J. Dairy Sci. 107:1334–1354
https://doi.org/10.3168/jds.2023-23940


# Gram-negative postpasteurization contamination patterns of single-serve fluid milk produced in 4 different processing facilities

**T. T. Lott, A. N. Stelick, M. Wiedmann, and N. H. Martin***
Department of Food Science, Cornell University, Ithaca, NY 14853

## ABSTRACT

An analysis of historic data on high temperature, short time (HTST) fluid milk quality showed higher total bacterial counts and lower sensory defect judging scores at d 14 postprocessing for milk packaged in single-serve containers as compared with milk packaged in half-gallon containers from the same processing facilities. As postpasteurization contamination with gram-negative bacteria is likely a major contributor to an increased spoilage risk associated with milk packaged in single-serve containers, we performed a comprehensive assessment of the microbial quality and shelf life of 265 commingled single-serve HTST fluid milk samples (including white [unflavored] skim, white [unflavored] 1%, chocolate skim, and chocolate 1%) collected over 2 visits to 4 commercial fluid milk processing facilities. Over 2 initial sampling visits, the frequency of gram-negative spoilage ranged from 14 to 79% of the product collected from the 4 facilities, with significant differences of gram-negative spoilage frequency between sampling visits, facilities (sampling visit 1, sampling visit 2, and both sampling visits combined), milk types (sampling visit 2), and filler lanes (sampling visit 2). We found no significant differences in the frequency of gram-negative spoilage between sampling time points (e.g., beginning, middle, and end of production run). Across facilities, single-serve containers of milk with gram-negative contamination showed significantly higher bacterial counts on d 7 and 14 and significantly lower sensory scores as compared with those without gram-negative contamination. Follow-up investigations, based on in-facility surveys that identified carton forming mandrels as filler components that frequently failed quality assurance ATP swab checks, found that bacterial genera, including *Pseudomonas* and *Bacillus*, isolated from single-serve milk samples were also frequently isolated from mandrels. Although interventions aimed at improving cleaning and sanitation of mandrels did not lead to significant reduction of gram-negative spoilage frequency in a comparison of 398 control and 400 intervention samples, our data still suggest that the unhygienic design of single-serve fillers is likely a root cause of gram-negative contamination of single-serve milk.

**Key words:** single-serve milk, fluid milk, milk carton, postpasteurization contamination, gram-negative contamination

## INTRODUCTION

Over a century ago, postpasteurization contamination (**PPC**) was identified as an issue for the microbial quality of fluid milk (Smith, 1920). Although it has been demonstrated that facilities can reduce PPC frequency, PPC has been reported to occur in roughly 50% of fluid milk samples in the United States (Martin et al., 2012, 2018). Although gram-negative bacteria are highly susceptible to destruction by pasteurization (Champagne et al., 1994; Villamiel and de Jong, 2000), they represent the most frequently identified group of postpasteurization microbial contaminants in fluid milk, including *Pseudomonas*, which frequently causes spoilage due to its ability to grow at low temperatures (Juffs, 1973; Schröder et al., 1982; Ternström et al., 1993; Eneroth et al., 1998; Deeth et al., 2002; Stevenson et al., 2003; Schroeder et al., 2008; Ranieri and Boor, 2009; Martin et al., 2011a). In the United States, shelf life is typically defined by the Food and Drug Administration (**FDA**) Pasteurized Milk Ordinance (**PMO**) standard plate count (**SPC**) threshold of 20,000 cfu/mL for grade "A" milk (FDA, 2019), and it has been found that samples contaminated with *Pseudomonas* can reach this concentration in HTST pasteurized milk in as little as 8 d postprocessing (Ranieri and Boor, 2009). Other than the ability of gram-negatives to grow to spoilage levels at low temperatures, they can also cause sensory defects (e.g., fruity, rotten, cheesy) in fluid milk by the production of extracellular enzymes and via bacterial proteolysis (Hayes et al., 2002; Alvarez, 2009).

Although there is a wide breadth of research on PPC and microbial shelf life of fluid milk, there has been limited specific focus on PPC of HTST single-serve fluid

Received July 8, 2023.
Accepted September 8, 2023.
*Corresponding author: nhw6@cornell.edu

milk. The quality of single-serve milk is of particular interest because milk packaged in single-serve half-pint paper cartons is the most common type of milk purchased for school lunch programs due to the low cost of paperboard (Sipple et al., 2021). Further, it has been reported that children have negative associations of milk brands or packages (e.g., single-serve cartons) if they previously had a negative sensory experience with that product (Sipple et al., 2021) and that milk consumption during adolescence is positively correlated with milk consumption during adulthood (McCarthy et al., 2017). Thus, microbial-induced sensory defects may lead to decreased consumption of milk among school children and may have lasting effects into adulthood.

Historical data from an ongoing, long-term fluid milk monitoring program (the Voluntary Shelf-Life [**VSL**] program) revealed that for HTST fluid milk packaged in half-pint (473 mL) cartons, SPC were significantly higher and sensory scores (based on defect judging) were significantly lower when compared with HTST fluid milk packaged in half-gallon (1.9 L) containers from the same facilities, suggesting that milk packaged in single-serve cartons has lower quality than milk packaged in half-gallon containers. As gram-negative bacteria have been found to be the driver of higher SPC and lower sensory scores (Alles et al., 2018; Reichler et al., 2018), the goals of this study were to (1) assess gram-negative contamination frequency in single-serve milk, (2) determine potential root causes of increased gram-negative contamination frequency in single-serve milk, (3) assess the bacterial genera isolated from single-serve milk and single-serve milk filler equipment, and (4) develop an intervention aimed at reducing gram-negative contamination in single-serve milk. Although some studies have analyzed PPC in milk packaged in single-serve cartons (Schröder, 1984; Gruetmacher and Bradley, 1999) and performed interventions aimed at reducing PPC on one single-serve fluid milk filler (Reichler et al., 2020), previous studies did not collect the large number of samples that we did in this study, which was important not only for understanding gram-negative contamination frequency trends between facilities and production time points, but also allowed us to determine if interventions significantly reduced gram-negative spoilage frequency.

## MATERIALS AND METHODS

### *Initial Comparative Assessment of PPC in Single-Serve Milk*

Commercial single-serve half-pint HTST fluid milk and half-gallon HTST fluid milk SPC data and sensory defect judging data collected between 2013 and 2017 through the Cornell VSL program (Martin et al., 2012) were assessed to identify differences in microbial or sensory quality on d 14 of shelf life. Only facilities that produced both single-serve and half-gallon milk were included in the comparison. Total SPC were determined by spiral plating 50 μL of sample in duplicate onto standard methods agar (**SMA**), and dilutions were performed, if necessary, by pipetting 1 mL of sample into 99 mL of sterile buffer water as previously described (Duncan et al., 2004). Plates were incubated at 32°C for 48 h before enumeration using an automated colony counter (IUL, S.A.). The limit of detection for SPC was 10 (1.00 $\log_{10}$) cfu/mL, with samples below the detection limit being assigned a value of 25% of the detection limit (2.5 [0.40 $\log_{10}$] cfu/mL). Any SPC that were too numerous to count were assigned values of the upper countable limit for spiral plates (400,000 cfu/mL) multiplied by the dilution factor. Sensory defect judging was performed according to Martin et al. (2012). Briefly, using randomized 3-digit codes, panelists were blinded to sample information, and were asked to assign defects (e.g., light oxidized, acid) and rate overall sensory score on a continuous scale of 0.0 to 10.0. Generally, severe defects (e.g., rancid, unclean) result in lower overall sensory score and less severe defects (e.g., cooked, flat) or no defects result in higher overall sensory scores. A sensory defect judging score of 9.0 to 10.0 is considered excellent quality milk, 8.0 to 8.9 for good quality, 6.1 to 7.9 for fair quality, and scores 6.0 and below were considered unacceptable as previously described (Martin et al., 2012).

### *In-Depth Evaluation of HTST Single-Serve Milk Produced in 4 Processing Facilities*

For further in-depth evaluation of HTST single-serve fluid milk quality, we recruited 4 processing facilities that each produce a substantial amount of fluid milk in single-serve cartons (as supported by the fact that HTST single-serve fluid milk from these facilities is distributed to multiple school districts). Each of these facilities fills single-serve milk cartons using an N-8 filler (Pactiv Evergreen, Lake Forest, IL). Each of these 4 fluid milk processing facilities was visited on 2 separate occasions (hereby referred to as "sampling visits") between November 2019 and April 2021, for a total of 8 sampling visits (Table 1). During these sampling visits, observations and surveys were performed and samples were collected over a single day of production, which was estimated to be an average of 14 h with a range of 9 to 22 h. As it was usually not feasible to observe the entire production run, therefore, any production times reported here are estimates. At the first visit to each facility, 6 single-serve (1/2-pint [473-mL]) cartons of each of 4 milk types (white skim, white 1%, choco-

**Table 1.** Sampling scheme and sample collection timing for the 4 enrolled facilities that produce single-serve milk

| | Month and year | | | |
|---|---|---|---|---|
| Facility code | Sampling visit 1 (estimated length of production) | Sampling visit 2 (estimated length of production) | Preintervention follow-up visit | Intervention implementation visit |
| 1 | November 2019 (13 h) | November 2020 (17 h) | February 2022 | April 2022 |
| 2 | January 2020 (9 h) | January 2021 (13 h) | December 2021 | May 2022 |
| 3 | October 2020 (22 h) | February 2021 (12 h) | January 2022 | April 2022 |
| 4 | March 2021 (14 h) | April 2021 (12 h) | December 2021 | June 2022 |

late skim, and 1% chocolate milks) were collected at the beginning, middle, and end of production (3 time points; hereby referred to as "sampling time points") for each of 4 lanes of the single-serve filler; for a total of up to 288 cartons per facility (6 samples × 4 milk types × 3 sampling time points × 4 lanes; see Supplemental Figure S1a; https://github.com/FSL-MQIP/single_serve_milk). The (1) beginning sampling time point was defined as taking samples within the first 1,000 cartons of a product run (e.g., white skim milk), (2) the middle sampling time was defined as the approximate mid-point of a product run based on each facility's estimate for the number of cartons that were to be produced (e.g., if the estimated number of cartons produced was 50,000, middle samples were taken immediately after approximately 25,000 cartons were produced), and (3) the end sampling time was defined as within the last 1,000 cartons of a product run. For sampling visit 1, samples representing a given milk type (e.g., white skim) obtained at a given sampling time point (e.g., beginning of sampling) were comingled to create a half-gallon (1.9-L) sample that included 2 cartons from each of the 4 lanes. As 6 samples of each milk type and sampling time point were collected per lane, 3 commingled half-gallon milk samples were collected per milk type and sampling time point. With 4 milk types, 3 sampling time points, and collection of samples in triplicate, this generated up to 36 commingled samples per facility. Commingling of samples from sampling visit 1 (and 2) was necessary to have enough sample for both microbial analyses and sensory defect judging.

For sampling visit 2, samples were collected using a different scheme that would allow for comparisons of gram-negative spoilage patterns across the 4 lanes of a given filler. More specifically, 16 single-serve cartons of each milk type (i.e., white skim, white 1%, chocolate skim, and 1% chocolate) were collected from each of the 4 lanes at each of 2 sampling time points (beginning and end of production), for a total of up to 512 cartons per facility (16 samples × 4 milk types × 4 lanes × 2 sampling times), as shown in Supplemental Figure S1b (https://github.com/FSL-MQIP/single_serve_milk). The beginning and end sample time points were defined the same as for sampling visit 1 (i.e., first 1,000 and last 1,000 cartons). Comingling for sampling visit 2 samples was performed such that 16 samples for each milk type collected from a given lane and time point were commingled, yielding a single 1-gallon (3.8-L) sample representing a given milk type from a given lane obtained at a given sample time point. With 4 milk types, 4 lanes, and 2 time points, up to 32 samples were generated per facility.

For both sampling visits 1 and 2, each commingled sample was used first to create (1) one 60-mL sample for d 0 microbial analysis and (2) sensory samples. The remaining volume was then used to pour 600-mL aliquots into two 1-L Pyrex (Corning, NY) glass, screw-top bottles for d 7 and d 14 sensory defect judging and microbial analyses, with a separate bottle for each day of testing. For microbial analysis, SPC were performed by spiral plating 50 μL of sample onto SMA; total gram-negative counts were performed by spiral plating 50 μL on crystal violet tetrazolium agar (**CVTA**) plates. The SPC and CVTA plates were incubated for 48 h at 32°C and 21°C, respectively. Following incubation, enumeration was performed using an automated colony counter (IUL, S.A.). Only typical (i.e., red) colonies were counted on CVTA plates. For both sampling visits 1 and 2, some samples were not collected if a particular lane of the filler was down (i.e., cartons were not filled in this lane) during production. As samples from multiple lanes were commingled into a single sample for sampling visit 1, this did not affect the overall number of commingled samples tested. However, for sampling visit 2, commingled samples represented a single lane; thus, some samples were not able to be tested. The overall number of commingled samples is reported in the Results section.

