CCQM Microbial Identity 16S rRNA Interlaboratory Study

Supplemental Results

Nathan Olson

July 8, 2014

Contents

List of Figures

List of Tables

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	140744.23	2.84	1.25	3.00
Ecoli	454	LGC	2	67980.73	2.85	0.63	2.98
Ecoli	454	LGC	3	128734.23	2.93	1.16	2.99
Ecoli	454	NMIA	1	11458.23	2.51	0.31	2.97
Ecoli	ION	NIMC	1	1162.23	2.78	0.59	3.14
Ecoli	ION	NIST	1	1109.23	2.49	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	103735.23	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	0.36	3.14
Lmono	ION	NIST	1	1271.23	2.56	0.32	2.92
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	4.66
Lmono	Sanger	NIST	1	242.23	3.41	2.18	4.10

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 postive variants were called due to contaminants. A low level of containating 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. Stand 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 as 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).

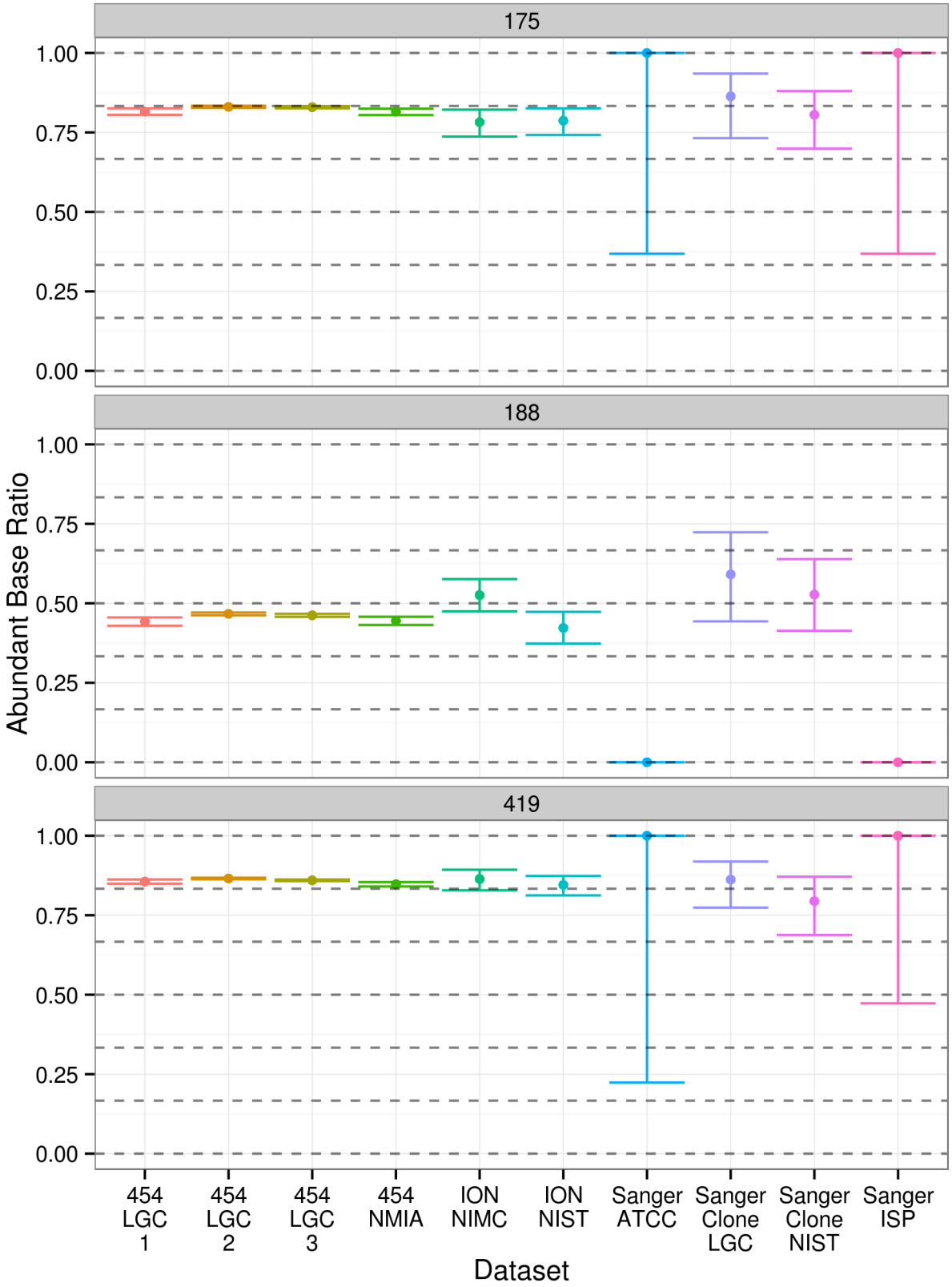

Figure S1: Base ratios at three biologically variable positions (175, 188 and 419) in *L. monocytogenes*. Variable positions shown in grey box above each graph. Error bars represent the 95 % posterior credibility interval estimated from a beta binomial distribution where is the major base count + 1 and is the minor base count + 1. One sided credibile intervals were calcualted 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).

