INTRODUCING

Genpax

A new era of connected pathogen genomics

An annotated overview featuring
Publications, Poster Presentations,
and selected supporting materials.

Confidential. Copyright and IP owned by Genpax.

Genpax

Introducing Genpax

- Genpax is a research and development-based company building novel solutions for clinical pathogen genomics. We are focused on **the needs of infection prevention and control (IPC)**, providing optimal information to recognize and respond to the transmission of strains in the healthcare system and beyond.

- Since 2021, we have established a team of specialist bioinformaticians focused upon bacterial pathogen genomics, building upon more than a century of prior collective experience to develop species-specific toolkits with new analytical capabilities. These are being made available through our **IDEM** platform, addressing over 120 healthcare-associated, public health, and food-associated pathogens – covering 99.9% of pathogens causing human disease by prevalence.

- Considering pathogens can change at a rate of 0 to 5 SNPs per genome per year, the actionable information needed to proactively detect (and exclude) outbreaks, infer transmission, and effectively direct IPC is beyond what other analysis pipelines can reliably deliver.

- Compromises, such as Sequence Typing (e.g. cgMLST) achieve scalability and error tolerance at the expense of sensitivity and specificity. In contrast, SNP-solutions (e.g. wgSNP) cannot be scaled and have reference-dependent accuracy (and good references do not exist for many strains and species).

- Genpax exists to **eliminate these constraints**; to deliver a new generation of pathogen analysis capabilities which address the IPC challenges of emergent pathogens and AMR.

- This brochure highlights a selection of **key differentiating capabilities** in our quest to make the best possible genomic pathogen analysis accessible to everyone.

Confidential. Copyright and IP owned by Genpax.

Genpax offers an automated cloud-based solution with the following features:

- **A SNP-throughout analysis** that makes maximal use of the safely interpretable genome sequencing information.
- An analysis solution that **does not require typing or other steps** to select a reference genome.
- **Performance equivalent to wgSNP under its most optimal conditions**, delivering equally optimal and comparable results for all species and strains.
- Entirely consistent findings from clinical replicates, identical strains from common sources, and samples in clinical ring-trials.
- A **near-zero error rate** meaning results from the multiple labs can be reliably combined and compared.
- **Unprecedented accuracy** combined with **addressing more of the genome sequence** information than either Sequence Typing or previous whole genome SNP comparisons.
- Unmatched capabilities to identify strains and **infer their likely membership of outbreaks (or not) and order of transmission**, even with only two isolates within a transmission cluster.
- Effective analysis that can be used **as part of a clinical solution** to help optimize infection prevention and control workflows for improved patient care and safety at the same time as reducing the costs of healthcare.
- **Open scalability**, so that each newly analyzed strain can be compared with all strains previously processed, within and between sites that choose to openly display their results.
- A **user-friendly platform** with interactive and continuously updated communication of findings.
- A **rapid turnaround time**, whether analyzing one or hundreds of isolates from a sequencing run, while comparing them with hundreds or tens of thousands (or more) previously analyzed strains.

Confidential. Copyright and IP owned by Genpax.

Translating Science into Impact: Validation Through Case Studies

As a commercial entity, Genpax cannot share its code and solutions. However, we are a team, comprised largely of former academics with hundreds of publications in the field between us, who want to share our system and its capabilities as openly and clearly as possible. To do this, we have performed a set of validation and demonstrational analyses using information from the highest quality publications and studies that we could identify, selecting those with the best evidence for 'ground truth' against which to be measured.

- In tests of **reproducibility** and near-zero error, genuine biological replicates from clinical ring trials (*E. coli*, *Staph.*) and large well-documented outbreaks and re-isolation studies have been addressed (*E. coli* and *K. pneumoniae*).
- In tests of **accuracy** (*E. coli*, *Campylobacter jejuni*, and *P. aeruginosa*), exceptional situations in which published or specifically generated almost identical high-quality reference genomes were used in the original studies were selected and then re-tested as the most stringent findings to measure performance against that we could identify.
- In tests of **natural reference-free** performance and transmission-chain re-structuring, species were selected that represent extremes of highly recombining panmictic (*Campylobacter jejuni*) and deeply rooted, highly-clonally diverse (*P. aeruginosa*) population structures.
- In the test of **scalable** comparisons, using *Listeria monocytogenes*, we processed data generated in large studies from different European laboratories.
- The test of MRSA **gene finding** (ASM, Staph. aureus) used sequencing and MRSA/MSSA data from two published studies from an EU reference laboratory.
- Likewise, our **economic modeling** adopts conservative assumptions, taking a cautious approach towards outbreak sizes and containment speed, in contrast to the assumptions of the published models it is built upon. Additionally, our analysis includes up-to-date costs of sequencing and analyses ensuring accurate financial impacts.

Each poster primarily addresses one or more aspects of our platform's performance for IPC applications: accuracy, low noise, high resolution, comparability, species applicability, and scalability. In combination they represent a real enhancement in what sequencing analysis can offer infection prevention and control.

Confidential. Copyright and IP owned by Genpax.

Calling Zero: A new foundation for diagnostic bacterial genomics

James C. Littlefair, Benedict J. Uttley, Dan G. Frampton, Gareth M. Linsmith, John F. Peden, & Nigel J. Saunders: **Genpax, London, United Kingdom** jlittlefair@genpax.co +44 7473 871 175

Introduction

- Bacterial strains typically diversify at rates below 10 SNPs per year and thresholds to recognize source-linked and transmission-associated strains typically range from 10 to 20 SNPs. Therefore, even low noise levels impact outbreak cluster detection and analysis for source attribution and transmission inference.
- Reproducibility, low noise, and the ability to call true zeros from resequencing of the same DNA, culture, patient, or closely-linked isolates, is essential for cluster recognition and accurate branching structures of within-cluster dendrograms.
- The most effective outbreak surveillance requires reproducibility within and between laboratories as it facilitates multi-site surveillance and comparability.
- A cloud-based solution which works directly from FASTQ, delivers SNP-resolution information, and requires no clonal reference can facilitate this enhanced infection control and prevention.

Results

- The Genpax analysis pipeline consistently obtained 0 SNP-distances within all 20 replicate groups across all five participating laboratories, producing 17 clusters from 110 samples.
- The replicate groups that clustered together at 0 SNP-distance, in concordance with the original study, were:
  - NGSRT01& NGSRT02
  - NGSRT03 & NGSRT05
  - NGSRT14 & NGSRT15
- However, replicate groups NGSRT18 & NGSRT19, identical in the original cgMLST analysis, were separated by a single intergenic SNP not addressed by the sequences used in cgMLST.
- Replicates within 0 SNP clusters typically shared >80% of their genome length from which variants could be called (based on a 2.8Mb genome), over 500kb more sequence than cgMLST.
- This compares to reported whole genome SNP resolutions of 72% within Sequence Type and 57% across the species [2].

Results

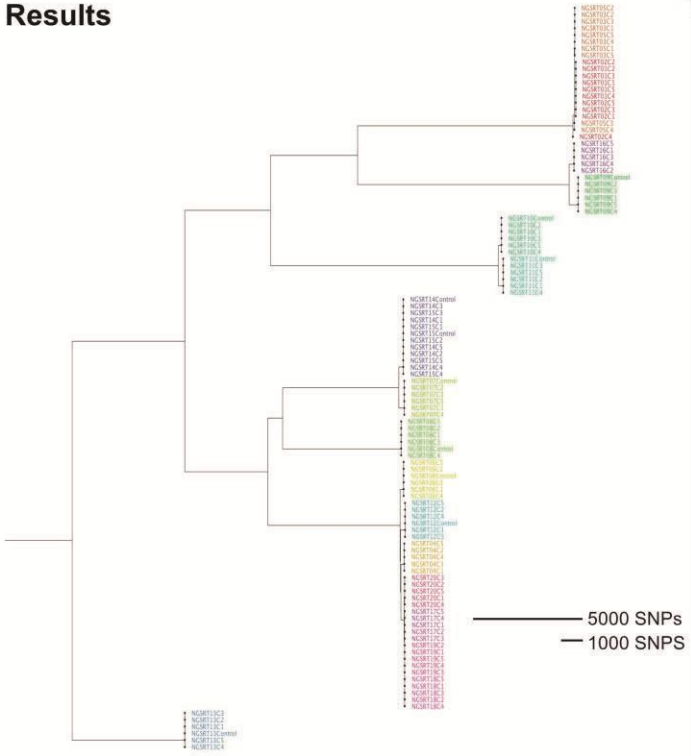


FIGURE 1 – NJ dendrogram showing all 110 Mellman ring trial *S. aureus* replicates as analyzed with the Genpax pipeline. Colours represent 0 SNP clusters.

Methods

- 110 readsets from the Mellmann ring trial [1] published in 2017 were processed through the Genpax analysis pipeline in the version as of Q1 2023.
- The ring trial included *S. aureus* sequences from diverse sources sequenced on Illumina MiSeq platforms by five different laboratories across three European countries (Denmark, Germany and the Netherlands) following the same protocol (Nextera XT Library Prep and 250-bp paired end).