### Sensory Defect Judging

For samples from sampling visit 1, each panelist evaluated the first replicate (labeled "R1") of the 3 replicates from each combination of sampling time point (i.e., beginning, middle, or end) and milk type (i.e., white skim, white 1%, chocolate skim, or choco-

late 1%) for sensory defect judging of a total of up to 12 commingled samples. For sampling visit 2 samples, panelists assessed up to 32 samples from each facility as replicates were not collected for each combination of time point (i.e., beginning or end), milk type (i.e., skim, 1%, chocolate skim, or chocolate 1%), and lane number (i.e., 1, 2, 3, or 4); to avoid sensory fatigue, panelists were asked to first evaluate half of the samples from sampling visit 2 and then evaluate the second half of the samples after waiting at least 1 h.

Sensory defect judging was performed by a total of 14 panelists (64% female, 36% male) with 6 to 7 panelists performing sensory for d 0, 7, and 14 of shelf life per sampling (e.g., sampling visit 1 to facility 1). These panelists were selected through a prescreening process as previously described (Reichler et al., 2018). Briefly, panelists were trained on sensory defect judging and demonstrated the ability to correctly identify common defects (e.g., rancid, bitter) as previously described (Alvarez, 2009) in ≥70% of reference samples. Sensory defect judging was performed as previously described (Martin et al., 2012). For each sample, each panelist received a 200-mL plastic cup containing a 50-mL aliquot of milk, with panelists receiving samples from the same commingled replicate. Panelists were blinded to samples by using randomly generated 3-digit codes for each sample. Panelists briefly heated samples in a microwave to warm products to approximately 15°C, as heating the samples facilitates the detection of volatile compounds by the trained sensory panelists (Francis et al., 2004). Then, panelists assessed samples by using the survey provided to them, which included reporting perceived defects (e.g., rancid, unclean), the intensity of those defects (i.e., slight, definite, or pronounced), and an overall score (using a continuous 0.0–10.0 scale). The panelist defect judging sensory scores were used to calculate the mean overall flavor score for each sample. Panelists were told the fat level and flavor of each sample (e.g., "this is a white skim milk") to compare the samples to "gold standards" that panelists were exposed to during their training as described above. Sensory defect judging performed here has exempt status as granted by the Cornell Institutional Review Board for Human Participant Research.

### Surveys for Collection of Single-Serve Milk Filler Information

In addition to sample collection during the initial visits to each facility, surveys and observations were also conducted to collect more information on issues related to the processing of single-serve milk, facility quality practices, and cleaning and sanitation practices, with a focus on practices and issues that may be relevant to postprocessing gram-negative contamination. Surveys were conducted face-to-face and included questions for quality management staff (e.g., "what part of the single-serve filler fails ATP swab verification most often?"), single-serve filler operators (e.g., "what is the most frequent reason for downtime?"), and cleaning and sanitation staff (e.g., "what is the hardest part to clean on the N-8 filler?"). This survey also included an "observations" section for recording events during single-serve milk production (e.g., spraying of hoses). These observations were performed by the first author (T. T. Lott), and in the Observations section, violations of good manufacturing practices (**GMP**) and downtime events were also recorded. The surveys used are available at (https://github.com/FSL-MQIP/single_serve_milk).

### Preintervention Assessment of Mandrels and Root-Cause Analysis

Following the identification of mandrels as potential sources of gram-negative contamination, each of the 4 facilities were revisited for targeted sampling and root-cause analysis (**RCA**) of gram-negative contamination in HTST fluid milk packaged in single-serve cartons. To assess the potential of mandrels to facilitate cross contamination, we collected, following the end of production (but before cleaning and sanitation), empty cartons (i.e., formed and sealed cartons without milk, hereby referred to as "carton samples") as well as environmental sponge samples of the mandrels (hereby referred to as "mandrel sponge samples"). Milk samples were also collected at the end of production to allow for a comparison of bacterial genera found on mandrels, cartons, and in packaged milk samples. As there are 6 mandrels for each of the 4 lanes (24 total mandrels), 6 mandrel sponge samples were taken from each lane. In addition, 6 consecutive milk samples and 6 consecutive carton samples were collected from each of the 4 lanes, representing 24 total milk and 24 total carton samples; samples were collected consecutively to ensure that one sample was collected to represent each mandrel. As the 6 single-serve milk samples were collected before cleaning and sanitation at the end of all production (defined as the time when the last 1,000 cartons were processed for a given production day), only a single milk type (e.g., white skim, chocolate 1%) was collected at each facility depending on the production schedules. After collection of empty cartons, the outside of each carton was sanitized with 70% ethanol. Then, 50 mL of sterile brain heart infusion (**BHI**) broth was injected into the sealed, empty cartons by using a sterile syringe pump and needle. Following the addition of BHI, the small hole created in each package by the syringe needles was

taped over using duct tape to prevent leakage. Cartons were then inverted as previously described (Duncan et al., 2004), to ensure that BHI comes into contact with all internal carton surfaces. Mandrel sponge samples were collected using sterile sponges in 10 mL of Dey and Engley neutralizing broth (3M); separate sponges were used for each mandrel, and all mandrel parts that come in contact with internal carton surfaces were scrubbed with the sponge. Additional sponge samples were collected by swabbing the mandrel hubs (the area where the axle of mandrels connects to the drive shaft) when excessive grease build-up was observed (see Supplemental Figure S2; https://github.com/FSL-MQIP/single_serve_milk); 1 or 2 mandrel hub sponge samples were collected from facilities 1, 2, and 4, but no mandrel hub sponge samples were collected from facility 3, as the cleaning and sanitation schedule did not allow for us to access the mandrel hubs to take a sample.

All samples were kept on ice and transported to Cornell University within 4 h of sampling. For sponges, 90 mL of phosphate buffer solution was added to each sponge bag, followed by stomaching for 1 min. Subsequently, two 50-μL aliquots from each mandrel or mandrel hub sponge sample, milk sample, or BHI broth from carton samples were taken to perform one total SPC and one total gram-negative counts by spiral plating on SMA and CVTA plates, respectively. For each sample, approximately 50 mL were also transferred to a 60-mL vial, which was then incubated at 21°C for 24 h, which represents an enrichment approach, or "stress test," to facilitate detection of low levels of bacteria. Following incubation of the samples, two 50-μL aliquots of each sample were again taken for plating on one SMA and one CVTA plate. After incubation of SPC and CVTA plates for 48 h at 32°C and 21°C, respectively, plates were qualitatively assessed for growth by assigning plates "positive" if any growth was detected (including plates that were TNTC) and assigning plates "negative" if no growth was detected. If samples were positive on SPC and CVTA, we considered this to be evidence of gram-negative contamination.

For samples (i.e., milk, carton, mandrel sponge, or mandrel hub) with evidence of gram-negative contamination, up to 5 colonies with unique morphologies, representing the most frequently observed colony morphologies, were isolated per plate type (i.e., SPC, CVTA); these isolates were frozen at −80°C in 15% glycerol for further characterization.

### Design and Implementation of Interventions

We performed RCA using our initial observations, employee survey responses, and our initial findings from the preintervention assessment that gram-negative bacteria were frequently found in both milk and mandrel sponge samples. An RCA fishbone diagram specifically designed for this study (Supplemental Figure S3; https://github.com/FSL-MQIP/single_serve_milk) was used during the preintervention visits and was presented during a discussion with quality management staff from all 4 facilities. This RCA identified the carton forming mandrels, including excessive grease build-up, as representing a hygienic design issue and hence, a challenge for cleaning and sanitation, which may be a key factor in the observed gram-negative contamination in single-serve milk. Thus, we selected the mandrels as targets for a cleaning and sanitation intervention aimed at reducing gram-negative spoilage in single-serve milk. An RCA similar to this study, which uses a fishbone diagram, has been performed previously to identify *Listeria* in apple packinghouses (Belias et al., 2021).

The intervention used a randomized control trial design, which involved applying a standardized cleaning intervention at the N-8 filler level to all 6 mandrels per filler lane associated with 2 out of the 4 N-8 filler lanes at each of the 4 facilities. For each facility, lanes were numbered 1–4 and a set of 2 lanes was randomly selected to serve as an intervention group by running a script in RStudio that repeatedly selects 4 pairs of lanes until each lane is represented exactly 2 times (e.g., 1 and 2, 1 and 3, 2 and 4, 3 and 4). Due to scheduling and time constraints, the pairs were assigned conveniently to facilities by the order in which pairs were selected and the order in which facilities were visited. For example, if the first random pair of mandrels was lanes 1 and 3, the intervention was applied to lanes 1 and 3 of the N-8 filler at the first facility visited during the intervention part of the study.

For the intervention, a mandrel hand cleaning standard operating procedure (**SOP**) was developed based on expert elicitation from industry professionals at the Cornell Dairy facility and the Cornell Food Processing and Development Laboratory, building upon a generic hand cleaning SOP for dairy processing equipment. This SOP was modified and adopted to be specific to N-8 fillers, based on our observations of N-8 filler operations and the cleaning and sanitation procedures that existed in the 4 study facilities; the final mandrel hand cleaning SOP used for the interventions is available at (https://github.com/FSL-MQIP/single_serve_milk). Briefly, a solution of each facility's preferred cleaner (HC-10 Chlorinated Kleer-Mor, EcoLab for all 4 facilities) was prepared according to the manufacturer's instructions. After the solution was prepared, a 3M Scotch-Brite Cleansing Pad was soaked in the solution and used to clean each of the 12 mandrels (2 lanes, 6 mandrels each) assigned to the intervention, with the cleansing pad being re-soaked in the solution

between the cleaning of each mandrel. To maintain consistency, one of the co-authors (T. T. Lott) cleaned all mandrels assigned to the intervention at all 4 facilities. Additionally, a timer was used to ensure each mandrel was cleaned for 1 min. Mandrels not assigned to the intervention were cleaned by cleaning and sanitation staff according to pre-existing standard practices at each facility. The mandrel lanes cleaned by sanitation staff are hereby referred to as "controls."

To evaluate whether the mandrel cleaning intervention was effective, we collected single-serve milk samples at both the beginning and end of production (as defined previously), which allowed us to test the hypothesis that the intervention reduces contamination frequency throughout the full shift. The type of milk (e.g., white skim, 1% chocolate) collected at the beginning and end of production was dependent on each facility's production schedule. A sample size calculation was performed for Fisher's exact test to detect a 50% reduction of gram-negative contamination with an $\alpha$ of 0.05 and power of 0.80; the contamination proportion for the control was assumed to be 0.34 as the frequency of gram-negative spoilage was 34% of single-serve milk collected from all 4 facilities during our initial 2 sampling visits. As the initial sample size calculation indicated that 178 total samples were needed across 4 lanes, yielding 24 samples (per a single lane) after accounting for the design effect, we decided to collect 25 single-serve milk samples per lane at the beginning and end of production (50 samples per lane). This resulted in a total of 800 samples (4 facilities $\times$ 4 lanes $\times$ 2 sampling time points $\times$ 25 samples) collected. Of these 800 samples, a single sample was not plated on SPC and for another sample the SPC plate result was not recorded. Thus, these 2 sample SPC were recorded as laboratory errors and were not included in the analysis (which resulted in n = 798). The initial sample size was calculated using G*Power, and the design effect was calculated in RStudio. The G*Power calculation and code used in RStudio are available at (https://github.com/FSL-MQIP/single_serve_milk).

### 16S rRNA Gene Sequencing Analysis of Selected Bacterial Isolates

Isolates that were selected from the preintervention assessment and the intervention study, as detailed previously, were characterized by 16S rRNA gene PCR and subsequent sequencing, both performed as previously described (Huck et al., 2007). The partial consensus 16S rRNA gene sequences generated were compared against the Ribosomal Database Project (Center for Microbial Ecology, Michigan State University, East Lansing, MI) using the SeqMatch tool (Cole et al., 2014). An isolate was assigned to a genus if the top 10 matches returned by the search had >95% identity to the isolate and shared the same genus.

For isolates where the top 10 matches returned were not the same genus, the top 10 matches were reviewed, and family level was assigned if all 10 matches shared the same family. If family level could not be assigned, isolates were assigned to the same order if the top 10 matches shared the same order, or if order could not be assigned, the remaining isolates (n = 58) were assigned to the same class as all remaining isolates had top 10 matches that shared the same class. Although we recognize the limitations of assigning taxonomic classification above the genus level, a minority of isolates were not able to be assigned a genus and this methodology allowed us to provide some baseline identification for isolates. This identification can benefit future studies that aim to understand persistent and transient dairy spoilage organisms of concern (e.g., through whole-genome sequencing [**WGS**] of gram-negative bacteria).

### Data Analysis

Raw data were organized in Microsoft Excel (Microsoft Excel for Microsoft 365 MSO [Version 2303 Build 16227.20280] 64-bit; Microsoft Corp.) and data wrangling was performed in OpenRefine (version 3.4.1, https://openrefine.org/). Data manipulation, statistical analyses, and creation of plots were all performed using R (version 3.6.2, The R Foundation for Statistical Computing; R Core Team, 2019) in RStudio (version 2022.02.3+492, RStudio PBC; RStudio Team, 2022). All data and code are available at (https://github.com/FSL-MQIP/single_serve_milk).

