

Article

Probing Genomic Diversity of *Cronobacter sakazakii* in the United States by Single Nucleotide Polymorphisms

Wei Zhang ^{1,*} , Catherine W. Y. Wong ¹ , Richard Zhang ², Renmao Tian ¹, Behzad Imanian ¹, Yan Li ³ 
and Hongmei Jiang ⁴ 

¹ Institute for Food Safety and Health, Illinois Institute of Technology, Bedford Park, IL 60501, USA; rtian@illinoistech.edu (R.T.); bimanian@illinoistech.edu (B.I.)

² Hinsdale Central High School, Hinsdale, IL 60521, USA; rzhang7036@stu.hinsdale86.org

³ Center for Research Informatics, Bioinformatics Core, University of Chicago, Chicago, IL 60637, USA; liyan@uchicagomedicine.org

⁴ Department of Statistics and Data Science, Northwestern University, Evanston, IL 60208, USA; hongmei@northwestern.edu

* Correspondence: zhangw@illinoistech.edu; Tel.: +1-708-563-2980

Abstract

Cronobacter sakazakii is an opportunistic pathogen commonly associated with powdered infant formula and causes severe neonatal infections. While whole-genome sequencing (WGS)-based single nucleotide polymorphism (SNP) analysis has revolutionized surveillance and outbreak investigations, comprehensive population-level analyses remain limited, and establishing proper thresholds for detecting epidemiologically related *C. sakazakii* isolates requires assessment using large-scale genomic datasets. We analyzed 1870 *C. sakazakii* genomes from the United States (1970–2025) to examine pan- and core-genomic structure, analyze SNP distance matrices encompassing 1,747,515 unique pairwise comparisons, and reconstruct population phylogeny. Our analyses revealed exceptional genomic diversity with a large pan-genome of 24,035 gene families and an average of $29,442 \pm 13,097$ SNPs between genome pairs. Phylogenetic reconstruction identified 22 major clusters encompassing 89.3% of genomes, including environmental complexes demonstrating persistent contamination spanning multiple years. Using 209 monophyletic genome pairs with concordant metadata, we propose a tiered SNP threshold framework (≤ 234 to 506 SNPs) for detecting potentially epidemiologically-related genomes with improved sensitivity. As genomes from Michigan comprised 39.3% of the dataset, these thresholds should be interpreted with caution when applied to other US regions. This study provides population genomics infrastructure to enhance *C. sakazakii* surveillance and traceback studies for improving powdered infant formula safety.

Keywords: *Cronobacter sakazakii*; whole-genome sequencing (WGS); single nucleotide polymorphism (SNP); pan genome; phylogenetic reconstruction; genomic diversity; SNP thresholds; outbreak surveillance; population genomics



Academic Editor: Denis Roy

Received: 5 March 2026

Revised: 2 April 2026

Accepted: 7 April 2026

Published: 9 April 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Cronobacter sakazakii is an opportunistic Gram-negative bacterial pathogen in the family *Enterobacteriaceae*, and can cause severe neonatal infections, including necrotizing enterocolitis, bacteremia, and meningitis, with case-fatality rates of about 40% [1–6]. This organism is frequently isolated from powdered infant formula (PIF) manufacturing environments and

has been implicated in multiple high-profile product recalls, most notably the 2022 PIF recall that led to a nationwide shortage for months [7–9]. *C. sakazakii* possesses a ~4.4–4.6 Mb genome (56–57% G + C) [10,11] encoding virulence factors including outer membrane proteins, iron acquisition systems, and a type VI secretion system [10,12]. Sequence type 4 (ST4) predominates among neonatal meningitis cases [13], while ST1 is dominant in PIF facilities carrying extensive resistance gene repertoires [14,15].

Population genomics, the large-scale analysis of genomic variation across a species, is essential for understanding bacterial evolution, transmission dynamics, and calibrating epidemiological thresholds. Large-scale SNP-based studies of *Salmonella enterica* have elucidated some lineage-specific epidemiological patterns [16,17], population genomic frameworks for *Klebsiella pneumoniae* have mapped the global distribution of high-risk clones [18,19], and analyses of *Escherichia coli* and *Listeria monocytogenes* have revealed deep population structure and persistent contamination spanning years in food environments [20–22]. These studies demonstrate that understanding within-species diversity is essential for distinguishing outbreak clusters from background variations.

Despite some advances across *Enterobacteriaceae*, SNP-based genomic studies of *C. sakazakii* remain limited to small collections from individual outbreaks or specific facilities [23–25], without the population-scale sampling needed to characterize intraspecies diversity to establish robust SNP thresholds. This gap is critical given this pathogen's ubiquitous nature, environmental persistence, extensive horizontal gene transfer, and ecological versatility that may lead to population dynamics distinct from other pathogens [12,13,26,27]. Molecular typing has advanced from MLST (>700 sequence types) [13] to cgMLST schemes [28] and some pan-genome analyses [29,30], yet comprehensive population-scale analysis is still lacking.

Whole-genome sequencing has revolutionized foodborne disease investigations, with US and international agencies now promoting WGS integration into routine disease surveillance and environmental monitoring [31–33]. However, the accuracy of SNP-based inference depends critically on the analytical method chosen, as different approaches can yield substantially different results from the same genomic data. Reference-based pipelines, including the CFSAN SNP Pipeline/CSP [34,35], Snippy [36], Lyve-SET [37], and Parsnp [38] are used to map query genomes against a single reference genome to identify high-confidence variant calls, but are inherently limited to genomic regions present in that reference [39–43]. In contrast, reference-free methods such as kSNP [44] and SKA2 [45] use k-mer decomposition strategies to eliminate reference genome bias and can detect variants across the entire pan-genome. Whole-genome alignment tools like MUMmer [46] offer additional approaches with less stringent computational frameworks. Each method involves different sensitivity-specificity trade-offs that can significantly influence SNP detection rates, phylogenetic reconstruction, and ultimately, epidemiological conclusions about strain relationships [41,47].

Benchmarking studies have consistently shown that different SNP discovery methods produce substantially different absolute counts from identical datasets [41,47]. These methods often vary by several-fold, but generally preserve the relative relationships between isolates [47,48]. This methodological variation has critical implications for surveillance practice: SNP thresholds calibrated with one method may not be reliably applied to results from another method [37,49,50]. Operational thresholds (e.g., <20–25 core-genome SNPs) have been applied in regulatory investigations of several foodborne bacterial pathogens, though these values were derived from limited outbreak datasets and have not been validated across diverse global populations [51–53]. While tiered threshold frameworks offering graded confidence levels have gained support [50], and recent modeling studies have proposed dynamic, species-specific approaches incorporating mutation rates and temporal

parameters [54–56], no consensus exists on optimal methods or thresholds for detecting epidemiologically related strains [50,57]. This lack of standardization is compounded by the diversity of computational pipelines and analytical frameworks employed across laboratories worldwide [42,47,58].

To address these knowledge gaps, this study presents a comprehensive population genomic analysis of 1870 *C. sakazakii* genomes from the United States, which is the largest dataset analyzed for this species to date. We pursued four specific objectives: (1) characterize the pan-genome structure at the species level; (2) perform population-scale total SNP discovery to assess genomic diversity; (3) reconstruct phylogenetic population structure; and (4) develop population-based SNP thresholds for detecting epidemiologically related strains. Here we define epidemiologically related isolates as genome pairs that cluster phylogenetically and share concordant epidemiological metadata across three categories (isolation source, geographic location, and collection timeframe), indicating potential transmission links or common sources of contamination. By integrating phylogenetic evidence with epidemiological metadata, we establish context-appropriate thresholds designed to enhance *C. sakazakii* outbreak surveillance, source attribution studies, and root cause investigations.

2. Materials and Methods

2.1. Genome Dataset and Quality Control

A total of 1870 *C. sakazakii* genome assemblies from the United States were retrieved from NCBI GenBank (<https://www.ncbi.nlm.nih.gov/genbank/>, accessed 1 December 2025) in FASTA format with GCA_ accession numbers. Metadata including isolation source, geographic location, and collection date were extracted from BioSample records. Isolation sources were categorized as Environmental, Infant Formula, Infant Cereal, Clinical, Food/Other, or Unknown. Geographic locations were classified using state names, such as Michigan, Maryland, California, New York, Tennessee, Other USA, or Unknown. Collection dates spanned 1970–2025. Genome quality was assessed through a multi-step filtering pipeline applied to the full initial retrieval of 2974 *C. sakazakii* genome assemblies. As a first-pass filter, assembly-level statistics were evaluated for each genome, including total assembly size, GC content, contig count, and N50 value; assemblies falling outside the expected range for *C. sakazakii* (genome size 3.35–5.13 Mb; GC content 56.0–57.5%) were flagged for exclusion. Genome completeness and contamination were subsequently assessed using PPanGGOLiN v2.2.1 single-copy marker gene analysis [59], which evaluates gene content completeness, identifies redundant or chimeric sequences, and performs gene prediction validation across all input assemblies. This QC workflow was applied uniformly to all 2974 genomes. Assemblies failing completeness thresholds, exhibiting elevated contamination estimates, or showing evidence of fragmentation or assembly artifacts were excluded. Following this filtering process, 1104 low-quality genomes were removed, yielding a final curated dataset of 1870 high-quality genome assemblies that were used for all subsequent pan-genome, phylogenetic, and SNP analyses. For comparative reference, CheckM assessment data were available for 152 genomes as reported by individual submitters to NCBI; these independently confirmed high completeness ($97.8 \pm 1.2\%$) and low contamination ($1.9 \pm 1.1\%$), consistent with the quality criteria applied across the full dataset.

2.2. Pan-Genome Analysis

Pan-genome analysis was conducted using PPanGGOLiN v2.2.1 [59]. Protein-coding sequences were predicted using Prodigal v2.6.3 [60] with bacterial-optimized parameters. Predicted genes were clustered into homologous families using MMseqs2 v14.7e284 [61] with 80% identity and coverage thresholds. Gene families were partitioned using the

Neighbors-Expectation Maximization algorithm into persistent or soft core (95–100% presence), shell (3–95% presence), and cloud (<3% presence) categories. Regions of genomic plasticity (RGPs) were identified as contiguous regions enriched in shell and cloud genes. Pan-genome openness was assessed using rarefaction analysis with 30 random subsampling iterations, and the growth parameter (γ) was calculated using Heaps' law fitting.

2.3. SNP Discovery and Distance Matrix Construction

Total genome single nucleotide polymorphisms were identified using kSNP4.1 [44] with k-mer length $k = 19$, determined by Kchooser optimization. The reference-free k-mer approach eliminated potential reference selection bias. Pairwise SNP distances were calculated for all genome combinations, generating an 1870×1870 symmetric distance matrix. Matrix integrity was validated through symmetry verification and diagonal checks. Pairwise distances were categorized as Very Close (≤ 100 SNPs), Close (101–500 SNPs), Moderate (501–1000 SNPs), Distant (1001–5000 SNPs), or Very Distant (> 5000 SNPs).

2.4. Phylogenetic Reconstruction

Maximum likelihood phylogenetic trees were constructed using RAxML v8.2.12 [62] with the GTRGAMMA substitution model [63]. Bootstrap confidence was assessed using 1000 replicates. Trees were midpoint-rooted for branch length balance. Major phylogenetic clusters were defined as monophyletic groups containing ≥ 50 genomes with bootstrap support ≥ 0.85 .

Principal coordinate analysis (PCoA) was performed on the SNP distance matrix. Optimal population clustering was determined using K-means with silhouette score optimization. PERMANOVA testing evaluated associations between genomic structure and metadata variables. Temporal persistence was assessed by analyzing collection date ranges within clusters. Mantel tests evaluated correlations between genetic distance and temporal/spatial variables.

2.5. Evidence-Based Threshold Development

Epidemiologically related genome pairs were defined using three mandatory criteria: (1) monophyletic relationship (sister taxa in phylogenetic tree), (2) strong bootstrap support ($> 90\%$), and (3) metadata concordance (shared isolation source, geographic location, or collection year). Related pairs were stratified into confidence tiers: Tier 1 (all three metadata criteria concordant), Tier 2 (≥ 2 criteria concordant), Tier 3 (≥ 1 criterion concordant). Optimal SNP thresholds were determined using 95th percentile analysis for each tier to balance 95% sensitivity with specificity. Framework performance was compared against the FDA-proposed threshold (< 21 SNPs) using sensitivity calculations.