TABLE 1 Characteristics of the 20 human *S. aureus* isolates that were sent as DNA samples to the five participating laboratories in a blinded fashion and used as controls

Sample ID Ring trial replicate group	Original strain	Spa type (based on Sanger sequencing)	Comment/reference	Genpax 0 SNP cluster
NGSRT01	468	t011	Livestock-associated MRSA	**1**
NGSRT02	551	t011	Livestock-associated MRSA, identical cgMLST genotype as NGSRT01	**1**
NGSRT03	1346	t011	Livestock-associated MRSA	**2**
NGSRT04	1354	t010	Classical hospital-acquired MRSA	3
NGSRT05	1360	t011	Livestock-associated MRSA, identical cgMLST genotype as NGSRT03	**2**
NGSRT06[a]	2180	t002	Central European community-acquired PVL[b]-positive MRSA	4
NGSRT07[a]	2482	t008	US typical community-acquired PVL-positive MRSA	5
NGSRT08[a]	2560	t044	Central European community-acquired PVL-positive MRSA	6
NGSRT09[a]	2638	t012	Classical hospital-acquired MRSA	7
NGSRT10[a]	2786	t843	*mecC*-positive MRSA	8
NGSRT11[a]	2949	t843	*mecC*-positive MRSA	9
NGSRT12[a]	2994	t003	Classical hospital-acquired MRSA	10
NGSRT13[a]	3039	t032	Classical hospital-acquired MRSA	11
NGSRT14[a]	COL	t008	MRSA strain COL	**12**
NGSRT15[a]	COL	t008	Duplicate of MRSA reference strain COL	**12**
NGSRT16	ATCC 25923	t021	MSSA quality control strain ATCC 25923	13
NGSRT17	P1	t001	Isolate P1 from reference 23	14
NGSRT18	P3	t001	Isolate P3 from reference 23	15
NGSRT19	P4	t001	Isolate P4 from reference 23, identical cgMLST genotype as NGSRT18	16
NGSRT20	P12	t001	Isolate P12 from reference 23	17

[a]These samples were separately cultivated, and DNA was extracted and sequenced as controls.
[b]PVL, Panton-Valentine leukocidin.
Genpax 0 SNP clusters in bold font comprised more than one ring trial replicate group

Results

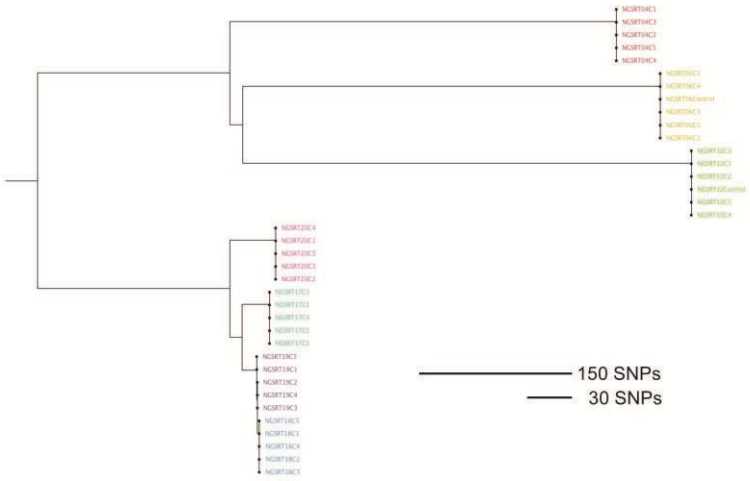


FIGURE 2 – NJ dendrogram showing 37 Mellman ring trial *S. aureus* replicates within the same clonal complex (CC5). Colours represent 0 SNP clusters.

Conclusions

- Because of limitations in comparability and scalability, it has previously been necessary to use typing methods such as cgMLST, which do not make full use of the available WGS information when performing large-scale and multi-site public health surveillance.
- The Genpax analysis pipeline, despite the multi-site nature of the study with variability in sequence coverage and quality, demonstrates increased resolution whilst simultaneously reducing noise, giving accurate and comprehensive SNP-resolution information.
- By using a strategy that does not depend upon a clonal reference, which is necessary in the absence of prior knowledge and typing, the pipeline is applicable to all strains within the species diversity, including MSSA and MRSA, which is essential for optimal infection prevention and control.

References

1. Mellmann A, Andersen PS, Bletz S, Friedrich AW, Kohl TA, Lilje B, et al. High Interlaboratory Reproducibility and Accuracy of Next-Generation-Sequencing-Based Bacterial Genotyping in a Ring Trial. J Clin Microbiol. 2017;55(3):908-13
2. Gorrie CL, Da Silva AG, Ingle DJ, Higgs C, Seemann T, Stinear TP, et al. Key parameters for genomics-based real-time detection and tracking of multidrug-resistant bacteria: a systematic analysis. Lancet Microbe. 2021;2(11):e575-e83

Declaration

This research was entirely funded by Genpax.

Genpax is a bioinformatics company founded in 2020 seeking to develop novel solutions that overcome the limitations of established analysis strategies to maximize the usefulness of bacterial genome sequences in infection control and prevention.

Check out our website:

A novel genome comparison tool producing near-zero error for same-patient isolates of *E. coli* ST131

Poster No. 265
Date: 6/17/2023

Rebecca J. Bengtsson, John F. Peden, Dan Frampton, Gareth Linsmith, Benedict J. Uttley, Arthur Poivet, Luis Montemayor & Nigel J. Saunders
Genpax, London, United Kingdom

research@genpax.co
+44 203 603 6869

Introduction

- Escherichia coli ST131 is a globally disseminated clone and a significant contributor to the global burden of urinary tract infections (UTIs).
- Whole genome sequencing (WGS) based analysis provides higher resolution than traditional typing methods and its use in public health settings can significantly help improve surveillance and outbreak investigation of pathogens.
- Currently, WGS is primarily applied retrospectively and is limited by its speed and scalability.
- To achieve its potential impact, it needs to be used proactively to direct patient care and healthcare resource management.
- WGS analysis also requires expert knowledge, including selecting appropriate reference genomes and tools for analysis of the sequence data as well as interpreting the resulting data to clinically actionable results.
- To tackle this limitation, Genpax has developed a novel species-centric, automated genome comparison tool that does not require the selection of an appropriate reference genome to perform SNP-resolution analysis.
- In this study, the performance and accuracy of this method is compared to traditional core genome SNP analysis with UTI E. coli isolates from a published study [1].

Check out our website and other supporting information

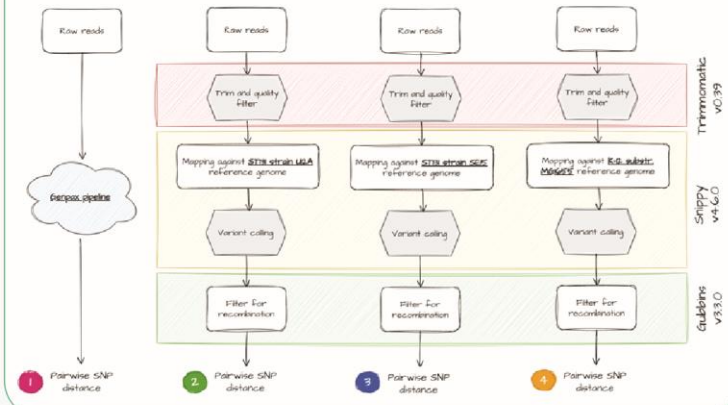
Declaration

This research was entirely funded by Genpax.

Genpax is a bioinformatics company developing novel solutions that overcome the limitations of established analysis strategies to maximize the usefulness of bacterial genome sequences in infection prevention and control.

Methods

- The published study consisted of 65 *E. coli* ST131 faecal and urinary isolates sampled from a single patient with a long-term UTI were analyzed.
- This study was selected because it contained similar biological replicates, and had generated a reference genome (U12A, using PacBio) from one of the clonal isolates.
- The use of a nearly identical reference provides a best case for analysis that minimizes the introduction of systematic errors during mapping and maximizes the accuracy and resolution of pairwise SNP (pw-SNP) distance determination [2].
- The dataset was processed using two methods.
- The first is the Genpax analysis pipeline which does not require the selection of a reference genome and is referred to as analysis 1.
- For the second method, we used the current versions of the industry standard tools used for bacterial genomic analysis. This included Trimmomatic for short reads adapter trimming and quality filter, Snippy for reference mapping and variant calling, and Gubbins for recombination filtering. This is similar to the strategy used in the original publication [1].
- Three analyses using the second method were carried out with different reference genomes for mapping. Analysis 2 used the intra-clonal U12A study-generated reference genome. Analysis 3 used a standard reference genome for this clonal complex: *E. coli* ST131 strain SE15. Analysis 4 used standard *E. coli* strain K-12 substr. MG1655.
- To assess the accuracy of the developed pipeline, the pairwise-SNP distances from the four analyses were compared to the results from the original publication.



References

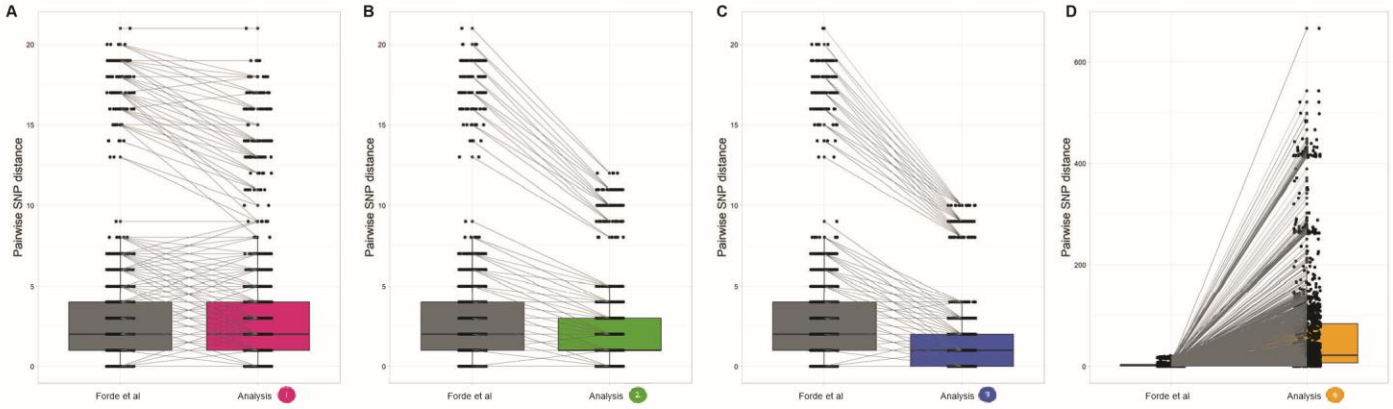
1. Forde, B.M., Roberts, L.W., Phan, M.D., Peters, K.M., Fleming, B.A., Russell, C.W., Lenherr, S.M., Myers, J.B., Barker, A.P., Fisher, M.A. and Chong, T.M., 2019. Population dynamics of an Escherichia coli ST131 lineage during recurrent urinary tract infection. Nature communications, 10(1), p.3643.
2. Gorrie, C.L., Da Silva, A.G., Ingle, D.J., Higgs, C., Seemann, T., Stinear, T.P., Williamson, D.A., Kwong, J.C., Grayson, M.L., Sherry, N.L. and Howden, B.P., 2021. Key parameters for genomics-based real-time detection and tracking of multidrug-resistant bacteria: a systematic analysis. The Lancet Microbe, 2(11), pp.e575-e583.