The R "stats" package (version 3.6.2) was used for (1) Wilcoxon rank-sum tests for assessing differences in the VSL SPC and sensory defect judging data, (2) all logistic regression for assessing gram-negative contamination and spoilage frequencies, (3) ANOVA for assessing differences between sensory scores and SPC between samples with or without gram-negative spoilage. Version 1.6.3 of the package "emmeans" (Lenth, 2021) was used for summarizing logistic regression model results. Version 0.5.2 of the package "lsr" (Navarro, 2015) and version 0.7.0.5 of the package "effectsize" (Ben-Shachar et al., 2020) were used for determining partial eta squared ($\eta^2$) values as a measure of effect size, with $\leq$0.01, 0.06–0.13, and >0.14 indicating small, medium, and large effects, respectively (Cohen, 1988). To assess differences between bacterial populations, nonmetric multidimensional scaling (**NMDS**) and analysis of similarities (**ANOSIM**) were performed using version 2.5.7 of the "vegan" package (Oksanen et al., 2020), and multipattern analyses were performed using ver-

sion 1.7.12 of the "indicspecies" package(De Cáceres and Legendre, 2009). Data manipulation, including filtering and grouping of data, was performed using the "rstatix" (Kassambara, 2021) and "dplyr" (Wickham et al., 2021) packages. Figures were created with "ggplot2" (Wikcham, 2016).

### Definitions of Gram-Negative Spoilage and Gram-Negative Contamination

Whereas previous studies have used the term post-pasteurization contamination (PPC) to refer to post-processing contamination caused by gram-negative bacteria as well as gram-positive bacteria, here the term PPC (including "PPC frequency") solely refers to PPC due to gram-negative bacteria. Furthermore, the term "gram-negative spoilage" refers to samples that had (1) detectable growth (at least one typical [i.e., red] colony) on CVTA, a selective medium for gram-negative bacteria, and (2) SPC of >20,000 cfu/mL, based on the FDA PMO limit for grade "A" pasteurized milk products (FDA, 2019). It should be noted that although 20,000 cfu/mL is the standard regulatory limit, fluid milk defects are typically not detected by consumers until SPC surpass 1,000,000 cfu/mL (Carey et al., 2005). Gram-negative spoilage frequency is the percentage of total samples that met this criterion; this outcome was used for assessing PPC in samples collected in our initial 2 sampling visits and the intervention part of our study. Second, the term "gram-negative contamination" refers to samples that had (1) detectable growth (at least one typical [i.e., red) colony]) on CVTA, and (2) detectable growth (at least one colony) on SPC. This criterion was used for assessing initial visit SPC and sensory scores between samples with or without gram-negative contamination. Gram-negative contamination frequency is the percent of total samples that met this criterion; this was used for assessing PPC in milk, carton, and mandrel sponge samples collected in the preintervention assessment of our study.

## RESULTS

### Historical Data Show Reduced Microbial and Sensory Quality of Fluid Milk in Single-Serve Cartons as Compared with Half-Gallon Containers

A query of our database of microbial and sensory defect judging fluid milk quality data, collected between 2013 and 2017, for fluid milk processing facilities located in the Northeast United States (representing data routinely collected as part of the Cornell VSL program; Martin et al., 2012) that produced milk packaged in both half-gallon containers and half-pint cartons, yielded 12 fluid milk processing facilities. Data from these facilities represented a total of 326 and 88 samples of half-gallon and half-pint milk, respectively. For all facilities, we specifically retrieved microbial and sensory defect judging data for d 14 of shelf life. As data were found to not be normally distributed, we performed a Wilcoxon rank-sum test to identify differences between SPC and sensory scores of half-gallon and half-pint milk samples. Bacterial counts for d 14 (medians of 3.53 and 5.22 $\log_{10}$ cfu/mL for half-gallon and half-pint milk samples, respectively) were found to be significantly higher for half-pint milk compared with half-gallon milk ($P = 0.002$; Figure 1a). Sensory scores for d 14 (medians of 8.7 and 8.2 for half-gallon and half-pint milk samples, respectively) were found to be significantly lower for half-pint milk compared with half-gallon milk ($P < 0.001$; Figure 1b).

### Gram-Negative Bacterial Spoilage Is Frequent in Single-Serve Milk

Overall, for the 144 commingled samples obtained from sampling visit 1, mean SPC were 1.79, 2.37, and 5.05 $\log_{10}$ cfu/mL for d 0, 7, and 14, respectively (Table 2). Further, for a given day, a sample was considered spoiled by gram-negative bacteria if (1) typical growth was detected on CVTA (i.e., ≥1 colony); and (2) if the SPC was >20,000 cfu/mL (the FDA PMO limit). Based on these criteria, 0%, 11%, and 24% of all commingled samples obtained during sampling visit 1 showed evidence of spoilage by gram-negative bacteria on d 0, 7, and 14, respectively. Across all sampling visit 1 samples, mean sensory scores on d 0, 7, and 14 for the commingled samples were 8.7, 8.7, and 7.3, respectively; these mean sensory scores were based on evaluation of 1 out of 3 replicates for each combination of milk type and sampling time point (e.g., white skim milk, middle of processing run).

For the 121 commingled samples from sampling visit 2, mean SPC were 1.96, 3.83, and 6.21 $\log_{10}$ cfu/mL for d 0, 7, and 14, respectively. Based on the criteria detailed above, 0%, 33%, and 52% of commingled samples obtained during sampling visit 2 were spoiled by gram-negative bacteria on d 0, 7, and 14, respectively. Across all sampling visit 2 commingled samples, mean sensory scores were 9.0, 8.6, and 5.6 on d 0, 7, and 14, respectively.

For both sampling visits combined, mean SPC were 1.87, 3.03, 5.58 $\log_{10}$ cfu/mL, respectively. Overall, the frequencies of gram-negative spoilage of single-serve milk were 0%, 21%, and 37% for d 0, 7, and 14, respectively. Lastly, mean sensory scores of both sampling visits were 8.9, 8.6, and 6.0 on d 0, 7, and 14, respectively.

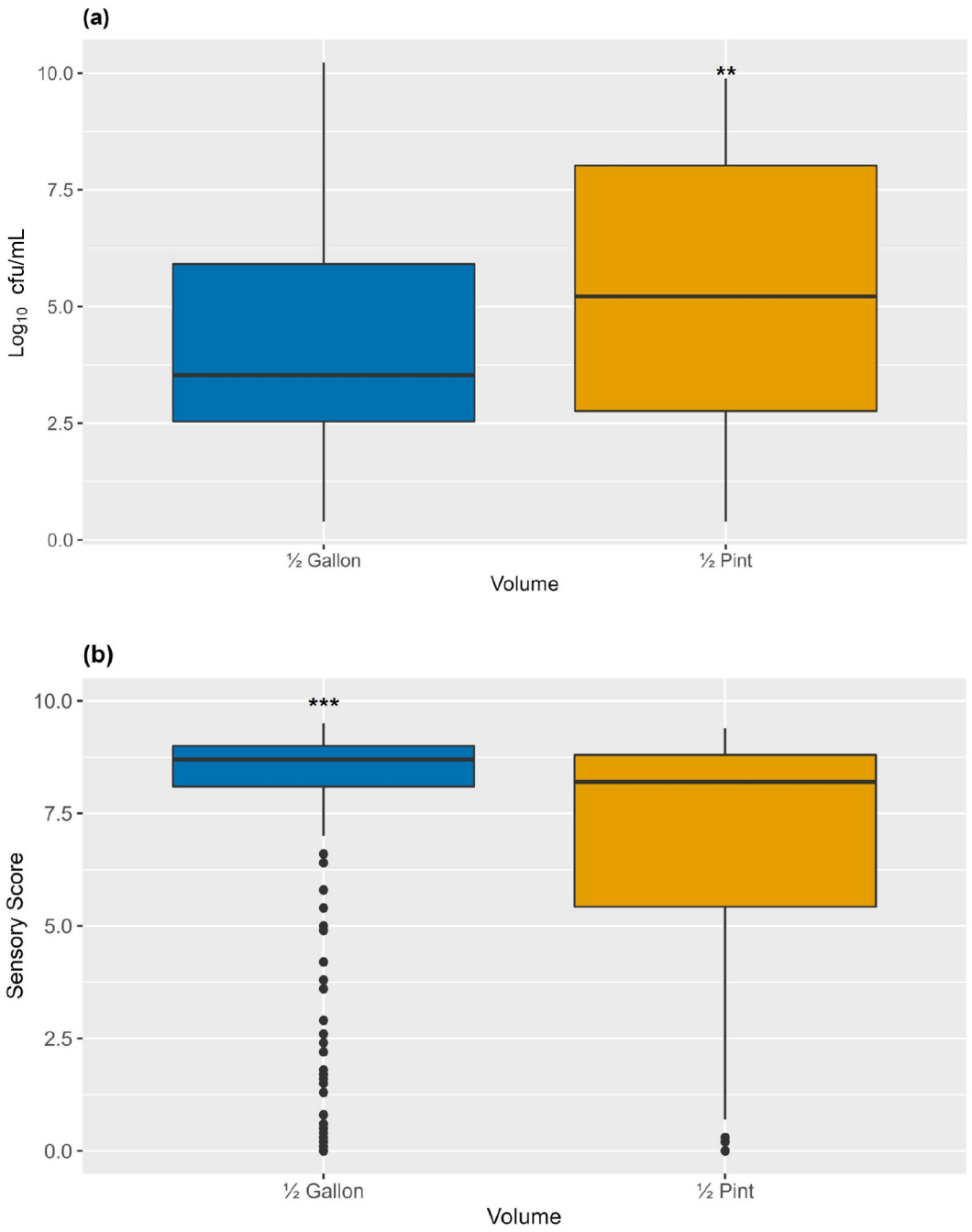


**Figure 1.** Half-gallon (n = 326) and half-pint (n = 88) fluid milk sample (a) standard plate counts (SPC) and (b) sensory scores on d 14 of shelf-life. Data are from 12 fluid milk plants that are enrolled in the Cornell Voluntary Shelf-Life Program and produced both half-gallon and half-pint volumes of fluid milk between 2013 and 2017. The bar within each box represents the median. The upper and lower ends of each box represent the 25th and 75th percentiles. The upper and lower whiskers extend to the highest and lowest values, respectively. The whiskers do not extend past 1.5 multiplied by the interquartile range (IQR), which is defined as the difference between the 25th and 75th percentiles. Individual points are outside the range defined by the IQR multiplied by 1.5. The limit of detection for SPC was 10 (1.00 $\log_{10}$) cfu/mL, with samples below the detection limit being assigned (and plotted) a value of 25% of the detection limit of 2.5 (0.40 $\log_{10}$) cfu/mL. Any SPC that were too numerous to count were assigned values of the upper countable limit for spiral plating (400,000 cfu/mL) multiplied by the dilution factor. **Indicates SPC of half-pint fluid milk samples are significantly higher than SPC of half-gallon milk ($P < 0.01$) based on Wilcoxon rank-sum. ***Indicates flavor scores of half-gallon fluid milk samples are significantly higher than flavor scores of half-pint fluid milk samples ($P < 0.001$) based on Wilcoxon rank-sum.

**Table 2.** Day 0, 7, and 14 standard plate counts (SPC), sensory scores, and gram-negative spoilage frequency of single-serve milk collected from 4 different processing facilities on 2 separate sampling visits

| | Mean SPC ($log_{10}$ cfu/mL) on day[1] | | | Mean sensory score (0.0–10.0) on day[2] | | | % Gram-negative spoilage on day | | |
|---|---|---|---|---|---|---|---|---|---|
| Facility (no. of commingled samples) | 0 | 7 | 14 | 0 | 7 | 14 | 0 | 7 | 14 |
| Sampling visit 1 | | | | | | | | | |
| Facility 1 (n = 36) | 2.09 | 2.90 | 5.10 | 8.5 | 8.8 | 6.1 | 0 | 30 | 41 |
| Facility 2 (n = 36) | 0.99 | 1.55 | 4.35 | 8.8 | 8.8 | 7.8 | 0 | 3 | 19 |
| Facility 3 (n = 36) | 2.45 | 3.04 | 5.24 | 8.7 | 8.9 | 7.9 | 0 | 8 | 23 |
| Facility 4 (n = 36) | 1.57 | 2.02 | 5.54 | 8.8 | 8.2 | 7.5 | 0 | 6 | 14 |
| All facilities | 1.79 | 2.37 | 5.05 | 8.7 | 8.7 | 7.3 | 0 | 11 | 24 |
| Sampling visit 2 | | | | | | | | | |
| Facility 1 (n = 31) | 1.21 | 4.03 | 6.70 | 9.2 | 8.9 | 5.3 | 0 | 44 | 79 |
| Facility 2 (n = 32) | 1.12 | 3.07 | 5.47 | 9.0 | 8.7 | 5.5 | 0 | 25 | 42 |
| Facility 3 (n = 30) | 3.88 | 4.76 | 6.29 | 8.9 | 8.6 | 5.5 | 0 | 27 | 33 |
| Facility 4 (n = 28) | 1.71 | 3.51 | 6.45 | 8.7 | 8.2 | 6.0 | 0 | 37 | 54 |
| All facilities | 1.96 | 3.83 | 6.21 | 9.0 | 8.6 | 5.6 | 0 | 33 | 52 |
| Both sampling visits | 1.87 | 3.03 | 5.58 | 8.9 | 8.6 | 6.0 | 0 | 21 | 37 |

[1]Detection limit of 20 cfu/mL (1.30 $log_{10}$ cfu/mL); plates below the detection limit were assigned 25% of the detection limit (5 cfu/mL [0.70 $log_{10}$ cfu/mL]) for computing mean plate counts; thus, some mean plate counts are <1.30 $log_{10}$ cfu/mL.