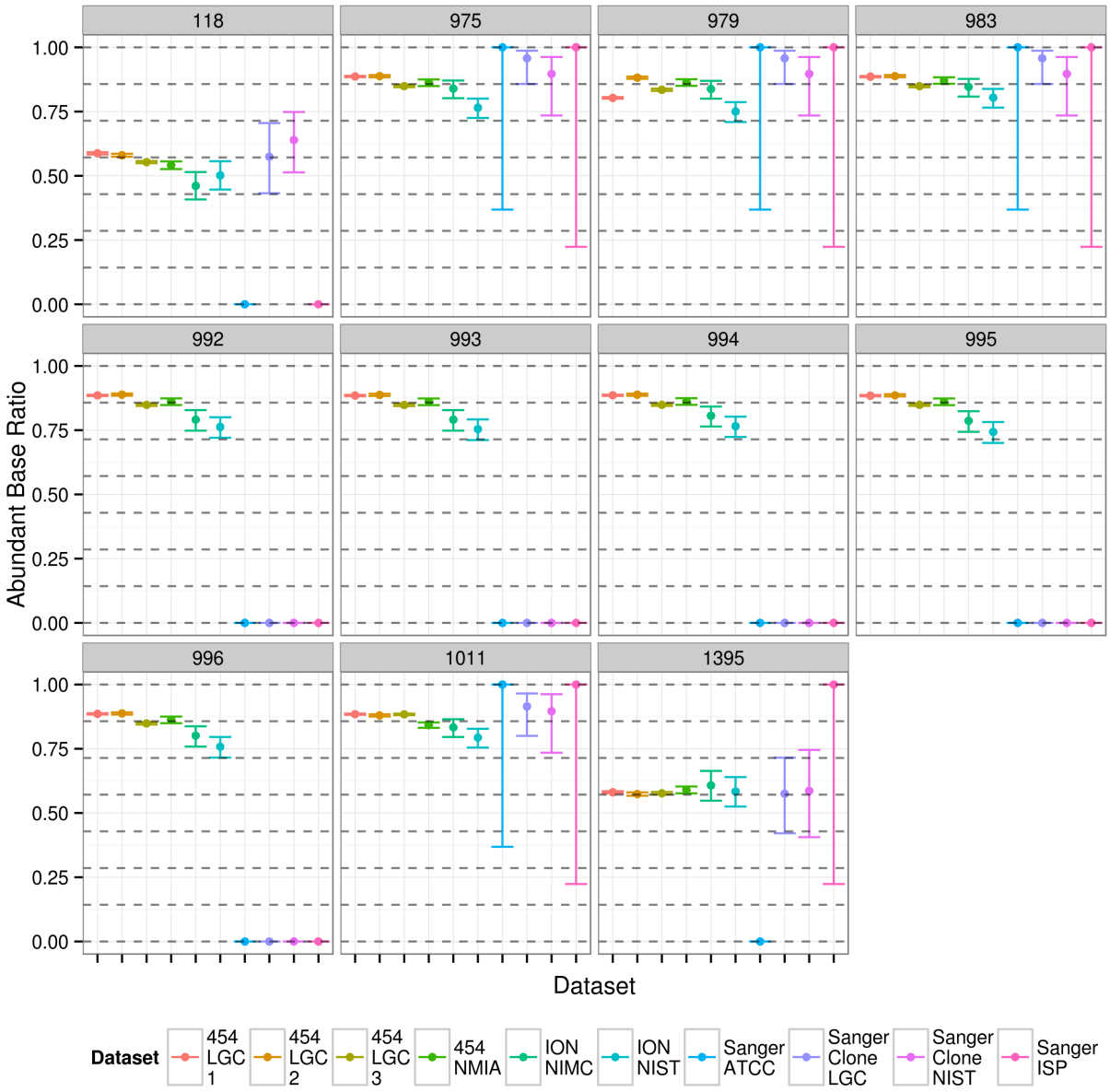

Figure S2: Base ratios at eleven biologically variable positions in *E. coli*. Variable positions shown in grey box above each graph. Error bars represent the 95 % posterior credibility interval estimated from a beta binomial distribution where is the major base count + 1 and is the minor base count + 1. One sided credibile intervals were calcualted 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	346.17	3	3	1
Ecoli-454-LGC-2	0.03	272.21	3	3	1
Ecoli-454-LGC-3	0.04	225.29	3	3	1
Ecoli-454-NMIA-1	0.06	32.83	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	0.03	787.63	3	3	1

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

dataset	likelihood	chimera	GCG	GTA	GTG	TCG
Lmono-454-LGC-1	0.00	47.77	2	1	2	1
Lmono-454-LGC-2	0.01	541.90	2	1	2	1
Lmono-454-LGC-3	0.01	309.23	2	1	2	1
Lmono-454-NMIA-1	0.00	56.36	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	826.40	2	1	2	1

4 Appendix

Full List of False Positive Variants

All variants called by the 8 piplines 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 my 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	281.77	60.00	40.22	End of read
Ecoli	454	LGC	1	TMAP	gatk	324	250	647.77	88.36	67.97	End of read
Ecoli	454	LGC	1	bwa	gatk	325	250	318.77	60.00	57.70	End of read
Ecoli	454	LGC	1	TMAP	gatk	325	250	265.77	88.36	46.75	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	1482.77	60.00	9.28	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	632.77	59.95	15.31	End of read