Results

Table 1. Number of samples processed and variants genotyped

	Genpax 1	U12A 2	SE15 3	K-12 4
Samples processed	65	63	63	61
Variants genotyped	59	24	22	1,640

All 65 samples were analyzed by the Genpax pipeline. However, using the Snippy/Gubbins pipeline, 2 samples were rejected using the U12A and SE15 reference genome, and 4 samples were rejected using the K-12 reference genome. Therefore, we present the 61 samples that are common to all analyses. Table 1 describes the total number of variants genotyped across the 61 samples analyzed.



Comparison of pairwise SNP distance distribution of 61 *E. coli* ST131 isolates from the four analyses under investigation. Each box plot demonstrates the distribution of pairwise-SNP distances between all isolate pairs compared to pairwise-SNP distances from the original publication. (A) Comparison of pairwise-SNP distances between isolate pairs from Forde et al relative to isolates pairs from analysis 1, (B) isolate pairs from analysis 2, (C) isolate pairs from analysis 3 and (D) isolate pairs from analysis 4.

- There are two main populations from the original publication, isolate pairs with distance between 0 and 9, and pairs with distance between 13 and 21.
- The boxplots revealed that analysis 1 processed using the Genpax pipeline produced pairwise-SNP distances that shared the most similarities to distances from the published results, for both populations.
- The current version of Snippy and Gubbins with both the intra-clonal U12A (analysis 2) and intra-ST SE15 (analysis 3) reference genomes detected lower pairwise-SNP distances relative to the original publication, with analysis 3 detecting even lower comparative pairwise-SNP distances than analysis 2.
- Analysis 4 using the general *E. coli* K-12 reference genome resulted in a dramatic decrease in SNP detection precision, by which all isolate pairs displayed a significant increase in pairwise-SNP distance compared to the published results, reflecting the widely recognized impact of selecting distantly related reference genomes for mapping-based SNP analyses.

Conclusions

- The Genpax reference-independent method produced isolate pairs with similar pairwise-SNP distances to those reported using a gold standard method with an optimal intra-clonal reference genome.
- Using the standard reference-based method, the intra-clonal strain reference U12A improves SNP detection sensitivity compared to using a reference of the same ST; but neither matched the performance of the Genpax pipeline.
- With a standard reference-based method, use of a species-specific reference genome outside of the clonal complex resulted in a dramatic drop in SNP precision and over-estimation of SNPs identified.
- The Genpax reference-independent method demonstrated substantially better performance when compared to using an industry standard methodology.

Near zero error using large-scale hospital outbreak whole genome sequence data for *Klebsiella pneumoniae*

Poster No. 267
Date: 6/17/2023

Ramiro Morales-Hojas, James C. Littlefair, Daniel Frampton, John F. Peden & Nigel J. Saunders: **Genpax**, London, United Kingdom

research@genpax.co
+44 203 603 6869

Introduction

- *Klebsiella pneumoniae* is a common nosocomial pathogen responsible for a range of severe and life-threatening infections [1].
- *K. pneumoniae* is a leading cause of extended-spectrum beta-lactamase (ESBL)-producing, and carbapenem-resistant healthcare-associated infections and is considered a critical public health threat by the WHO [2].
- Although the overall prevalence is lower than *E. coli*, *K. pneumoniae* is notably linked with higher rates of hospital transmission [3].
- Several genomic studies indicate that up to 1/3 of hospital infections are linked to within-hospital transmission events [4,5], and proactive genomic surveillance is becoming the gold standard infection control practice to detect nosocomial infections and to prevent the transmission of highly resistant strains.
- Prospective surveillance requires accurate and high-resolution genomic analysis solutions to detect transmission events when bacterial species typically evolve at a rate of between 1 and 10 SNPs per year, which is challenging to accomplish using existing methods.

Objective: To test the resolving power of Genpax's WGS analysis, through the reanalysis of clinical outbreaks of *K. pneumoniae* using data from published studies with established epidemiological contexts.

Methods

- The performance of the Genpax pipeline was evaluated using a recently published dataset obtained from two intensive care units (ICUs) in Vietnam [6]. This study was chosen due to the number of isolates associated with transmission events and low genomic distances, providing an ideal opportunity to assess the pipeline's performance.
- The novel pipeline was used to call genome-wide SNPs for 1314 *K. pneumoniae* isolates from the study.
- Cluster analysis was conducted based on informative SNPs. This analysis assigned samples to zero-SNP transmission clusters if they exhibited a distance of 0 SNPs to any other sample within the cluster.

References

1. Pendleton, J. N., Gorman, S. P. & Gilmore, B. F. (2013). Clinical relevance of the ESKAPE pathogens. *Expert. Rev. Anti Infect. Ther.*, 11: 297-308.
2. World Health Organization. (2017). Global Priority List of Antibiotic-Resistant Bacteria to Guide Research, Discovery, and Development of New Antibiotics.
3. Hilty M *et al.* (2012). Transmission dynamics of extended-spectrum β-lactamase-producing Enterobacteriaceae in the tertiary care hospital and the household setting. *Clin. Infect. Dis.*, 55: 967-975.
4. Marimuthu, K., *et al.* (2022). Whole genome sequencing reveals hidden transmission of carbapenemase-producing *Enterobacterales*. *Nat. Commun.*, 13: 3052-3062.
5. Tosas Auguet, O., *et al.* (2018). Frequent undetected ward-based methicillin-resistant *Staphylococcus aureus* transmission linked to patient sharing between hospitals. *Clin. Infect. Dis.*, 66: 840-848.
6. Roberts, L. W. *et al.* (2022). Genomic characterisation of multidrug-resistant *Escherichia coli*, *Klebsiella pneumoniae*, and *Acinetobacter baumannii* in two intensive care units in Hanoi, Viet Nam: a prospective observational cohort study. *The Lancet Microbe*, 3: e857-866.

Results

- The isolates had an average genome length of 5.2 Mbp, of which 89% (4.6 Mbp) was suitable for SNP calling using the Genpax pipeline. This represents a 24% increase in genome length available for variation analysis compared to the 3.76 Mbp core genome described in the original publication.
- The reanalysis revealed a higher number of 0-SNP clusters for all *K. pneumoniae* samples compared to the original publication. This difference was primarily due to the increased number of clusters identified in the reanalysis of the ST15 isolates (Table 1).
- ST15 Cluster 15: The cluster revealed a revised 0-SNP cluster with 78 members. Among these, 67 (85%) were shared with the previously reported cluster. The reanalysis identified an additional 11 isolates, while 12 originally included isolates had 1 to 4 verified SNPs (Figure 1A). Notably, four of the samples originally reported to be part of the ST15 0-SNP cluster were assigned to a different 0-SNP group from the original publication (Table 2).
- ST16 Cluster 12: 13 out of the 14 ST16 isolates reported to be in a 0-SNP cluster were included in a redefined 0-SNP cluster. One isolate originally assigned to cluster 12 differed by 2 verified SNPs (Figure 1B). Six additional isolates were placed in this redefined cluster (Table 2), making a final group of 19 rather than 14.
- Cluster reanalysis of ST15 isolates in 0-SNP clusters utilizing a 5-SNP threshold was performed, as in the original publication. The largest cluster obtained in the reanalysis comprised 130 isolates of the 138 in the largest 5-SNP cluster in the publication (Table 3).
- Cluster reanalysis of ST16 isolates in 0-SNP clusters using a 5-SNP threshold identified 3 clusters instead of the single 5-SNP cluster reported previously (Table 3). Isolates excluded from the redefined 5-SNP clusters differed by 6 to 12 identified (and verified) SNPs.

Figure 1: IGV screenshots of SNP positions identified in the present reanalysis, highlighting the differentiation among isolates originally assigned to the 0-SNP clusters 12 and 15. **A)** Reference and alternative SNP alleles in two positions separating two samples of the ST15 0-SNP cluster (cluster 15). **B)** Reference and alternative SNP alleles in two positions identified in the reanalysis, differentiating ERR3585327 from the other samples in the ST16 0-SNP cluster 12.

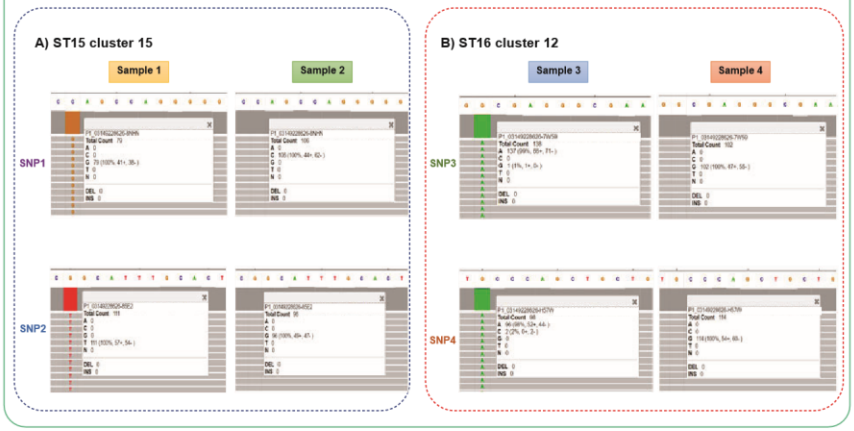


Table 1: Number of 0-SNP clusters reported by Roberts et al. for each of the main sequence types (STs) and those obtained in this reanalysis with the Genpax pipeline.

	Roberts et al.	Genpax
K. pneumoniae	71	84
ST15	21	28
ST16	17	16
ST11	11	14
ST656	13	11

Table 2: The number of isolates in two 0-SNP clusters reported by Roberts et al. for ST15 and ST16 and the corresponding clusters obtained with the Genpax pipeline.

