76	0.00	Non-target region
Ecoli	454	NMIA	1	TMAP	gatk	938	250	8620.77	71.71	0.00	Non-target region
Ecoli	454	NMIA	1	TMAP	sam	938	2747	9.54	56.00		Non-target region
Ecoli	454	NMIA	1	TMAP	gatk	939	250	8619.77	71.71	0.00	Non-target region

Ecoli	454	NMIA	1	TMAP	sam	939	2747	15.20	60.00		Non-target region
Ecoli	454	NMIA	1	TMAP	sam	941	2747	9.52	55.00		Non-target region
Ecoli	ION	NIMC	1	TMAP	sam	1463	169	22.50	60.00		End of reference
Ecoli	Sanger	NIST	1	TMAP	sam	1463	29	139.00	60.00		End of reference
Ecoli	Sanger	NIST	1	TMAP	sam	1464	29	214.00	60.00		End of reference

Table S5: *L. monocytogenes* Positions Pipeline Comparison  
Characteristics of variant calls for different bioinformatic pipelines.

Org	Plat	Lab	Rep	Map	Var	POS	DP	QUAL	MQ	FS	Cause
Lmono	454	LGC	1	bwa	gatk	315	250	4752.77	45.92	101.16	Strand bias
Lmono	454	LGC	1	bwa	gatk	328	250	4865.77	45.96	107.67	Strand bias
Lmono	454	LGC	1	TMAP	gatk	334	250	4700.77	68.98	300.59	Strand bias
Lmono	454	LGC	1	bwa	gatk	354	250	62.77	57.95	38.10	End of read
Lmono	454	LGC	1	bwa	gatk	366	248	47.77	57.94	40.36	End of read
Lmono	454	LGC	1	bwa	gatk	508	250	1386.77	51.13	7.03	Contaminants
Lmono	454	LGC	1	bwa	sam	508	7744	10.40	55.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	533	166	1407.77	46.04	0.00	Contaminants
Lmono	454	LGC	1	bwa	sam	533	1763	156.00	38.00		Contaminants
Lmono	454	LGC	1	TMAP	gatk	533	250	1768.77	48.73	0.00	Contaminants
Lmono	454	LGC	1	TMAP	gatk	536	250	1779.77	48.73	0.00	Contaminants
Lmono	454	LGC	1	bwa	gatk	537	166	1394.77	46.04	1.78	Contaminants
Lmono	454	LGC	1	bwa	sam	537	1763	88.00	38.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	538	166	1634.77	46.04	0.83	Contaminants
Lmono	454	LGC	1	bwa	sam	538	1623	128.00	38.00		Contaminants
Lmono	454	LGC	1	TMAP	gatk	539	250	1798.77	48.73	0.00	Contaminants
Lmono	454	LGC	1	bwa	gatk	548	166	2211.77	46.04	0.00	Contaminants
Lmono	454	LGC	1	bwa	sam	548	1762	201.00	37.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	549	166	2217.77	46.04	0.00	Contaminants
Lmono	454	LGC	1	bwa	sam	549	1763	186.00	37.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	550	166	2247.77	46.04	0.00	Contaminants
Lmono	454	LGC	1	bwa	sam	550	1763	175.00	37.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	555	167	2077.77	46.14	0.00	Contaminants
Lmono	454	LGC	1	bwa	sam	555	1764	222.00	37.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	559	167	2201.77	46.14	0.00	Contaminants
Lmono	454	LGC	1	bwa	sam	559	1763	222.00	37.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	574	168	2288.77	46.03	0.00	Contaminants
Lmono	454	LGC	1	bwa	sam	574	1765	189.00	37.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	585	168	1737.77	46.03	0.00	Contaminants
Lmono	454	LGC	1	bwa	sam	585	1736	213.00	37.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	587	168	1980.77	46.03	0.00	Contaminants
Lmono	454	LGC	1	bwa	sam	587	1741	212.00	37.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	595	168	2352.77	46.03	0.00	Contaminants
Lmono	454	LGC	1	bwa	sam	595	1741	188.00	38.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	677	250	5127.77	58.81	4.08	Contaminants
Lmono	454	LGC	1	bwa	sam	677	4525	222.00	58.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	700	249	5372.77	56.79	14.63	Contaminants
Lmono	454	LGC	1	bwa	sam	700	4604	222.00	58.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	703	249	4820.77	56.79	4.37	Contaminants
Lmono	454	LGC	1	bwa	sam	703	4604	222.00	58.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	712	249	4896.77	56.79	13.22	Contaminants
Lmono	454	LGC	1	bwa	sam	712	4602	222.00	58.00		Contaminants

