NCBI Pathogen Detection updates on Oxford Nanopore sequence

Arjun Prasad



U.S. National Library of Medicine
National Center for Biotechnology Information

Genomic Surveillance

Oxford Nanopore (ONT) for Bacterial Isolates: Benefits & Challenges

STRENGTHS

Large contigs

Spans repetitive regions to often achieve complete, single-contig assemblies.

Plasmid & AMR

Accurately resolves plasmids and locates resistance genes within mobilome structures.

Low Entry Cost

Low hardware capital barrier enables decentralized adoption in public health labs.

CHALLENGES

Higher Base Error Rates

Errors can obscure fine-scale SNP cluster matching in outbreak tracing.

Rapid Evolution

Frequent model and flow-cell revisions complicate methods development and quality assessments.

Platform-Agnostic Specification

5 FIELDS • ALL REQUIRED

Long read metadata fields for SRA Also used for PacBio

flowcell_type	Flow cell (ONT) or SMRT Cell (PacBio) used for sequencing	Required
library_preparation_kit	Kit used to build the sequencing library	Required
library_preparation_version	Version / chemistry / release of the prep kit or protocol	Required
basecalling_software_version	Version of the basecalling software used to generate submitted reads	Required
on_instrument_basecalling	Whether basecalling ran on-instrument (MinKNOW/SMRT Link) vs. external pipeline	Required

Developed by the GenFS long-reads working groups and Jess Chen

Platform Specific

3 FIELDS

Oxford Nanopore Technologies (ONT)

ont_basecalling_software	Converts raw ionic current signals to sequence Specific basecaller algorithm (Dorado, Guppy, or Bonito)	Required
ont_basecalling_model_string	Encodes pore type, chemistry, translocation speed, and model version Full model descriptor from FASTQ header	Required
ont_modification_model	Documents epigenetic callers when applied Modified-base detection model used (e.g., 5mC, 6mA)	Conditional

Developed by the GenFS long-reads working groups and Jess Chen

What we're testing so far

Assembler

- Stylo assembly pipeline from CDC
  - https://github.com/ncezid-biome/stylo
  - `nanoq`, `rasusa`, `flye`, `dnaapler`, `dorado`, `quast`, `multiqc`

Two datasets provided by GenFS collaborators

- `narjol` – provided by Narjol Gonzalez-Escalona at FDA and used for his validation
- `biggel` – Suggested by Jess Chen at CDC
  - From bioRxiv preprint Biggel et al. 2026
  - From University of Zurich, CDC are co-authors

biggel Institute for Food Safety and Hygiene, University of Zurich

From Biggel et al., 2026 - Standalone nanopore sequencing for foodborne pathogen surveillance: a large-scale evaluation and quality control framework https://www.biorxiv.org/content/10.64898/2026.03.20.713089v1

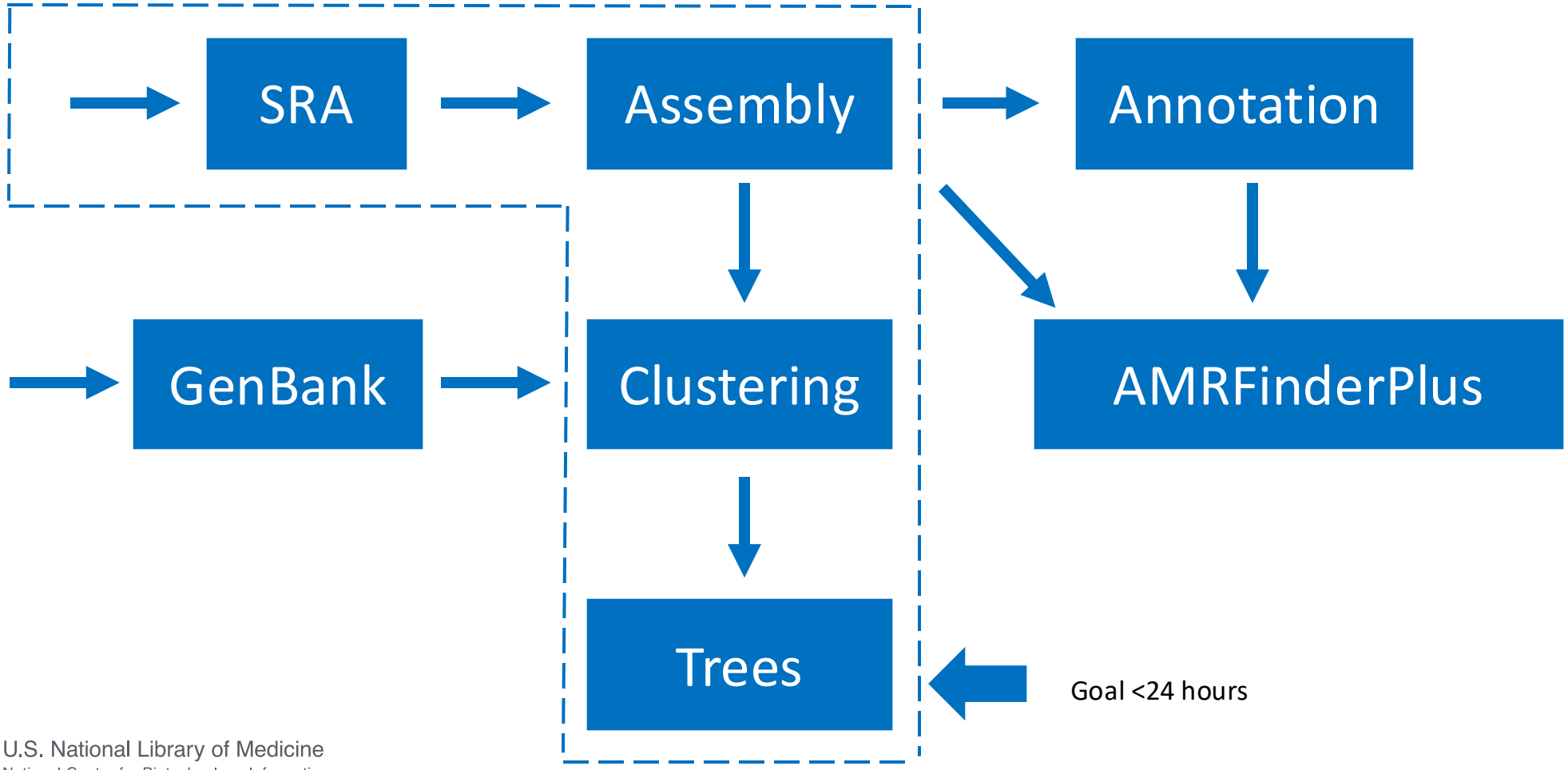
Paired Oxford Nanopore (ONT) and Illumina runs