2.6. Statistical Analysis and Computational Resources

Statistical analyses were performed using Python v3.10 [64] with NumPy v1.24 [65], pandas v1.5 [66], SciPy v1.11 [67], and scikit-learn v1.3 [68]; specialized correlation and agreement testing used pingouin v0.5.3 [69]. Descriptive statistics such as mean, median, standard deviation, range, coefficient of variation, and interquartile range, were computed for pairwise SNP distance distributions across all 1,747,515 genome comparisons. Because data distributions were non-normal, the Kruskal-Wallis H-test was used to compare SNP distances across sequencing/alignment methods, followed by pairwise Mann-Whitney U tests with Bonferroni correction (adjusted $\alpha = 0.005$). Both Pearson and Spearman correlation coefficients were calculated to assess linear and monotonic relationships between methods, and agreement across methods was quantified using the Intraclass Correlation Coefficient (ICC, two-way random effects model). Population structure was characterized by Principal Coordinate Analysis (PCoA) on the full SNP distance matrix, and optimal

genomic cluster number was determined by K-means clustering with silhouette score optimization. Associations between genomic structure and metadata variables (source, geography, collection year) were evaluated using PERMANOVA, while Mantel tests assessed correlations between genetic distance and collection date. Phylogenetic support was assessed using bootstrap values, with clusters requiring $\geq 85\%$ bootstrap support for structural inference and $>90\%$ for high-confidence pair classification. SNP threshold performance for outbreak detection was evaluated by 95th-percentile analysis within confidence tiers, benchmarked against the FDA-proposed ≤ 21 SNP threshold using sensitivity calculations. All bioinformatics workflows were executed on a high-performance computing system (20-core Intel Xeon, 512 GB RAM), with complete source code, intermediate matrices, and metadata archived under version control to ensure full reproducibility.

Zero-Inflated Mixture Model for SNP Distance Distribution

To characterize the distribution of pairwise SNP distances among closely related genome pairs (<5000 SNPs), a zero-inflated three-component log-normal mixture model was fitted. This approach was necessary because the SNP distance data exhibited two features incompatible with standard parametric distributions: (1) an excess of exact zero values (~ 900 pairs, corresponding to identical genomes), and (2) a right-skewed, multimodal positive distribution reflecting discrete phylogenetic subpopulations within the closely related fraction of the dataset. A standard log-normal or Gaussian mixture without zero-inflation would misrepresent the true density at low SNP distances and thereby distort percentile-based threshold estimates. The model specifies that a pairwise SNP distance X follows:

$$X \sim \pi_0 \cdot \delta_0 + (1 - \pi_0) \cdot \sum_{k=1}^3 w_k \cdot \text{LogNormal}(\mu_k, \sigma_k^2)$$

where δ_0 is a degenerate point mass at zero (assigning probability 1 to $X = 0$), π_0 is the estimated probability of an exact zero, w_k are the mixture weights for the three log-normal components (with $\sum w_k = 1$), and μ_k and σ_k^2 are the mean and variance of the log-transformed distances within each component.

Parameter estimation was carried out in two steps. First, the zero-inflation probability π_0 was estimated empirically as the proportion of exact zero distances in the dataset. Second, conditional on positive distances ($X > 0$), the three-component log-normal mixture parameters $\{w_k, \mu_k, \sigma_k^2\}_{k=1}^3$ were estimated using the Expectation–Maximization (EM) algorithm. All analyses were performed in R v4.5.2 [70] using the mclust package [71], which implements maximum likelihood estimation for Gaussian mixture models via the EM algorithm.

Although the Bayesian Information Criterion (BIC) is the conventional criterion for selecting the number of mixture components, it consistently favored models with a larger number of components in this dataset. This behavior is expected given the very large number of pairwise distances ($n > 300,000$ pairs with SNP distance < 5000) and the well-known tendency of mixture models to exploit additional components to approximate fine-grained distributional features rather than biologically meaningful subgroups. Accordingly, BIC was not used as the sole criterion for model selection. Instead, a parsimonious three-component model was adopted to provide an interpretable and biologically grounded stratification of SNP distances among closely related genome pairs. The three components capture increasing levels of genetic divergence within this closely related fraction: pairs that are nearly identical (very small SNP distances), extremely close but not identical, and moderately close. This stratification directly aligns with the operational goal of characterizing fine-scale genetic similarity in the context of outbreak and contamination source attribution, while avoiding overfitting to noise in the distance distribution.

To confirm that the key quantitative summaries are robust to the choice of the number of components, we compared percentile estimates under $K = 3$ and $K = 5$. Under $K = 3$, approximately 1% of pairwise SNP distances among closely related pairs fall below 2 SNPs, 5% below 22 SNPs, and 10% below 207 SNPs, with a median of 757 SNPs and a 75th percentile of 1138 SNPs. Under $K = 5$, the corresponding estimates are: 1% below 2 SNPs, 5% below 23 SNPs, 10% below 215 SNPs, a median of 754 SNPs, and a 75th percentile of 1150 SNPs. The negligible differences between these two models confirm that the reported percentile thresholds are not sensitive to the choice of K , supporting the validity of the three-component model as the basis for the tiered SNP threshold framework.

3. Results

3.1. Dataset Composition and Characteristics

A comprehensive dataset of 1870 *C. sakazakii* genomes was assembled from NCBI databases, representing the largest *C. sakazakii* collections analyzed to date (Figure 1). All genomes originated from the United States with broad geographic representation across 42 states, though distribution was heterogeneous with Michigan contributing the largest proportion (735 genomes, 39.3%) followed by unspecified U.S. locations (794 genomes, 42.5%). The temporal span covered 55 years (1970–2025), with dramatic increases in genome availability after 2020 coinciding with widespread adoption of whole genome sequencing for surveillance.

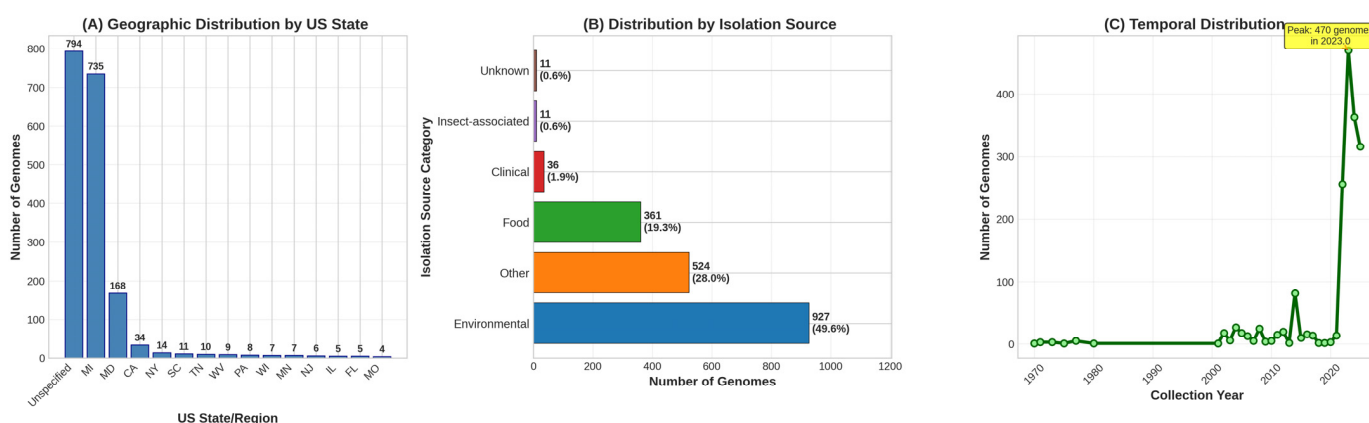


Figure 1. Geographic and source distribution of 1870 *Cronobacter sakazakii* Genomes. **(A)** Geographic distribution by US state. Michigan contributed the largest number of genomes (735, 39.3%), followed by unspecified U.S. locations (794, 42.5%) and Maryland (168, 9.0%). The remaining genomes were distributed across 39 additional states. **(B)** Distribution by isolation source. Environmental samples dominated (927, 49.6%), followed by other/unspecified sources (524, 28.0%), food-related sources (361, 19.3%), clinical specimens (36, 1.9%), and insect-associated isolates (11, 0.6%). **(C)** Temporal distribution. Genome collection from 1970–2025 showing sparse early sampling followed by exponential growth after 2020 with widespread adoption of whole genome sequencing. Peak collection occurred in 2023 (470 genomes).

Assembly quality was assessed across all 1870 genomes using the PPanGGOLiN-based QC pipeline described in Section 2.1. These 1870 assemblies represent the high-quality subset retained from an initial pool of 2974 genomes, following the exclusion of 1104 assemblies that failed completeness, contamination, or assembly integrity thresholds. Mean assembly size was 4.57 ± 0.13 Mb (range: 3.35–5.13 Mb) and GC content was $56.9 \pm 0.2\%$ (range: 56.0–57.5%), consistent with published *C. sakazakii* genome characteristics. The majority of assemblies were contig-level (1833, 98.0%), with scaffold-level (33, 1.8%) and complete genomes (4, 0.2%) also represented. For 152 genomes for which CheckM assessment data were available from NCBI submitters, mean completeness was $97.8 \pm 1.2\%$ and mean

contamination was $1.9 \pm 1.1\%$, independently corroborating the quality of the curated dataset. The taxonomic homogeneity, extensive temporal and geographic coverage, and diverse epidemiological contexts provide an ideal foundation for investigating *C. sakazakii* evolutionary patterns and population structure.

3.2. Pan-Genome Architecture and Core Genome Characteristics

Pan-genome analysis of 1870 *C. sakazakii* genomes yielded 24,035 gene families from 7,855,496 predicted protein-coding sequences (Figure 2). PPanGGOLiN's statistical partitioning classified these into three compartments: persistent (3405 families, 14.2%), shell (3351 families, 13.9%), and cloud (17,279 families, 71.9%) (Table 1). The strict core genome (present in 100% of genomes) comprised only 1033 gene families (4.3%) and represented one of the smallest core genomes documented among *Enterobacteriaceae*. The percentage of soft core genome (14.3%) of *C. sakazakii* was comparable to those of *Salmonella enterica* (12.6%) and *Escherichia coli* (12%) (Table 2). Using the soft core definition ($\geq 95\%$ presence), the core expanded to 3437 families (14.3%), suggesting approximately 2400 additional families are nearly universal but absent in only a small subset of strains due to gene loss, pseudogenization, or assembly artifacts. The percentage of soft core genome estimation in this study (14.3%) based on 1870 *C. sakazakii* genomes was smaller than those in two previous studies (19.5% and 25.2%) based on 237 [29] and 748 [30] genomes, respectively, largely due to the different pan-genomes identified (Table 2).

Table 1. Pan-genome partition statistics and gene family distribution.

Partition	Gene Families	Percentage	Description
Total pan-genome	24,035	100.0%	All gene families
Exact core	1033	4.3%	Present in 100% of genomes
Soft core	3437	14.3%	Present in $\geq 95\%$ of genomes
Shell	3351	13.9%	Accessory genes (3–95%)
Cloud	17,279	71.9%	Rare genes ($<3\%$)

Table 2. Comparative pan-genome architecture of select *Enterobacteriaceae* species.

Species	N Genomes	Pan-Genome	Soft Core	Core %	Reference
<i>K. pneumoniae</i>	328	29,000	1743	6.0%	[72]
<i>C. sakazakii</i>	237	17,158	3346	19.5%	[29]
<i>C. sakazakii</i>	748	13,763	3475	25.2%	[30]
<i>C. sakazakii</i>	1870	24,035	3437	14.3%	This study
<i>E. coli</i>	1324	25,000	3000	12%	[73]
<i>E. coli</i>	2377	7580	2398	31.6%	[74]
<i>S. enterica</i>	339	15,096	3368	22.3%	[75]
<i>S. enterica</i>	4839	25,300	3200	12.6%	[76]

The dominant accessory genome (85.8%) reflects extensive genomic plasticity driven by horizontal gene transfer. Analysis identified 81,495 Regions of genomic plasticity (RGPs) across all genomes (mean: 43.6 ± 8.7 per genome), with 153 integration hotspots where $\geq 50\%$ of RGPs from different genomes overlapped, indicating conserved preferential sites for horizontal gene acquisition. Additionally, 922 functional modules containing 5525 gene families were identified, representing co-localized gene clusters potentially transferred as coordinated units.