Lmono	454	LGC	1	bwa	gatk	716	249	4009.77	56.79	0.00	Contaminants
Lmono	454	LGC	1	bwa	sam	716	4602	222.00	59.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	729	226	3856.77	58.93	0.72	Contaminants
Lmono	454	LGC	1	bwa	sam	729	4510	222.00	59.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	731	226	2873.77	58.93	0.00	Contaminants
Lmono	454	LGC	1	bwa	sam	731	4487	201.00	59.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	733	226	3028.77	58.93	0.00	Contaminants
Lmono	454	LGC	1	bwa	sam	733	4511	182.00	60.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	738	226	4461.77	58.93	0.00	Contaminants
Lmono	454	LGC	1	bwa	sam	738	4493	222.00	59.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	740	226	4588.77	58.93	0.00	Contaminants
Lmono	454	LGC	1	bwa	sam	740	4499	222.00	59.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	741	227	4498.77	58.91	0.00	Contaminants
Lmono	454	LGC	1	bwa	sam	741	4503	222.00	59.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	742	227	4566.77	58.91	0.00	Contaminants
Lmono	454	LGC	1	bwa	sam	742	4509	222.00	59.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	743	227	4633.77	58.91	0.00	Contaminants
Lmono	454	LGC	1	bwa	sam	743	4509	222.00	59.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	753	250	3750.77	58.29	0.78	Contaminants
Lmono	454	LGC	1	bwa	sam	753	4617	222.00	59.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	757	250	4258.77	58.29	34.57	Contaminants
Lmono	454	LGC	1	bwa	sam	757	4617	222.00	59.00		Contaminants
Lmono	454	LGC	1	TMAP	gatk	924	189	2175.77	43.35	0.00	Contaminants
Lmono	454	LGC	1	TMAP	gatk	926	189	2221.77	43.35	0.00	Contaminants
Lmono	454	LGC	1	TMAP	gatk	928	189	2184.77	43.35	0.00	Contaminants
Lmono	454	LGC	1	TMAP	gatk	930	189	2184.77	43.35	0.00	Contaminants
Lmono	454	LGC	1	TMAP	gatk	953	250	4356.77	55.87	0.00	Contaminants
Lmono	454	LGC	1	TMAP	gatk	955	250	4334.77	55.87	0.00	Contaminants
Lmono	454	LGC	1	TMAP	gatk	957	250	4172.77	55.91	0.00	Contaminants
Lmono	454	LGC	1	TMAP	gatk	958	250	4210.77	55.91	0.00	Contaminants
Lmono	454	LGC	1	TMAP	gatk	959	250	4251.77	55.91	0.00	Contaminants
Lmono	454	LGC	1	TMAP	gatk	961	250	4334.77	55.91	0.00	Contaminants
Lmono	454	LGC	1	TMAP	gatk	963	250	4217.77	55.91	0.00	Contaminants
Lmono	454	LGC	1	bwa	gatk	982	250	1899.77	60.00	687.88	Contaminants
Lmono	454	LGC	1	bwa	gatk	1047	250	3824.77	59.30	35.71	Contaminants
Lmono	454	LGC	1	bwa	sam	1047	8006	225.00	60.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	1055	250	3067.77	59.16	2.72	Contaminants
Lmono	454	LGC	1	bwa	sam	1055	8011	225.00	60.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	1072	250	3129.77	58.75	1.51	Contaminants
Lmono	454	LGC	1	bwa	sam	1072	8022	225.00	60.00		Contaminants
Lmono	454	LGC	1	bwa	sam	1077	7975	225.00	60.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	1192	250	5493.77	58.56	55.04	Contaminants
Lmono	454	LGC	1	bwa	sam	1192	8008	225.00	60.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	1201	250	4530.77	58.56	63.08	Contaminants
Lmono	454	LGC	1	bwa	sam	1201	8006	225.00	60.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	1208	250	3580.77	58.56	23.54	Contaminants
Lmono	454	LGC	1	bwa	sam	1208	8009	162.00	60.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	1213	250	4330.77	58.56	64.05	Contaminants
Lmono	454	LGC	1	bwa	sam	1213	8010	216.00	60.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	1304	249	4878.77	59.84	69.84	Contaminants
Lmono	454	LGC	1	bwa	sam	1304	7998	225.00	60.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	1307	250	5334.77	59.84	74.39	Contaminants
Lmono	454	LGC	1	bwa	sam	1307	7999	225.00	60.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	1318	250	5106.77	59.78	91.22	Contaminants