- 78 *Listeria*
- 62 *Campylobacter*
- 57 *Salmonella*
- 39 *Yersinia enterocolitica*
- 19 *Bacillus cereus*
- 16 *E. coli*
- 15 *Cronobacter*
- 8 *Vibrio parahaemolyticus*
- **Total: 294 isolates** with matched Illumina and ONT

`narjol` From Narjol Gonzalez-Escalona at FDA-HFP

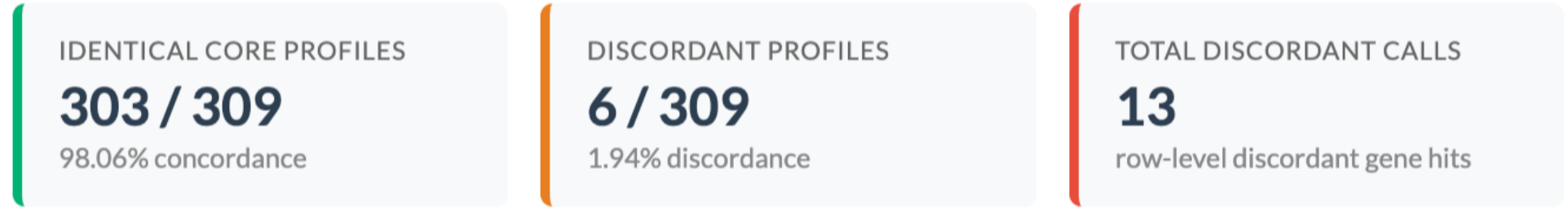
- 6 *Listeria* isolates
  - Nanopore sequencing involved 15 runs.
- 7 *Salmonella* isolates
  - 13 runs were completed with Oxford Nanopore.
- 2 *E. coli* isolates
  - Oxford Nanopore was used for 4 runs.
- Every isolate has one Illumina run.
- Overall, the **15 isolates** generated **32 Oxford Nanopore runs**.