ST	Cluster	N (6)	N Genpax	Common	Unique (6)	Unique Genpax
15	15	79	78	67	12	11*
16	12	14	19	13	1	6**

* 4 of these samples were reported in a different 0-SNP cluster by Roberts et al. (6); 7 were not in any cluster
** 1 sample reported on a different 0-SNP cluster by Roberts et al. (6); 5 were not in any cluster

Table 3: The number of 5-SNP clusters reported by Roberts et al. [6] for ST15 and ST16 and redefined with the Genpax pipeline; and number of shared isolates between the largest original and redefined 5-SNP clusters.

ST	N (6)	N Genpax	Shared isolates in largest cluster
15	5	8	130/138
16	1	3	114/117

Conclusions

- This reanalysis found that the Genpax pipeline accurately identified and differentiated transmission-linked samples. With lower noise, higher accuracy, higher resolution, and an improved determination of strain relationships compared to the standard tools used in the original study.
- The high coverage and resolution, reflected by the substantial (24%) increase in addressed genome space, was achieved while maintaining high accuracy and low noise (near-zero error).
- The resolution exhibited by this novel pipeline enables precise evaluation of transmission networks, offering more accurate insights into hospital outbreaks of *K. pneumoniae*.

Check out our website and other supporting information

Declaration

This research was entirely funded by Genpax.

Genpax is a bioinformatics company developing novel solutions that overcome the limitations of established analysis strategies to maximize the usefulness of bacterial genome sequences in infection prevention and control.

Confidential. Copyright and IP owned by Genpax.

Reference-free WGS SNP-resolution analysis of *Campylobacter jejuni*

Poster No. 264
Date: 6/17/2023

Karolina Jackowska, Dan Frampton, John F. Peden, Nigel J. Saunders: **Genpax**, London, United Kingdom

research@genpax.co
+44 203 603 6869

Introduction

- Analyzing highly recombinogenic species with panmictic populations poses challenges for whole genome sequence analysis due to the absence of reference genomes for accurate strain analysis and generation of universally applicable and comparable results.
- This means that analysis of these species is often limited to Sequence Typing methods that have sub-optimal resolution and provide poor source-attribution and inferred transmission information.
- *Campylobacter jejuni* exemplifies this issue, as outbreaks can be spread through food distribution networks, resulting in transmission-linked isolates being potentially distributed across different laboratories.
- The objective was to evaluate the performance of Genpax's WGS analysis pipeline by applying it to sequence data from Pascoe's study [1] and compare the results.
- This study was selected because of its clearly defined transfer history of strains, low SNP differences, but counter-intuitive findings with respect to history and rates of evolutionary change - even when using a close to optimal WGS analysis strategy with a well established highly similar ancestral reference genome.
- It also presents a well controlled and documented simulation of the processes that occur in the context of a clinical outbreak, in which mutations are tracked to determine strain identity, and to infer transmission. It also represents a stringent test against which to compare a reference-independent analysis.

Methods

- Genome-wide SNPs for the 22 publicly available *C. jejuni* strain 11168 derivative isolates from various UK laboratories were called using the Genpax pipeline.
- All metadata was taken from the Pascoe study [1].
- Dendrograms were created using a neighbour-joining method [2].
- Missing data was excluded from further analysis.

References

1. Pascoe B, Williams LK, Calland JK, Meric G, Hitchings MD, Dyer M, Ryder J, Shaw S, Lopes BS, Chintoan-Uta C, Allan E, Vidal A, Feamley C, Everest P, Pachebat JA, Cogan TA, Stevens MP, Humphrey TJ, Wilkinson TS, Cody AJ, Colles FM, Jolley KA, Maiden MCJ, Strachan N, Pearson BM, Linton D, Wren BW, Parkhill J, Kelly DJ, van Vliet AHM, Forbes KJ, Sheppard SK. Domestication of Campylobacter jejuni NCTC 11168. Microb Genom. 2019 Jul;5(7):e000279. doi: 10.1099/mgen.0.000279. Epub 2019 Jul 16. PMID: 31310201; PMCID: PMC6700657.
2. N Saitou, M Nei, The neighbor-joining method: a new method for reconstructing phylogenetic trees., *Molecular Biology and Evolution*, Volume 4, Issue 4, Jul 1987, Pages 406–425, https://doi.org/10.1093/oxfordjournals.molbev.a040454

Results

- Results were obtained addressing an average of 1.38 Mbp of *C. jejuni*, equivalent to 85.3% of the published genome.
- The AL11168.1 sequence which is the closest to the original isolate is clearly ancestral, as it should be, unlike in the original analysis.
- The strains strongly correlate with the sources and laboratories between which they were circulated, unlike in the original analysis.
- Genome-wide SNP analysis identified an average of 9 SNP differences to the reference (AL11168.1) per sample; in contrast to 29 SNP differences in the Pascoe et al. (2019) study.
- Two 0-SNP distance clusters were identified, totaling 12 of the 22 samples, of 5 and 7 members that are separated by 3 SNPs.

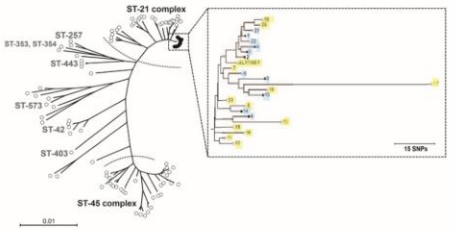


Figure 1: Phylogenetic tree from Pascoe et al. (2019), created using maximum-likelihood. Phylogeny was reconstructed in FastTree2 with the generalized time reversible substitution model.

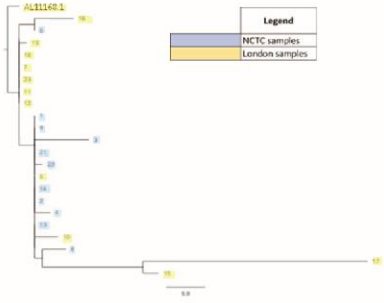


Figure 2: Phylogenetic tree created using reanalyzed data from the Genpax pipeline. Two near 0-distance clusters are visualized

A

ID		AL11168.1	5920	5921	5922	5923	5925	5926	5927	5928	5929	5930	5931	5932	5933	5934	5935	5936	5937	5938	5939	5940	5941	5942	5943
	Sender	London	Aberystwyth	Aberdeen	Bristol	Bristol	Glasgow	Glasgow	Glasgow	Norwich	Norwich	LSHTM	LSHTM	Manchester	PHE	Oxford	Sheffield	Sheffield	Sheffield	Sheffield	UCL	Surrey	Surrey	Edinburgh
ID	Sender	Distribution hub	LSHTM	PHE	PHE	PHE	LSHTM	LSHTM	LSHTM	PHE	PHE	LSHTM	LSHTM	LSHTM	PHE	PHE	LSHTM	LSHTM	LSHTM	LSHTM	LSHTM	PHE	PHE	LSHTM
AL11168.1	London	LSHTM		12	16	26	26	19	14	17	35	16	28	17	16	17	14	54	22	106	19	24	21	18	16
5920	Aberystwyth	PHE	4		9	19	18	8	13	13	29	8	18	13	14	8	7	48	23	98	13	20	10	8	14
5921	Aberdeen	PHE	6	1		28	20	14	20	15	30	13	21	15	18	12	9	51	25	101	16	22	14	13	16
5922	Bristol	PHE	13	8	9		26	19	26	25	35	22	28	24	28	24	22	63	36	104	29	31	22	20	27
5923	Bristol	PHE	8	3	2	11		18	28	25	35	19	27	25	23	23	22	52	36	104	28	33	16	20	23
5925	Glasgow	LSHTM	5	0	1	8	3		19	19	28	12	22	16	18	14	11	47	28	101	21	23	16	14	17
5926	Glasgow	LSHTM	4	4	5	13	8	5		11	34	16	27	11	14	16	17	54	14	101	14	21	19	18	14
5927	Glasgow	LSHTM	4	3	4	11	6	3	2		33	14	29	6	9	13	13	54	22	98	12	18	16	15	11
5928	Norwich	PHE	9	5	6	11	8	5	9	7		23	37	30	31	30	27	51	41	106	35	38	33	31	31
5929	Norwich	PHE	4	0	1	6	3	0	4	2	5		19	13	15	9	9	48	24	94	18	21	15	14	15
5930	LSHTM	LSHTM	8	4	8	11	7	4	7	6	8	4		29	27	22	20	56	37	106	29	31	22	22	27
5931	LSHTM	LSHTM	5	4	5	12	7	4	3	0	8	3	7		9	16	13	52	19	99	12	18	19	16	8
5932	Manchester	LSHTM	5	4	5	12	7	4	3	0	8	3	7	0		17	15	53	21	102	8	19	16	17	9
5933	PHE	PHE	5	0	1	8	3	0	5	3	5	0	3	3	3		5	51	26	98	18	22	11	10	18
5934	Oxford	PHE	5	0	0	8	2	0	5	3	5	0	3	4	4	0		50	23	97	19	20	13	12	19
5935	Sheffield	LSHTM	22	15	18	23	19	15	22	19	18	17	20	21	21	17	16		58	69	58	60	52	50	53
5936	Sheffield	LSHTM	10	10	11	19	14	11	8	8	15	9	14	9	9	11	11	28		108	53	27	30	29	20
5937	Sheffield	LSHTM	51	45	50	52	44	45	45	49	45	45	44	47	49	43	47	31	52		104	107	102	101	102
5938	Sheffield	LSHTM	4	3	4	11	6	3	2	0	7	2	6	0	0	3	3	20	7	47		18	17	19	8
5939	UCL	LSHTM	5	4	5	12	7	4	3	1	8	3	7	0	0	3	4	18	9	46	1		23	23	19
5941	Surrey	PHE	6	1	2	7	4	1	6	4	6	0	5	5	5	1	1	18	12	48	4	5		6	17
5942	Surrey	PHE	7	2	3	8	5	2	7	5	7	1	6	6	6	2	2	16	13	48	5	6	1		18
5943	Edinburgh	LSHTM	5	4	5	12	7	4	3	0	8	3	8	0	0	3	4	21	9	49	0	2	6	6	

B

Table 1: A) SNP distance matrix showing the pairwise comparisons between samples from Pascoe *et al.* (2019). B) SNP distance matrix showing pairwise comparisons between samples using the Genpax pipeline.