[2]Sensory scores were only determined for 1 out of 3 commingled samples for each combination of milk type and sampling time point for sampling visit 1.

### *Frequency of Gram-Negative Spoilage Differed Between Processing Facilities, Milk Types, Sampling Visits, and Lanes, But Not Between Processing Sample Time Points*

A logistic regression analysis of sampling visit 1 data revealed no significant differences in gram-negative spoilage frequency between any milk type or any sampling time point (all pairwise comparison $P > 0.05$; Table 3); however, samples from facility 1 had higher gram-negative spoilage frequency than samples from facilities 2 ($P = 0.007$), 3 ($P = 0.045$), and 4 ($P = 0.003$), while gram-negative spoilage frequency of samples from 2, 3, and 4 did not significantly differ (all $P > 0.05$). Logistic regression analysis of sampling visit 2 data also revealed no significant differences in gram-negative spoilage frequency between sampling time points (only beginning and end; $P > 0.05$) but found significant differences in gram-negative spoilage frequencies between facilities, milk types, and lanes (Table 3). More specifically, for visit 2, facility 1 showed the highest gram-negative spoilage frequency, which was significantly higher as compared with facilities 2 ($P = 0.003$) and 3 ($P < 0.001$) but did not differ significantly from facility 4 ($P > 0.05$). For gram-negative spoilage frequency by milk type, white skim, white 1%, and chocolate 1% milk products did not significantly differ (all $P > 0.05$), but all did have higher gram-negative spoilage frequency as compared with chocolate skim milk (white skim, $P < 0.001$; white 1%, $P < 0.001$; chocolate 1%, $P = 0.010$). Finally, lanes 2 and 3 had significantly higher frequencies of gram-negative spoilage as compared with lane 4 ($P = 0.027$, $P = 0.022$, respectively), while lane 1 gram-negative spoilage frequency was not different from lanes 2, 3, and 4 (all $P > 0.05$).

To initially evaluate overall trends between both sampling visits, combined d 14 gram-negative spoilage frequencies were plotted for facilities and milk types by facility (Supplemental Figure S4a–b; https://github.com/FSL-MQIP/single_serve_milk), while sampling time points and lanes were plotted by facility separately for sampling visits 1 and 2, respectively (Supplemental Figure S4c–d). We also performed an overall logistic regression of combined data from sampling visits 1 and 2 but excluding data for the middle time point from sampling visit 1 (as this time point was not assessed during sampling visit 2). These data revealed that the frequency of single-serve containers with gram-negative spoilage was significantly different ($P < 0.001$) between sampling visit 1 and 2 (24 and 52% of samples had gram-negative spoilage on d 14, respectively), which may be explained by the different sampling schemes, as 8 cartons were used to create a single commingled sample for sampling visit 1, compared with 16 cartons for a single commingled sample for sampling visit 2. Thus, the chance of a single carton introducing gram-negative contamination into a commingled sample is higher for sampling visit 2 compared with sampling visit 1. As shown in Table 3, facility 1 showed the highest gram-negative spoilage frequency (59%), which was significantly higher than the gram-negative spoilage frequency of samples from facilities 2 (30%; $P < 0.001$), 3 (28%; $P < 0.001$), and 4 (31%; $P < 0.001$). Among milk types, gram-negative spoilage frequency did not significantly differ (all $P > 0.05$). Beginning and end sampling time points also did not

**Table 3.** Day 14 gram-negative spoilage frequency of single-serve milk samples separated by facility and milk type for both sampling visits and sampling time point for sampling visit 1 and lane for sampling visit 2

| | Gram-negative spoilage sample proportion and frequency (%)[1,2] | | |
|---|---|---|---|
| Item | Sampling visit 1 | Sampling visit 2 | Both visits |
| Overall | 34/141 (24%)[a] | 61/118 (52%)[b] | 95/259 (37%) |
| Facility | | | |
| 1 | 14/34 (41%)[a] | 23/29 (79%)[a] | 37/63 (59%)[a] |
| 2 | 7/36 (19%)[b] | 13/31 (42%)[b] | 20/67 (30%)[b] |
| 3 | 8/35 (23%)[b] | 10/30 (33%)[b] | 18/65 (28%)[b] |
| 4 | 5/36 (14%)[b] | 15/28 (54%)[ab] | 20/64 (31%)[b] |
| Milk type | | | |
| White skim | 7/34 (21%) | 20/29 (69%)[a] | 27/63 (43%) |
| White 1% | 7/36 (19%) | 20/30 (67%)[a] | 27/66 (41%) |
| Chocolate skim | 11/36 (31%) | 7/31 (23%)[b] | 18/67 (27%) |
| Chocolate 1% | 9/35 (26%) | 14/28 (50%)[a] | 23/63 (37%) |
| Sampling time point | | | |
| Beginning | 13/47 (28%) | 33/60 (55%) | 46/107 (43%) |
| Middle | 7/48 (15%) | NA | NA |
| End | 14/46 (30%) | 28/58 (48%) | 42/104 (40%) |
| Lane | | | |
| 1 | NA | 15/30 (50%)[ab] | NA |
| 2 | NA | 17/30 (57%)[a] | NA |
| 3 | NA | 20/30 (67%)[a] | NA |
| 4 | NA | 9/28 (32%)[b] | NA |

[a,b]Gram-negative spoilage frequencies with differing letters for the "Overall" row or within the same column and same category (i.e., facility, milk type, sampling time point, or lane) are significantly different ($P < 0.05$) based on logistic regression analysis that included both d 7 and 14 gram-negative spoilage frequencies.

[1]Gram-negative spoilage frequency is defined as a sample having a standard plate count of >20,000 cfu/mL and showing at least 1 typical colony (i.e., red colony [detectable limit 20 cfu/mL]) on crystal violet tetrazolium agar.

[2]NA is defined as "not applicable." For sampling visit 1, no sampling of separate lanes was performed, and for sampling visit 2, the middle time point was not sampled; corresponding lanes and appropriate columns were thus designated as NA.

show significant differences between gram-negative spoilage frequencies ($P > 0.05$).

### *Milk with Gram-Negative Bacterial Contamination Shows Significantly Higher Microbial Counts and Lower Sensory Scores During Shelf-Life as Compared with Milk Without Evidence of Gram-Negative Bacterial Contamination*

Mean microbial counts at d 7 and 14 and sensory scores at d 7 and 14 revealed differences between samples with gram-negative contamination (at least 1 colony on both SPC and CVTA) and without gram-negative contamination (Table 4). Separate ANOVA for SPC and sensory scores showed that gram-negative contamination had a significant effect on (1) SPC ($P < 0.001$) with a large effect size (as measured by partial eta squared; $\eta^2 = 0.25$) and (2) sensory scores ($P < 0.001$) with a medium effect size ($\eta^2 = 0.11$). Subsequent Tukey pairwise comparisons revealed that samples with gram-negative contamination had significantly higher SPC ($P < 0.001$) and lower sensory scores ($P < 0.001$) compared with samples without gram-negative contamination. Day of shelf life also had a significant effect on (1) SPC ($P < 0.001$) with a large effect size ($\eta^2 = 0.42$) and (2) sensory scores ($P < 0.001$) with a large effect size ($\eta^2 = 0.31$). Tukey pairwise comparisons revealed that across all samples, d 14 SPC were significantly higher ($P < 0.001$), and d 14 sensory scores were significantly lower ($P < 0.001$) than d 7 SPC and d 7 sensory scores, respectively.

Separate ANOVA for SPC and sensory scores for assessing the effect of milk type (i.e., white skim, white 1%, chocolate skim, chocolate 1%), showed that milk type had a significant effect on SPC ($P < 0.001$) with a large effect size $\eta^2 = 0.22$) but did not have a significant effect on sensory scores ($P > 0.05$; Table 5). Tukey pairwise comparisons for assessing differences in SPC among milk types revealed that chocolate skim milk had significantly higher SPC than white skim ($P < 0.001$) and white 1% ($P < 0.001$) milks. Chocolate 1% milk also had significantly higher SPC than white skim ($P < 0.001$) and white 1% ($P < 0.001$) milks. The SPC of white skim and white 1% did not significantly differ ($P > 0.05$) and the SPC of chocolate skim and chocolate 1% did not significantly differ ($P > 0.05$)

**Table 4.** Sampling visit 1 and 2 estimated marginal means and Tukey pairwise comparisons of d 7 and 14 standard plate counts (SPC) and d 7 and 14 sensory scores of samples with or without gram-negative contamination[1]

| Item | Samples with evidence of gram-negative contamination | Samples without evidence of gram-negative contamination |
|---|---|---|
| SPC ($\log_{10}$ cfu/mL) | | |
| d 7 | 4.74 ± 0.16[a] | 2.15 ± 0.12[b] |
| d 14 | 6.57 ± 0.15[a] | 4.88 ± 0.12[b] |
| Sensory scores (0.0–10.0) | | |
| d 7 | 7.9 ± 0.2[a] | 9.0 ± 0.2[b] |
| d 14 | 5.4 ± 0.2[a] | 6.5 ± 0.2[b] |

[a,b]Different superscripts indicate values that differ significantly based on ANOVA (all $P < 0.001$) between samples with and without gram-negative contamination.

[1]Values reported are estimated marginal means (±SE) of a general linear model with SPC or sensory score as response variables and gram-negative contamination (yes or no), day of shelf life (7, 14), and milk type (white skim, white 1%, chocolate skim, and chocolate 1%) as predictor variables.

### *In-Facility Surveys and Observations Identified Mandrels as a Potential Source of Contamination and Showed Correlations Between Frequency of Downtime Events and Gram-Negative Contamination Events*

An initial simple survey, administered to N-8 (single-serve) filler operators, employees that clean N-8 fillers, employees who perform maintenance on N-8 fillers, and quality management at the 4 participating facilities, was used to identify potential areas of concern regarding the N-8 fillers used to produce single-serve fluid milk (Table 6). From these surveys, the "infeed" (where flat cartons are placed to be formed for filling) was most commonly identified as an "issue" when operating the N-8 fillers (11/13 respondents), and the infeed was also most commonly identified as an area for improvement (7/13 respondents). Additionally, "ovens" (where the top of the carton gets sealed after filling) were identified as the hardest part to clean (7/10 respondents), and mandrels were identified as both the second hardest parts to clean (3/10 respondents) as well as the part that most often failed ATP tests (5/7 respondents).

To further assess operations of the N-8 fillers, one of the authors (T. T. Lott), who was present for at least 50% of the run where samples were collected, tracked downtime events, including (1) overall downtime; (2) downtimes due to filler specific issues (i.e., infeed issues, forming and sealing, milk temperature deviations, cabinet downtimes), and (3) downtimes due to issues not related to the filler (e.g., employee breaks, code date printer, milk crate issues such as no available crates). Among the 8 initial sampling visits (4 processing facilities × 2 sampling visits), an average of 13 overall downtimes were observed (by T. T. Lott), with estimated mean downtimes per run of 5.5 and 7.5 observed due to filler specific issues and nonfiller issues, respectively (see Supplemental Figure S5a–b; https://github.com/FSL-MQIP/single_serve_milk). Although we were not able to observe and track downtimes over 100% of a given production run, we believe that the high frequency of filler related downtimes suggests that there are mechanical processing issues associated with the production of single-serve milk.

**Table 5.** Day 7 and 14 combined estimated marginal means and Tukey pairwise comparisons of standard plate counts (SPC) and sensory scores of different single-serve milk types

| | Shelf-life measure[1] | |
|---|---|---|
| Milk type | SPC ($\log_{10}$ cfu/mL) | Sensory score |
| White skim | 4.00 ± 0.13[a] | 7.33 ± 0.2 |
| White 1% | 3.56 ± 0.13[a] | 7.18 ± 0.2 |
| Chocolate skim | 5.29 ± 0.13[b] | 7.02 ± 0.2 |
| Chocolate 1% | 5.49 ± 0.14[b] | 7.43 ± 0.2 |

[a,b]Mean values in the same column with different superscripts significantly differ (all $P < 0.01$) between single-serve milk types based on ANOVA.

[1]Means reported are estimated marginal means (±SE) of a general linear model with SPC or sensory score as response variables and gram-negative contamination (yes or no), day of shelf life (7, 14), and milk type (white skim, white 1%, chocolate skim, and chocolate 1%) as predictor variables.