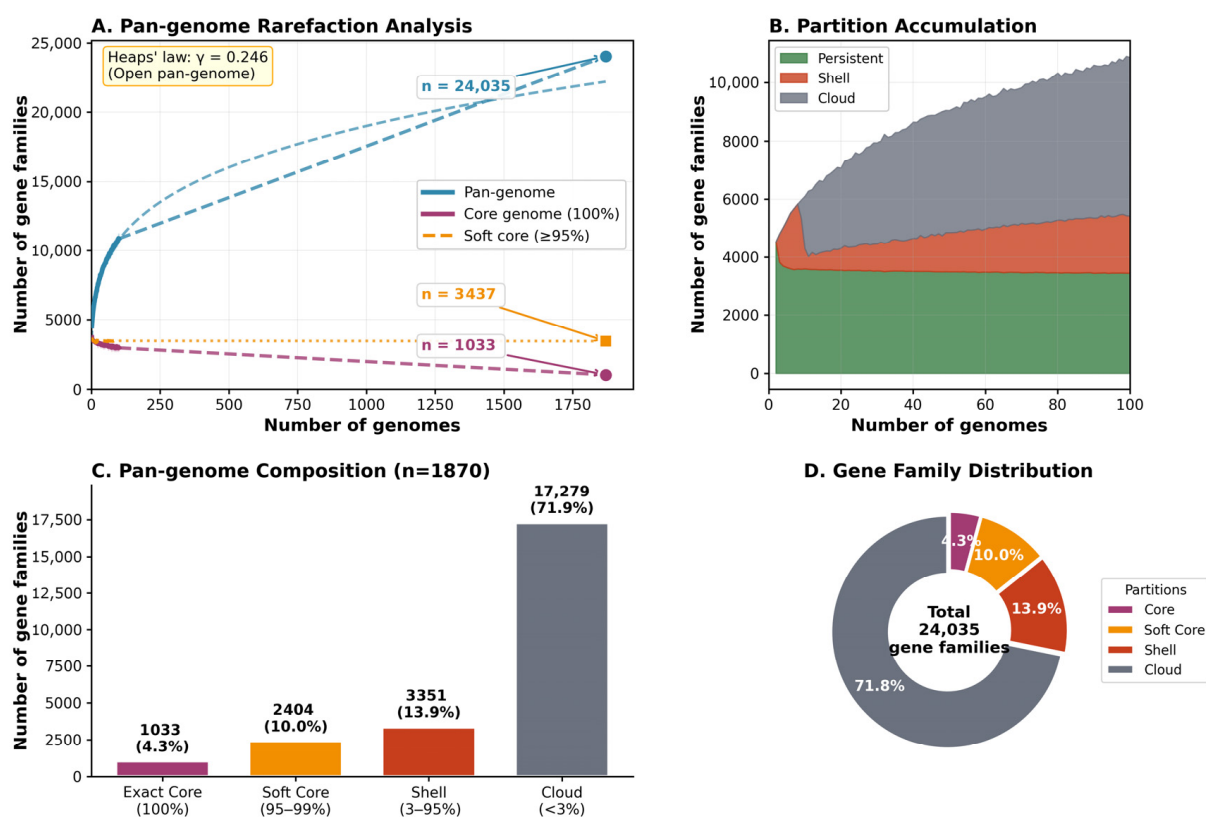


Figure 2. Pan-genome analysis of *Cronobacter sakazakii* (n = 1870 genomes) (A) Pan-genome rarefaction analysis shows pan-genome expansion (blue line, n = 24,035 gene families) and core genome contraction with increasing sample size. Heaps' law fitting ($\gamma = 0.246$) confirms an open pan-genome structure with continued gene discovery. Core genome (100% presence) contains 1033 families, while soft core ($\geq 95\%$ presence) contains 3437 families. (B) Partition accumulation. Stacked area plot showing the accumulation of PPanGGOLiN partitions (persistent, shell, cloud) as genomes are added, demonstrating the dominance of cloud genes in driving pan-genome expansion. (C) Pan-genome composition. Bar chart showing the distribution of 24,035 gene families across partitions: exact core (1033, 4.3%), soft core (2404, 10.0%), shell (3351, 13.9%), and cloud (17,279, 71.9%). (D) Gene family distribution. Pie chart visualizing the proportional composition of pan-genome partitions, highlighting the predominance of cloud genes (71.8%) representing rare, strain-specific content.

Rarefaction analysis confirmed an open pan-genome following Heaps' law ($\gamma = 0.246$), indicating that sequencing additional genomes would continue to reveal novel gene families. Compared with two previous studies [29,30], our expanded dataset revealed 6877 to 10,272 additional gene families while the soft core genome remained largely unchanged. Individual genomes contained an average of 4096 gene families (SD: 167), distributed as ~3380 persistent, ~580 shell, and ~140 cloud genes per genome.

3.3. SNP-Based Population Genomic Analysis

Pan-genome SNP analysis using kSNP4.1 identified a total of 891,621 SNP sites across the 1870-genome dataset (Figure 3). Of these, 159,020 SNP sites (17.83%) were located in the soft core genome (present in $\geq 95\%$ of genomes). The resulting SNP distance matrix contained 1,747,515 unique pairwise comparisons, with SNP distances ranging from 0 to 38,926 substitutions (mean: $29,442 \pm 13,097$; median: 35,278) (Figure 3). Distance distribution was highly skewed (Table 3). Combined, epidemiologically relevant comparisons (≤ 500 SNPs) represented only 4.35% of all relationships, indicating limited clonal expansion within the sampled population.

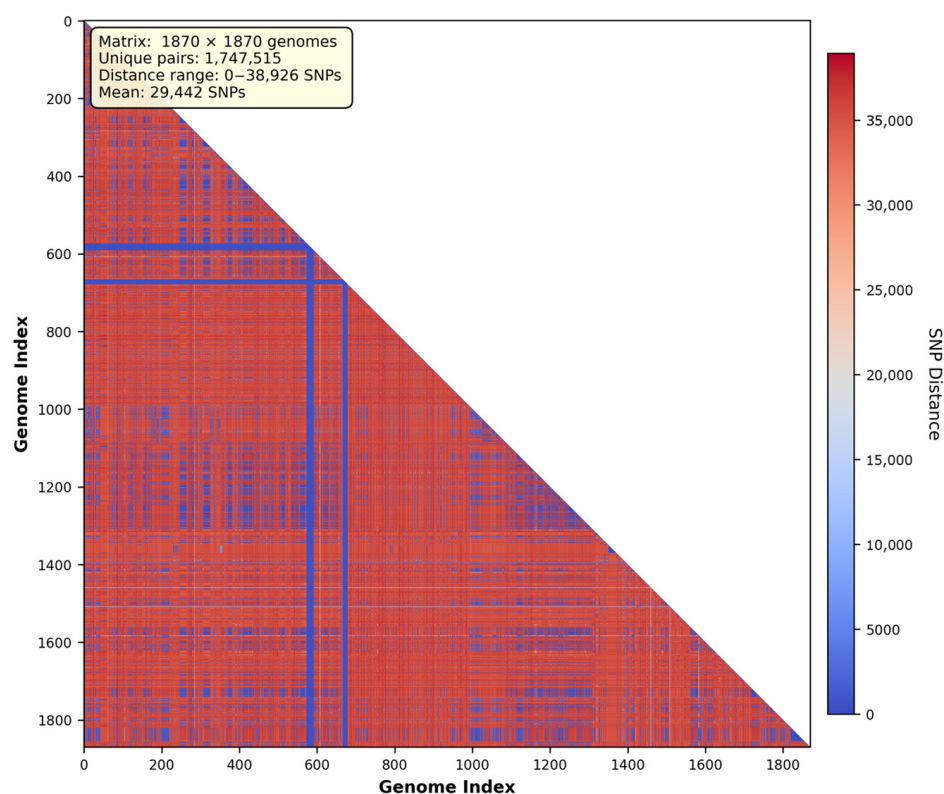


Figure 3. SNP distance matrix—triangular view. Triangular heatmap displaying pairwise SNP distances among 1870 *Cronobacter sakazakii* genomes. The matrix contains 1,747,515 unique pairwise comparisons with SNP distances ranging from 0 to 38,926 substitutions (mean: 29,442 SNPs). Color scale represents SNP distance from blue (low divergence, 0 SNPs) to red (high divergence, >35,000 SNPs). Distinct blue blocks indicate closely related genome clusters, while the predominant red coloration reflects extensive genetic diversity within the species. Vertical and horizontal blue lines at approximately genome index 600 highlight a major phylogenetic division, corresponding to the Michigan environmental complex identified in subsequent phylogenetic analysis.

Table 3. SNP distance distribution categories and population structure statistics.

Category	SNP Range	Pairs	Percentage	Interpretation
Identical	0	927	0.1%	Duplicates or same strains
Very Close	1–100	24,084	1.4%	Potentially recent transmission/contact
Close	101–500	48,910	2.8%	Potentially epidemiologically related—requires context
Moderate	501–1000	134,366	7.7%	Distant relationships
Distant	1001–5000	92,586	5.3%	Distinct lineages
Very Distant	>5000	1,447,569	82.8%	Highly divergent lineages

The histogram of pairwise SNP distances in Figure 4 shows a clear bimodal structure, indicating the presence of two distinct clusters. The cluster concentrated near zero corresponds to pairs of closely related samples with small genetic distances, while the cluster on the right corresponds to pairs of genetically distant samples with large SNP distances.

We further focused on the analysis of genome pairs with <5000 SNPs. Let X denote the pairwise distance. Noticing that there are about 900 values of X equal to 0, we model X using a zero-inflated mixture distribution to accommodate excess zeros and heterogeneous positive values. Specifically, X is assumed to follow a mixture of a point mass at zero

with probability π_0 and a three-component log-normal mixture for positive values with probability $1 - \pi_0$:

$$X \sim \pi_0 \delta_0 + (1 - \pi_0) \sum_{k=1}^3 w_k \text{LogNormal}(\mu_k, \sigma_k^2).$$

Here δ_0 denotes a degenerate distribution at zero, i.e., it assigns probability 1 to the value $X = 0$ and 0 elsewhere. The mixture weights satisfy $\sum_{k=1}^3 w_k = 1$, and μ_k and σ_k^2 represent the mean and variance of the log-transformed positive observations within each component. This formulation allows flexible modeling of skewed positive values while explicitly accounting for zero inflation. Using this mixture model, among the closely related samples, 1% of pairwise SNP distances fall below 2, 5% below 22, 10% below 207, with a median of 757 and a 75th percentile of 1138. Sensitivity analysis comparing $K = 3$ and $K = 5$ component models confirmed that these percentile estimates are robust to the choice of the number of components, with differences of ≤ 8 SNPs across all reported quantiles (see Section “Zero-Inflated Mixture Model for SNP Distance Distribution” for details).

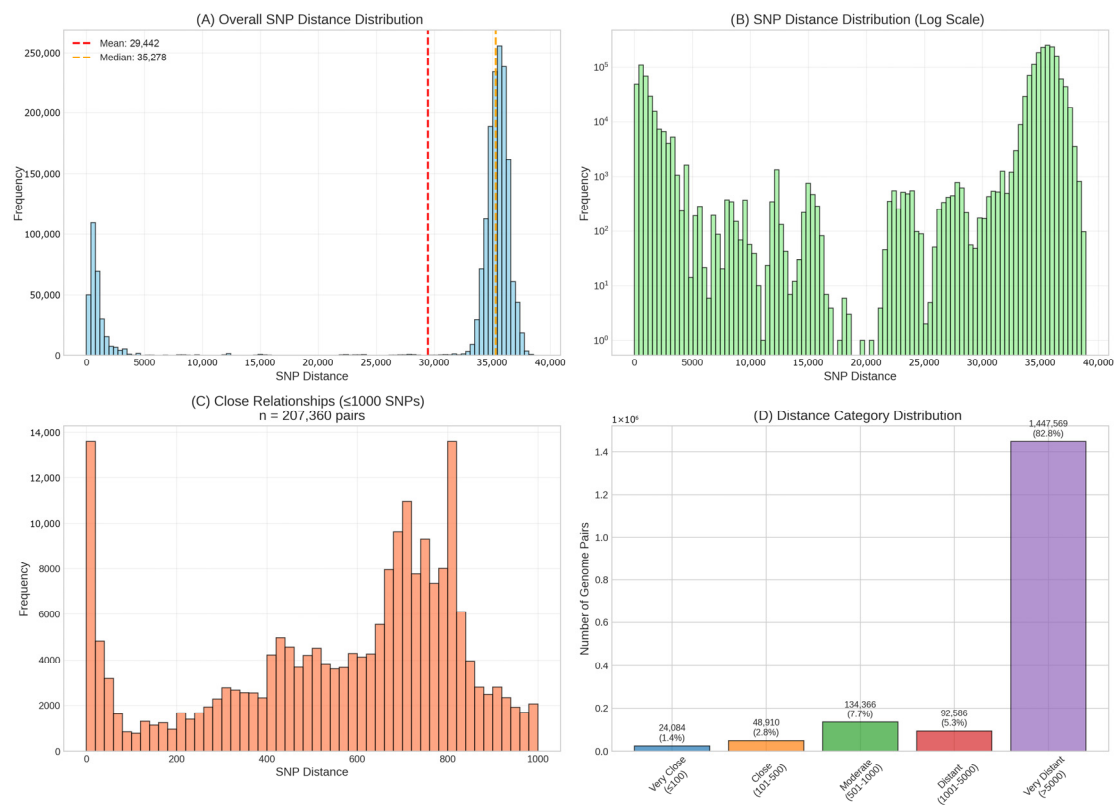


Figure 4. SNP distance distribution analysis. (A) Histogram shows the distribution of 1,747,515 unique pairwise SNP distances with mean 29,442 SNPs (red dashed line) and median 35,278 SNPs (orange dashed line). The bimodal distribution reflects distinct population structure with peaks representing intra- and inter-lineage comparisons. (B) Log-transformed frequency distribution reveals the full range of genetic diversity, from identical genomes (0 SNPs) to highly divergent strains (>35,000 SNPs). The log scale highlights the relative frequency of close relationships amid the predominantly distant comparisons. (C) Detailed view of 207,360 genome pairs with SNP distances ≤ 1000 . The multimodal pattern indicates discrete phylogenetic clusters with varying internal diversity. (D) Categorical breakdown of all pairwise comparisons: identical pairs, very close relationships 1–100 SNPs, close relationships 101–500 SNPs, moderate distances 501–1000 SNPs, and distant or very distant relationships >1000 SNPs.