Lmono	454	LGC	1	bwa	sam	1318	8002	225.00	60.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	1321	250	4735.77	59.78	81.66	Contaminants
Lmono	454	LGC	1	bwa	sam	1321	8002	225.00	60.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	1329	250	4976.77	59.81	85.25	Contaminants
Lmono	454	LGC	1	bwa	sam	1329	8005	225.00	60.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	1356	249	3918.77	59.73	50.60	Contaminants
Lmono	454	LGC	1	bwa	sam	1356	8010	191.00	60.00		Contaminants
Lmono	454	LGC	2	bwa	gatk	315	250	3384.77	51.77	1.33	Contaminants
Lmono	454	LGC	2	bwa	gatk	328	250	3396.77	51.77	1.33	Contaminants
Lmono	454	LGC	2	bwa	gatk	346	250	149.77	59.51	37.09	End of read
Lmono	454	LGC	2	bwa	gatk	347	250	413.77	59.51	51.04	End of read
Lmono	454	LGC	2	TMAP	gatk	347	250	130.77	91.19	40.53	End of read
Lmono	454	LGC	2	bwa	gatk	555	144	69.77	57.88	2.20	Contaminants
Lmono	454	LGC	2	bwa	gatk	587	144	83.77	57.88	2.17	Contaminants
Lmono	454	LGC	2	bwa	gatk	677	145	122.77	59.80	2.17	Contaminants
Lmono	454	LGC	2	bwa	gatk	700	145	119.77	59.80	2.17	Contaminants
Lmono	454	LGC	2	bwa	gatk	703	145	119.77	59.80	2.17	Contaminants
Lmono	454	LGC	2	bwa	gatk	712	145	105.77	59.80	2.17	Contaminants
Lmono	454	LGC	2	bwa	gatk	716	145	94.77	59.80	2.17	Contaminants
Lmono	454	LGC	2	bwa	gatk	729	145	118.77	59.80	2.17	Contaminants
Lmono	454	LGC	2	bwa	gatk	738	145	122.77	59.80	2.17	Contaminants
Lmono	454	LGC	2	bwa	gatk	740	145	122.77	59.80	2.17	Contaminants
Lmono	454	LGC	2	bwa	gatk	741	145	161.77	59.80	2.12	Contaminants
Lmono	454	LGC	2	bwa	gatk	742	145	119.77	59.80	2.17	Contaminants
Lmono	454	LGC	2	bwa	gatk	743	145	119.77	59.80	2.17	Contaminants
Lmono	454	LGC	2	bwa	gatk	753	145	114.77	59.80	2.17	Contaminants
Lmono	454	LGC	2	bwa	gatk	757	145	120.77	59.80	2.17	Contaminants
Lmono	454	LGC	2	bwa	gatk	963	122	701.29	60.00	0.00	Non-target region
Lmono	454	LGC	2	TMAP	gatk	963	21	82.31	74.82	0.00	Non-target region
Lmono	454	LGC	2	bwa	gatk	1047	250	3748.77	59.89	4.09	Contaminants
Lmono	454	LGC	2	bwa	gatk	1055	250	3571.77	59.89	7.83	Contaminants
Lmono	454	LGC	2	bwa	gatk	1072	250	2637.77	59.92	18.48	Contaminants
Lmono	454	LGC	2	bwa	gatk	1077	249	2161.77	59.92	18.66	Contaminants
Lmono	454	LGC	2	bwa	gatk	1192	250	4830.77	59.65	8.22	Contaminants
Lmono	454	LGC	2	bwa	gatk	1201	250	4741.77	59.65	5.67	Contaminants
Lmono	454	LGC	2	bwa	gatk	1208	250	4714.77	59.65	8.23	Contaminants
Lmono	454	LGC	2	bwa	gatk	1213	250	4701.77	59.65	8.18	Contaminants
Lmono	454	LGC	2	bwa	gatk	1304	250	4652.77	60.00	6.46	Contaminants
Lmono	454	LGC	2	bwa	gatk	1307	250	4789.77	60.00	8.27	Contaminants
Lmono	454	LGC	2	bwa	gatk	1318	250	4710.77	60.00	8.42	Contaminants
Lmono	454	LGC	2	bwa	gatk	1321	250	4641.77	60.00	8.42	Contaminants
Lmono	454	LGC	2	bwa	gatk	1329	250	4555.77	60.00	10.65	Contaminants
Lmono	454	LGC	2	bwa	gatk	1356	250	4453.77	60.00	8.63	Contaminants
Lmono	454	LGC	3	bwa	gatk	346	250	102.77	60.00	32.12	End of read
Lmono	454	LGC	3	TMAP	gatk	346	250	294.77	91.54	40.01	End of read
Lmono	454	LGC	3	bwa	gatk	347	250	255.77	60.00	42.21	End of read
Lmono	454	LGC	3	bwa	gatk	370	250	67.77	60.00	8.83	End of read
Lmono	454	LGC	3	bwa	gatk	963	111	302.48	60.00	0.00	Non-target region
Lmono	454	LGC	3	TMAP	gatk	963	10	78.77	67.25	0.00	Non-target region
Lmono	454	NMIA	1	TMAP	gatk	330	250	5990.77	79.12	466.25	Strand bias
Lmono	454	NMIA	1	TMAP	gatk	334	250	5973.77	79.12	514.24	Strand bias
Lmono	454	NMIA	1	TMAP	gatk	335	250	5199.77	79.12	512.93	Strand bias
Lmono	454	NMIA	1	bwa	gatk	381	250	37.77	60.00	13.82	End of read
Lmono	454	NMIA	1	TMAP	gatk	533	249	1662.77	67.08	0.00	End of read