### *Environmental Samples from Fillers Frequently Exhibited Evidence of Gram-Negative Contamination*

For 101 out of the 288 (35%) total samples (milk, carton, and mandrel environmental sponge samples) collected during a follow-up visit to each facility, we found detectable growth on SPC (direct plating without stress test), with all counts being <500 cfu/mL. All 288 samples were subjected to a "stress test" (i.e., incubation at 21°C for 24 h, followed by plating on SPC and CVTA) to test (with higher sensitivity) for gram-negative contamination; samples were considered to be contaminated with gram-negatives if there was detectable growth on both CVTA and SPC plates for a given

**Table 6.** Questions administered and responses from facility personnel on survey questions regarding operation and sanitation issues specific to single-serve fillers[1]

| | Survey question | | | |
|---|---|---|---|---|
| Item | What specific issues do you have with this filler? (respondents, n = 13) | If you could improve one thing, what would it be? (respondents, n = 13) | What is the hardest part to clean? (respondents, n = 10) | What section fails ATP[2] most often? (respondents, n = 7) |
| Response (number) | Infeed (11)<br>Forming/sealing (9)<br>Other (5) | Infeed (7)<br>Consistency (3)<br>Other (7)[3] | Oven (7)<br>Mandrels (3)<br>Other (3) | Mandrels (5)<br>Filler valves (1)<br>Nothing (1) |

[1]Some respondents gave multiple answers for the same question; thus, the total number of responses is greater than the number of respondents for some questions.

[2]Adenonsine triphosphate testing (postcleaning and sanitation) is performed to verify cleaning and assess surface cleanliness.

[3]Responses (e.g., supplier carton packaging consistency) were grouped into "Other" if 2 or fewer responses matched a specific category.

sample, as described above. Based on these criteria, 123 out of the 288 (43%) samples were positive for gram-negative contamination (Table 7). Logistic regression revealed significant differences in gram-negative contamination frequency between facilities and between sample types (i.e., milk, cartons, mandrel sponges), but no significant difference between the 4 different lanes (all pairwise $P > 0.05$). Pairwise comparisons revealed that carton samples had significantly lower gram-negative contamination frequency (13%) than milk (49%) and mandrel sponge samples (67%) (both $P < 0.001$), and that mandrel sponge samples had significantly higher gram-negative contamination frequency than milk samples ($P = 0.012$). Additionally, samples collected from facility 4 had significantly lower gram-negative contamination frequency (13%) than facility 1 (38%; $P = 0.002$) and facilities 2 (46%) and 3 (75%) (both $P < 0.001$), and facility 3 had significantly higher gram-negative contamination frequency compared with facility 1 ($P < 0.001$) and facility 2 ($P < 0.001$; Table 7). Consistent with the observation that facility 3 had the highest contamination frequency, only for facility 3, some samples tested by direct plating tested positive for gram-negative contamination (i.e., detectable growth on SPC and CVTA plates), including milk (2/24 samples; 8%), carton (7/24 samples; 29%), and mandrel sponge samples (4/24 samples; 17%).

### 16S rRNA Gene Sequencing Analysis Indicates That Bacterial Genera Obtained From Finished Products Were Also Frequently Found on Mandrels

Among a total of 702 isolates collected, 664 were successfully characterized; these 664 isolates were obtained from milk (n = 372), mandrel sponges (n = 202), carton samples (n = 66) as well as environmental sponge samples taken from the mandrel hub (an axle-like unit where the mandrels are attached; n = 24). Samples from the mandrel hub were taken because excessive grease build-up was observed on this part of the equipment, as noted above.

Among the 664 characterized isolates, we were able to assign genus level classification to 554 isolates, including 323, 158, 53, and 20 isolates from milk, mandrel sponge,

**Table 7.** Gram-negative contamination data for milk, mandrel sponge, and carton samples collected at follow-up visits before intervention[1]

| | Gram-negative contamination sample proportion and frequency (%) | | | |
|---|---|---|---|---|
| Facility | Milk | Mandrel sponges | Cartons | All samples |
| 1 | 2/24 (8%) | 24/24 (100%) | 1/24 (4%) | 27/72 (38%)[b] |
| 2 | 18/24 (75%) | 13/24 (54%) | 2/24 (8%) | 33/72 (46%)[b] |
| 3 | 24/24 (100%) | 21/24 (88%) | 9/24 (38%) | 54/72 (75%)[a] |
| 4 | 3/24 (13%) | 6/24 (25%) | 0/24 (0%) | 9/72 (13%)[c] |
| Overall | 47/96 (49%)[b] | 64/96 (67%)[a] | 12/96 (13%)[c] | 123/288 (43%) |

[a–c]Based on logistic regression analysis, differing letters in the "Overall" row indicate significant differences ($P < 0.05$) of gram-negative contamination frequencies between sample type (i.e., milk, mandrel sponges, cartons) and differing letters in the "All Samples" column indicate significant differences ($P < 0.01$) of gram-negative contamination frequencies between facilities (i.e., 1, 2, 3, and 4). Sample type was not assessed (by logistic regression) at the level of individual facility.

[1]Gram-negative contamination frequency is defined here as plated samples that had (1) detectable growth (at least one typical [i.e., red] colony on crystal violet tetrazolium agar), a selective medium for gram-negative bacteria, and (2) detectable growth (at least one colony) for standard plate count.

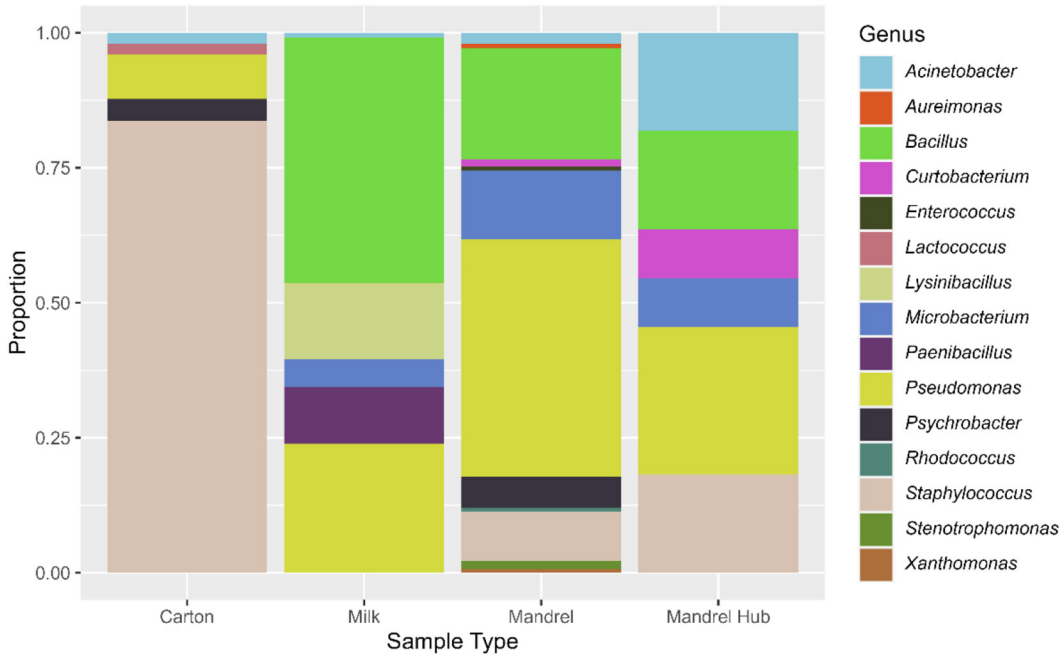


**Figure 2.** Genus-level classification based on 16S rRNA gene sequencing data for isolates (n = 419) collected from carton (49 isolates), milk (218 isolates), mandrel sponge (mandrel; 141 isolates), and mandrel hub sponge (mandrel hub; 11 isolates) samples collected during a follow-up visit to each of 4 facilities and then stored at stored at 21°C for 24 h (representing a "stress test") to facilitate detection of low levels of bacteria. Among the total 664 isolates characterized as part of this study, a subset of 245 isolates is not represented here, including 135 isolates that were collected from direct plating of samples (i.e., before incubation of samples at 21°C for 24 h), 65 isolates that were classified to the family level, 4 classified to the order level, and 41 to the class level.

carton samples, and mandrel hubs, respectively. Among the remaining 110 isolates, 41, 4, and 65 isolates were characterized at the class, order, and family taxonomic levels, respectively. Of the 554 isolates with genus classification, a total of 419 isolates were from stress tests as described above; this subset of 419 isolates was used for comparing populations between sample types. The most frequent genera among these isolates were (1) *Bacillus* from milk samples (99/218), (2) *Pseudomonas* from mandrel sponge samples (62/141), (3) *Staphylococcus* from cartons (41/49), and (4) *Pseudomonas* from mandrel hubs (3/11; Figure 2). Genera isolated from all 4 sample types included *Pseudomonas* (121/419) and *Acinetobacter* (8/419).

To determine if certain genera were associated with a given sample type (i.e., carton, milk, mandrel sponge samples), facility, or lane number, we used a dissimilarity matrix to create NMDS plots and perform ANOSIM analyses; this analysis was performed on 408 isolates with genus assignment obtained from stress tests (excluding isolates from mandrels hubs due to the small number of samples in this category). The NMDS plots showed no obvious clustering by lane or facility (Supplemental Figure S6a; https://github.com/FSL-MQIP/single_serve_milk), but there appeared to be some clustering by sample type (i.e., carton, milk, and mandrel sponge), as shown in Supplemental Figure S6b. For example, for facility 1, mandrel sponge samples clustered together, whereas milk samples clustered separately. Also, carton isolates from facilities 2 and 3 clustered together (in the top right; see Supplemental Figure S4b), consistent with the fact that the majority of carton isolates from these facilities represented the genus *Staphylococcus* (7/10 and 29/41 carton isolates from facilities 2 and 3 were classified as *Staphylococcus*). A nonparametric, ANOVA-like test for dissimilarity matrices (**ANOSIM**) confirmed a significant ($P < 0.001$) dissimilarity of genera between sample types (R = 0.55). ANOSIM also found a significant ($P = 0.033$) dissimilarity of genera between facilities, even though the effect size was small (R = 0.10). We found no significant dissimilarity of genera between lane numbers ($P > 0.05$).

Given that the dissimilarity of genera between sample types and between facilities was significant, we followed-up the ANOSIM test with multilevel pattern analyses to determine if any genera were significantly associated with a given sample type or facility. At the level of sample type, (1) *Staphylococcus* was significantly associated with carton and mandrel sponge samples ($P < 0.001$); (2) *Bacillus* was significantly associated with milk and mandrel sponge samples ($P < 0.001$);

and (3) *Lysinibacillus* and *Paenibacillus* were significantly associated with milk samples (both $P < 0.001$). At the level of facility, (1) *Microbacterium* and *Paenibacillus* were significantly associated with facilities 1, 2, and 4 ($P = 0.026$, $P = 0.048$, respectively); and (2) *Pseudomonas* was significantly associated with facilities 1, 3, and 4 ($P = 0.003$).

### Interventions Aimed at Improving Sanitation of Mandrels Did Not Lead to Significant Reductions in Frequency of Gram-Negative Spoilage in Single-Serve Milk

A logistic regression that includes intervention/control lanes, facility, beginning or end time points, and lane number as predictors and frequency of samples with evidence of gram-negative spoilage (defined as typical growth detected on CVTA, i.e., ≥1 colony; and SPC was >20,000 cfu/mL) after 14 d of storage as the outcome, did not reveal a significant difference between intervention and control lanes, as 154/400 (39%) samples from intervention lanes and 139/398 (35%) samples for control lanes showed evidence of gram-negative spoilage ($P > 0.05$; Table 8). However, we did find significant differences between facilities and time points. Facility 4 had significantly higher gram-negative spoilage frequency (177/199; 89%) than facilities 1 (87/199; 44%), 2 (10/200; 5%), and 3 (19/200; 10%; all $P < 0.001$), and facility 1 had significantly higher gram-negative spoilage frequency than facilities 2 and 3 (both $P < 0.001$). For time points, samples collected at the beginning time point had significantly higher gram-negative spoilage (167/399; 42%) than samples collected at the end time point (126/399; 32%; $P < 0.001$). For gram-negative spoilage frequencies between lanes, lane 1 (65/199; 33%), lane 2 (77/200; 39%), lane 3 (74/199; 37%), and lane 4 (77/200; 39%) were not significantly different (all $P > 0.05$).

A total of 159 isolates were collected across control and intervention samples, and 151 of these isolates were successfully characterized by 16S rRNA gene sequencing. The 151 isolates were characterized by the class (n = 17), order (n = 3), family (n = 16), and genus level (n = 115). Given we were able to characterize the majority of isolates at the genus level, the 115 isolates characterized at the genus level were used for comparing populations between control (n = 55) and intervention (n = 60) samples (Figure 3a), between facility 1 (n = 27), 2 (n = 31), 3 (n = 22), and 4 (n = 35) samples (Figure 3b), between white milk (n = 61) and chocolate milk (n = 54) samples (Figure 3c), and between beginning (n = 66) and end (n = 49) sample time points (Figure 3d).