Principal coordinate analysis explained 65.74% of total variance in the first five dimensions (PC1: 29.33%, PC2: 16.10%, PC3: 7.18%, PC4: 7.05%, PC5: 6.08%), revealing distinct population structure with clear genetic differentiation among major lineages. K-means clustering optimization identified $k = 6$ as optimal with a silhouette score of 0.895, indicating well-defined population structure.

PERMANOVA testing revealed significant associations between genomic variation and epidemiological metadata: isolation source ($F = 19.70$, $\rho = 0.001$, Cramér's $V = 0.380$), geographic state ($F = 12.08$, $\rho = 0.001$, Cramér's $V = 0.359$), and host category ($F = 3.03$, $\rho = 0.003$, Cramér's $V = 0.427$), all with large effect sizes. Temporal analysis via Mantel testing showed no correlation between genetic distance and collection time ($r = 0.0139$, $\rho = 0.316$). Cluster validation identified one cluster with a within/between cluster distance ratio of 3.44 (Mann-Whitney U test: $\rho < 0.001$), indicating significantly tighter genetic clustering within the cluster compared to external genomes.

3.4. Phylogenetic Reconstruction and Population Structure

Maximum likelihood phylogenetic analysis using 891,621 total SNPs revealed 22 major clusters containing ≥ 50 genomes each, encompassing 1669 isolates (89.3% of the dataset) (Figure 5). Bootstrap support exceeded 0.90 for 18 of 22 clusters (81.8%), with 14 clusters (63.6%) achieving perfect support (1.00), confirming robust phylogenetic relationships (Table 4).

Table 4. Major phylogenetic clusters.

Cluster ID	No of Genomes	Max SNP Diameter	Bootstrap Support	Dominant Source (%)	Dominant Location (%)	Date Range
1	150	470,347	1.00	Food/Other (27.3%)	Other USA (72.7%)	1970–N/A
2	136	444,562	1.00	Environment (51.5%)	Other USA (52.9%)	2005–N/A
3	124	510,399	0.56	Environment (51.6%)	Other USA (64.5%)	2009–2025
4	111	38,910	1.00	Environment (88.3%)	Michigan (84.7%)	2002–2025
5	105	58,329	0.86	Environment (68.6%)	Michigan (64.8%)	2001–N/A
6	101	8827	0.98	Environment (98.0%)	Michigan (94.1%)	2022–2025
7	84	53,604	1.00	Environment (58.3%)	Other USA (67.9%)	1973–N/A
8	77	26,409	1.00	Environment (42.9%)	Other USA (75.3%)	2002–N/A
9	67	458,578	1.00	Environment (47.8%)	Other USA (61.2%)	2002–N/A
10	64	35,985	1.00	Environment (76.6%)	Other USA (60.9%)	1971–N/A
11	62	40,122	0.96	Inf. Formula (66.1%)	Other USA (75.8%)	2008–2025
12	59	71,151	1.00	Environment (47.5%)	Other USA (62.7%)	2013–2025
13	57	48,771	0.90	Food/Other (29.8%)	Other USA (66.7%)	2004–N/A
14	56	22,522	1.00	Food/Other (37.5%)	Other USA (92.9%)	1970–N/A
15	55	9362	0.89	Environment (100%)	Michigan (100%)	2022–2025
16	53	117,212	1.00	Environment (56.6%)	Other USA (47.2%)	2004–N/A
17	53	16,387	1.00	Environment (100%)	Michigan (100%)	2022–2025
18	52	39,391	1.00	Environment (48.1%)	Michigan (61.5%)	2001–N/A
19	51	81,529	1.00	Inf. Formula (41.2%)	Other USA (86.3%)	2008–2025
20	51	35,326	1.00	Environment (52.9%)	Other USA (54.9%)	2022–N/A
21	51	37,055	1.00	Environment (66.7%)	Michigan (66.7%)	2004–N/A
22	50	5242	0.94	Environment (100%)	Michigan (100%)	2022–2025

Notes: Michigan Environmental Complex clusters (4, 5, 6, 15, 17, 18, 21, 22) comprising 475 genomes (25.4% of dataset) with 24-year temporal persistence (2001–2025). Infant formula-associated clusters showing distinct epidemiological patterns with broad geographic distribution. N/A indicates missing collection year information on some genomes in the cluster.

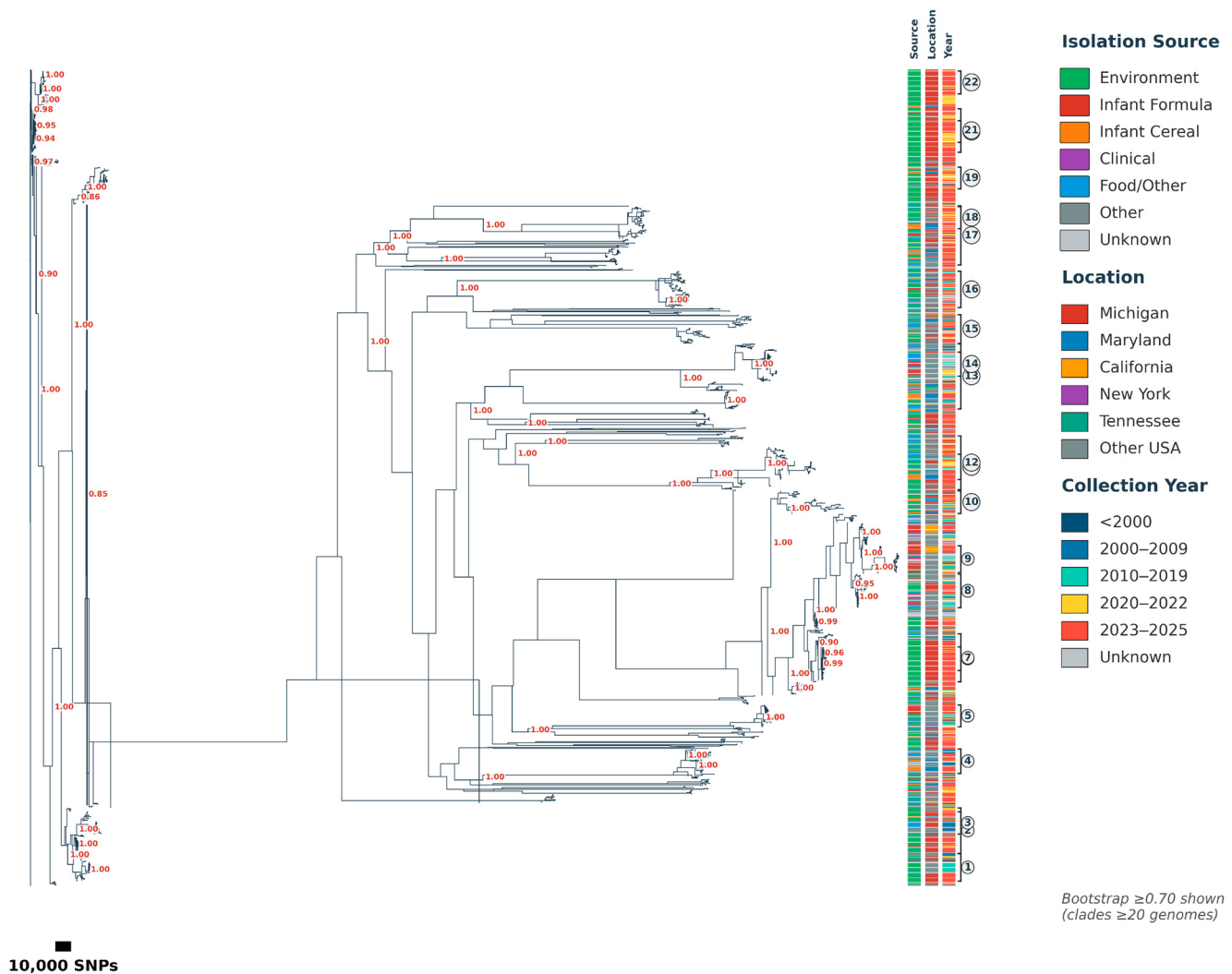


Figure 5. Maximum likelihood phylogenetic tree of 1870 *Cronobacter sakazakii* genomes constructed from 891,621 total SNPs, revealing 22 major clusters (≥ 50 genomes each, numbered 1–22). Bootstrap support values ≥ 0.85 are shown in red at key nodes. Colored strips represent isolation source (green: environmental, red: infant formula, orange: infant cereal, purple: clinical, blue: food/other, gray: other/unknown), geographic location (red: Michigan, blue: Maryland, orange: California, purple: New York, teal: Tennessee, gray: other USA), and collection year (dark blue: <2000, teal: 2000–2009, light blue: 2010–2019, yellow: 2020–2022, orange/red: 2023–2025, gray: unknown). Scale bar represents 10,000 SNPs.

The most prominent finding was a Michigan environmental complex comprising five major clusters ($>80\%$ of genomes from Michigan in clusters 4, 6, 15, 17, 22) containing 475 genomes (25.4% of all isolates), demonstrating temporal persistence spanning 24 years (2001–2025) and strong geographic clustering (475/735 Michigan isolates, 64.6%). Three representative clusters are shown in Figure 6. Cluster 6 ($n = 101$, max diameter = 8827 SNPs) consisted of predominantly environmental isolates (98.0%) from Michigan (94.1%) spanning 2022–2025, showing recent emergence with tight genetic clustering and sustained contamination over multiple years. Cluster 15 ($n = 55$, max diameter = 9362 SNPs) consisted of exclusively environmental isolates (100%) from Michigan (100%) collected 2022–2025, representing a geographically and temporally focused contamination event with moderate internal genetic diversity. Cluster 22 ($n = 50$, max diameter = 5242 SNPs) consisted of all environmental isolates (100%) from Michigan (100%) spanning 2022–2025, showing the most genetically homogeneous cluster with the smallest maximum diameter and in-

dicating recent common ancestry and ongoing transmission. These clusters collectively demonstrate the persistence and genetic stability of *C. sakazakii* populations in Michigan PIF production environments. Infant formula-associated clusters showed distinct patterns: Cluster 11 (62 genomes, 66.1% formula association, broad U.S. distribution) and Cluster 19 (51 genomes, 41.2% formula association). Clinical isolates (n = 42) distributed across multiple clusters rather than forming distinct clades, supporting diverse infection sources and multi-reservoir exposure pathways.