Ecoli	454	LGC	3	bwa	gatk	324	250	484.77	59.83	51.04	End of read
Ecoli	454	LGC	3	TMAP	gatk	324	250	437.77	89.03	45.86	End of read
Ecoli	454	LGC	3	bwa	gatk	325	250	649.77	59.83	61.88	End of read
Ecoli	454	LGC	3	TMAP	gatk	325	250	341.77	89.03	42.14	End of read
Ecoli	454	LGC	3	bwa	gatk	348	250	803.77	59.92	11.45	End of read
Ecoli	454	LGC	3	bwa	sam	417	1032	33.00	60.00		Homopolymer
Ecoli	454	LGC	3	TMAP	sam	417	1020	62.00	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	5243.77	79.46	438.22	Strand bias
Ecoli	454	NMIA	1	TMAP	gatk	508	250	1207.77	83.99	0.00	End of read
Ecoli	454	NMIA	1	TMAP	gatk	509	250	1252.77	83.99	0.00	End of read
Ecoli	454	NMIA	1	TMAP	gatk	510	250	1318.77	83.99	0.00	End of read
Ecoli	454	NMIA	1	TMAP	gatk	514	250	1246.77	83.99	0.00	End of read
Ecoli	454	NMIA	1	TMAP	sam	514	6337	42.00	59.00		End of read
Ecoli	454	NMIA	1	TMAP	gatk	901	208	8071.77	84.39	0.00	Non-target region
Ecoli	454	NMIA	1	TMAP	gatk	904	208	8053.77	84.39	0.00	Non-target region
Ecoli	454	NMIA	1	TMAP	gatk	934	250	8172.77	69.19	0.00	Non-target region
Ecoli	454	NMIA	1	TMAP	gatk	935	250	8172.77	69.19	0.00	Non-target region
Ecoli	454	NMIA	1	TMAP	gatk	938	250	8038.77	69.39	0.00	Non-target region
Ecoli	454	NMIA	1	TMAP	sam	938	2746	9.54	56.00		Non-target region
Ecoli	454	NMIA	1	TMAP	gatk	939	250	7913.77	69.39	0.00	Non-target region