Lmono	454	NMIA	1	TMAP	gatk	932	94	3555.77	75.44	0.00	Non-target region
Lmono	454	NMIA	1	TMAP	gatk	936	92	3557.77	75.89	0.00	Non-target region
Lmono	454	NMIA	1	TMAP	gatk	954	250	8461.77	77.83	0.00	Non-target region
Lmono	454	NMIA	1	TMAP	gatk	957	250	8402.77	77.83	0.00	Non-target region
Lmono	454	NMIA	1	TMAP	gatk	961	250	8422.77	77.83	0.00	Non-target region
Lmono	454	NMIA	1	TMAP	gatk	962	250	8332.77	77.83	0.00	Non-target region
Lmono	454	NMIA	1	TMAP	gatk	963	250	8246.77	77.83	0.00	Non-target region
Lmono	ION	NIST	1	bwa	gatk	792	259	132.77	60.00	54.38	Strand bias
Lmono	ION	NIST	1	TMAP	gatk	792	275	323.77	85.79	69.53	Strand bias
Lmono	Sanger	LGC	1	bwa	sam	390	81	25.50	60.00		End of read
Lmono	Sanger	LGC	1	bwa	sam	1409	44	13.70	60.00		End of read
Lmono	Sanger	LGC	1	TMAP	sam	1505	41	71.20	60.00		End of reference
Lmono	Sanger	LGC	1	TMAP	sam	1506	41	71.20	60.00		End of reference
Lmono	Sanger	NIST	1	bwa	sam	865	74	76.50	60.00		End of read
Lmono	Sanger	NIST	1	TMAP	sam	865	68	77.50	59.00		End of read
Lmono	Sanger	NIST	1	bwa	gatk	867	67	264.77	60.00	0.00	End of read
Lmono	Sanger	NIST	1	bwa	sam	867	67	10.40	60.00		End of read
Lmono	Sanger	NIST	1	TMAP	gatk	867	64	249.77	96.41	0.00	End of read
Lmono	Sanger	NIST	1	TMAP	sam	867	64	12.30	59.00		End of read
Lmono	Sanger	NIST	1	TMAP	sam	1504	35	214.00	60.00		End of read

## Contaminants - BLAST results

BLAST reports for representative sequences of reads responsible for false positive variant calls in the LGC *L. monocytogenes* "454" rep 1 dataset.

BLASTN 2.2.29+

Reference: Zheng Zhang, Scott Schwartz, Lukas Wagner, and Webb Miller (2000), "A greedy algorithm for aligning DNA sequences", J Comput Biol 2000; 7(1-2):203-14.

RID: KH9SY3U8014

Database: Representative Chromosomes

2,857 sequences; 5,609,140,793 total letters

Query=

Length=558

		Score	E
Sequences producing significant alignments:		(Bits)	Value
ref NC_000913.3	Escherichia coli str. K-12 substr. MG1655, c...	979	0.0
ref NC_018658.1	Escherichia coli O104:H4 str. 2011C-3493 chr...	979	0.0
ref NC_017634.1	Escherichia coli O83:H1 str. NRG 857C chromo...	979	0.0
ref NC_011751.1	Escherichia coli UMN026 chromosome, complete...	979	0.0
ref NC_011750.1	Escherichia coli IAI39 chromosome, complete ...	979	0.0
ref NC_011740.1	Escherichia fergusonii ATCC 35469 chromosome...	979	0.0
ref NC_007384.1	Shigella sonnei Ss046 chromosome, complete g...	979	0.0
ref NC_002695.1	Escherichia coli O157:H7 str. Sakai chromoso...	979	0.0
ref NC_004337.2	Shigella flexneri 2a str. 301 chromosome, co...	974	0.0

ref|NC\_007613.1| Shigella boydii Sb227 chromosome, complete g... 974 0.0

#### ALIGNMENTS

>ref|NC\_000913.3| Escherichia coli str. K-12 substr. MG1655, complete genome  
Length=4641652