Isolate	ID	Source laboratory	Original source	Archived
1	5920	Aberystwyth	NCTC	2015
2	5921	Aberdeen	NCTC	2002
3	5922	Bristol	NCTC	2002
4	5923	Bristol	NCTC	2005
5	5925	Glasgow	London	2000
6	5926	Glasgow	NCTC (via Martin Skirrow)	2000
7	5927	Glasgow	London	2000
8	5928	Norwich	NCTC	2004
9	5929	Norwich	NCTC	2004
10	5930	London	London	2000
11	5931	London	London	2000
12	5932	Manchester	London	2003
13	5933	Swansea	NCTC	2016
14	5934	Oxford	NCTC	2014
15	5935	Sheffield	London	2015
16	5936	Sheffield	London	2013
17	5937	Sheffield	London (via Birmingham)	2000
18	5938	Sheffield	London	2010
19	5939	London	London	2002
20	5940	London	London	2002
21	5941	Surrey	NCTC (via Cambridge)	2000
22	5942	Surrey	NCTC (via Cambridge)	2000
23	5943	Edinburgh	London (via Sheffield)	2013
Reference	AL11168	NCTC	London	2000

Table 2: Metadata table from the Pascoe study [1]. London sourced samples cluster together and NCTC sourced samples cluster together (with the exception of isolate 5). Full table available in the Pascoe study.

- Phylogenetic trees reflect and are consistent with the relationship between the 23 isolates in terms of their known origin and distribution histories and can be used to trace the transfer of samples across different locations (See Figure 1 and Table 2).
- Matrices show a significant reduction in pairwise SNP differences. The Genpax method exhibits reduced noise, as well as enhanced resolution and sensitivity.
- NCTC sourced samples tend to cluster together. This is also the case for London sourced samples.

Check out our website and other supporting information

Declaration

This research was funded by Genpax.

Genpax is a bioinformatics company developing novel solutions that overcome the limitations of established analysis strategies to maximize the usefulness of bacterial genome sequences in infection prevention and control.

Conclusions

- The re-analysis generates an inferred set of relationships that is more parsimonious with respect to both the ancestral strain, and the pattern of distribution of strains between laboratories.
- The re-analysis does not confirm the original conclusions as to the substantial diversity and non-comparability of studies conducted in different laboratories using derivatives of *C. jejuni* strain 11168.
- The Genpax method shows increased sensitivity and decreased noise compared to previous analyses, and more accurate SNP-calling enables better identification and near-distance determination between isolates and the relationships between them, thereby enabling detection and definition of transmission and outbreaks in future applications.
- These results strongly validate the accuracy, resolution, and performance of a reference-free WGS analysis that is applicable to any strain, when measured against the performance of established methods with an ideal reference genome.

Reference-free whole genome SNP analysis of *Pseudomonas aeruginosa*, with the restructuring of outbreaks analyzed with established methods

Poster No. 227
Date: 6/18/2023

Arthur Poivet, Rebecca J. Bengtsson, Dan Frampton, John F. Peden & Nigel J. Saunders: **Genpax**, London, United Kingdom

research@genpax.co
+44 203 603 6869

Introduction

- *Pseudomonas aeruginosa* is one of the main nosocomial pathogens with a high prevalence in burns units, Intensive Care Units (ICUs), and patients with cystic fibrosis [1].
- *Ps. aeruginosa* has been classified as one of three critical priority pathogens, and is considered a major threat by the WHO and CDC due to the emergence of multidrug- and extended-drug-resistant clinical isolates [2,3].
- *Ps. aeruginosa* is an ancient species with a diverse clonal population structure, for which reference genome solutions are unsuitable.
- The lack of suitable reference genomes typically restricts analysis to Sequence Typing (MLST, cgMLST, wgMLST).
- A general solution that does not depend upon local, high-quality references is needed to deliver clinical genomics and proactive sequencing for infection prevention and control for this AMR priority species.

Objective: To test a reference-free whole genome SNP analysis for *Ps. aeruginosa*, and to compare its performance against published studies.

Methods

- The performance of the Genpax analysis pipeline was evaluated using datasets from two published studies, and compared to the original findings:

➢ 156 isolates from Magalhães et al [4].
  - Clinical and environmental isolates were originally typed and separated in 3 groups mainly corresponding to ST1076, ST253, and ST17.
  - For each ST group, a complete reference genome was created by sequencing a clinical isolate with both PacBio and Illumina HiSeq technologies, and whole genome SNP distances were obtained using these references.

➢ 38 isolates from Cunningham et al., which were originally analyzed using two cgMLST methods. One of those was an in-house method which addressed 4,041 alleles based on the PAO1 reference genome [5].

References

1. Stover, C. K., et al. "Complete Genome Sequence of Pseudomonas aeruginosa PAO1, an Opportunistic Pathogen." *Nature* 406, no. 6799 (August 2000): 959–64. https://doi.org/10.1038/35023079.
2. Reyes, J., et al. "Global Epidemiology and Clinical Outcomes of Carbapenem-Resistant Pseudomonas aeruginosa and Associated Carbapenemases (POP): A Prospective Cohort Study." *The Lancet Microbe* 4, no. 3 (March 1, 2023): e159–70. https://doi.org/10.1016/S2666-5247(22)00329-9.
3. Karruli, A., et al. "Evidence-Based Treatment of Pseudomonas aeruginosa Infections: A Critical Reappraisal." *Antibiotics* 12, no. 2 (February 2023): 399. https://doi.org/10.3390/antibiotics12020399.
4. Magalhães, B., et al. "Combining Standard Molecular Typing and Whole Genome Sequencing to Investigate Pseudomonas aeruginosa Epidemiology in Intensive Care Units." *Frontiers in Public Health* 8 (January 28, 2020): 3. https://doi.org/10.3389/fpubh.2020.00003.
5. Cunningham, B. A., et al. "Core Genome Multilocus Sequence Typing and Antibiotic Susceptibility Prediction from Whole-Genome Sequence Data of Multidrug-Resistant Pseudomonas aeruginosa Isolates." *Microbiology Spectrum* 10, no. 6 (November 9, 2022): e03920-22. https://doi.org/10.1128/spectrum.03920-22.

Results – SNP analysis

- The mean sequence length analyzed using the Genpax pipeline represents around 85% of the average genome length.
- The Genpax analysis replicated the results obtained from whole genome SNP analysis (which used optimal bespoke reference genomes), both for clonal outbreaks (figure 1) and more diverse isolates (figure 2).
- Previously identified clusters of closely related strains were identified by the Genpax pipeline, and the performance was assessed by comparing within-cluster SNP distances (Table 1).
  - In the ST1076 outbreak (Figure 1), all samples fall within 14 SNPs of each other, and local transmission chains can be inferred (examples in colored boxes).
  - In some cases, the environmental source of the infection could be identified (e.g. Figure 2 – blue cluster).

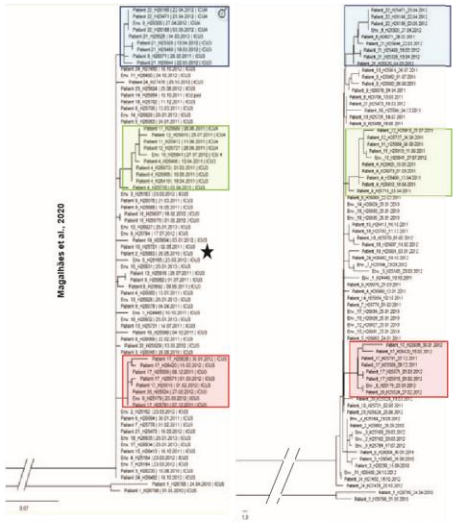


Figure 1: Comparison of whole genome SNP analysis of all ST1076 isolates, between a tree from Magalhães et al. (left), and the reference-free Genpax solution (right) (* = reference).

ST	target	Magalhães et al. (SNPs)	Genpax (SNPs)
1076	*All iso.*	≤ 14	≤ 14
1076	All same-patient iso.	≤ 10	≤ 11
1076	Patient 4 iso.	≤ 7	≤ 5
1076	Patient 24 iso.	≤ 2	≤ 4
253	All same-patient iso.	≤ 4	≤ 3
253	Burns unit	≤ 11	≤11
253	ICU 5 cluster	11 ≤ 14	10 ≤ 14
253	ICU 2 cluster (suspected outbreak)	≤ 1	≤ 0
17	All same-patient iso.	≤ 6	≤ 12
17	Suspected outbreak (figure 2 - green)	≤ 13	≤ 12
17	Patient 11 cluster (figure 2 - blue)	≤ 7	≤ 2

Table 1: Comparing the performance of the Genpax pipeline by comparing whole genome SNP distances of closely related isolates.

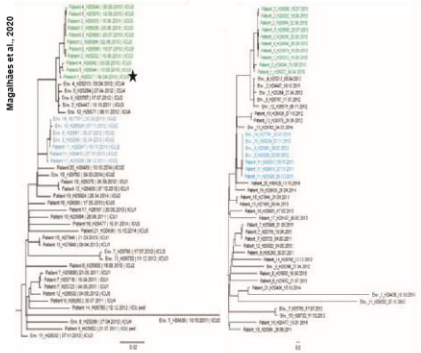


Figure 2: Comparison of whole genome SNP analysis of all ST17 isolates, between a tree from Magalhães et al. (left), and the reference-free Genpax solution (right).
A suspected outbreak in the burns unit is coloured in green. A second subclade composed of isolates from patient 11 and environmental isolates from the same ICU is coloured in blue (* = reference).