Ecoli	454	NMIA	1	TMAP	sam	939	2746	15.20	60.00	Non-target region
Ecoli	454	NMIA	1	TMAP	sam	941	2746	12.30	58.00	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	4958.77	45.76	80.63	Strand bias
Lmono	454	LGC	1	bwa	gatk	328	250	5020.77	45.71	83.42	Strand bias
Lmono	454	LGC	1	TMAP	gatk	334	250	4722.77	70.27	298.10	Strand bias
Lmono	454	LGC	1	bwa	gatk	366	249	44.77	57.87	23.14	End of read
Lmono	454	LGC	1	bwa	gatk	508	250	1408.77	51.12	6.96	Contaminants
Lmono	454	LGC	1	bwa	sam	508	7744	6.98	55.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	533	166	1429.77	45.53	0.00	Contaminants
Lmono	454	LGC	1	bwa	sam	533	1763	147.00	38.00		Contaminants
Lmono	454	LGC	1	TMAP	gatk	533	250	2057.77	50.24	2.90	Contaminants
Lmono	454	LGC	1	TMAP	gatk	536	250	2097.77	50.24	2.53	Contaminants
Lmono	454	LGC	1	bwa	gatk	537	166	1502.77	45.53	5.46	Contaminants
Lmono	454	LGC	1	bwa	sam	537	1763	76.00	38.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	538	166	1842.77	45.53	1.89	Contaminants
Lmono	454	LGC	1	bwa	sam	538	1624	187.00	38.00		Contaminants
Lmono	454	LGC	1	TMAP	gatk	539	250	2100.77	50.24	2.54	Contaminants
Lmono	454	LGC	1	bwa	gatk	548	166	2202.77	45.53	0.00	Contaminants
Lmono	454	LGC	1	bwa	sam	548	1762	201.00	38.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	549	166	2187.77	45.53	0.00	Contaminants
Lmono	454	LGC	1	bwa	sam	549	1763	177.00	38.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	550	166	2189.77	45.53	0.00	Contaminants
Lmono	454	LGC	1	bwa	sam	550	1763	182.00	38.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	555	167	2212.77	45.63	0.00	Contaminants
Lmono	454	LGC	1	bwa	sam	555	1764	222.00	37.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	559	167	2110.77	45.63	0.00	Contaminants
Lmono	454	LGC	1	bwa	sam	559	1763	222.00	38.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	574	168	2337.77	45.53	0.00	Contaminants
Lmono	454	LGC	1	bwa	sam	574	1765	211.00	38.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	585	167	1711.77	45.61	0.00	Contaminants
Lmono	454	LGC	1	bwa	sam	585	1737	209.00	38.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	587	167	1872.77	45.61	0.00	Contaminants
Lmono	454	LGC	1	bwa	sam	587	1742	194.00	37.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	595	167	2305.77	45.61	0.00	Contaminants
Lmono	454	LGC	1	bwa	sam	595	1742	199.00	38.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	677	250	5060.77	58.63	0.76	Contaminants
Lmono	454	LGC	1	bwa	sam	677	4544	222.00	58.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	700	249	5431.77	56.74	10.29	Contaminants
Lmono	454	LGC	1	bwa	sam	700	4623	222.00	58.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	703	249	4798.77	56.74	0.71	Contaminants
Lmono	454	LGC	1	bwa	sam	703	4623	222.00	58.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	712	249	4843.77	56.74	7.74	Contaminants
Lmono	454	LGC	1	bwa	sam	712	4621	222.00	58.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	716	249	4065.77	56.74	0.00	Contaminants