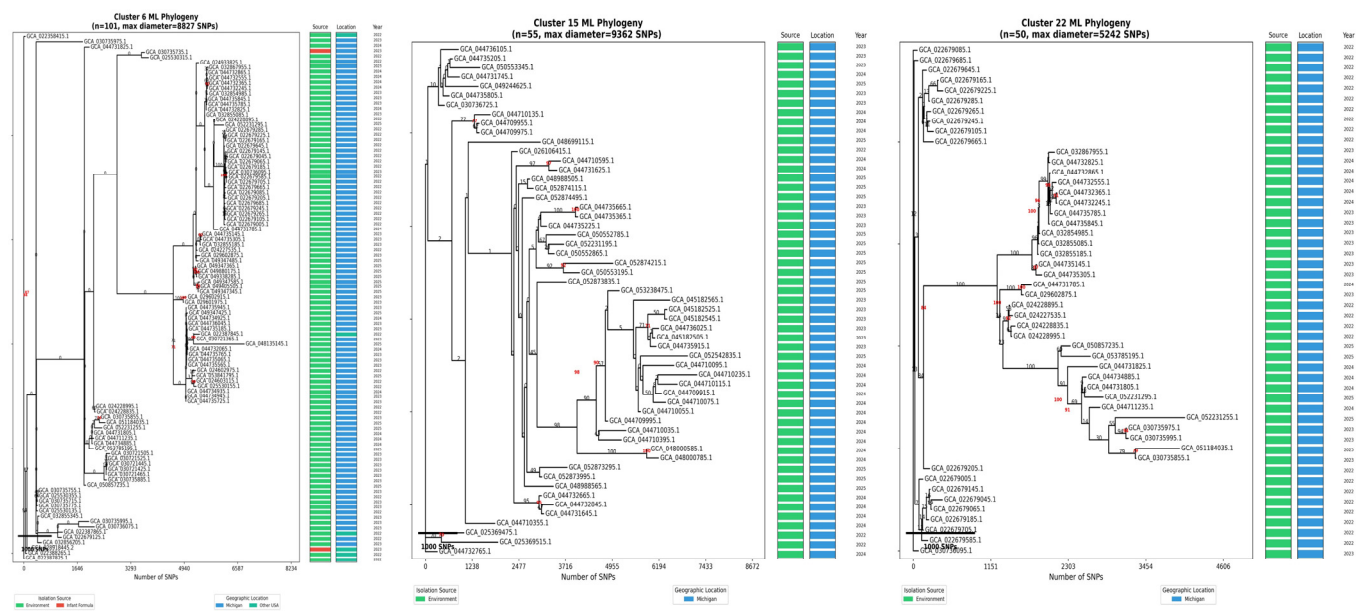


Figure 6. Maximum likelihood phylogenetic trees of three representative clusters from the Michigan environmental complex, demonstrating persistent environmental contamination patterns. Bootstrap support values ≥ 0.7 are shown in red at nodes. Colored bars indicate isolation source (green: environmental; red: infant formula), geographic location (blue: Michigan), and collection year. Scale bars represent SNP distances.

This phylogenetic framework establishes baseline diversity expectations for genomic surveillance, enabling rapid contextualization of new isolates within the population structure and identification of potential transmission routes and cross contamination scenarios.

3.5. Evidence-Based SNP Threshold Framework Development

From 1870 *C. sakazakii* genomes, we identified 372 monophyletic pairs with bootstrap support $>90\%$. Of these, 209 pairs (56.2%) met at least one metadata concordance criterion (shared isolation source, geographic location, or collection year) (Table 5). Stratification by concordant criteria yielded three tiers: Tier 1 (met all three criteria, n = 110), Tier 2 (met two criteria, n = 59), and Tier 3 (met one criterion, n = 40).

Table 5. Tiered SNP threshold framework and performance validation for *Cronobacter sakazakii*.

Tier	Metadata Criteria Match	Pairs (n)	Median SNPs	95th Percentile Threshold	Sensitivity (%)	FDA Sensitivity	Additional Pairs
1	All 3	110	3	≤ 234	95	76.4%	20 pairs
2	≥ 2	169	7	≤ 506	95	66.3%	48 pairs
3	≥ 1	209	10	≤ 498	95	58.9%	76 pairs

Notes: Metadata criteria based on concordance of isolation source, geographic location, and collection year. All three tiers share the same upper range maximum of 1128 SNPs. FDA comparison uses <21 SNP threshold. Additional pairs represent isolates correctly identified by tiered framework, but missed by FDA threshold.

SNP distance distributions differed significantly between tiers (Kruskal-Wallis $p < 0.001$), with Tier 1 showing the lowest median distance (3 SNPs, 95th percentile: 234), followed by Tier 2 (median 7 SNPs, 95th percentile: 506) and Tier 3 (median 10 SNPs, 95th percentile: 498). All tiers displayed substantial ranges extending to 1128 SNPs, reflecting the complex relationship between epidemiological concordance and genetic similarity.

Performance comparison (Figure 7) revealed insufficient sensitivity of the FDA threshold (<21 SNPs): Tier 1 (76.4%), Tier 2 (66.3%), and Tier 3 (58.9%). The proposed tiered framework using 95th percentile thresholds (≤ 234 to 506 SNPs) achieved 95% sensitivity across all categories. Validation using the Michigan environmental cluster complex confirmed framework performance: all within-cluster comparisons from Clusters 15, 17, and 22 fell below Tier 1 thresholds, while the FDA threshold would have missed 67–89% of these relationships.

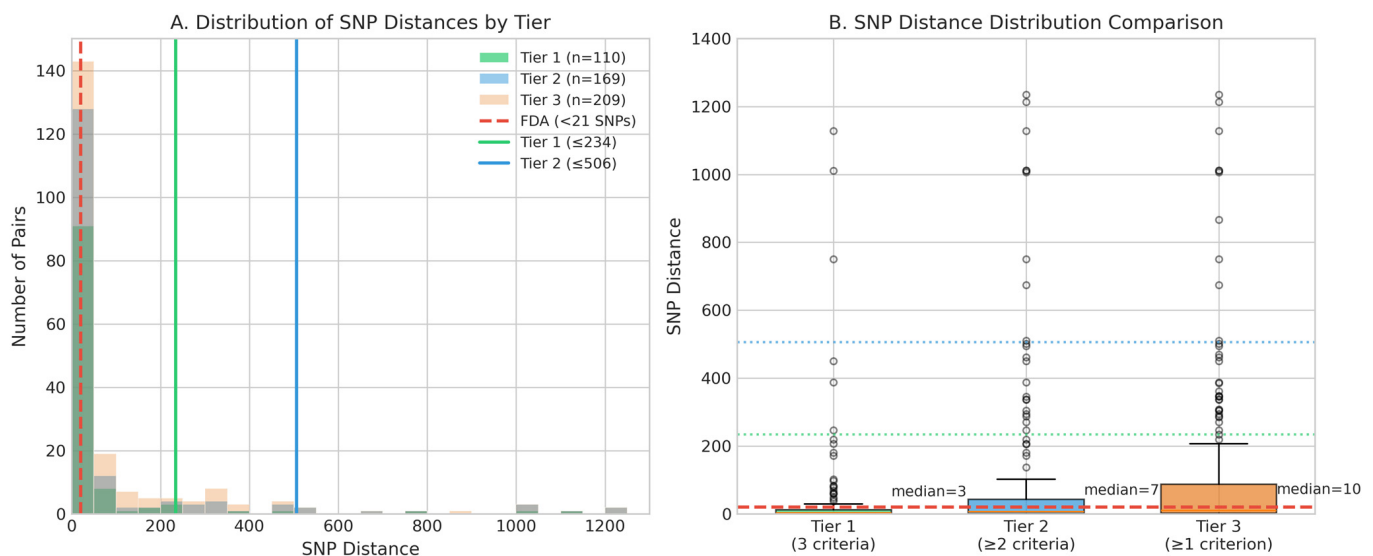


Figure 7. SNP distance distributions and tiered threshold framework. **(A)** Histogram showing SNP distance distributions for monophyletic genome pairs classified into three tiers based on metadata concordance. Tier 1 (green, $n = 110$) represents pairs with all three criteria matching (source + location + year), Tier 2 (blue, $n = 169$) has ≥ 2 criteria matching, and Tier 3 (orange, $n = 209$) has ≥ 1 criterion matching. Vertical lines indicate current FDA threshold (<21 SNPs, red dashed line) and proposed tiered thresholds: Tier 1 (≤ 234 SNPs, green solid line) and Tier 2 (≤ 506 SNPs, blue solid line). Most epidemiologically concordant pairs cluster at low SNP distances, but substantial ranges extend to >1000 SNPs across all tiers. **(B)** Box plots comparing SNP distance distributions across the three tiers. Tier 1 shows the lowest median (3 SNPs), followed by Tier 2 (median = 7 SNPs) and Tier 3 (median = 10 SNPs). Gray dotted lines indicate the 95th percentile thresholds used for the tiered framework (234, 506, and 498 SNPs respectively). The FDA threshold (<21 SNPs, red dashed line) captures only a fraction of epidemiologically related pairs, while the proposed tiered thresholds achieve 95% sensitivity across all confidence levels.

4. Discussion

4.1. Genomic Architecture and Evolutionary Implications

The pan-genome of *C. sakazakii* in this study comprised 24,035 gene families from 7,855,496 predicted genes, and was significantly larger than previously reported pan-genomes of this species [29,30]. The dominance of cloud genes (71.9%) indicates extensive strain-specific content reflecting ongoing horizontal gene transfer and niche adaptation, while the open pan-genome structure ($\gamma = 0.246$) confirms continued evolutionary diversification consistent with ecological plasticity across diverse environmental and clinical niches.

The identification of 81,495 RGPs with 153 integration hotspots provides mechanistic insight into genomic plasticity, aligning with the organism's remarkable environmental persistence and pathogenic versatility. This genomic flexibility likely facilitates survival across conditions ranging from powdered infant formula manufacturing environments to neonatal bloodstream infections. Comparative analysis positions *C. sakazakii* as exhibiting similar genomic diversity to *E. coli* and *S. enterica*, potentially reflecting dual lifestyle requirements for environmental persistence and virulence with the small core genome representing essential functions while the extensive accessory genome enables niche-specific adaptation.

4.2. Population Structure and Epidemiological Significance

The extensive genetic diversity observed (mean $29,442 \pm 13,097$ SNPs) substantially exceeds values reported for many foodborne bacterial pathogens, indicating deep evolutionary divergence within the species. The skewed distance distribution, with 94.4% of comparisons showing >1000 SNPs, demonstrates limited recent clonal expansion and supports SNP-based outbreak investigation utility.

Principal coordinate analysis revealed structured population genetics with 65.74% variance explained in five dimensions, reflecting non-random genetic associations. Significant PERMANOVA associations with isolation source ($F = 19.70, p = 0.001$) and geographic location ($F = 12.08, p = 0.001$) confirm that genetic structure correlates with epidemiological metadata, validating the biological relevance of phylogenetic clustering.

The identification of a Michigan environmental complex comprising 475 genomes (25.4% of the dataset) with 24-year persistence represents compelling evidence for long-term environmental contamination in food production facilities. This finding has immediate regulatory implications, suggesting that environmental monitoring and facility decontamination protocols may require enhancement to address persistent *C. sakazakii* populations.

4.3. Population-Based Tiered Threshold Framework

Current foodborne pathogen outbreak surveillance relies predominantly on single-threshold genetic distance approaches, exemplified by the FDA's < 21 SNP cutoff derived from a limited set of acute outbreak investigations [51]. Such thresholds may lack sufficient sensitivity to detect epidemiologically related isolates that have diverged beyond the threshold boundary despite sharing a common contamination source [50,55]. In the present study, the FDA's < 21 SNP criterion captured only a fraction of epidemiologically related isolates, failing to identify 67–89% of putative relationships within the Michigan environmental complex. This limited sensitivity is an expected consequence of deriving thresholds from acute outbreak scenarios, which may not adequately represent the broader genomic diversity characterizing persistent environmental contamination, particularly in facilities where multiple genetically distinct strains co-circulate over extended timeframes, a pattern well-documented for *C. sakazakii* in powdered infant formula manufacturing environments.

A uniform single-threshold framework is further constrained by its inability to accommodate the varying epidemiological evidence and operational objectives inherent to different surveillance contexts [50,55]. Outbreak confirmation, investigational follow-up, and broad environmental monitoring represent distinct applications with fundamentally different requirements for sensitivity and specificity; applying identical genetic distance criteria across these contexts conflates their distinct objectives and risks both over- and under-attribution of strain relatedness [57].

Traditional criteria for classifying isolates as “epidemiologically related” have been shaped predominantly by acute outbreak investigations, in which relationships are established through confirmed illness clusters, case-control studies, and product traceback [51,77]. While this framework is appropriate for acute public health responses, its

outbreak-centric focus can create systematic surveillance blind spots because it recognizes strain relationships only when linked to detected cases of human illness [78–80]. This framework can also overlook relatedness among isolates sharing a common contamination event that has not yet or may never manifest as a recognized outbreak [31,81,82].

To address this limitation, we propose expanding the definition of epidemiological relatedness to encompass broader patterns of epidemiological concordance. Our metadata-concordance framework operates on the premise that isolates sharing coherent characteristics across isolation source, geographic location, and temporal window may represent epidemiologically meaningful relationships even in the absence of confirmed outbreak linkage, including common contamination events, persistent environmental reservoirs, or transmission pathways not yet associated with detected illness.