Pathogen Detection Pipeline



AMRFinderPlus *core* AMR symbols

- AMR genes of primary interest and/or expected effect on phenotype



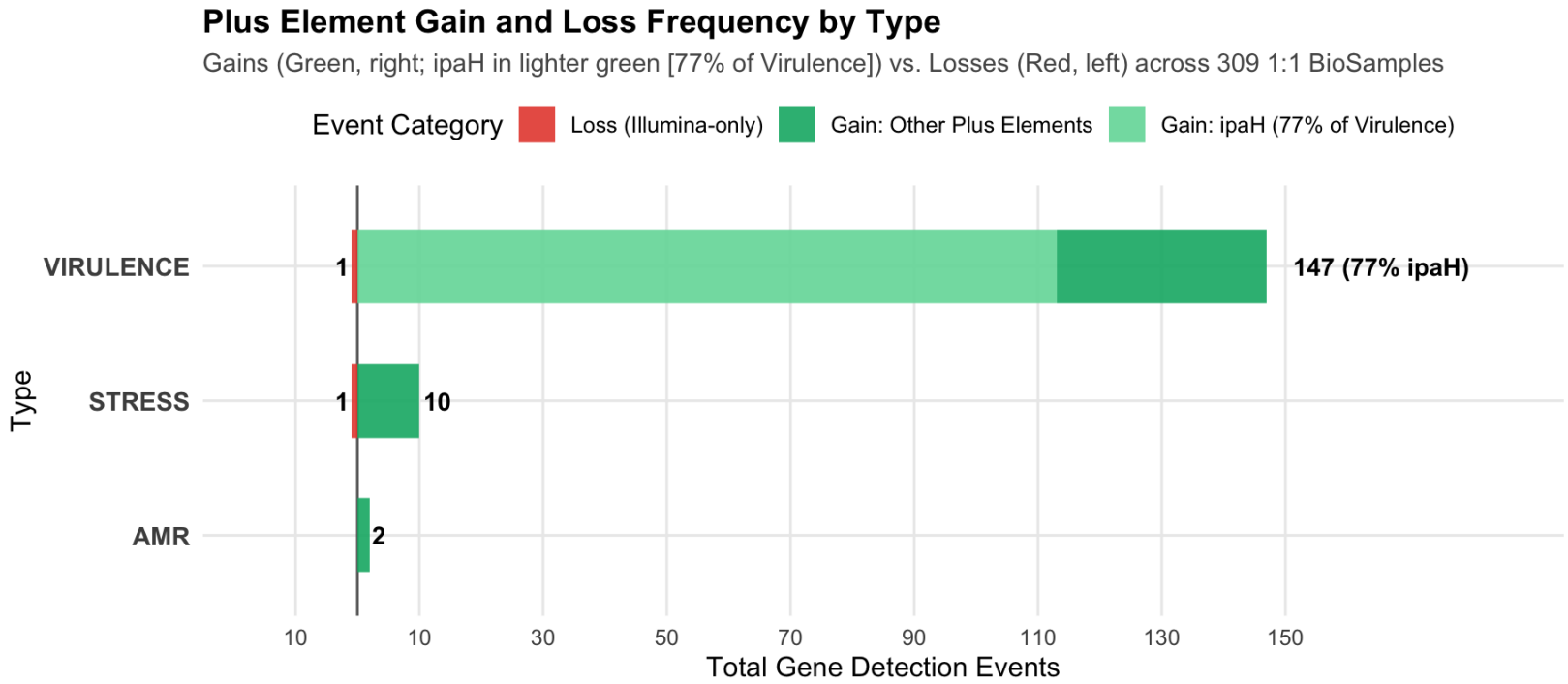
- 3 isolates with genes missing in Stylo
  - low-quality, low–coverage, or partial of full length that was found in both assemblies
- 3 isolates where Stylo assemblies had nfsA_G126GGTer mutation
  - SKESA assembly breaks at the 2-bp insertion GGGG -> GGGGGG

AMRFinderPlus *plus* element symbols

- There are no SAUTE guided assembly targets for these
- Lower priority targets

100% CONCORDANT PAIRS	TOTAL GAINED ELEMENTS	TOTAL LOST ELEMENTS
261 / 309	**159**	**2**
84.47% concordance	Stylo-only detection events	Illumina-only detection events