*Paenibacillus* and *Pseudomonas* represented the most (44/115) and second most (32/115) frequently isolated genera, respectively, and combined for more than 65% of both control sample isolates (20 and 16 isolates, respectively) and intervention sample isolates (24 and 16 isolates, respectively). *Bacillus* was the third most frequently isolated genus from both control (n = 7, 13%) and intervention (n = 8, 13%) samples. No other genus represented more than 10% of isolates in either control or intervention samples. Other genera isolated from both control and intervention samples included

**Table 8.** Gram-negative spoilage frequency of single-serve milk samples separated by lanes with control and intervention cleanings, facility, sampling time point, and lane

| | Proportion and frequency (%) of samples spoiled by gram-negative bacteria[1] | | |
|---|---|---|---|
| Item | Control | Intervention | Combined |
| Overall | 139/398 (35%) | 154/400 (39%) | 293/798 (37%) |
| Facility | | | |
| 1 | 35/99 (35%) | 52/100 (52%) | 87/199 (44%)[a] |
| 2 | 2/100 (2%) | 8/100 (8%) | 10/200 (5%)[b] |
| 3 | 15/100 (15%) | 4/100 (4%) | 19/200 (10%)[b] |
| 4 | 87/99 (88%) | 90/100 (92%) | 177/199 (89%)[c] |
| Sampling time point | | | |
| Beginning | 84/199 (42%) | 83/200 (42%) | 167/399 (42%)[a] |
| End | 55/199 (28%) | 71/200 (36%) | 126/399 (32%)[b] |
| Lane | | | |
| 1 | 15/99 (15%) | 50/100 (50%) | 65/199 (33%) |
| 2 | 42/100 (43%) | 35/100 (35%) | 77/200 (39%) |
| 3 | 53/99 (54%) | 21/100 (21%) | 74/199 (37%) |
| 4 | 29/100 (29%) | 48/100 (48%) | 77/200 (39%) |

[a–c]Based on logistic regression analysis, gram-negative spoilage frequencies with differing letters within the same category (i.e., facility, sampling time point, or lane) are significantly different ($P < 0.05$).

[1]Gram-negative spoilage frequency is defined here as a sample having (1) detectable growth (at least one typical [i.e., red] colony on crystal violet tetrazolium agar), a selective medium for gram-negative bacteria, and (2) standard plate count of >20,000 cfu/mL.

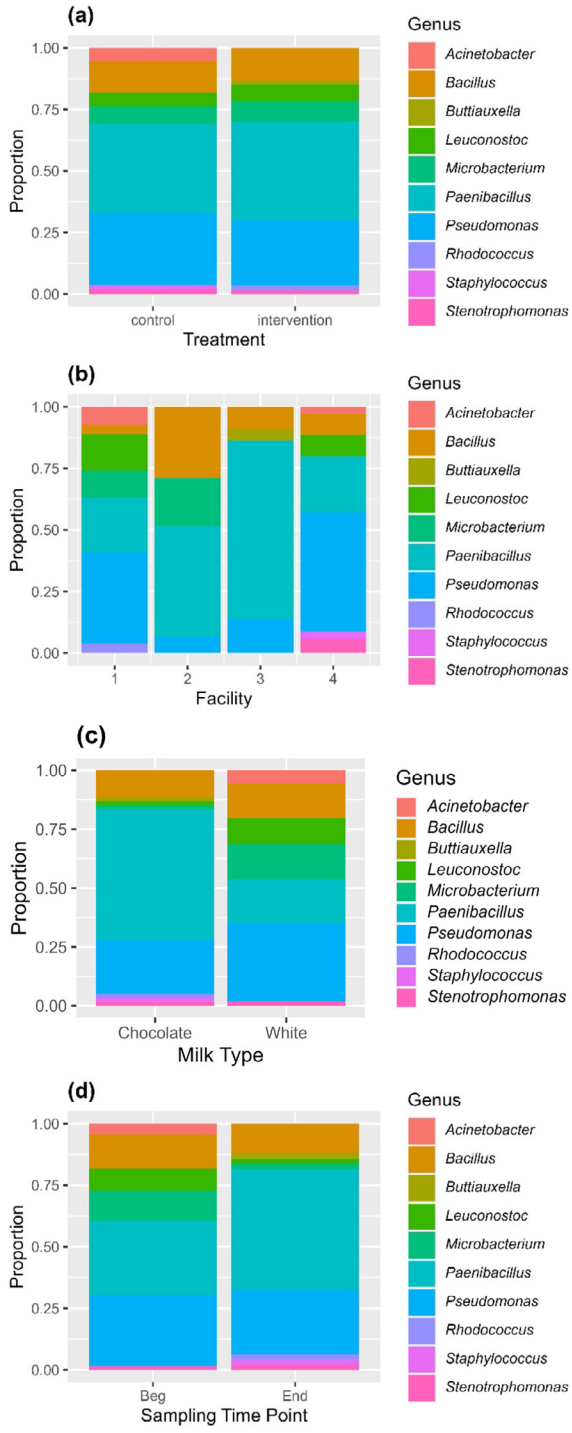


**Figure 3.** Genus-level classification based on 16S rRNA gene sequencing data for 115 isolates collected from the intervention part of the study separated by (a) samples collected from control lanes (55 isolates) or lanes subjected to a cleaning intervention (60 isolates); (b) samples collected from processing facilities 1 (27 isolates), 2 (31 isolates), 3 (22 isolates), or 4 (35 isolates); (c) chocolate (54 isolates) or white milk samples (61 isolates); and (d) samples collected at the beginning (Beg; 66 isolates) or end of production (49 isolates). Among 151 isolates, 16 were classified to the family level, 3 isolates to the order level, and 17 isolates to the class level. These 36 isolates are not represented here.

*Microbacterium* (4 and 5 isolates, respectively), *Leuconostoc* (3 and 4 isolates, respectively), and *Stenotrophomonas* (1 isolate each). Some genera were isolated only from control and intervention samples, but these genera represented a small proportion of the total number of isolates. *Acinetobacter* (n = 3) and *Staphylococcus* (n = 1) were only isolated from control samples, and *Buttiauxella* (n = 1) and *Rhodococcus* (n = 1) were only isolated from intervention samples.

The most frequently isolated genera from each facility were *Pseudomonas* at facilities 1 (n = 10) and 4 (n = 17) and *Paenibacillus* at facilities 2 (n = 14) and 3 (n = 16). *Bacillus*, *Paenibacillus*, and *Pseudomonas* were the only 3 genera isolated from all 4 facilities, and represented a combined 63, 81, 95, and 80% of total isolates from facilities 1, 2, 3, and 4. The percentage of isolates that were classified as gram-negative was 56, 94, 82, and 43% at facilities 1, 2, 3, and 4, respectively. *Bacillus*, *Paenibacillus*, and *Pseudomonas* were also frequently found at both sampling time points, representing 14, 30, and 29%, respectively, of isolates collected from samples collected at the beginning, and 12, 49, and 27%, respectively, of isolates collected at the end. The most frequently isolated genus from each milk type was *Paenibacillus* from chocolate milk (n = 34) and *Pseudomonas* from white milk (n = 18). ANOSIM was performed to determine if populations differed between milk types, facilities, sample time points, or treatments (i.e., intervention, control); however, no significant differences were found (all $P > 0.05$).

## DISCUSSION

Although assuring the microbial and sensory quality of HTST pasteurized fluid milk represents a broad concern for the dairy industry, there is specific concern about assuring the quality of fluid milk in single-serve containers, as milk in this type of packing is often distributed to schools and, thus, frequently consumed by adolescents. This concern is at least partially driven by the fact that it is well established that negative experiences with fluid milk quality in children can negatively influence their consumption of fluid milk long-term (McCarthy et al., 2017; Sipple et al., 2020). Although several studies have specifically assessed the sensory quality of school milk, particularly as influenced by different flavors (Fayet-Moore, 2016), packaging (Sipple et al., 2021), and fat levels (Keefer et al., 2022), there is a need to further and specifically understand microbial spoilage of HTST pasteurized fluid milk that is served in schools and to children, and this knowledge needs to be translated into practical strategies to improve the quality of these fluid milk products. In this study, we aimed to thoroughly investigate microbial

spoilage of milk packaged in single-serve cartons with an emphasis on gram-negative contamination and spoilage, as it is well known that gram-negative bacteria can cause premature spoilage in fluid milk. An initial analysis of historic data collected as part of our group's VSL program revealed that SPC are significantly higher and that sensory scores are significantly lower in single-serve milk (i.e., half-pint) as compared with milk from the same facilities packaged in larger volume containers (i.e., half-gallon). Our subsequent detailed study on 4 fluid milk processing facilities (1) further confirmed a high frequency of gram-negative spoilage and contamination in single-serve milk, and (2) indicated that specific equipment components of single-serve milk fillers (e.g., mandrels) are difficult to clean and may represent likely sources of microbial PPC. We also showed that focused filler-level cleaning interventions targeting putative high-risk sources of contamination were not effective at reducing PPC in the 4 study facilities, suggesting the need for more comprehensive interventions to improve the quality of single-serve HTST fluid milk.

### *Single-Serve HTST Fluid Milk Appears to Show a High Frequency of Gram-Negative Bacterial Spoilage Across Production Facilities*

An initial analysis of our group's historic VSL data for 12 processing facilities, which produced both single-serve milk and half-gallon milk, revealed that single-serve milk had significantly higher SPC and significantly lower sensory scores as compared with milk packaged in half-gallon containers. These findings were further corroborated by milk quality data collected for single-serve milk samples collected over 2 initial visits to 4 different processing facilities; data collected as this part of our study showed that 37% of single-serve fluid milk packages showed evidence of gram-negative spoilage after 14 d of storage at 6°C. This evidence for frequent PPC is important, as gram-negative organisms (e.g., *Pseudomonas*), which are typically introduced through PPC, are well established to (1) rapidly grow in milk at refrigeration temperatures (Ternström et al., 1993), which can surpass the sensory threshold of 1,000,000 cfu/mL in as little as 14 d postprocessing (Martin et al., 2012) and (2) produce sensory defects that are severe and easily perceived by consumers and hence, lead to bad experiences with fluid milk (Hayes et al., 2002). Our data collected here also further confirmed previous studies (Schröder, 1984; Martin et al., 2012; Alles et al., 2018; Reichler et al., 2018) regarding the specific importance of PPC with gram-negative spoilage organisms as milk with gram-negative PPC showed significantly lower sensory scores as compared with milk without evidence for gram-negative PPC. In terms of shelf life, it is important to note that shelf life samples were stored at 6°C with minimal variation (±0.5°C), though a greater variation in refrigeration temperature would be expected along the supply chain (i.e., from processor to school). To predict shelf life more accurately, future research could include gathering comprehensive supply-chain refrigeration and storage time data and pair it with modeling tools, such as the Monte Carlo model previously described for predicting spoilage due to PPC in fluid milk (Lau et al., 2022).

Although our data indicated a higher frequency of PPC with gram-negative organisms in single-serve fluid milk, there is an overall substantial issue with gram-negative PPC across all fluid milk packaging sizes and types, which is consistent with several previous studies. For example, in a study of 3.8-L (1 gallon) fluid milk samples from New York State, 26% of samples had evidence of PPC (Martin et al., 2011a), whereas 51% of samples from Norway and Sweden had evidence of spoilage by gram-negative bacteria (Ternström et al., 1993). Although we are not aware of any other studies that specifically assessed the microbial quality of single-serve HTST milk, it is important to point out that most single-serve fluid milk products in the United States, particularly those that go into price-sensitive markets, such as schools, are filled on gable top fillers (such as the N-8 filler used in all 4 study facilities included here), while larger size (e.g., half-gallon or one-gallon high-density polyethylene containers) are frequently filled on different types of fillers (e.g., rotary fillers). One issue with the design of gable top fillers is that many of the filler components (e.g., mandrels, mandrel hubs, carton infeed) are not cleanable by clean-in-place (**CIP**) systems, when compared with rotary fillers. The lack of a hygienic design for gable top fillers is further exacerbated by the fact that the majority of these components (not subject to CIP) are not easily accessible for hand cleaning. In this case, these factors are consistent with 2 categories on our RCA (Supplemental Figure S3; https://github.com/FSL-MQIP/single_serve_milk): (1) cleaning and sanitation and (2) milk processing equipment, which are contributors to contamination of school milk.

Interestingly, from our initial 2 visits to each of the 4 facilities, we did see a higher level of gram-negative spoilage frequency in white milks than chocolate milks, yet the mean SPC for chocolate milk were significantly higher (it should be noted that the estimated marginal means reported in Table 5 are combined values of samples with or without evidence of gram-negative contamination). Although this may seem counterintuitive,

given we did see a significant effect of gram-negative contamination on SPC, it has been well documented that chocolate milk may be more susceptible to microbial spoilage due to spore-formers (Douglas et al., 2000; Lima et al., 2011; Orleans 2011; Aouadhi et al., 2014) and has also been reported that psychrotolerant spore-formers, particularly *Paenibacillus*, are capable of growing to >4.00 $\log_{10}$ cfu/mL in milk stored at 6°C for 14 d (Ranieri and Boor, 2009). More specifically, it has been reported that some strains of *Paenibacillus odorifer* (a commonly isolated fluid milk spoilage organism) are capable of proliferating faster in chocolate milk than in white milk, likely due to several factors, including the addition of sucrose and the introduction of spores with the addition of cocoa powder (Rush et al., 2022). The hypothesis that these psychrotolerant spore-formers resulted in the higher SPC observed in chocolate milk here is supported by the fact that we found *Paenibacillus* more frequently than any other genera from the chocolate milk samples collected in the intervention part of our study, although *Paenibacillus* was less frequent among isolates collected from white milk. These findings serve as a reminder that psychrotolerant spore-forming bacteria present an additional challenge to the microbial shelf life of single-serve milk, particularly chocolate milk, and that comprehensive programs to improve the quality of single-serve and school milk also need to consider spoilage due to spore-forming bacteria.