Lmono	454	LGC	1	bwa	sam	716	4621	222.00	59.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	729	228	3820.77	58.71	0.00	Contaminants
Lmono	454	LGC	1	bwa	sam	729	4529	222.00	59.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	731	228	2738.77	58.71	0.00	Contaminants
Lmono	454	LGC	1	bwa	sam	731	4506	141.00	59.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	733	228	2828.77	58.71	0.00	Contaminants
Lmono	454	LGC	1	bwa	sam	733	4530	222.00	60.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	738	228	4372.77	58.71	0.76	Contaminants
Lmono	454	LGC	1	bwa	sam	738	4512	222.00	59.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	740	228	4603.77	58.71	0.00	Contaminants
Lmono	454	LGC	1	bwa	sam	740	4518	222.00	59.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	741	229	4531.77	58.69	0.00	Contaminants
Lmono	454	LGC	1	bwa	sam	741	4522	222.00	59.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	742	229	4660.77	58.69	0.00	Contaminants
Lmono	454	LGC	1	bwa	sam	742	4528	222.00	59.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	743	229	4733.77	58.69	0.00	Contaminants
Lmono	454	LGC	1	bwa	sam	743	4528	222.00	59.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	753	250	3642.77	58.17	6.29	Contaminants
Lmono	454	LGC	1	bwa	sam	753	4636	222.00	59.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	757	250	4227.77	58.17	44.27	Contaminants
Lmono	454	LGC	1	bwa	sam	757	4636	222.00	59.00		Contaminants
Lmono	454	LGC	1	TMAP	gatk	924	190	2130.77	42.92	0.00	Contaminants
Lmono	454	LGC	1	TMAP	gatk	926	190	2181.77	42.92	0.00	Contaminants
Lmono	454	LGC	1	TMAP	gatk	928	190	2183.77	42.92	0.00	Contaminants
Lmono	454	LGC	1	TMAP	gatk	930	189	2183.77	43.02	0.00	Contaminants
Lmono	454	LGC	1	TMAP	gatk	953	250	4314.77	55.28	0.00	Contaminants
Lmono	454	LGC	1	TMAP	gatk	955	250	4314.77	55.28	0.00	Contaminants
Lmono	454	LGC	1	TMAP	gatk	957	250	4169.77	56.03	0.00	Contaminants
Lmono	454	LGC	1	TMAP	gatk	958	250	4147.77	56.03	0.00	Contaminants
Lmono	454	LGC	1	TMAP	gatk	959	250	4187.77	56.03	0.00	Contaminants
Lmono	454	LGC	1	TMAP	gatk	961	250	4354.77	56.03	0.00	Contaminants
Lmono	454	LGC	1	TMAP	gatk	963	250	4224.77	56.03	0.00	Contaminants
Lmono	454	LGC	1	bwa	gatk	982	250	1902.77	60.00	683.65	Contaminants
Lmono	454	LGC	1	bwa	gatk	1047	250	3661.77	59.63	45.12	Contaminants
Lmono	454	LGC	1	bwa	sam	1047	8006	225.00	60.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	1055	250	3124.77	59.13	2.86	Contaminants
Lmono	454	LGC	1	bwa	sam	1055	8011	225.00	60.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	1072	250	3266.77	58.75	8.12	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	5389.77	58.40	61.75	Contaminants
Lmono	454	LGC	1	bwa	sam	1192	8008	225.00	60.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	1201	250	4675.77	58.40	68.21	Contaminants
Lmono	454	LGC	1	bwa	sam	1201	8006	225.00	60.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	1208	250	3351.77	58.40	15.18	Contaminants
Lmono	454	LGC	1	bwa	sam	1208	8009	154.00	60.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	1213	250	4232.77	58.40	71.00	Contaminants
Lmono	454	LGC	1	bwa	sam	1213	8010	213.00	60.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	1304	249	4726.77	59.76	82.01	Contaminants
Lmono	454	LGC	1	bwa	sam	1304	7998	225.00	60.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	1307	250	5253.77	59.76	92.48	Contaminants
Lmono	454	LGC	1	bwa	sam	1307	7999	225.00	60.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	1318	250	4842.77	59.77	101.08	Contaminants
Lmono	454	LGC	1	bwa	sam	1318	8002	174.00	60.00		Contaminants