- 2 lost
  - 1 partial in Stylo
  - 1 missing
- 159 gained
  - 71% ipaH



Clustering and phylogenetic trees

Clustering and Phylogenetic trees

- Kmer / SNP clustering
- wgMLST clustering
- SNP differences

Clustering results for 325 assemblies

K-mer and SNP clustering

Used where we don't have a wgMLST scheme

1. Coarse clustering by K-mer pairwise distance
2. Fine clustering by pairwise 50-SNP single-linkage

320 / 325 (98.5%)
Clustered with production assemblies

wgMLST

Requires wgMLST scheme

250 / 325 assemblies were in organism groups with schemes

1. Max 25 wgMLST allele differences within cluster

247 / 250 (98.8%)
Clustered with production assemblies

Kmer / SNP clustering

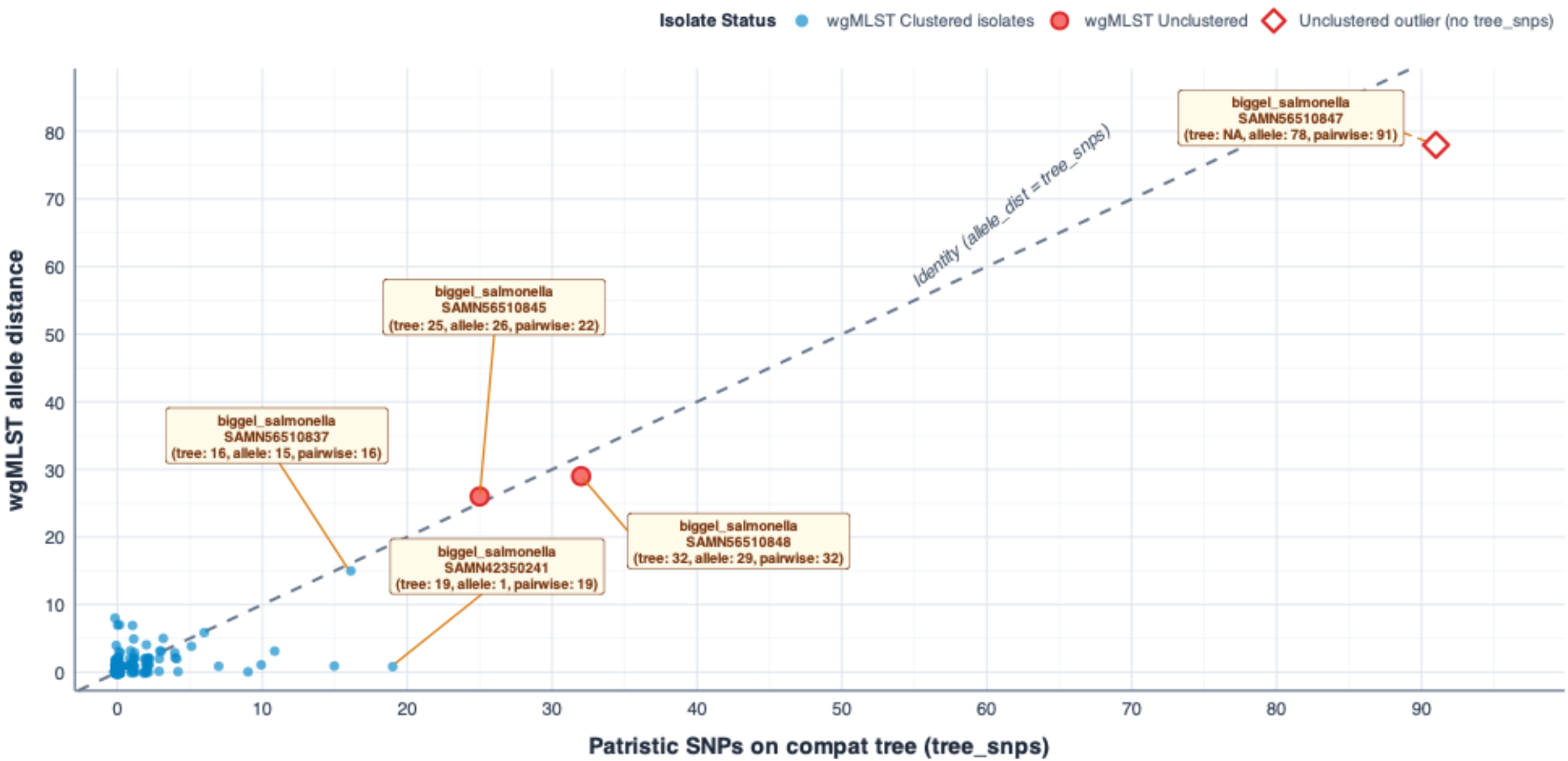
(All 32 Narjol set assemblies clustered with the production assembly)

- 78 *Listeria*
- 62 *Campylobacter*
- 57 *Salmonella*
  - 1 Stylo assembly excluded because distance of 91 SNPs exceeds 50 SNP threshold
- 39 *Yersinia enterocolitica*
- 19 *Bacillis cereus*
  - 2 Illumina runs failed assembly QC for SAMN56511884 and SAMN56511876
  - 2 submitted unicycler (nanopore) assemblies were removed from kmer clustering for violating triangle inequality SAMN21214333 and SAMN21214338
- 16 *E. coli*
- 15 *Cronobacter*
- 8 *Vibrio parahaemolyticus*
- **Total:** SNP distances for 289 isolates with matched Illumina and ONT