Lastly, sensory defect judging revealed significant differences between samples with and without gram-negative contamination. As a sensory defect judging score of 6.1 to 7.9 is considered to be fair quality milk and scores below 6.0 are considered to be unacceptable quality milk (Martin et al., 2012), on average, samples without gram-negative contamination were fair quality (6.5) and samples with gram-negative contamination were unacceptable (5.4) on d 14 of shelf life. Sensory defect judging does have its limitations as point reductions are not linear (Bodyfelt et al., 2008) and results cannot be directly correlated with consumer acceptability (Drake, 2007). Nonetheless, sensory defect judging can be a valuable shelf life assessment tool for processors by identifying severe defects, some of which can be caused by some *Pseudomonas* (i.e., bitter; Hayes et al., 2002; Alvarez, 2009), and it can be reasonably assumed that consumers would likely reject these defects. Future research could lead to the development of a risk assessment model to assess the risk that school children are exposed to sensory defects in single-serve fluid milk caused by gram-negative bacteria, such as the one previously described for assessing mold in yogurt (Buehler et al., 2018).

### *Filler Related Factors Are the Most Likely Root Causes for a High Frequency of Gram-Negative Bacterial Contamination in Single-Serve Milk*

Survey responses from employees at all levels revealed that mandrels were hard to clean and most often failed ATP quality checks following cleaning and sanitation. The RCA performed with quality management also revealed excessive grease buildup on mandrel hubs; this grease could be a vehicle for transfer of spoilage organisms to the mandrels and may also lead to reduced effectiveness of cleaning procedures. Our overall findings are consistent with previous studies that have suggested fluid milk fillers as likely sources of contamination (Schröder, 1984; Ralyea et al., 1998; Reichler et al., 2020). Additionally, Reichler at al. (2020) also isolated the same gram-negative bacterial subtypes from N-8 rubber goods and fluid milk processed on the same N-8 filler, specifically supporting filler-related sources of gram-negative organisms for the same type of filler investigated in this study. Importantly, carton forming mandrels have also previously been reported as specific sources of PPC (Gruetmacher and Bradley, 1999). However, sources other than fillers have also been reported to contribute to contamination of fluid milk. For example, in an older study, Schröder (1984) reported that pasteurized milk holding tanks were more likely to be the source of contamination of fluid milk as compared with fillers; however, even in this study, fillers were reported to more likely cause contamination at higher levels, supporting the specific importance of fillers as sources of contamination events that are likely to accelerate fluid milk spoilage. Given the potential of gram-negative contamination of fluid milk via filler components, the hygienic design challenges of single-serve fillers may result in an increased risk of gram-negative contamination. Anecdotally, the observation that the single-serve fillers have an unhygienic design is supported by data, collected here, that some processing facility employees mentioned the single-serve carton filler as the most difficult to clean filler in their facilities.

The likely importance of filler-associated sources for PPC of single-serve fluid milk in the 4 study facilities was also supported by the 16S rRNA gene sequencing data generated here on isolates from single-serve fluid milk, cartons, and environmental samples (i.e., mandrel and mandrel hub sponge samples). Our 16S rRNA gene sequencing data on isolates collected during the follow-up visits to the 4 facilities studied specifically found that *Pseudomonas* represented at least 23% of isolates from milk, mandrel sponge, and mandrel hub sponge samples, suggesting that mandrels and mandrel hubs

are 2 potential sources of gram-negative contamination of single-serve milk. More specifically, *Pseudomonas* was isolated from carton (n = 2), milk (n = 18), and mandrel sponge samples (n = 10) from lane 3 of the filler at facility 3. It is however, important to note that we only performed high-level characterization of isolates to the genus level, rather than more discriminatory subtyping, as could be achieved by methods such as pulse field gel electrophoresis (Martin et al., 2011b), multilocus sequence typing (Nakamura et al., 2021), or WGS (Nastasijevic et al., 2017), which all have been previously applied to track bacterial contamination sources. Future use of highly discriminatory subtyping methods such as WGS may thus be valuable for improved RCA and tracking of spoilage contamination sources.

Importantly, 16S rRNA gene sequencing data generated here also identified a number instances where gram-positive organisms (e.g., *Bacillus*, *Staphylococcus*, *Microbacterium*) were identified from milk and carton samples and internal filler surfaces such as mandrel and mandrel hub sponge samples. More specifically, *Staphylococcus* was frequently isolated from carton samples; the most likely source of these organisms is employee hands, as cartons need to be manually loaded onto the single-serve filler. Importantly, although *Bacillus* and other aerobic spore-formers are typically assumed to originate from the raw milk supply and survive pasteurization (Boor et al., 2017), evidence suggests that these organisms can also be introduced into fluid milk from environmental sources (Ranieri et al., 2009; Doll et al., 2017). For example, Doll et al., (2017) found that *Bacillus cereus* sensu lato was found in finished product but was not isolated from the raw milk source. Overall, this suggests that fillers can be sources of gram-positive contamination in single-serve milk in addition to gram-negative contamination; hence, improved strategies to control PPC originating from fillers may also reduce risks of gram-positive contamination events.

Focused filler-level cleaning interventions targeting putative high-risk sources of contamination were not effective at reducing PPC in the 4 study facilities, suggesting the need for more comprehensive interventions

Based on the data and RCA efforts detailed above, we reasoned that a targeted hand cleaning intervention, aimed at the mandrels of single-serve milk fillers, may reduce gram-negative PPC of the single-serve HTST milk products produced in the 4 plants studied. However, the intervention study did not result in a reduction of gram-negative spoilage frequency in samples compared with the standard cleaning practices at each facility. These findings are consistent with other studies (Etter et al., 2017; Reichler et al., 2020) that indicate that relatively narrow targeted interventions often do not yield measurable and significant quality improvements, which, for example, could be due to widespread issues, including hygienic design issues affecting multiple pieces of equipment. For example, one study (Reichler et al., 2020) reported that neither GMP training nor interventions aimed at CIP systems reduced PPC. Reichler et al. (2020), however, did report a reduction of samples spoiled due to PPC following the replacement of rubber gaskets and O-rings on a single-serve filler in one facility, but these results need to be treated with caution as the findings were based on only 3 samples each for pre- and postintervention. In another study, Ralyea et al. (1998) found that replacement of worn-out rubber filler nozzles on a fluid milk rotary filler was able to resolve a severe PPC issue, indicating that RCA and targeted interventions in individual facilities can be successful in reducing PPC.

Importantly, our study, along with other previous studies (Evanowski et al., 2020; Reichler et al., 2020) suggest that more comprehensive interventions are often needed to improve fluid milk quality. For one, comprehensive interventions need to consider different possible contamination sources, as for example, supported by Schröder (1984) who reported that pasteurized milk holding tanks may be the source of contamination for up to 92% of samples, highlighting that gram-negative contamination needs to be addressed at different points after pasteurization. Importantly, our review of the cleaning and sanitation SOP at each facility revealed that (1) most lacked details on how to hand clean parts, including the mandrels, which were cited as a difficult part to clean, (2) the SOP were highly variable between facilities, despite the single-serve fillers being the same make and model at each facility, and (3) employees did not usually follow the SOP. These reports are also consistent with previous recalls (e.g., undeclared allergens) and foodborne illness outbreaks that have been linked to failures in SOP or lack of adherence to SOP (Schmidt and Pierce 2016). These findings suggest that a comprehensive intervention approach may also need to include efforts to develop a stronger food safety/quality culture, including efforts to broadly ensure hygienic equipment design. Further, the fact that gram-negative spoilage frequency was significantly different between facilities further suggests that facility-specific approaches may be required for improved control of PPC. Overall, it is likely that in many cases a combination of factors may be responsible for gram-negative PPC, with one of more sources contributing various concentrations of contamination. Thus, interventions for reducing gram-negative PPC are nontrivial and should consider multiple potential interventions.

## CONCLUSIONS

Although microbial fluid milk quality issues, particularly PPC, remain a major issue across many facilities and packaging sizes, our data suggest that the types of quality challenges may be particularly pronounced with single-serve HTST fluid milk, which is commonly distributed and consumed in schools. With the importance of school milk programs for setting up lifelong fluid milk and dairy product consumption, efforts to improve quality of school and single-serve milk are essential. Our study indicates that comprehensive approaches to milk quality improvement are necessary, addressing issues as diverse as different possible contamination sources, hygienic equipment design, and food quality culture, as well as specific design and cleaning and sanitation issues, such as mandrels and mandrel hubs which are specific to certain fillers.

## ACKNOWLEDGMENTS

This project was funded by the New York Dairy Promotion Order (Albany, NY) through NY State's Sponsor Award Number C012388. The authors thank the participating dairy processing facilities for allowing us to investigate single-serve HTST fluid milk quality, the Cornell Statistical Consulting Unit (Ithaca, NY), especially Erika Mudrak, for guidance on statistical analyses, and Sarah Murphy for help with the development of the survey used at processing facilities. The authors thank the laboratory members from both the Milk Quality Improvement Program (Ithaca, NY) and the Food Safety Laboratory (Ithaca, NY), who assisted with sampling visits, especially Rachel Evanowski and Zoe Wasserlauf who also directed incoming sample processing. The authors acknowledge that N. Martin is a section editor for the *Journal of Dairy Science*. The authors have not stated any other conflicts of interest.

## REFERENCES

Alles, A. A., M. Wiedmann, and N. H. Martin. 2018. Rapid detection and characterization of postpasteurization contaminants in pasteurized fluid milk. J. Dairy Sci. 101:7746–7756. https://doi.org/10.3168/jds.2017-14216.

Alvarez, V. B. 2009. Fluid milk and cream products. Pages 73–133 in The Sensory Evaluation of Dairy Products. S. Clark, M. Costello, M. Drake, and F. Bodyfelt, ed. Springer US, New York, NY.

Aouadhi, C., Z. Rouissi, S. Mejri, and A. Maaroufi. 2014. Inactivation of *Bacillus sporothermodurans* spores by nisin and temperature studied by design of experiments in water and milk. Food Microbiol. 38:270–275. https://doi.org/10.1016/j.fm.2013.10.005.

Belias, A., G. Sullivan, and M. Wiedmann. 2021. Root cause analysis can be used to identify and reduce a highly diverse *Listeria* population in an apple packing house: A case study. Food Protection Trends. Accessed Jun. 1, 2023. https://www.foodprotection.org/publications/food-protection-trends/archive/2021-11-root-cause-analysis-can-be-used-to-identify-and-reduce-a-highly-diverse-listeria-population-/purchase/2021-11-root-cause-analysis-can-be-used-to-identify-and-reduce-a-highly-diverse-listeria-population-/.

Ben-Shachar, M., D. Lüdecke, and D. Makowski. 2020. effectsize: Estimation of effect size indices and standardized parameters. J. Open Source Softw. 5:2815. https://doi.org/10.21105/joss.02815.

Bodyfelt, F. W., M. A. Drake, and S. A. Rankin. 2008. Developments in dairy foods sensory science and education: From student contests to impact on product quality. Int. Dairy J. 18:729–734. https://doi.org/10.1016/j.idairyj.2008.03.011.

Boor, K. J., M. Wiedmann, S. Murphy, and S. Alcaine. 2017. A100-year review: Microbiology and safety of milk handling. J. Dairy Sci. 100:9933–9951. https://doi.org/10.3168/jds.2017-12969.

Buehler, A., N. Martin, K. Boor, and M. Wiedmann. 2018. Psychrotolerant spore-former growth characterization for the development of a dairy spoilage predictive model. J. Dairy Sci. 101:6964–6981. https://doi.org/10.3168/jds.2018-14501.

Carey, N. R., S. C. Murphy, R. N. Zadoks, and K. J. Boor. 2005. Shelf lives of pasteurized fluid milk products in New York State: A ten-year study. Food Prot. Trends 25:102–113.

Champagne, C. P., R. R. Laing, D. Roy, A. A. Mafu, M. W. Griffiths, and C. White. 1994. Psychrotrophs in dairy products: Their effects and their control. Crit. Rev. Food Sci. Nutr. 34:1–30. https://doi.org/10.1080/10408399409527648.

Cohen, J. 1988. Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Routledge, New York, NY.

Cole, J. R., Q. Wang, J. A. Fish, B. Chai, D. M. McGarrell, Y. Sun, C. T. Brown, A. Porras-Alfaro, C. R. Kuske, and J. M. Tiedje. 2014. Ribosomal database project: Data and tools for high throughput rRNA analysis. Nucleic Acids Res. 42(D1):D633–D642. https://doi.org/10.1093/nar/gkt1244.

De Cáceres, M., and P. Legendre. 2009. Associations between species and groups of sites: indices and statistical inference. Ecology 90:3566–3574. https://doi.org/10.1890/08-1823.1.

Deeth, H. C., T. Khusniati, N. Datta, and R. B. Wallace. 2002. Spoilage patterns of skim and whole milks. J. Dairy Res. 69:227–241. https://doi.org/10.1017/S0022029901005301.

Doll, E. V., S. Scherer, and M. Wenning. 2017. Spoilage of microfiltered and pasteurized extended shelf life milk is mainly induced by psychrotolerant spore-forming bacteria that often originate from recontamination. Front. Microbiol. 8:135. https://doi.org/10.3389/fmicb.2017.00135.