Results – SNP vs cgMLST

- This new SNP-based approach successfully identified outbreak clusters from clustered strains of diverse clonal complexes.
- The analysis demonstrates superior resolution and more informative sub-structuring of the outbreak than the minimum spanning trees derived from either of the two cgMLST schema under comparison.
- In Figure 3a the SNP-level analysis identified differences in 3 strains previously found to be identical by cgMLST.
- In Figure 3b the analysis of cluster 2 confirmed that 3 isolates near the root of the outbreak were within 3 SNPs and yielded different transmission inferences:
  - MBRL-2439 and MBRL-2456 are no longer sequential in the transmission chain.
  - MBRL-2429 is not tangential to the outbreak structure.

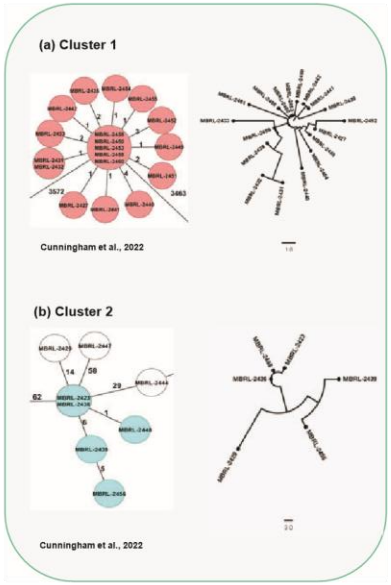


Figure 3: Comparison between cgMLST minimum spanning trees (left) and Genpax dendrograms (right). The distances on the minimum spanning tree refer to allelic differences [5].

Conclusions

- The novel Genpax methodology accurately determined strain identity without the need for a closely related reference genome, prior knowledge of strain types, or clonal clusters.
- The achieved resolution in this study matched that of an expensive and inscalable customized high-quality approach and surpassed the resolution achieved with cgMLST approaches.
- The utilization of this innovative analysis tool can enable real-time phylogenetic analysis in clinical settings.

Check out our website and other supporting information

Declaration

This research was entirely funded by Genpax.

Genpax is a bioinformatics company developing novel solutions that overcome the limitations of established analysis strategies to maximize the usefulness of bacterial genome sequences in infection prevention and control.

Confidential. Copyright and IP owned by Genpax.

Openly comparable and scalable SNP-resolution analysis for *Listeria monocytogenes* using a novel genome comparison tool

Poster No. 269
Date: 6/17/2023

Georgina R. Russell, Benedict J. Uttley, Dan Frampton, John F. Peden, Nigel J. Saunders: **Genpax,** London, United Kingdom

research@genpax.co.uk
+44 203 603 6869

Introduction

- *Listeria monocytogenes* is a food-borne pathogen with symptomatic infections resulting in a highhospitalization and mortality rate.
- *L. monocytogenes* persists in food processing environments for extended periods and can be widely distributed through food transportation networks.

•Public health and food safety laboratories need to accurately and iteratively compare strains.

•It has a diverse clonal population structure and lacks the comprehensive set of reference genomes necessary to underpin traditional whole genome SNP analysis with optimum accuracy and resolution.

•It is therefore an ideal candidate for novel reference-free and scalable solutions that work at SNP-level resolution.

Objective: To assess the ability of the Genpax pipeline to integrate and compare findings from three previously published studies [1-3] that use a range of cgMLSTand reference-genome SNP analyses from three different laboratories in Germany and Austria.

Methods

•Sequences(n=587)fromthreedifferentstudies[1-3]spanningisolatesfrommultipleEuropeancountries,mostly overthelast15yearsweredownloaded.

•ThesewereanalyzedwiththelatestGenpaxdevelopedpipelineandclusteranalysisofSNPpairwisedistances wasconducted.

Results

•Themedianof each input genome that was analyzed was 91.5% of the average genome length and389,933 variant positions were called. In contrast the 1,701 cgMLSTschemefor *L. monocytogenes* analyzes 53.5% of the genome ~1.5Mb

•At a 20-SNP threshold, 54 clusters were identified ranging in size from 2 to 51 isolates: ten contained isolatesfr om multiple studies.

•At a 2-SNP threshold, 41 clusterswere identified ranging in size from 2 to 42 isolates: four contained insolates from multiple studies.

•We found previously unrecognized relationships spanning laboratories and countries of isolation.

•This analysis also clustered human and food isolates together (3-SNP threshold, not shown) providing links betw een source and patient.

	Halbedel 2018	Hyden 2016	Schmid 2014	SourceSerogroupCountryLineage
Cluster 1	13	1	0	HumanIIaGermanyLineage II
Cluster 2	0	9	5	FoodIIbAustriaLineage I
Cluster 3	1	1	0	HumanIIaGermany/AustriaLineage II
Cluster 4	1	0	1	HumanIIbGermanyLineage I

Table 1: Clusters of isolates within 2SNPs containing isolates from multiple studies with selected metadata

Results-continued

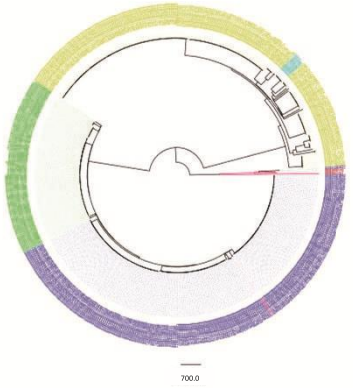


Figure1:
Concordance between serogroups and clustering

Lineagel: IIb, Ivb, IVb-v1
LineageII: IIa, IIc
LineageIII: IIIa, Ivb

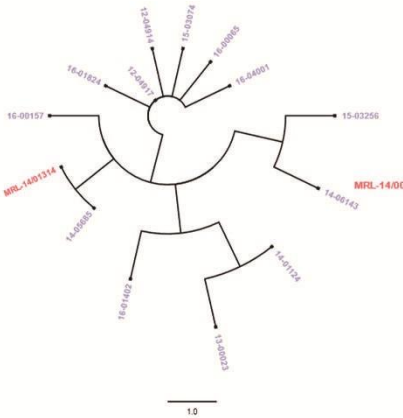


Figure 2:
Dendrogram ofCluster 1 isolates, colour indicates source see Table 1; isolate from the Hydenstudy is in red).

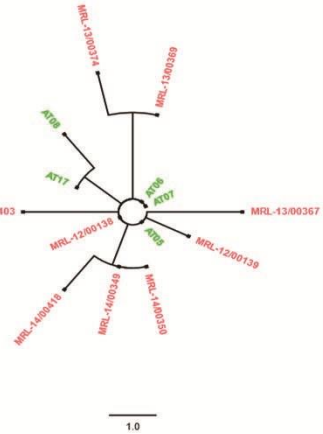


Figure 3:
Dendrogram ofCluster 2 isolates, colour indicates source see Table 1; isolates from the Schmid study are in green. Isolates from the Hyden study are in red).

Conclusions

•Clusters were found with isolates separated by time, source, and location.

•Our reference-free SNP-level resolution provided additional population structuring and transmission inference to cgMLST, and traditional whole genomeSNP.

•These findings show the value and importance of being able to meaningfully compare strains over temporal and geographical space with SNP resolution at scale.

Declaration

This research was entirely funded by Genpax.

Genpax is a bioinformatics company developing novel solutions that overcome the limitations of established analysis strategies to maximize the usefulness of bacterial genome sequences in infection prevention and control.

Check out our website and other supporting information

References

.Halbedel, S., Prager, R., Fuchs, S., Trost, E., Werner, G. and Flieger, A., 2018. Whole-genome sequencing of recent Listeria monocytogenes isolates from Germany reveals population structure and disease clusters. *Journal of clinical microbiology*, *56*(6), pp.e00119-18.

.Hyden, P., Pietzka, A., Lennkh, A., Murer, A., Springer, B., Blaschitz, M., Indra, A., Huhulescu, S., Allerberger, F., Ruppitsch, W. and Sensen, C.W., 2016. Whole genome sequence-based serogroupingof Listeria monocytogenes isolates. *Journal of biotechnology*, *235*, pp.181-186.

.Schmid, D., Allerberger, F., Huhulescu, S., Pietzka, A., Amar, C., Kleta, S., Prager, R., Preussel, K., Aichinger, E. and Mellmann, A., 2014. Whole genome sequencing as a tool to investigate a cluster of seven cases of listeriosis in Austria and Germany, 2011–2013. *Clinical Microbiology and Infection*, *20*(5), pp.431-436.

Novel genome comparison tool reveals both false-positive and false-negative MRSA and MSSA strain identification and a failure to detect transmission-linked strains using phenotypic, PCR, and previous genomic strategies

Poster No. 266
Date: 6/17/2023

James C. Littlefair, Benedict J. Uttley, Gareth Linsmith, Dan Frampton, John F. Peden, & Nigel J. Saunders: **Genpax**, London, United Kingdom

research@genpax.co
+44 203 603 6869

Introduction

- *Staphylococcus aureus* is associated with >1 million deaths a year globally and is an AMR priority pathogen [1].
- Methicillin resistance in *S. aureus* is typically conferred by the presence of the *mecA* gene carried on the mobile SCCmec cassette, which is often spontaneously lost during culture, leading to discordant phenotypes [2].
- Transmission links may be overlooked due to the common approach of partitioning strains into MRSA and MSSA, especially where the determination of methicillin resistance can be unreliable using current methodologies such as indicator media, susceptibility testing, and PCR.
- Thus, it is necessary to use a genome comparison tool coupling accurate SNP-resolution strain identification with optimised gene detection.

Methods

- All publicly-deposited readsets (n = 369, as of April 2023) from two published studies [3, 4] which collected MRSA and MSSA isolates in tandem within the same acute hospital between May 2017 and March 2019 were processed through the Genpax analysis pipeline. This included SNP-resolution strain identification and gene detection (which assigns gene detection confidence and identifies putative degeneracy).
- Isolates which met inclusion criteria (107/111 previously identified as MRSA and 243/258 previously identified as MSSA using screening methods including EUCAST susceptibility testing and PCR) were screened for *mecA/mecC* and *mupA/mupB*.
- The genes *lukS-PV* / *lukF-PV* (PVL), *tsst-1*, *eta*, *etb*, *etd* and *ete* were also screened for as they are important for the clinical management of *S. aureus*.