wgMLST clustering results

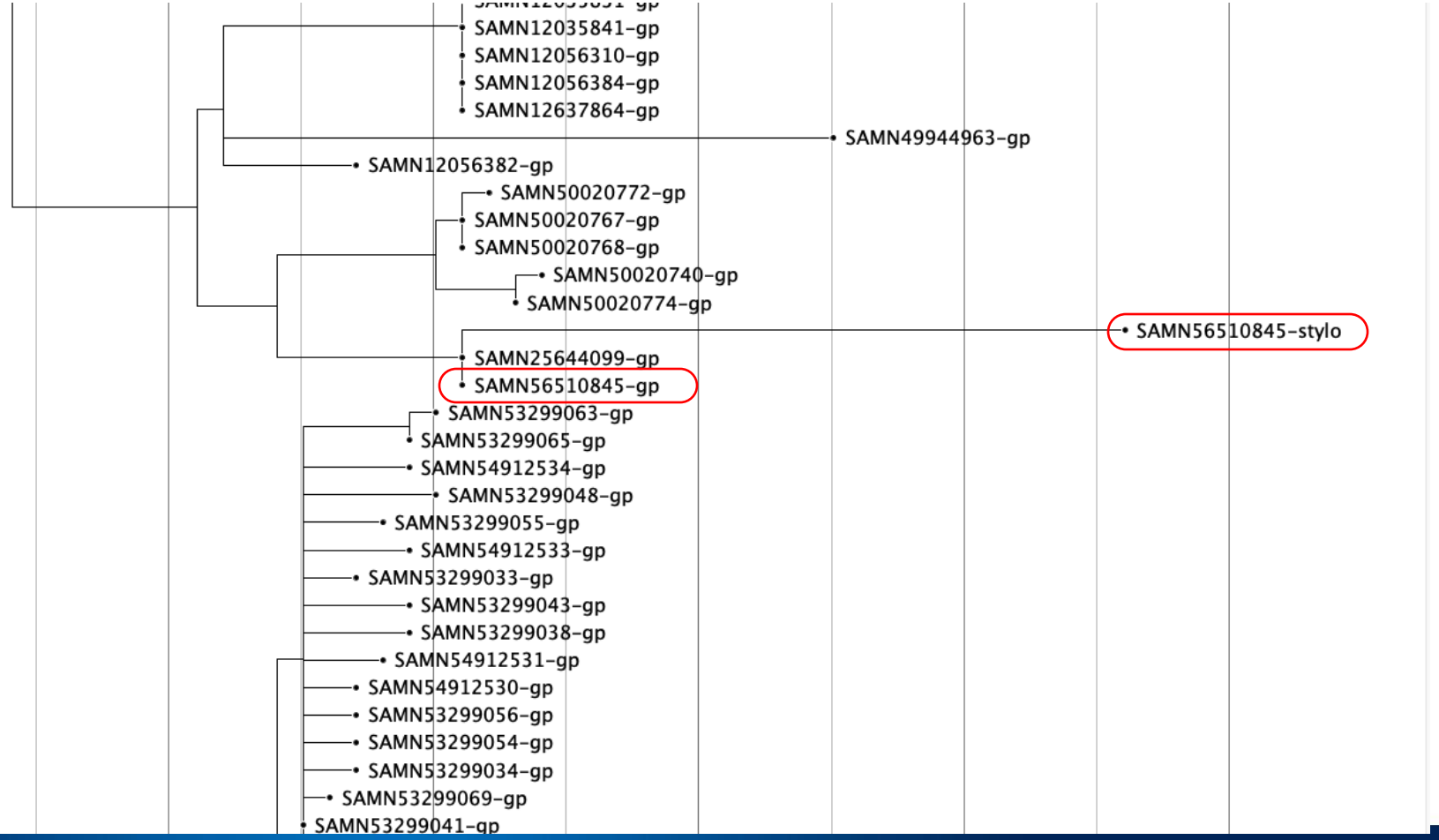
- Could only test with TaxGroups that have wgMLST schemes (250 assemblies)
  - *Salmonella*
  - *Campylobacter*
  - *E. coli* and *Shigella*
  - *Vibrio parahaemolyticus*
- 3 *Salmonella enterica* failed to cluster with the production assembly
  - All were assemblies that had high numbers of SNPs when compared a matching SKESA assembly

Comparison of wgMLST allele distance vs. patristic tree SNPs to production assembly



SAMN56510847 was not in a Kmer/SNP cluster, so no tree_snps were computed; plotted at its pairwise alignment SNP distance (91) vs. allele_dist (78).

Example of highly divergent stylo assembly



Acknowledgements

Richa Agarwala
Victor Ananiev
Azat Badretdin
Slava Brover
Joshua Cherry
Jinna Choi
Vyacheslav Chetvernin
Pavan Chilamakuri
Robert Cohen
DeAnne Cravaritis
Michael DiCuccio
Olga Ermoleava
Boris Fedorov
Mike Feldgarden
Lewis Geer
Renata Geer
Dan Haft
Lianyi Han
Avi Kimchi
Michel Kimelman
William Klimke
Alex Kotliarov
Valerii Lashmanov
Aleksandr Morgulis
Eyal Moses
Chris O'Sullivan
Arjun Prasad
Edward Rice
Kirill Rotmistrovskyy
Alejandro A. Schaffer
Nadya Serova
Sergey Shiryev
Martin Shumway
Oleg Shutov
Douglas Slotta
Alexandre Souvorov
Tatiana Tatusova
Francoise Thibaud-Nissen
Igor Tolstoy
Ed Wachtel
Lukas Wagner
Jay Worley
Hlavina Wratko
Chunlin Xiao
Alexander Zasypkin
Eugene Yaschenko
Mingzhang Yang