Douglas, S. A., M. J. Gray, A. D. Crandall, and K. J. Boor. 2000. Characterization of chocolate milk spoilage patterns. J. Food Prot. 63:516–521. https://doi.org/10.4315/0362-028X-63.4.516.

Drake, M. A. 2007. Invited review: Sensory analysis of dairy foods. J. Dairy Sci. 90:4925–4937. https://doi.org/10.3168/jds.2007-0332.

Duncan, S. E., B. R. Yaun, S. S. Sumner, and J. Bruhn. 2004. Chapter 9: Microbiological methods for dairy products. In Standard Methods for the Examination of Dairy Products. 17th ed. H. M. Wehr and J. F. Frank, ed. American Public Health Association, Washington, DC. https://doi.org/10.2105/9780875530024ch09.

Eneroth, Å., A. Christiansson, J. Brendehaug, and G. Molin. 1998. Critical contamination sites in the production line of pasteurised milk, with reference to the psychrotrophic spoilage flora. Int. Dairy J. 8:829–834. https://doi.org/10.1016/S0958-6946(98)00123-X.

Etter, A. J., S. R. Hammons, S. Roof, C. Simmons, T. Wu, P. W. Cook, A. Katubig, M. J. Stasiewicz, E. Wright, S. Warchocki, J. Hollingworth, H. S. Thesmar, S. A. Ibrahim, M. Wiedmann, and H. F. Oliver. 2017. Enhanced sanitation standard operating procedures have limited impact on *Listeria monocytogenes* prevalence in retail delis. J. Food Prot. 80:1903–1912. https://doi.org/10.4315/0362-028X.JFP-17-112.

Evanowski, R. L., D. J. Kent, M. Wiedmann, and N. H. Martin. 2020. Milking time hygiene interventions on dairy farms reduce spore counts in raw milk. J. Dairy Sci. 103:4088–4099. https://doi.org/10.3168/jds.2019-17499.

Fayet-Moore, F. 2016. Effect of flavored milk vs plain milk on total milk intake and nutrient provision in children. Nutr. Rev. 74:1–17. https://doi.org/10.1093/nutrit/nuv031.

FDA. 2019. Grade "A" Pasteurized Milk Ordinance. Food and Drug Administration (FDA), Washington, DC.

Francis, L. L., S. H. Kong, D. H. Chambers, and I. J. Jeon. 2004. Serving temperature effects on milk flavor, milk aftertaste, and volatile-compound quantification in nonfat and whole milk. Kansas Agric. Exp. Stn. Res. Reports 16–21. https://doi.org/10.4148/2378-5977.3176.

Gruetmacher, T. J., and R. L. Bradley Jr.. 1999. Identification and control of processing variables that affect the quality and safety of fluid milk. J. Food Prot. 62:625–631. https://doi.org/10.4315/0362-028X-62.6.625.

Hayes, W., C. H. White, and M. A. Drake. 2002. Sensory aroma characteristics of milk spoilage by *Pseudomonas* species. J. Food Sci. 67:448–454. https://doi.org/10.1111/j.1365-2621.2002.tb11427.x.

Huck, J. R., B. H. Hammond, S. C. Murphy, N. H. Woodcock, and K. J. Boor. 2007. Tracking spore-forming bacterial contaminants in fluid milk-processing systems. J. Dairy Sci. 90:4872–4883. https://doi.org/10.3168/jds.2007-0196.

Juffs, H. S. 1973. Identification of *Pseudomonas* spp. isolated from milk produced in south eastern Queensland. J. Appl. Bacteriol. 36:585–598. https://doi.org/10.1111/j.1365-2672.1973.tb04145.x.

Kassambara, A. 2021. rstatix: Pipe-friendly framework for basic statistical tests. https://CRAN.R-project.org/package=rstatix.

Keefer, H. M., L. R. Sipple, B. G. Carter, D. M. Barbano, and M. A. Drake. 2022. Children's perceptions of fluid milk with varying levels of milkfat. J. Dairy Sci. 105:3004–3018. https://doi.org/10.3168/jds.2021-20826.

Lau, S., A. Trmcic, N. H. Martin, M. Wiedmann, and S. I. Murphy. 2022. Development of a Monte Carlo simulation model to predict pasteurized fluid milk spoilage due to post-pasteurization contamination with gram-negative bacteria. J. Dairy Sci. 105:1978–1998. https://doi.org/10.3168/jds.2021-21316.

Lenth, R. V. 2021. emmeans: Estimated marginal means, aka least-squares mean. https://CRAN.R-project.org/package=emmeans.

Lima, L. J. R., H. J. Kamphuis, M. J. R. Nout, and M. H. Zwietering. 2011. Microbiota of cocoa powder with particular reference to aerobic thermoresistant spore-formers. Food Microbiol. 28:573–582. https://doi.org/10.1016/j.fm.2010.11.011.

Martin, N. H., K. J. Boor, and M. Wiedmann. 2018. Symposium review: Effect of post-pasteurization contamination on fluid milk quality. J. Dairy Sci. 101:861–870. https://doi.org/10.3168/jds.2017-13339.

Martin, N. H., N. R. Carey, S. C. Murphy, M. Wiedmann, and K. J. Boor. 2012. A decade of improvement: New York State fluid milk quality. J. Dairy Sci. 95:7384–7390. https://doi.org/10.3168/jds.2012-5767.

Martin, N. H., M. L. Ranieri, S. C. Murphy, R. D. Ralyea, M. Wiedmann, and K. J. Boor. 2011a. Results from raw milk microbiological tests do not predict the shelf-life performance of commercially pasteurized fluid milk. J. Dairy Sci. 94:1211–1222. https://doi.org/10.3168/jds.2010-3915.

Martin, N. H., S. C. Murphy, R. D. Ralyea, M. Wiedmann, and K. J. Boor. 2011b. When cheese gets the blues: *Pseudomonas fluorescens* as the causative agent of cheese spoilage. J. Dairy Sci. 94:3176–3183. https://doi.org/10.3168/jds.2011-4312.

McCarthy, K. S., M. Parker, A. Ameerally, S. L. Drake, and M. A. Drake. 2017. Drivers of choice for fluid milk versus plant-based alternatives: What are consumer perceptions of fluid milk? J. Dairy Sci. 100:6125–6138. https://doi.org/10.3168/jds.2016-12519.

Nakamura, A., H. Takahashi, M. Arai, T. Tsuchiya, S. Wada, Y. Fujimoto, Y. Shimabara, T. Kuda, and B. Kimura. 2021. Molecular subtyping for source tracking of *Escherichia coli* using core genome multilocus sequence typing at a food manufacturing plant. PLoS One 16:e0261352. https://doi.org/10.1371/journal.pone.0261352.

Nastasijevic, I., D. Milanov, B. Velebit, V. Djordjevic, C. Swift, A. Painset, and B. Lakicevic. 2017. Tracking of *Listeria monocytogenes* in meat establishment using whole genome sequencing as a food safety management tool: A proof of concept. Int. J. Food Microbiol. 257:157–164. https://doi.org/10.1016/j.ijfoodmicro.2017.06.015.

Navarro, D. 2015. Learning statistics with R: A tutorial for psychology students and other beginners. https://learningstatisticswithr.com.

Oksanen, J., F. G. Blanchet, M. Friendly, R. Kindt, P. Legendre, D. McGlinn, P. R. Minchin, R. B. O'Hara, G. L. Simpon, P. Solymos, H. H. Stevens, E. Szoecs, and H. Wagner. 2020. vegan: Community ecology package. https://CRAN.R-project.org/package=vegan.

Orleans, K. 2011. Microbiological and chemical changes during shelf-life in regular and chocolate milk. Master's thesis. Ohio State University, Columbus, OH. http://rave.ohiolink.edu/etdc/view?acc_num=osu1308253657.

R Core Team. 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Ralyea, R. D., M. Wiedmann, and K. J. Boor. 1998. Bacterial tracking in a dairy production system using phenotypic and ribotyping methods. J. Food Prot. 61:1336–1340. https://doi.org/10.4315/0362-028X-61.10.1336.

Ranieri, M. L., and K. J. Boor. 2009. Short communication: Bacterial ecology of high-temperature, short-time pasteurized milk processed in the United States. J. Dairy Sci. 92:4833–4840. https://doi.org/10.3168/jds.2009-2181.

Ranieri, M. L., J. R. Huck, M. Sonnen, D. M. Barbano, and K. J. Boor. 2009. High temperature, short time pasteurization temperatures inversely affect bacterial numbers during refrigerated storage of pasteurized fluid milk. J. Dairy Sci. 92:4823–4832. https://doi.org/10.3168/jds.2009-2144.

Reichler, S. J., S. I. Murphy, A. W. Erickson, N. H. Martin, A. B. Snyder, and M. Wiedmann. 2020. Interventions designed to control postpasteurization contamination in high-temperature, short-time-pasteurized fluid milk processing facilities: A case study on the effect of employee training, clean-in-place chemical modification, and preventive maintenance programs. J. Dairy Sci. 103:7569–7584. https://doi.org/10.3168/jds.2020-18186.

Reichler, S. J., A. Trmčić, N. H. Martin, K. J. Boor, and M. Wiedmann. 2018. *Pseudomonas fluorescens* group bacterial strains are responsible for repeat and sporadic postpasteurization contamination and reduced fluid milk shelf life. J. Dairy Sci. 101:7780–7800. https://doi.org/10.3168/jds.2018-14438.

RStudio Team. 2022. RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA. http://www.rstudio.com/.

Rush, C. E., J. Johnson, S. Burroughs, B. Riesgaard, A. Torres, L. Meunier-Goddik, and J. Waite-Cusic. 2022. Evaluating *Paenibacillus odorifer* for its potential to reduce shelf life in reworked high-temperature, short-time fluid milk products. JDS Commun. 3:91–96. https://doi.org/10.3168/jdsc.2021-0168.

Schmidt, R. H., and P. D. Pierce. 2016. The use of standard operating procedures (SOPs). Pages 221–233 in Handbook of Hygiene Control in the Food Industry. 2nd ed. H. Lelieveld, J. Holah, and D. Gabrić, ed. Woodhead Publishing.

Schröder, M. J. A. 1984. Origins and levels of post pasteurization contamination of milk in the dairy and their effects on keeping quality. J. Dairy Res. 51:59–67. https://doi.org/10.1017/S0022029900023323.

Schröder, M. J. A., C. M. Cousins, and C. H. McKinnon. 1982. Effect of psychrotrophic post-pasteurization contamination on the keeping quality at 11 and 5°C of HTST-pasteurized milk in the UK. J. Dairy Res. 49:619–630. https://doi.org/10.1017/S0022029900022767.

Schroeder, D. L., S. S. Nielsen, and K. D. Hayes. 2008. The effect of raw milk storage temperature on plasmin activity and plasminogen activation in pasteurized milk. Int. Dairy J. 18:114–119. https://doi.org/10.1016/j.idairyj.2007.08.003.

Sipple, L. R., D. M. Barbano, and M. Drake. 2020. Invited review: Maintaining and growing fluid milk consumption by children in school lunch programs in the United States. J. Dairy Sci. 103:7639–7654. https://doi.org/10.3168/jds.2020-18216.

Sipple, L. R., A. N. Schiano, D. C. Cadwallader, and M. A. Drake. 2021. Child preferences and perceptions of fluid milk in school meal programs. J. Dairy Sci. 104:5303–5318. https://doi.org/10.3168/jds.2020-19546.

Smith, R. S. 1920. Bacterial control in milk plants. J. Dairy Sci. 3:540–554. https://doi.org/10.3168/jds.S0022-0302(20)94297-2.

Stevenson, R. G., M. T. Rowe, G. B. Wisdom, and D. Kilpatrick. 2003. Growth kinetics and hydrolytic enzyme production of *Pseudomonas* spp. isolated from pasteurized milk. J. Dairy Res. 70:293–296. https://doi.org/10.1017/S0022029903006204.

Ternström, A., A.-M. Lindberg, and G. Molin. 1993. Classification of the spoilage flora of raw and pasteurized bovine milk, with special reference to *Pseudomonas* and *Bacillus*. J. Appl. Bacteriol. 75:25–34. https://doi.org/10.1111/j.1365-2672.1993.tb03403.x.

Villamiel, M., and P. de Jong. 2000. Inactivation of *Pseudomonas fluorescens* and *Streptococcus thermophilus* in Trypticase® soy broth and total bacteria in milk by continuous-flow ultrasonic treatment and conventional heating. J. Food Eng. 45:171–179. https://doi.org/10.1016/S0260-8774(00)00059-5.

Wickham, H., R. Francois, L. Henry, and K. Muller. 2021. dplyr: A grammar of data manipulation. https://CRAN.R-project.org/package=dplyr.

Wikcham, H. 2016. Ggplot2: Elegant graphics for data analytics. Springer-Verlag, New York, NY. https://ggplot2.tidyverse.org.

## ORCIDS

T. T. Lott https://orcid.org/0000-0003-2281-2977
A. N. Stelick https://orcid.org/0000-0001-5490-9349
M. Wiedmann https://orcid.org/0000-0002-4168-5662
N. H. Martin https://orcid.org/0000-0003-1704-0634