This expanded definition serves three critical surveillance needs that outbreak-centric frameworks cannot address [77–79]. First, it enables identification of contamination sources before they manifest as clinical cases, supporting preventive intervention [77,83]. Many contamination events represent “silent” food safety risks that produce no detected illness due to low infectious doses, limited exposure of vulnerable populations, or surveillance gaps [79]. Second, it facilitates detection of persistent environmental contamination within production facilities, particularly relevant for *C. sakazakii* given its well-characterized desiccation tolerance and biofilm formation capacity [84,85]. Third, it accommodates incomplete epidemiological metadata by leveraging available information rather than requiring perfect documentation for a relationship to be recognized [31,49,86,87].

Our population genomics-based tiered framework addresses these limitations by providing context-appropriate guidance calibrated to different levels of epidemiological confidence. The framework stratifies relationships into three tiers: Tier 1 (≤ 234 SNPs) for detecting potential epidemiologically-related genomes requiring further epidemiological investigation and support when isolates share all three metadata criteria; Tier 2 (≤ 506 SNPs) and Tier 3 (≤ 498 SNPs) for supporting broader surveillance monitoring when isolates share limited criteria. Validation using phylogenetically confirmed clusters demonstrated substantial improvements over current approaches, with sensitivity gains of 18.6–36.1% across all confidence levels relative to FDA thresholds. Validation against the Michigan environmental complex was particularly informative: all within-cluster comparisons fell below Tier 1 thresholds, confirming the framework’s capacity to capture epidemiological relationships that single-threshold approaches systematically miss.

This tiered structure balances sensitivity and specificity across the full range of surveillance applications, from acute outbreak response to long-term environmental monitoring. By providing epidemiologically informed, statistically validated thresholds, the framework supports more effective contamination source identification, facility persistence monitoring, and regulatory decision-making support, pending prospective validation with confirmed outbreak data, with direct implications for protecting vulnerable infant populations.

4.4. Study Limitations and Implications for Food Safety Surveillance

This study, while representing the largest *C. sakazakii* genomic dataset to date ($n = 1870$), has several limitations affecting generalizability. The dataset contains exclusively U.S. isolates, which may not fully represent global genomic diversity, and geographic concentration in Michigan (39.3%) and Maryland (9.0%) introduces potential bias in population structure interpretations. The disproportionate representation of Michigan isolates is largely due to recently enhanced environmental monitoring conducted in that region, and threshold estimates may therefore be influenced by this sampling structure.

We used kSNP4 as a reference-free approach for large-scale phylogenomic analysis. The k-mer based methodology, though effective at eliminating reference bias, may classify

positions containing gaps, ambiguous nucleotides (Ns), or assembly artifacts as legitimate SNPs when k-mers are detected in sufficient genomes [20,44]. In large, diverse datasets such as our 1870 *C. sakazakii* genomes spanning multiple highly divergent lineages and various assembly qualities, 77.4% of SNP positions exhibited some missing data across the dataset, reflecting the substantial genomic heterogeneity inherent to intra-species collections [88]. These missing data may inflate evolutionary distances between distantly related genomes and introduce noise into SNP distance calculations, particularly in inter-lineage comparisons where assembly differences are prevalent [41,47]. Consequently, phylogenetic tree topology and epidemiologically relevant SNP thresholds should be interpreted with these dataset characteristics in mind [37,50]. Core genome approaches that restrict SNP calling to universally present, high-quality genomic regions, or the application of stricter data completeness filters, represent complementary strategies for obtaining more conservative estimates of evolutionary relationships [88,89].

Limited metadata availability presents additional constraints. Many isolates in the NCBI database lack detailed source, location, or temporal information, a gap that is particularly pronounced for historical isolates and that reduces precision in epidemiological relationship assessments [57,86,90,91]. Most critically, the absence of validated outbreak linkages between environmental, food, and clinical isolates necessitates reliance on phylogenetic inference and metadata concordance as proxies for true epidemiological relatedness [55]. While the tiered framework provides statistically robust thresholds, this absence of ground-truth validation reduces confidence in threshold precision and represents an inherent limitation of any genomic surveillance framework developed in the absence of prospectively documented outbreak data [37,50,57,82].

This reflects a broader challenge in *C. sakazakii* surveillance: documented outbreaks are rare relative to apparent environmental prevalence, creating fundamental difficulties in establishing evidence-based thresholds through conventional epidemiological validation [26,55,74,92,93]. Future studies incorporating real-time outbreak investigations, ideally with matched environmental, food product, and clinical isolates, would enable empirical threshold refinement and strengthen the evidentiary basis for regulatory adoption.

The present framework relies on genome SNP distances, and integration of additional genomic features could further enhance discriminatory power. Plasmid content, antimicrobial resistance profiles, and virulence gene distributions may improve epidemiological relationship assessments, particularly for closely related isolates where SNP distance alone provides insufficient resolution. Machine learning approaches incorporating multiple genomic features could enable more nuanced classification than pairwise SNP distance. Notably, pan-genomic analyses of accessory gene content, including genes associated with desiccation tolerance, heavy metal resistance, and Type VI secretion systems, have recently been shown to correlate with *C. sakazakii* isolation source and geographic origin [30], suggesting that accessory genome profiles may complement genome SNP analysis in distinguishing environmental persistence from active transmission and in identifying isolates with elevated pathogenic potential.

Despite these limitations, the framework provides essential infrastructure for enhanced surveillance with immediate public health applications in the United States. The tiered thresholds may support more sensitive identification of putative epidemiological clusters and support broader food safety functions including traceback investigations and root cause analysis. Isolates separated by 100–500 SNPs may nonetheless reveal contamination sources, transmission routes, and environmental persistence patterns that would be invisible to single-threshold approaches, enabling identification of silent contamination events before clinical cases occur.

The documentation of persistent environmental contamination spanning decades, exemplified by the Michigan environmental complex, reveals inadequacies in current decontamination protocols [12,25]. The genetic stability of these populations suggests that established *C. sakazakii* communities are substantially more resilient to standard sanitation measures than previously recognized, necessitating enhanced environmental monitoring and more aggressive intervention strategies [27,57,82,85]. Traditional assumptions of effective decontamination following detection may be insufficient given the pathogen's demonstrated capacity for long-term facility persistence and biofilm-mediated survival [12,85,94].

5. Conclusions

This study presents the largest population genomics analysis of *C. sakazakii* conducted to date, encompassing 1870 genomes and establishing a statistically grounded framework for genomic surveillance of this pathogen. Comprehensive characterization of species-level diversity revealed extensive accessory genome plasticity consistent with *C. sakazakii*'s broad ecological range across food production environments, processing facility surfaces, and clinical infections.

The identification of persistent environmental contamination spanning 24 years, exemplified by the Michigan environmental complex, demonstrates that established *C. sakazakii* populations may be substantially more resilient to standard decontamination protocols than current regulatory frameworks assume.

While not intended as a standalone outbreak-confirmation tool, the tiered framework may support the full spectrum of food safety surveillance applications: traceback investigations, root cause analysis, and detection of silent contamination events prior to clinical case recognition. By formally accommodating the reality that epidemiologically related isolates may be separated by 100–500 SNPs as a consequence of environmental persistence and ongoing microevolution, the framework enables a shift from reactive outbreak response toward proactive risk identification, a shift of particular importance given the severity of *C. sakazakii* infections in premature neonates and other vulnerable populations.

Implementation of this population genomic infrastructure provides an actionable foundation for improving contamination source attribution, strengthening evidence-based regulatory decision-making, and ultimately enhancing protection of vulnerable infant populations through earlier detection and more comprehensive genomic risk assessment. However, as Michigan isolates comprised 39.3% of the dataset, prospective validation across geographically diverse US collections is recommended to confirm the generalizability of the proposed SNP thresholds before broader regulatory adoption.

Author Contributions: Conceptualization, W.Z.; methodology, C.W.Y.W., H.J. and W.Z.; software, R.Z. and Y.L.; validation, B.I. and R.T.; formal analysis, W.Z., C.W.Y.W., R.Z., H.J. and Y.L.; investigation, W.Z., R.Z., B.I. and R.T.; resources, W.Z.; data curation, C.W.Y.W., R.Z., R.T. and Y.L.; writing—original draft preparation, W.Z.; writing—review and editing, C.W.Y.W., R.Z., H.J., Y.L., B.I., R.T. and W.Z.; visualization, W.Z., H.J. and R.Z.; supervision, W.Z.; project administration, W.Z.; funding acquisition, W.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Bowen, A.B.; Braden, C.R. Invasive *Enterobacter sakazakii* Disease in Infants. *Emerg. Infect. Dis.* **2006**, *12*, 1185–1189. [CrossRef]
2. Holý, O.; Forsythe, S. *Cronobacter* spp. as Emerging Causes of Healthcare-Associated Infection. *J. Hosp. Infect.* **2014**, *86*, 169–177. [CrossRef] [PubMed]
3. Forsythe, S.J. Updates on the *Cronobacter* Genus. *Annu. Rev. Food Sci. Technol.* **2018**, *9*, 23–44. [CrossRef] [PubMed]
4. Stryko, J.; Cope, J.R.; Martin, H.; Tarr, C.; Hise, K.; Collier, S.; Bowen, A. Food Safety and Invasive *Cronobacter* Infections during Early Infancy, 1961–2018. *Emerg. Infect. Dis.* **2020**, *26*, 857–865. [CrossRef] [PubMed]
5. U.S. Center for Disease Control and Prevention. About *Cronobacter* Infection. Available online: <https://www.cdc.gov/cronobacter/about/index.html> (accessed on 23 March 2026).
6. Government of Canada. *Cronobacter*: For Health Professionals. Available online: <https://www.canada.ca/en/public-health/services/food-poisoning/cronobacter/health-professionals.html> (accessed on 22 February 2026).
7. U.S. Food and Drug Administration. FDA Investigation of *Cronobacter* Infections: Powdered Infant Formula. Available online: <https://www.fda.gov/food/outbreaks-foodborne-illness/fda-investigation-cronobacter-infections-powdered-infant-formula-february-2022> (accessed on 16 February 2026).
8. Centers for Disease Control and Prevention. *Cronobacter* Outbreak Linked to Powdered Infant Formula. Available online: <https://www.cdc.gov/cronobacter/outbreaks/formula-2022/index.html> (accessed on 16 February 2026).
9. Haston, J.C.; Miko, S.; Cope, J.R.; McKeel, H.; Walters, C.; Joseph, L.A.; Griswold, T.; Katz, L.S.; Andújar, A.A.; Tourdot, L.; et al. *Cronobacter sakazakii* Infections in Two Infants Linked to Powdered Infant Formula and Breast Pump Equipment—United States, 2021 and 2022. *MMWR Morb. Mortal. Wkly. Rep.* **2023**, *72*, 223–226. [CrossRef]
10. Kucerova, E.; Clifton, S.W.; Xia, X.-Q.; Long, F.; Porwollik, S.; Fulton, L.; Fronick, C.; Minx, P.; Kyung, K.; Warren, W.; et al. Genome Sequence of *Cronobacter sakazakii* BAA-894 and Comparative Genomic Hybridization Analysis with Other *Cronobacter* Species. *PLoS ONE* **2010**, *5*, e9556. [CrossRef]
11. Chen, Y.; Strain, E.A.; Allard, M.; Brown, E.W. Genome Sequence of *Cronobacter sakazakii* E899, a Strain Associated with Human Illness. *J. Bacteriol.* **2011**, *193*, 5861. [CrossRef]
12. Joseph, S.; Sonbol, H.; Hariri, S.; Desai, P.; McClelland, M.; Forsythe, S.J. Diversity of the *Cronobacter* Genus as Revealed by Multilocus Sequence Typing. *J. Clin. Microbiol.* **2012**, *50*, 3031–3039. [CrossRef]
13. Joseph, S.; Forsythe, S.J. Predominance of *Cronobacter sakazakii* Sequence Type 4 in Neonatal Infections. *Emerg. Infect. Dis.* **2011**, *17*, 1713–1715. [CrossRef]
14. Parra-Flores, J.; Flores-Soto, F.; Flores-Balboa, C.; Alarcón-Lavín, M.P.; Cabal-Rosel, A.; Daza-Prieto, B.; Springer, B.; Cruz-Córdova, A.; Leiva-Caro, J.; Forsythe, S.; et al. Characterization of *Cronobacter sakazakii* and *Cronobacter malonaticus* Strains Isolated from Powdered Dairy Products Intended for Consumption by Adults and Older Adults. *Microorganisms* **2023**, *11*, 2841. [CrossRef]
15. Mousavi, Z.E.; Koolman, L.; Macori, G.; Fanning, S.; Butler, F. Comprehensive Genomic Characterization of *Cronobacter sakazakii* Isolates from Infant Formula Processing Facilities Using Whole-Genome Sequencing. *Microorganisms* **2023**, *11*, 2749. [CrossRef]
16. Deng, X.; Desai, P.T.; den Bakker, H.C.; Mikoleit, M.; Tolar, B.; Trees, E.; Hendriksen, R.S.; Frye, J.G.; Porwollik, S.; Weimer, B.C.; et al. Genomic Epidemiology of *Salmonella enterica* Serotype Enteritidis Based on Population Structure of Prevalent Lineages. *Emerg. Infect. Dis.* **2014**, *20*, 1481–1489. [CrossRef]
17. Liu, C.C.; Hsiao, W.W.L. Large-Scale Comparative Genomics to Refine the Organization of the Global *Salmonella enterica* Population Structure. *Microb. Genom.* **2022**, *8*, 000906. [CrossRef]
18. Wyres, K.L.; Lam, M.M.C.; Holt, K.E. Population Genomics of *Klebsiella pneumoniae*. *Nat. Rev. Microbiol.* **2020**, *18*, 344–359. [CrossRef] [PubMed]
19. Budia-Silva, M.; Kostyanov, T.; Ayala-Montaña, S.; Guéneau, V.; Páslér, M.; García-Castillo, M.; Nüesch-Inderbinen, M.; Hawkey, J.; Goossens, H.; Cantón, R.; et al. International and Regional Spread of Carbapenem-Resistant *Klebsiella pneumoniae* in Europe. *Nat. Commun.* **2024**, *15*, 5092. [CrossRef]
20. Lees, J.A.; Harris, S.R.; Tonkin-Hill, G.; Gladstone, R.A.; Lo, S.W.; Weiser, J.N.; Corander, J.; Bentley, S.D.; Croucher, N.J. Fast and Flexible Bacterial Genomic Epidemiology with PopPUNK. *Genome Res.* **2019**, *29*, 304–316. [CrossRef]
21. Moura, A.; Tourdjman, M.; Leclercq, A.; Hamelin, E.; Laurent, E.; Fredriksen, N.; Van Cauteren, D.; Bracq-Dieye, H.; Thouvenot, P.; Vales, G.; et al. Real-Time Whole-Genome Sequencing for Surveillance of *Listeria monocytogenes*, France. *Emerg. Infect. Dis.* **2017**, *23*, 1462–1470. [CrossRef]
22. Zhang, W.; Qi, W.; Albert, T.J.; Motiwala, A.S.; Alland, D.; Hyttia-Trees, E.K.; Ribot, E.M.; Fields, P.I.; Whittam, T.S.; Swaminathan, B. Probing Genomic Diversity and Evolution of *Escherichia coli* O157 by Single Nucleotide Polymorphisms. *Genome Res.* **2006**, *16*, 757–767. [CrossRef]
23. Jang, H.; Addy, N.; Ewing, L.; Jean-Gilles Beaubrun, J.; Lee, Y.; Woo, J.; Negrete, F.; Finkelstein, S.; Tall, B.D.; Lehner, A.; et al. Whole-Genome Sequences of *Cronobacter sakazakii* Isolates Obtained from Foods of Plant Origin and Dried-Food Manufacturing Environments. *Genome Announc.* **2018**, *6*, e00223-18. [CrossRef] [PubMed]