David Lipman
James Ostell
Kim Pruitt
Stephen Sherry

Jess Chen, CDC
Narjol Gonzalez-Escalona, FDA
GenFS Long Reads Working Group

ASM
APHL
BWH
CDC
FDA/HFP
FDA/CVM
GenFS
USDA-FSIS
PHAC/CFIA
PHE/FERA
NIAID
Wadsworth/MDH

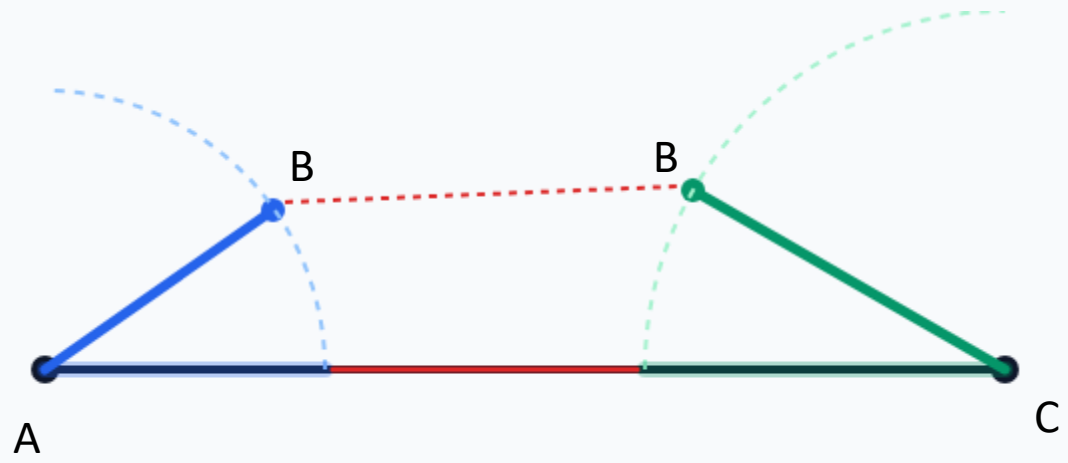
pd-help@ncbi.nlm.nih.gov

This work was supported by the National Center for Biotechnology Information of the National Library of Medicine (NLM) and the National Institute of Allergy and Infectious Diseases (NIAID), National Institutes of Health (NIH), and the Human Foods Program (HFP), Food and Drug Administration (FDA). The contributions of the NIH author(s) are considered Works of the United States Government. The findings and conclusions presented in this work are those of the author(s) and do not necessarily reflect the views of the NIH or the U.S. Department of Health and Human Services.

COMPARATIVE GENOMICS & METRIC SPACE

SNP Distance Bad Triples

$$d(A, B) + d(B, C) < d(A, C) \implies \text{Non-Metric Space}$$



Assembly A → B (20 SNPs) Discrepancy Gap (Violation) Assembly B → C (20 SNPs)

Metric Space Condition

Pairwise SNP counts must satisfy: **$d(A, B) + d(B, C) \geq d(A, C)$**.

Biological Breakdown

Missing call coverage or misassembly distorts direct pairwise comparisons.

Intermediate assembly B cannot co-exist with A and C in a Euclidean distance space.

Direct Distance d(A, C)