Lmono	454	LGC	1	bwa	gatk	1321	250	4577.77	59.77	86.50	Contaminants
Lmono	454	LGC	1	bwa	sam	1321	8002	225.00	60.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	1329	250	4754.77	59.63	100.20	Contaminants
Lmono	454	LGC	1	bwa	sam	1329	8005	225.00	60.00		Contaminants
Lmono	454	LGC	1	bwa	gatk	1356	249	3578.77	59.62	42.87	Contaminants
Lmono	454	LGC	1	bwa	sam	1356	8010	205.00	60.00		Contaminants
Lmono	454	LGC	2	bwa	gatk	315	250	3340.77	51.69	0.00	Contaminants
Lmono	454	LGC	2	bwa	gatk	328	250	3354.77	51.69	0.00	Contaminants
Lmono	454	LGC	2	bwa	gatk	346	250	98.77	59.51	34.89	End of read
Lmono	454	LGC	2	TMAP	gatk	346	250	165.77	91.15	40.63	End of read
Lmono	454	LGC	2	bwa	gatk	347	250	337.77	59.51	50.87	End of read
Lmono	454	LGC	2	TMAP	gatk	347	250	441.77	91.15	54.02	End of read
Lmono	454	LGC	2	bwa	gatk	548	141	47.77	57.80	2.14	Contaminants
Lmono	454	LGC	2	bwa	gatk	549	141	47.77	57.80	2.14	Contaminants
Lmono	454	LGC	2	bwa	gatk	550	141	48.77	57.80	2.15	Contaminants
Lmono	454	LGC	2	bwa	gatk	555	141	102.77	57.80	2.15	Contaminants
Lmono	454	LGC	2	bwa	gatk	559	141	65.77	57.80	2.15	Contaminants
Lmono	454	LGC	2	bwa	gatk	574	141	45.77	57.80	2.15	Contaminants
Lmono	454	LGC	2	bwa	gatk	587	141	89.77	57.80	2.15	Contaminants
Lmono	454	LGC	2	bwa	gatk	595	141	51.77	57.80	2.14	Contaminants
Lmono	454	LGC	2	bwa	gatk	677	141	131.77	59.92	2.15	Contaminants
Lmono	454	LGC	2	bwa	gatk	700	141	131.77	59.92	2.15	Contaminants
Lmono	454	LGC	2	bwa	gatk	703	141	128.77	59.92	2.15	Contaminants
Lmono	454	LGC	2	bwa	gatk	712	141	117.77	59.92	2.15	Contaminants
Lmono	454	LGC	2	bwa	gatk	716	141	106.77	59.92	2.15	Contaminants
Lmono	454	LGC	2	bwa	gatk	729	141	130.77	59.92	2.15	Contaminants
Lmono	454	LGC	2	bwa	gatk	738	141	131.77	59.92	2.15	Contaminants
Lmono	454	LGC	2	bwa	gatk	740	141	131.77	59.92	2.15	Contaminants
Lmono	454	LGC	2	bwa	gatk	741	141	131.77	59.92	2.15	Contaminants
Lmono	454	LGC	2	bwa	gatk	742	141	131.77	59.92	2.15	Contaminants
Lmono	454	LGC	2	bwa	gatk	743	141	131.77	59.92	2.15	Contaminants
Lmono	454	LGC	2	bwa	gatk	753	141	126.77	59.92	2.15	Contaminants
Lmono	454	LGC	2	bwa	gatk	757	141	129.77	59.92	2.15	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	3706.77	59.94	10.55	Contaminants
Lmono	454	LGC	2	bwa	gatk	1055	250	3448.77	59.94	13.47	Contaminants
Lmono	454	LGC	2	bwa	gatk	1072	250	2476.77	59.94	28.58	Contaminants
Lmono	454	LGC	2	bwa	gatk	1077	249	2035.77	59.94	27.79	Contaminants
Lmono	454	LGC	2	bwa	gatk	1192	250	4787.77	59.63	11.68	Contaminants
Lmono	454	LGC	2	bwa	gatk	1201	250	4703.77	59.63	6.44	Contaminants
Lmono	454	LGC	2	bwa	gatk	1208	250	4676.77	59.63	9.57	Contaminants
Lmono	454	LGC	2	bwa	gatk	1213	250	4659.77	59.63	9.53	Contaminants
Lmono	454	LGC	2	bwa	gatk	1304	250	4652.77	60.00	9.64	Contaminants
Lmono	454	LGC	2	bwa	gatk	1307	250	4789.77	60.00	11.73	Contaminants
Lmono	454	LGC	2	bwa	gatk	1318	250	4710.77	60.00	11.93	Contaminants
Lmono	454	LGC	2	bwa	gatk	1321	250	4649.77	60.00	11.93	Contaminants
Lmono	454	LGC	2	bwa	gatk	1329	250	4578.77	60.00	12.11	Contaminants
Lmono	454	LGC	2	bwa	gatk	1356	250	4462.77	60.00	9.87	Contaminants
Lmono	454	LGC	3	TMAP	gatk	346	250	341.77	91.27	42.30	End of read
Lmono	454	LGC	3	bwa	gatk	347	250	646.77	59.90	61.75	End of read
Lmono	454	LGC	3	TMAP	gatk	347	250	388.77	91.27	42.02	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	5960.77	79.34	470.81	Strand bias
Lmono	454	NMIA	1	TMAP	gatk	334	250	5893.77	79.34	511.60	Strand bias
Lmono	454	NMIA	1	TMAP	gatk	335	250	5183.77	79.34	498.23	Strand bias
Lmono	454	NMIA	1	bwa	gatk	381	250	295.77	60.00	24.26	End of read
Lmono	454	NMIA	1	TMAP	gatk	533	249	1724.77	67.93	0.00	End of read
Lmono	454	NMIA	1	TMAP	gatk	534	249	1679.77	67.93	0.00	End of read
Lmono	454	NMIA	1	TMAP	gatk	535	249	1710.77	67.93	0.00	End of read
Lmono	454	NMIA	1	TMAP	gatk	932	185	5634.77	69.68	0.00	Non-target region
Lmono	454	NMIA	1	TMAP	gatk	936	183	5636.77	69.86	0.00	Non-target region
Lmono	454	NMIA	1	TMAP	gatk	954	250	8532.77	78.14	0.00	Non-target region
Lmono	454	NMIA	1	TMAP	gatk	957	250	8490.77	78.14	0.00	Non-target region
Lmono	454	NMIA	1	TMAP	gatk	961	250	8545.77	78.17	0.00	Non-target region
Lmono	454	NMIA	1	TMAP	gatk	962	250	8456.77	78.17	0.00	Non-target region
Lmono	454	NMIA	1	TMAP	gatk	963	250	8435.77	78.17	0.00	Non-target region
Lmono	ION	NIST	1	bwa	gatk	792	264	302.77	60.00	67.67	Strand bias
Lmono	ION	NIST	1	TMAP	gatk	792	269	256.77	85.85	46.85	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	24.50	57.00		End of reference
Lmono	Sanger	LGC	1	TMAP	sam	1506	41	46.50	57.00		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	96.44	0.00	End of read
Lmono	Sanger	NIST	1	TMAP	sam	867	64	12.30	59.00		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  CCTGATGCAGCCATGCCGCGTGTATGAAGAAGGCCTTCGGGTTGTAAAGTAC-TTTCAGC  224213