24. Grim, C.J.; Kotewicz, M.L.; Power, K.A.; Gopinath, G.; Franco, A.A.; Jarvis, K.G.; Yan, Q.Q.; Jackson, S.A.; Sathyamoorthy, V.; Hu, L.; et al. Pan-Genome Analysis of the Emerging Foodborne Pathogen *Cronobacter* spp. Suggests a Species-Level Bidirectional Divergence Driven by Niche Adaptation. *BMC Genom.* **2013**, *14*, 366. [CrossRef] [PubMed]
25. Jang, H.; Chase, H.R.; Gangiredla, J.; Grim, C.J.; Patel, I.R.; Kothary, M.H.; Jackson, S.A.; Mammel, M.K.; Carter, L.; Negrete, F.; et al. Analysis of the Molecular Diversity Among *Cronobacter* Species Isolated from Filth Flies Using Targeted PCR, Pan Genomic DNA Microarray, and Whole Genome Sequencing Analyses. *Front. Microbiol.* **2020**, *11*, 561204. [CrossRef]
26. Iversen, C.; Forsythe, S. Risk Profile of *Enterobacter Sakazakii*, an Emergent Pathogen Associated with Infant Milk Formula. *Trends Food Sci. Technol.* **2003**, *14*, 443–454. [CrossRef]
27. Breeuwer, P.; Lardeau, A.; Peterz, M.; Joosten, H.M. Desiccation and Heat Tolerance of *Enterobacter Sakazakii*. *J. Appl. Microbiol.* **2003**, *95*, 967–973. [CrossRef]
28. Stevens, M.J.A.; Cernela, N.; Stephan, A. Comparative Genomics of *Cronobacter sakazakii* Strains from A Powdered Infant Formula Plant Reveals Evolving Populations. *LWT* **2023**, *184*, 115034. [CrossRef]
29. Lee, I.P.A.; Andam, C.P. Pan-Genome Diversification and Recombination in *Cronobacter Sakazakii*, an Opportunistic Pathogen in Neonates, and Insights to Its Xerotolerant Lifestyle. *BMC Microbiol.* **2019**, *19*, 306. [CrossRef]
30. Gao, M.; Pradhan, A.K.; Blaustein, R.A. Genomic Diversity of *Cronobacter sakazakii* across the Food System to Consumers at the Global Scale. *Int. J. Food Microbiol.* **2025**, *441*, 111335. [CrossRef]
31. Allard, M.W.; Strain, E.; Melka, D.; Bunning, K.; Musser, S.M.; Brown, E.W.; Timme, R. Practical Value of Food Pathogen Traceability through Building a Whole-Genome Sequencing Network and Database. *J. Clin. Microbiol.* **2016**, *54*, 1975–1983. [CrossRef]
32. Imanian, B.; Donaghy, J.; Jackson, T.; Gummalla, S.; Ganesan, B.; Baker, R.C.; Henderson, M.; Butler, E.K.; Hong, Y.; Ring, B.; et al. The Power, Potential, Benefits, and Challenges of Implementing High-Throughput Sequencing in Food Safety Systems. *npj Sci. Foods* **2022**, *6*, 35. [CrossRef] [PubMed]
33. Schadron, T.; Van Den Beld, M.; Mughini-Gras, L.; Franz, E. Use of Whole Genome Sequencing for Surveillance and Control of Foodborne Diseases: Status Quo and Quo Vadis. *Front. Microbiol.* **2024**, *15*, 1460335. [CrossRef]
34. Davis, S.; Pettengill, J.B.; Luo, Y.; Payne, J.; Shpuntoff, A.; Rand, H.; Strain, E. CFSAN SNP Pipeline: An Automated Method for Constructing SNP Matrices from next-Generation Sequence Data. *PeerJ Comput. Sci.* **2015**, *1*, e20. [CrossRef]
35. Litterman, R.; Gangiredla, J.; Rand, H.; Pettengill, J.B. CFSAN SNP Pipeline 2 (CSP2): A Pipeline for Fast and Accurate SNP Distance Estimation from Bacterial Genome Assemblies. *PeerJ Comput. Sci.* **2025**, *11*, e2878. [CrossRef]
36. Seemann, T. SNIPPY. Available online: <https://github.com/tseemann/snippy> (accessed on 17 February 2026).
37. Katz, L.S.; Griswold, T.; Williams-Newkirk, A.J.; Wagner, D.; Petkau, A.; Sieffert, C.; Van Domselaar, G.; Deng, X.; Carleton, H.A. A Comparative Analysis of the Lyve-SET Phylogenomics Pipeline for Genomic Epidemiology of Foodborne Pathogens. *Front. Microbiol.* **2017**, *8*, 375. [CrossRef]
38. Treangen, T.J.; Ondov, B.D.; Koren, S.; Phillippy, A.M. The Harvest Suite for Rapid Core-Genome Alignment and Visualization of Thousands of Intraspecific Microbial Genomes. *Genome Biol.* **2014**, *15*, 524. [CrossRef]
39. Croucher, N.J.; Page, A.J.; Connor, T.R.; Delaney, A.J.; Keane, J.A.; Bentley, S.D.; Parkhill, J.; Harris, S.R. Rapid Phylogenetic Analysis of Large Samples of Recombinant Bacterial Whole Genome Sequences Using Gubbins. *Nucleic Acids Res.* **2015**, *43*, e15. [CrossRef]
40. Bertels, F.; Silander, O.K.; Pachkov, M.; Rainey, P.B.; van Nimwegen, E. Automated Reconstruction of Whole-Genome Phylogenies from Short-Sequence Reads. *Mol. Biol. Evol.* **2014**, *31*, 1077–1088. [CrossRef]
41. Pightling, A.W.; Petronella, N.; Pagotto, F. Choice of Reference Sequence and Assembler for Alignment of *Listeria Monocytogenes* Short-Read Sequence Data Greatly Influences Rates of Error in SNP Analyses. *PLoS ONE* **2014**, *9*, e104579. [CrossRef] [PubMed]
42. Timme, R.E.; Rand, H.; Shumway, M.; Trees, E.K.; Simmons, M.; Agarwala, R.; Davis, S.; Tillman, G.E.; Defibaugh-Chavez, S.; Carleton, H.A.; et al. Benchmark Datasets for Phylogenomic Pipeline Validation, Applications for Foodborne Pathogen Surveillance. *PeerJ* **2017**, *5*, e3893. [CrossRef] [PubMed]
43. Didelot, X.; Bowden, R.; Wilson, D.J.; Peto, T.E.A.; Crook, D.W. Transforming Clinical Microbiology with Bacterial Genome Sequencing. *Nat. Rev. Genet.* **2012**, *13*, 601–612. [CrossRef] [PubMed]
44. Hall, B.G.; Nisbet, J. Building Phylogenetic Trees from Genome Sequences with kSNP4. *Mol. Biol. Evol.* **2023**, *40*, msad235. [CrossRef]
45. Derelle, R.; von Wachsmann, J.; Mäklin, T.; Hellewell, J.; Russell, T.; Lavani, A.; Chindelevitch, L.; Croucher, N.J.; Harris, S.R.; Lees, J.A. Seamless, Rapid, and Accurate Analyses of Outbreak Genomic Data Using Split k-Mer Analysis. *Genome Res.* **2024**, *34*, 1661–1673. [CrossRef]
46. Kurtz, S.; Phillippy, A.; Delcher, A.L.; Smoot, M.; Shumway, M.; Antonescu, C.; Salzberg, S.L. Versatile and Open Software for Comparing Large Genomes. *Genome Biol.* **2004**, *5*, R12. [CrossRef]