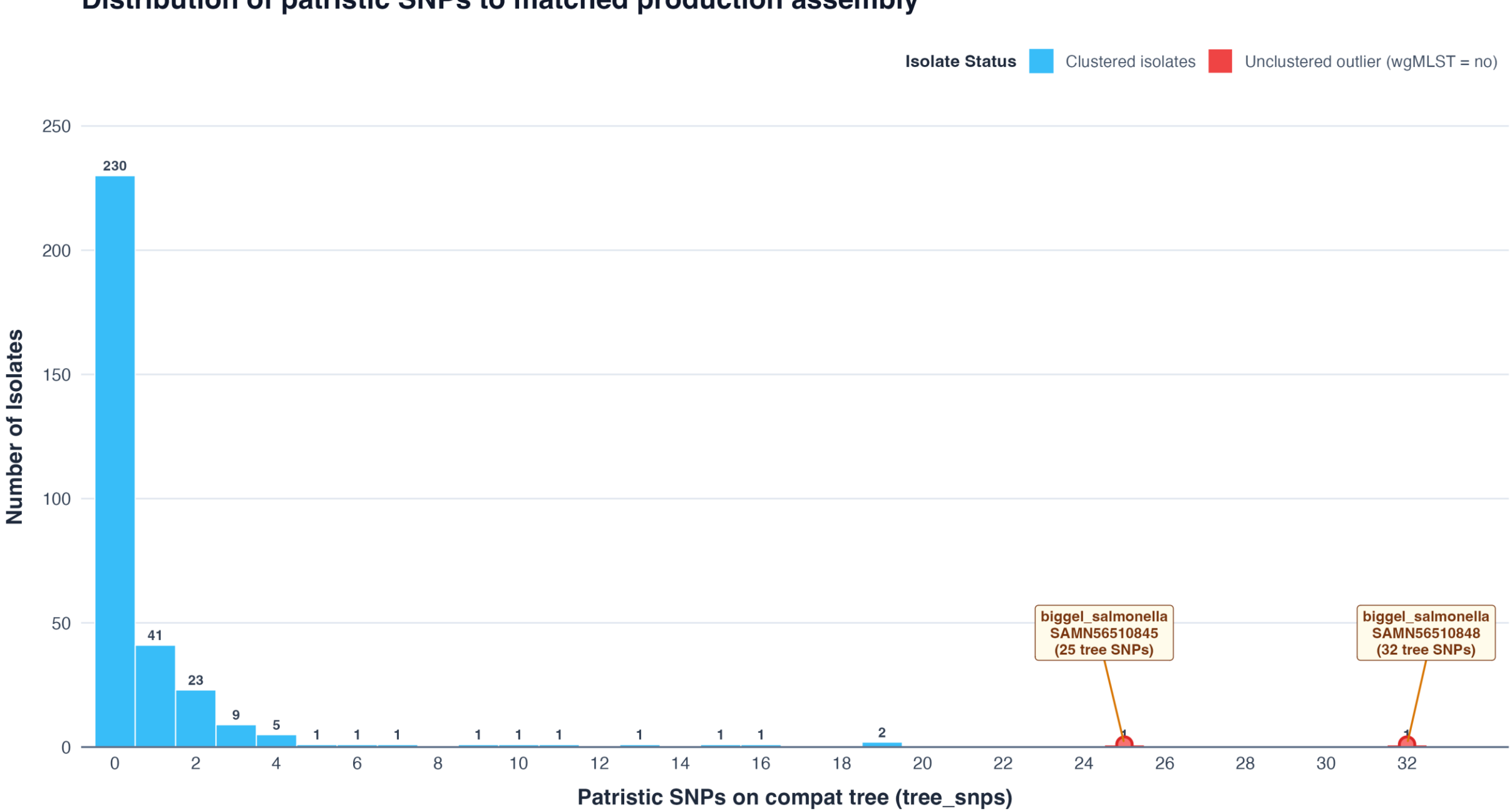
100 SNPs observed

Indirect Path d(A, B) + d(B, C)

40 SNPs total reach

Metric Violation

60 SNP unbridgeable deficit

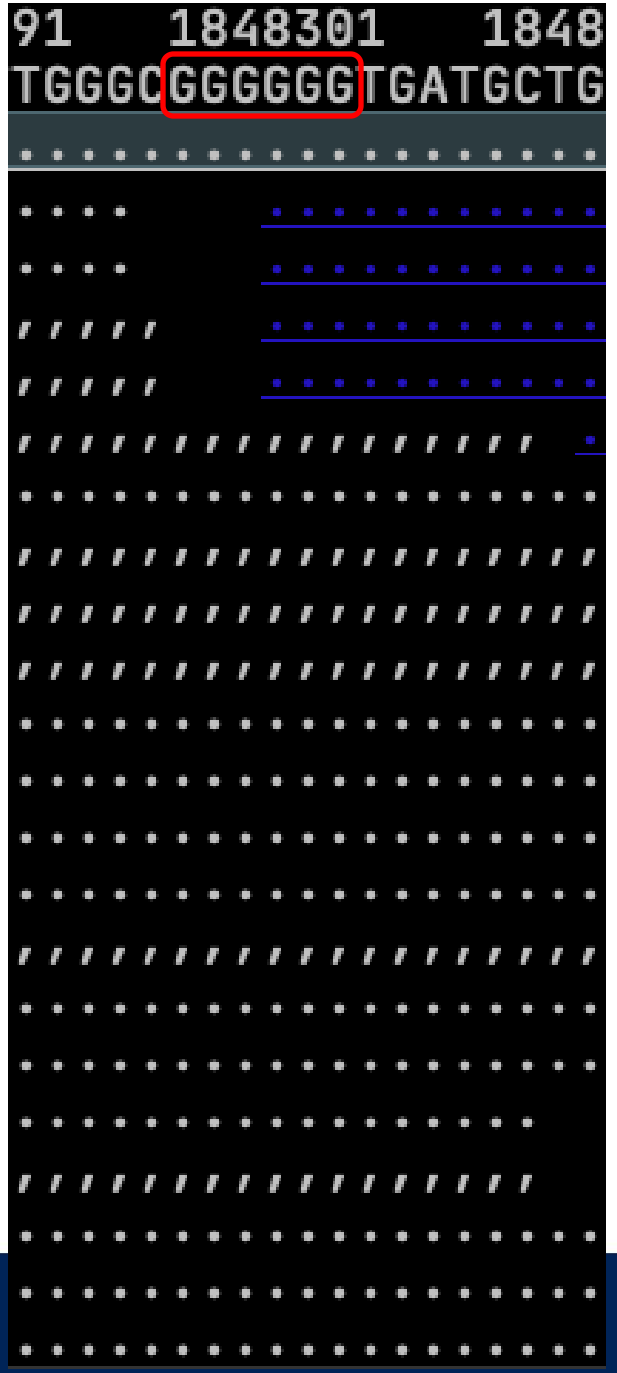

Distribution of patristic SNPs to matched production assembly
Isolate Status
Clustered isolates
Unclustered outlier (wgMLST = no)
Number of Isolates
250
200
150
100
50
0
230
41
23
9
5
1
1
1
1
1
1
1
1
1
2
1
1
biggel_salmonella SAMN56510845 (25 tree SNPs)
biggel_salmonella SAMN56510848 (32 tree SNPs)
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
Patristic SNPs on compat tree (tree_snps)