Query  62      GGGGAGGAAGGGAGTAAAGTTAATACCTTTGCTCATTGACGTTACCCGCAGAAGAAGCAC  121
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  224214  GGGGAGGAAGGGAGTAAAGTTAATACCTTTGCTCATTGACGTTACCCGCAGAAGAAGCAC  224273

Query  122     CGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCGTTAATCGGAATTA  181
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  224274  CGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCGTTAATCGGAATTA  224333

Query  182     CTGGGCGTAAAGCGCACGCAGGCGGTTTGTTAAGTCAGATGTGAAATCCCCGGGCTCAAC  241
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  224334  CTGGGCGTAAAGCGCACGCAGGCGGTTTGTTAAGTCAGATGTGAAATCCCCGGGCTCAAC  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     TCCACGCCGTAAACGATGTCGACTTGGAGGTTGTGCCCTTGAGGCGTGGCTTCCGGAGCT  481
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  224574  TCCACGCCGTAAACGATGTCGACTTGGAGGTTGTGCCCTTGAGGCGTGGCTTCCGGAGCT  224633

Query  482     AACGCGTTAAGTCGACCGCCTGGGGAGTACGGCCGCAAGGTTAAAACTCAAATGAATTGA  541
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  224634  AACGCGTTAAGTCGACCGCCTGGGGAGTACGGCCGCAAGGTTAAAACTCAAATGAATTGA  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)
```