47. Bush, S.J.; Foster, D.; Eyre, D.W.; Clark, E.L.; De Maio, N.; Shaw, L.P.; Stoesser, N.; Peto, T.E.A.; Crook, D.W.; Walker, A.S. Genomic Diversity Affects the Accuracy of Bacterial Single-Nucleotide Polymorphism–Calling Pipelines. *Gigascience* **2020**, *9*, giaa007. [CrossRef]
48. Yoshimura, D.; Kajitani, R.; Gotoh, Y.; Katahira, K.; Okuno, M.; Ogura, Y.; Hayashi, T.; Itoh, T. Evaluation of SNP Calling Methods for Closely Related Bacterial Isolates and a Novel High-Accuracy Pipeline: BactSNP. *Microb. Genom.* **2019**, *5*, e000261. [CrossRef]
49. Moura, A.; Criscuolo, A.; Pouseele, H.; Maury, M.M.; Leclercq, A.; Tarr, C.; Björkman, J.T.; Dallman, T.; Reimer, A.; Enouf, V.; et al. Whole Genome-Based Population Biology and Epidemiological Surveillance of *Listeria monocytogenes*. *Nat. Microbiol.* **2016**, *2*, 16185. [CrossRef]
50. Schürch, A.C.; Arredondo-Alonso, S.; Willems, R.J.L.; Goering, R.V. Whole Genome Sequencing Options for Bacterial Strain Typing and Epidemiologic Analysis Based on Single Nucleotide Polymorphism versus Gene-by-Gene–Based Approaches. *Clin. Microbiol. Infect.* **2018**, *24*, 350–354. [CrossRef]
51. Pightling, A.W.; Pettengill, J.B.; Luo, Y.; Baugher, J.D.; Rand, H.; Strain, E. Interpreting Whole-Genome Sequence Analyses of Foodborne Bacteria for Regulatory Applications and Outbreak Investigations. *Front. Microbiol.* **2018**, *9*, 1482. [CrossRef] [PubMed]
52. Masood, N.; Moore, K.; Farbos, A.; Paszkiewicz, K.; Dickins, B.; McNally, A.; Forsythe, S. Genomic Dissection of the 1994 *Cronobacter sakazakii* Outbreak in a French Neonatal Intensive Care Unit. *BMC Genom.* **2015**, *16*, 750. [CrossRef] [PubMed]
53. Allard, M.W.; Strain, E.; Rand, H.; Melka, D.; Correll, W.A.; Hintz, L.; Stevens, E.; Timme, R.; Lomonaco, S.; Chen, Y.; et al. Whole Genome Sequencing Uses for Foodborne Contamination and Compliance: Discovery of an Emerging Contamination Event in an Ice Cream Facility Using Whole Genome Sequencing. *Infect. Genet. Evol.* **2019**, *73*, 214–220. [CrossRef] [PubMed]
54. Didelot, X.; Croucher, N.J.; Bentley, S.D.; Harris, S.R.; Wilson, D.J. Bayesian Inference of Ancestral Dates on Bacterial Phylogenetic Trees. *Nucleic Acids Res.* **2018**, *46*, e134. [CrossRef]
55. Didelot, X.; Fraser, C.; Gardy, J.; Colijn, C. Genomic Infectious Disease Epidemiology in Partially Sampled and Ongoing Outbreaks. *Mol. Biol. Evol.* **2017**, *34*, msw075. [CrossRef]
56. Campbell, F.; Didelot, X.; Fitzjohn, R.; Ferguson, N.; Cori, A.; Jombart, T. Outbreaker2: A Modular Platform for Outbreak Reconstruction. *BMC Bioinform.* **2018**, *19*, 363. [CrossRef] [PubMed]
57. Nadon, C.; Van Walle, I.; Gerner-Smidt, P.; Campos, J.; Chinen, I.; Concepcion-Acevedo, J.; Gilpin, B.; Smith, A.M.; Kam, K.M.; Perez, E.; et al. PulseNet International: Vision for the Implementation of Whole Genome Sequencing (WGS) for Global Food-Borne Disease Surveillance. *Eurosurveillance* **2017**, *22*, 30544. [CrossRef] [PubMed]
58. De Maio, N.; Ly-Trong, N.; Martin, S.; Minh, B.Q.; Goldman, N. Assessing Phylogenetic Confidence at Pandemic Scales. *Nature* **2025**, *647*, 472–478. [CrossRef] [PubMed]
59. Gautreau, G.; Bazin, A.; Gachet, M.; Planel, R.; Burlot, L.; Dubois, M.; Perrin, A.; Médigue, C.; Calteau, A.; Cruveiller, S.; et al. PPanGGOLiN: Depicting Microbial Diversity via a Partitioned Pangenome Graph. *PLoS Comput. Biol.* **2020**, *16*, e1007732. [CrossRef]
60. Hyatt, D.; Chen, G.-L.; LoCascio, P.F.; Land, M.L.; Larimer, F.W.; Hauser, L.J. Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification. *BMC Bioinform.* **2010**, *11*, 119. [CrossRef]
61. Steinegger, M.; Söding, J. MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets. *Nat. Biotechnol.* **2017**, *35*, 1026–1028. [CrossRef]
62. Stamatakis, A. RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. *Bioinformatics* **2014**, *30*, 1312–1313. [CrossRef]
63. Barba-Montoya, J.; Tao, Q.; Kumar, S. Using a GTR+ Γ Substitution Model for Dating Sequence Divergence When Stationarity and Time-Reversibility Assumptions Are Violated. *Bioinformatics* **2020**, *36*, i884–i894. [CrossRef]
64. Python Release Python 3.10.10. Available online: <https://www.python.org/downloads/release/python-31010/> (accessed on 23 February 2026).
65. Python NumPy: The Fundamental Package for Scientific Computing with Python. Available online: <https://numpy.org/> (accessed on 23 February 2026).
66. The Pandas Development Team Pandas. Available online: <https://zenodo.org/records/10537285> (accessed on 24 February 2026).
67. Python SciPy: Fundamental Algorithms for Scientific Computing in Python. Available online: <https://scipy.org/> (accessed on 23 February 2026).
68. Cournapeau, D. Sci-Kit Learn v1.3. Available online: https://scikit-learn.org/stable/whats_new/v1.3.html (accessed on 22 March 2026).
69. Vallat Pingouin. Available online: <https://github.com/raphaelvallat/pingouin> (accessed on 23 February 2026).
70. R: The R Project for Statistical Computing. Available online: <https://www.r-project.org/> (accessed on 22 March 2026).
71. Scrucca, L.; Fraley, C.; Murphy, T.B.; Raftery, A.E. *Model-Based Clustering, Classification, and Density Estimation Using mclust in R*; Chapman & Hall/CRC: Boca Raton, FL, USA, 2023. Available online: <https://mclust-org.github.io/book/> (accessed on 1 December 2025).
72. Holt, K.E.; Wertheim, H.; Zadoks, R.N.; Baker, S.; Whitehouse, C.A.; Dance, D.; Jenney, A.; Connor, T.R.; Hsu, L.Y.; Severin, J.; et al. Genomic Analysis of Diversity, Population Structure, Virulence, and Antimicrobial Resistance in *Klebsiella pneumoniae*, an Urgent Threat to Public Health. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, E3574–E3581. [CrossRef]

73. Tantoso, E.; Eisenhaber, B.; Kirsch, M.; Shitov, V.; Zhao, Z.; Eisenhaber, F. To Kill or to Be Killed: Pangenome Analysis of *Escherichia coli* Strains Reveals a Tailocin Specific for Pandemic ST131. *BMC Biol.* **2022**, *20*, 146. [\[CrossRef\]](#)
74. Chauhan, S.M.; Ardalani, O.; Hyun, J.C.; Monk, J.M.; Phaneuf, P.V.; Palsson, B.O. Decomposition of the Pangenome Matrix Reveals a Structure in Gene Distribution in the *Escherichia coli* Species. *mSphere* **2025**, *10*, e00532-24. [\[CrossRef\]](#) [\[PubMed\]](#)
75. Coluzzi, C.; Pison, B.; Dérozier, S.; Chiapello, H.; Gal-Mor, O. Comparative Genomics of *Salmonella enterica* Serovars Paratyphi A, Typhi and Typhimurium Reveals Distinct Profiles of Their Pangenome, Mobile Genetic Elements, Antimicrobial Resistance and Defense Systems Repertoire. *Virulence* **2025**, *16*, 2504658. [\[CrossRef\]](#)
76. Laing, C.R.; Whiteside, M.D.; Gannon, V.P.J. Pan-Genome Analyses of the Species *Salmonella enterica*, and Identification of Genomic Markers Predictive for Species, Subspecies, and Serovar. *Front. Microbiol.* **2017**, *8*, 1345. [\[CrossRef\]](#)
77. Jackson, B.R.; Tarr, C.; Strain, E.; Jackson, K.A.; Conrad, A.; Carleton, H.; Katz, L.S.; Stroika, S.; Gould, L.H.; Mody, R.K.; et al. Implementation of Nationwide Real-Time Whole-Genome Sequencing to Enhance Listeriosis Outbreak Detection and Investigation. *Clin. Infect. Dis.* **2016**, *63*, 380–386. [\[CrossRef\]](#)
78. Scallan, E.; Hoekstra, R.M.; Angulo, F.J.; Tauxe, R.V.; Widdowson, M.-A.; Roy, S.L.; Jones, J.L.; Griffin, P.M. Foodborne Illness Acquired in the United States—Major Pathogens. *Emerg. Infect. Dis.* **2011**, *17*, 7–15. [\[CrossRef\]](#)
79. Hoffmann, S.; White, A.E.; McQueen, R.B.; Ahn, J.-W.; Gunn-Sandell, L.B.; Scallan, E.J. Economic Burden of Foodborne Illnesses Acquired in the United States. *Foodborne Pathog. Dis.* **2025**, *22*, 4–14. [\[CrossRef\]](#)
80. Painter, J.A.; Hoekstra, R.M.; Ayers, T.; Tauxe, R.V.; Braden, C.R.; Angulo, F.J.; Griffin, P.M. Attribution of Foodborne Illnesses, Hospitalizations, and Deaths to Food Commodities by Using Outbreak Data, United States, 1998–2008. *Emerg. Infect. Dis.* **2013**, *19*, 407–415. [\[CrossRef\]](#) [\[PubMed\]](#)
81. Carpentier, B.; Cerf, O. Review—Persistence of *Listeria monocytogenes* in Food Industry Equipment and Premises. *Int. J. Food Microbiol.* **2011**, *145*, 1–8. [\[CrossRef\]](#) [\[PubMed\]](#)
82. Struelens, M.J.; Brisse, S. From Molecular to Genomic Epidemiology: Transforming Surveillance and Control of Infectious Diseases. *Eurosurveillance* **2013**, *18*, 20386. [\[CrossRef\]](#)
83. Charleston, H.A.; Gerner-Smidt, P. Whole-Genome Sequencing Is Taking over Foodborne Disease Surveillance: Public Health Microbiology Is Undergoing Its Biggest Change in a Generation, Replacing Traditional Methods with Whole-Genome Sequencing. *Microbe Mag.* **2016**, *11*, 311–317. [\[CrossRef\]](#)
84. Iversen, C.; Lane, M.; Forsythe, S.J. The Growth Profile, Thermotolerance and Biofilm Formation of *Enterobacter sakazakii* Grown in Infant Formula Milk. *Lett. Appl. Microbiol.* **2004**, *38*, 378–382. [\[CrossRef\]](#) [\[PubMed\]](#)
85. Lehner, A.; Riedel, K.; Eberl, L.; Breeuwer, P.; Diep, B.; Stephan, R. Biofilm Formation, Extracellular Polysaccharide Production, and Cell-to-Cell Signaling in Various *Enterobacter sakazakii* Strains: Aspects Promoting Environmental Persistence. *J. Food Prot.* **2005**, *68*, 2287–2294. [\[CrossRef\]](#)
86. Grad, Y.H.; Lipsitch, M. Epidemiologic Data and Pathogen Genome Sequences: A Powerful Synergy for Public Health. *Genome Biol.* **2014**, *15*, 538. [\[CrossRef\]](#)
87. Revez, J.; Espinosa, L.; Albiger, B.; Leitmeyer, K.C.; Struelens, M.J. Survey on the Use of Whole-Genome Sequencing for Infectious Diseases Surveillance: Rapid Expansion of European National Capacities, 2015–2016. *Front. Public Health* **2017**, *5*, 347. [\[CrossRef\]](#)
88. Page, A.J.; Cummins, C.A.; Hunt, M.; Wong, V.K.; Reuter, S.; Holden, M.T.G.; Fookes, M.; Falush, D.; Keane, J.A.; Parkhill, J. Roary: Rapid Large-Scale Prokaryote Pan Genome Analysis. *Bioinformatics* **2015**, *31*, 3691–3693. [\[CrossRef\]](#)
89. Didelot, X.; Wilson, D.J. ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. *PLoS Comput. Biol.* **2015**, *11*, e1004041. [\[CrossRef\]](#)
90. Yozwiak, N.L.; Schaffner, S.F.; Sabeti, P.C. Data Sharing: Make Outbreak Research Open Access. *Nature* **2015**, *518*, 477–479. [\[CrossRef\]](#)
91. Sinha, R.; Abnet, C.C.; White, O.; Knight, R.; Huttenhower, C. The Microbiome Quality Control Project: Baseline Study Design and Future Directions. *Genome Biol.* **2015**, *16*, 276. [\[CrossRef\]](#) [\[PubMed\]](#)
92. World Health Organization. *Enterobacter sakazakii* (*Cronobacter* spp.) in Powdered Follow-Up Formula: Meeting Report; World Health Organization: Geneva, Switzerland, 2008.
93. World Health Organization. *Enterobacter sakazakii* and Other Microorganisms in Powdered Infant Formula: Meeting Report; World Health Organization: Geneva, Switzerland, 2004.
94. Bridier, A.; Briandet, R.; Thomas, V.; Dubois-Brissonnet, F. Resistance of Bacterial Biofilms to Disinfectants: A Review. *Biofouling* **2011**, *27*, 1017–1032. [\[CrossRef\]](#) [\[PubMed\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.