Results – *mecA*

- 15/107 (~14%) of isolates previously identified as MRSA were both *mecA*- and *mecC*-.
- 8/243 (~3%) of isolates previously identified as MSSA were *mecA*+, all high confidence.
- Segregating strains into MRSA and MSSA based on susceptibility testing and PCR prior to genomic distance determination in the original analyses, led to missing transmission links in 9/48 (~19%) of transmission clusters (maximum pairwise distance of 15 SNPs).
- In the largest transmission cluster (n = 24), the original analyses excluded 4 isolates, of which 3 were identified as MSSA despite being *mecA*+, and 1 was genuinely *mecA*- (Figure 2).

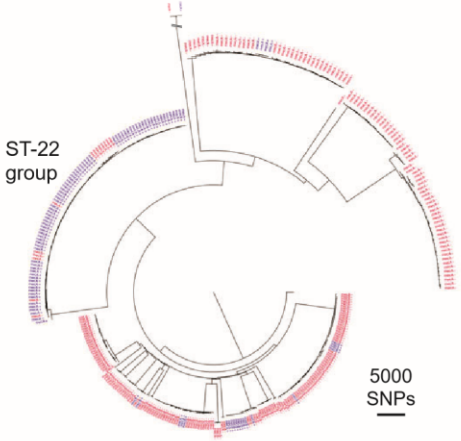


Figure 1: NJ dendrogram showing the distribution of *mecA* among all processed isolates. *mecA*+ isolates are colored blue and *mecA*- isolates are colored red. *mecA*+ isolates are found mostly within ST22-group.



Figure 2: NJ dendrogram showing the largest transmission cluster. MRSA (*mecA*+) and MSSA (*mecA*-) isolates are colored blue and red, respectively. The 3 MRSA isolates previously identified as MSSA are marked with an asterisk.

Results – other genes

- 2 out of 3 of the previously identified mupirocin-resistant isolates were confirmed to be *mupA*+, both of which were high confidence (Table 1).
- 6 isolates were found to be *PVL*+ (5 MSSA; 1 MRSA). Both constituent genes of this complex were found with high confidence in all isolates.
- 41 isolates were found to be *tsst-1*+ (38 MSSA; 3 MRSA). All were found with high confidence bar one which was found with medium confidence. These were overwhelmingly found in ST30/36-group isolates, even outside of transmission clusters (Figure 3).
- 12 isolates were found to be *eta*+ (12 MSSA), all of which were high confidence except one which was medium confidence.
- 2 isolates were found to be *etd*+ (2 MSSA), both of which were found with high confidence.

Table 1: Isolates identified as mupirocin resistant in the original analysis or *mupA*+ in the Genpax reanalysis

Sample	*mecA* status	Available?	EUCAST mupirocin phenotype	*mupA* status	*mupA*+ confidence
C1310091	*mecA*+	Yes	Resistant	*mupA*+	High
C165383	*mecA*+	Yes	Susceptible	*mupA*+ (frameshifted)	High
C76440	*mecA*-	Yes	Resistant	*mupA*+	High
HN0048.3	*mecA*-	Yes	Resistant	*mupA*-	N/A
PO0329	N/A	No	Resistant	N/A	N/A

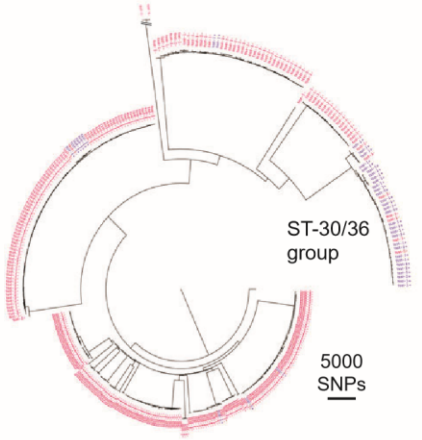


Figure 3: NJ dendrogram showing the distribution of *tsst-1* among all processed isolates. *tsst-1*+ isolates are colored blue and *tsst-1*- isolates are colored red.

Conclusions

- The reanalysis reveals putative false-positive and false-negative MRSA and MSSA determination by methods such as susceptibility testing and PCR, potential spontaneous gain/loss of *mecA* and *mupA*, and potential cryptic resistance.
- Partitioning strains into MRSA and MSSA leads to missed transmission-links, particularly when done unreliably, and indicates that transmission inference requires WGS of all clinically relevant *S. aureus* isolates supported by accurate, high-resolution, scalable genome analysis. To address the full diversity of isolates, a clonal reference is not required to get accurate and comprehensive results.
- The gene detection provided information on both resistance determinants and markers that indicate the requirements for different clinical management that were highly concordant with the strains underlying relationships, enabling confident interpretation of findings.

References

1. Ikuta KS, Swetschinski LR, Robles Aguilar G, Sharara F, Mestrovic T, Gray AP, et al. Global mortality associated with 33 bacterial pathogens in 2019: A systematic analysis for the global burden of disease study 2019. The Lancet. 2022;400(10369):2221–48. doi:10.1016/S0140-6736(22)02185-7
2. Tickler IA, Goering RV, Dowell S, Le VM, Johar L, Obradovich AE, et al. Mobile genetic elements responsible for discordant Staphylococcus aureus phenotypes and genotypes in the same blood culture bottle. Diagnostic Microbiology and Infectious Disease. 2020;98(4):115175. doi:10.1016/j.diagmicrobio.2020.115175
3. Kinnevey PM, Kearney A, Shore AC, Earls MR, Brennan GI, Pouwelbunmel TT, et al. Methicillin-resistant Staphylococcus aureus transmission among healthcare workers, patients and the environment in a large acute hospital under non-outbreak conditions investigated using whole-genome sequencing. Journal of Hospital Infection. 2021;118:99–107. doi:10.1016/j.jhin.2021.08.020
4. Kinnevey PM, Kearney A, Shore AC, Earls MR, Brennan GI, Pouwelbunmel TT, et al. Methicillin-susceptible Staphylococcus aureus transmission among healthcare workers, patients and the environment in a large acute hospital under non-outbreak conditions investigated using whole-genome sequencing. Journal of Hospital Infection. 2022;127:15–25. doi:10.1016/j.jhin.2022.05.004

Declaration

This research was entirely funded by Genpax.

Genpax is a bioinformatics company developing novel solutions that overcome the limitations of established analysis strategies to maximize the usefulness of bacterial genome sequences in infection prevention and control.

Check out our website and supporting information

Economic and health impact modelling of a whole genome sequencing-led intervention strategy for bacterial healthcare-associated infections for England and for the USA

Poster No. 290
Date: 6/17/2023

John M. Fox, Nigel J. Saunders & Susie H. Jerwood: **Genpax**, London, United Kingdom

research@genpax.co.uk
+44 203 603 6869

Introduction

- Bacterial HAIs are a substantial source of global morbidity and mortality, resulting in increased length of hospital stay and high healthcare costs.
- Costs associated with HAI ranges from $35 to $45 billion in the USA [1].
- WGS has been promoted as a new gold standard for outbreak detection, but widespread adoption to date is limited.
- The upfront costs of WGS implementation have been identified as obstacles to adoption.
- Previous models addressing the impact of WGS on bacterial HAI have predicted a wide range of clinical and financial impacts from various methodologies and scope [2,3].
- It is timely to determine the economic viability and impact of routine diagnostic bacterial genomics.
- The aim of building this model was to evaluate the clinical and economic impact of a prospective WGS-led track and trace system of eleven common healthcare associated and AMR priority bacterial pathogens in England and the USA compared to the current standard of care, without WGS.

Methods

- Using a synthesis of published models [2,3], inputs from national statistics, and peer-reviewed articles the clinical and financial impact models were created to address the most common nosocomial infections found in England and the USA.
- These are caused by *Staphylococcus aureus*, *Escherichia coli*, *Enterococcus* species, *Klebsiella* species, *Enterobacter* species, *Acinetobacter* species, *Stenotrophomonas maltophilia*, *Clostridioides difficile*, *Pseudomonas* species (mainly *Ps. aeruginosa*), *Citrobacter* species and *Serratia* species.
- All models were constructed, and analyses performed in Excel.
- Sensitivity analyses were conducted for each variable by varying the upper and lower limits within a wide range of available evidence.

Results

- The model shows bacterial HAI currently cost the NHS in England around £3 billion annually.
- WGS-based surveillance is predicted to cost £61.1 million associated with the prevention of 3,408 HAI and 1,257 deaths based on a cluster detection and intervention turnaround time of seven days.
- The net cost saving was £478.3 million, of which £65.8 million were from directly incurred savings (antibiotics, consumables etc.) and £412.5 million from opportunity cost savings due to re-allocation of hospital beds and healthcare professionals.
- The USA model indicates that the bacterial HAI care baseline costs are around $18.3 billion.
- WGS surveillance cost $169.2 million and resulted in a net saving of ca.$3.2 billion, while preventing 169,260 HAIs and 4,862 deaths, also based on a cluster detection and intervention turnaround time of seven days.
- The average reduction of total infections was 18% from using WGS.
- This clinical impact model estimated *S. aureus* to be the most common bacterial HAI in both England and the USA, and the cause of most deaths in the USA, with 17,176 deaths annually. In comparison, *E. coli* was responsible for the most nosocomial deaths in England, with 1,456 deaths.
- The model predicts a return to the hospitals of £7.83 per £1 invested in diagnostic WGS in the UK, and US$18.74 per $1 in the USA.