Features in this part of subject sequence:

rRNA-16S ribosomal RNA of rrnH operon

Score = 979 bits (530), Expect = 0.0  
Identities = 539/543 (99%), Gaps = 2/543 (0%)  
Strand=Plus/Plus

Query	3	CCTGATGCAGCCATGCCGCGTGTATGAAGAAGGCTTACGGGTTGT-AAGTACGTTTCAGC	61
Sbjct	224155	CCTGATGCAGCCATGCCGCGTGTATGAAGAAGGCTTACGGGTTGTAAAGTAC-TTTCAGC	224213
Query	62	GGGGAGGAAGGGAGTAAAGTTAATACCTTTGCTCATTGACGTTACCCGCAGAAGAAGCAC	121
Sbjct	224214	GGGGAGGAAGGGAGTAAAGTTAATACCTTTGCTCATTGACGTTACCCGCAGAAGAAGCAC	224273
Query	122	CGGCTAACTCCGTGCCAGCAGCCGCGTAATACGGAGGGTGCAAGCGTTAATCGGAATTA	181
Sbjct	224274	CGGCTAACTCCGTGCCAGCAGCCGCGTAATACGGAGGGTGCAAGCGTTAATCGGAATTA	224333
Query	182	CTGGGCGTAAAGCGCACGCAGGCGGTTTGTAAAGTCAGATGTGAAATCCCCGGGCTCAAC	241
Sbjct	224334	CTGGGCGTAAAGCGCACGCAGGCGGTTTGTAAAGTCAGATGTGAAATCCCCGGGCTCAAC	224393
Query	242	CTGGGAACTGCATCTGATACTGGCAAGCTTGAGTCTCGTAGAGGGGGGTAGAATTCCAGG	301
Sbjct	224394	CTGGGAACTGCATCTGATACTGGCAAGCTTGAGTCTCGTAGAGGGGGGTAGAATTCCAGG	224453
Query	302	TGTAGCGGTGAAATGCGTAGAGATCTGGAGGAATACCGGTGGCGAAGGCGGCCCCCTGGA	361
Sbjct	224454	TGTAGCGGTGAAATGCGTAGAGATCTGGAGGAATACCGGTGGCGAAGGCGGCCCCCTGGA	224513
Query	362	CGAAGACTGACGCTCAGGTGCGAAAGCGTGGGGAGCAAACAGGATTAGATACCCTGGTAG	421
Sbjct	224514	CGAAGACTGACGCTCAGGTGCGAAAGCGTGGGGAGCAAACAGGATTAGATACCCTGGTAG	224573
Query	422	TCCACGCCGTAAACGATGTGCGACTTGGAGGTTGTGCCCTTGAGGCGTGGCTTCCGGAGCT	481
Sbjct	224574	TCCACGCCGTAAACGATGTGCGACTTGGAGGTTGTGCCCTTGAGGCGTGGCTTCCGGAGCT	224633
Query	482	AACGCGTTAAGTCGACCGCCTGGGGAGTACGGCCGCAAGGTTAACTCAAATGAATTGA	541
Sbjct	224634	AACGCGTTAAGTCGACCGCCTGGGGAGTACGGCCGCAAGGTTAACTCAAATGAATTGA	224693
Query	542	CGG	544
Sbjct	224694	CGG	224696



Database: Representative Chromosomes  
Posted date: Mar 21, 2014 12:17 AM  
Number of letters in database: 5,609,140,793  
Number of sequences in database: 2,857

Lambda	K	H
1.33	0.621	1.12

Gapped

Lambda	K	H
1.28	0.460	0.850

Matrix: blastn matrix:1 -2

Gap Penalties: Existence: 0, Extension: 0

Number of Sequences: 2857

Number of Hits to DB: 6177

Number of extensions: 6

Number of successful extensions: 6

Number of sequences better than 10: 1

Number of HSP's better than 10 without gapping: 0

Number of HSP's gapped: 3

Number of HSP's successfully gapped: 3

Length of query: 558

Length of database: 5609140793

Length adjustment: 30

Effective length of query: 528

Effective length of database: 5609055083

Effective search space: 2961581083824

Effective search space used: 2961581083824

A: 0

X1: 13 (25.0 bits)

X2: 32 (59.1 bits)

X3: 54 (99.7 bits)

S1: 13 (25.1 bits)

S2: 21 (39.9 bits)