The three “missing” AMR Genes were all ok

- SAMN03002789 missing **blaR1** and **partial blaZ**, **fosB**, **qacB**
  - Guided contig filtered for low read-depth by PD pipeline
  - No evidence by BLAST of any of these genes in Oxford nanopore reads from 3 Oxford nanopore runs
- SAMN56511876 missing exact **aph(3”)-Ib**
  - Illumina assembly of this *Bacillus cereus* fails QC for genome size and contig #
  - SRA STAT shows evidence of *Pseudomonas* contamination
- SAMN56510874 ”missing” partial **blaOXA** at contig end
  - PARTIAL in SAUTE contig seeded by ant(3’’)-Ia
  - Matching full-length blaOXA-1 in SKESA, SAUTE, and Stylo assemblies
  - Artifact of our use of SAUTE guided assembler for Illumina data

Additional point mutations identified in 3 Stylo assemblies

- nfsA_G126GGTer – detected by POINT_DISRUPT in 3 *Shigella* assemblies
  - SAMN56510878, SAMN56510881, SAMN56510883
  - All have identical 2-bp insertion CGGGGT -> CGGGGGGT
  - SKESA assembly has a contig break here
  - Illumina reads align to the Stylo assembly with no mismatches
  - nfsA disruption alone does not cause clinical resistance to nitrofurantoin

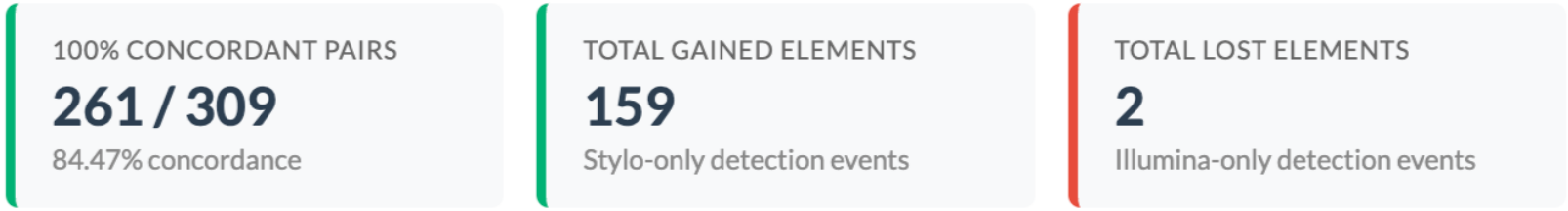


What does Stylo find that is missed by SKESA Illumina assemblies?

Rank	Element Symbol	Element Name	Type	Class	Stylo Assemblies	Illumina Assemblies	Difference (Stylo - Illumina)
1	ipaH3	T3SS effector E3 ubiquitin-protein ligase IpaH3/Hc	VIRULENCE	–	16	0	+16
2	ipaHa	T3SS effector E3 ubiquitin-protein ligase IpaHa	VIRULENCE	–	16	0	+16
3	ipaHb	T3SS effector E3 ubiquitin-protein ligase IpaHb	VIRULENCE	–	16	0	+16
4	ipaHd	T3SS effector E3 ubiquitin-protein ligase IpaHd	VIRULENCE	–	16	0	+16
5	ipaHe	T3SS effector E3 ubiquitin-protein ligase IpaHe	VIRULENCE	–	16	0	+16
6	ami	autolysin Ami	VIRULENCE	–	60	45	+15
7	ipaH1.4	T3SS effector E3 ubiquitin-protein ligase IpaH1.4	VIRULENCE	–	7	0	+7
8	ipaH4.5	T3SS effector E3 ubiquitin-protein ligase IpaH4.5	VIRULENCE	–	7	0	+7
9	ipaH7.8	T3SS effector E3 ubiquitin-protein ligase IpaH7.8	VIRULENCE	–	7	0	+7
10	ipaH9.8	T3SS effector E3 ubiquitin-protein ligase IpaH9.8	VIRULENCE	–	7	0	+7

Plus elements

- There are no SAUTE guided assembly targets for these
- Lower priority targets



- Lost in Stylo assembly
  - SAMN42394156 lost **cadA** in Stylo assembly
    - *Listeria* virulence gene
    - EXACTX in SKESA assembly
    - PARTIALX in Stylo assembly, misassembly with incorrect sequence insertion
  - SAMN42350247 lost **tcfA** in Stylo assembly
    - *Salmonella* virulence gene EXACTX in SKESA assembly
    - Low copy number (14x in Illumina, 5x in ONT)