Results –England sensitivity analysis

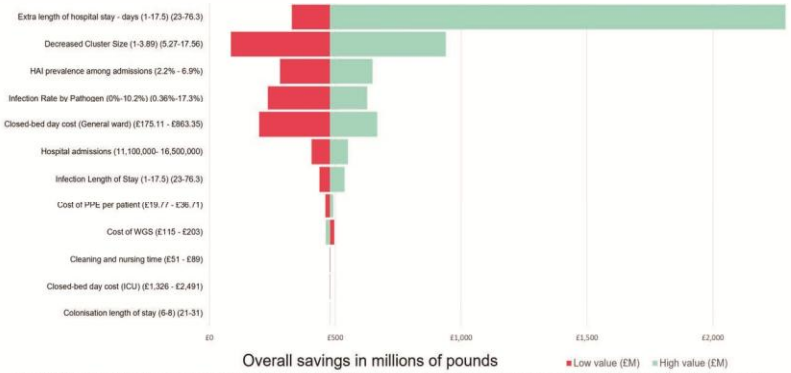


Figure 1: All savings are in millions of pounds relative to a base saving of £478.3 million. Low value (£M) represents the lower input for each variable and equally the High value (£M) represents the higher model input. The higher value for Cost of WGS was the only variable to reduce savings

Results -continued

Table 3: Estimated number of HAI and death for each pathogen group using current practice and estimated number of avoided infections and death with a WGS-based system

England	Current practice		WGS	
Organism	HAI	Death	HAI avoided	Death avoided
Staphylococcus aureus	100,044	1,009	10,765	109
Stenotrophomonas maltophilia	1,842	37	174	3
Escherichia coli	73,515	1,456	15,071	298
Enterococcus species	61,694	1,252	25,572	519
Klebsiella species	58,359	986	3,793	64
Enterobacter species	16,074	294	6,105	112
Acinetobacter species	9,171	128	2,201	31
Clostridioides difficile	41,685	375	7,816	70
Pseudomonas species (mainly *Ps. aeruginosa*)	8,337	375	787	35
Citrobacter species	14,173	68	1,338	6
Serratia species	8,337	92	787	9
Total	393,229	6,073	74,408	1,257

USA	Current practice		WGS	
Organism	HAI	Death	HAI avoided	Death avoided
Staphylococcus aureus	202,073	17,176	21,743	1,848
Stenotrophomonas maltophilia	3,722	74	351	7
Escherichia coli	148,490	1,336	30,440	274
Enterococcus species	124,612	2,530	51,652	1,049
Klebsiella species	117,876	2,299	7,662	149
Enterobacter species	32,466	1,623	12,331	617
Acinetobacter species	18,523	926	4,446	222
Clostridioides difficile	168,394	2,526	31,574	474
Pseudomonas species (mainly *Ps. aeruginosa*)	50,518	1,516	4,769	143
Citrobacter species	28,627	137	2,702	13
Serratia species	16,839	707	1,590	67
Total	912,141	30,851	169,260	4,862

Table 4: Estimated differences in costs in England and the USA with current practice and WGS surveillance

	England		USA	
Direct hospital resources	**Current practice**	**WGS**	**Current practice**	**WGS**
WGS	£ -	£ 61,118,073	$ -	$ 169,245,276
PPE	£ 266,750,089	£ 212,448,937	$ 740,286,599	$ 593,475,563
Consumables	£ 63,113,318	£ 51,170,819	$ 177,685,114	$ 144,713,345
Antibiotic treatment	£ 319,695,715	£ 259,060,468	$ 1,865,639,355	$ 1,353,822,910
Total cost of hospital resources	£ 649,559,122	£ 583,798,298	$ 2,583,611,068	$ 2,261,257,094
Allocation of hospital beds and healthcare professionals	**Current practice**	**WGS**	**Current practice**	**WGS**
Extra length of stay - General Ward	£ 2,220,992,705	£ 1,820,127,623	$ 15,278,211,667	$ 12,501,524,165
Extra length of stay – ICU	£ 4,234,642	£ 3,479,079	$ 168,858,625	$ 146,123,367
Cleaning and nursing time	£ 24,289,160	£ 19,696,643	$ 68,180,372	$ 55,534,825
Cost of infection prevention and control team	£ 33,200,330	£ 26,918,060	$ 203,127,450	$ 165,434,527
Total cost of allocation of hospital beds and healthcare professionals	£ 2,282,716,838	£ 1,870,221,405	$ 15,718,378,114	$ 12,868,616,885
Overall total	£ 2,932,275,960	£ 2,454,019,702	$ 18,301,989,182	$ 15,129,873,979
Overall cost savings with WGS surveillance	£	478,256,258	$	3,172,115,203

Conclusions

- This economic analysis indicates that substantial savings and improvements in clinical outcomes are generated using proactive clinical genomics of bacterial pathogens.
- Sensitivity analyses demonstrate that calculations involved in length of hospital stay affect the financial outcome the most.
- The largest savings are associated with improved use of healthcare resources, due to avoidance of prolonged patient stays and the ability to use facilities more effectively.
- Savings were retained when tested by sensitivity analyses, which showed small impacts from the costs of WGS.

References

. Storr, J., A. Twyman, W. Zingg, N. Damani, C. Kilpatrick, J. Reilly, L. Price, M. Egger, M. L. Grayson, E. Kelley, and B. Allegranzi. 2017. 'Core components for effective infection prevention and controlprogrammes: new WHO evidence-based recommendations',*AntimicrobResist Infect Control*, 6: 6.

. Gordon, Louisa G., Thomas M. Elliott, Brian Forde, Brett Mitchell, Philip L. Russo, David L. Paterson, and Patrick N. A. Harris. 2021. 'Budget impact analysis ofroutinelyusing whole-genomicsequencing ofsixmultidrug-resistant bacterial pathogensinQueensland, Australia', *BMJ Open*, 11: e041968.

. Kumar, P., A. J. Sundermann, E. M. Martin, G. M. Snyder, J. W. Marsh, L. H. Harrison, and M. S. Roberts. 2021. 'Method forEconomic Evaluation of Bacterial Whole Genome Sequencing SurveillanceCompared toStandard ofCare inDetecting HospitalOutbreaks', *Clin InfectDis*,73:e9–e18.

Check out our website and supporting information

Declaration

This research was entirely funded by Genpax.

Genpax is a bioinformatics company developing novel solutions that overcome the limitations of established analysis strategies to maximize the usefulness of bacterial genome sequences in infection prevention and control.

INTRODUCTION TO THE FOLLOWING HEALTH ECONOMICS PAPER

A clinical genomics solution must provide an increase in patient and public safety with improved patient care and improved management of healthcare resources, while also being economically viable. Several previous publications indicate that this is the case for proactive clinical genomics for IPC. (These are cited in the following paper.)

However, these papers do not always include all attributable costs of sequencing, analysis, staffing and other infrastructure costs, and others are dated or don't work with the most current epidemiological information or a more limited set of species. This new analysis draws upon the best of this previous work, updates it, addresses a core set of healthcare-associated hospital-acquired AMR priority pathogens, using inclusive current real-world costs, and a best-available set of epidemiological information. The publication also makes its model available in an Excel format to enable local hospitals or others to modify it to generate locally informed versions and ongoing updates. Key findings of this modelling include:

- Addressing NHS England as an example:
    - it would be possible to **save over 400,000 bed days per year**, which is the equivalent of building and fully equipping and staffing a new 1100-bed hospital with full occupancy.
    - it is possible to **prevent more than 1,200 avoidable hospital care associated deaths**; representing 10-20% of estimated avoidable hospital deaths per year.
    - it is possible to **save at least £480 million per year** in avoidable costs.
- There is **no economic obstacle to adoption**, because the savings to hospitals and healthcare delivery systems are considerably greater than the costs of adopting and delivering proactive bacterial genomics surveillance for IPC.
- **Improved patient safety and actions to contain and prevent the spread of AMR** within the hospital can be achieved at negative costs.
- The **hospital-level costs savings** are dominated by improved use of healthcare resources, such that large savings remain with wide variations in the costs of sequencing and analysis.
- Larger savings and proportionate returns on investment are available in the US than the UK.

The only remaining obstacles to adoption are sequencing, which is now available in-house to any modern laboratory capable of typical microbiology and pathology services or through external services, and the expertise and resources that are required for analysis and interpretation that are now openly available from Genpax.

Finally, it should be noted that **these models are intentionally conservative**. They do not include savings from other activities such as combined environmental, healthcare worker, and preadmission screening; the additional benefits of addressing non-AMR/antibiotic sensitive strains with similar transmission mechanisms and clinical consequences (e.g. MSSA which has a 20 to 30% mortality); nor additional species. They also do not include costs associated with exceptional responses such as ward closure, rebuild and refits, equipment replacement, insurance company non-payment or claw-backs, or legal liabilities for hospital transmitted infections. Nor the savings from avoidable responses to ‘non-outbreaks’ that suspected on epidemiological grounds are caused by strains are unrelated and not connected, or being able to demonstrate that infections were not caused by hospital-associated strains.

Confidential. Copyright and IP owned by Genpax.

MICROBIAL GENOMICS

RESEARCH ARTICLE

Fox *et al.*, *Microbial Genomics* 2023;9:001087

DOI 10.1099/mgen.0.001087

MICROBIOLOGY SOCIETY

OPEN DATA OPEN ACCESS

Economic and health impact modelling of a whole genome sequencing-led intervention strategy for bacterial healthcare-associated infections for England and for the USA

John M. Fox, Nigel J. Saunders and Susie H. Jerwood*

Abstract

Bacterial healthcare-associated infections (HAIs) are a substantial source of global morbidity and mortality. The estimated cost associated with HAIs ranges from $35 to $45 billion in the USA alone. The costs and accessibility of whole genome sequencing (WGS) of bacteria and the lack of sufficiently accurate, high-resolution, scalable and accessible analysis for strain identification are being addressed. Thus, it is timely to determine the economic viability and impact of routine diagnostic bacterial genomics. The aim of this study was to model the economic impact of a WGS surveillance system that proactively detects and directs interventions for nosocomial infections and outbreaks compared to the current standard of care, without WGS. Using a synthesis of published models, inputs from national statistics, and peer-reviewed articles, the economic impacts of conducting a WGS-led surveillance system addressing the 11 most common nosocomial pathogen groups in England and the USA were modelled. This was followed by a series of sensitivity analyses. England was used to establish the baseline model because of the greater availability of underpinning data, and this was then modified using USA-specific parameters where available. The model for the NHS in England shows bacterial HAIs currently cost the NHS around £3 billion. WGS-based surveillance delivery is predicted to cost £61.1 million associated with the prevention of 74 408 HAIs and 1257 deaths. The net cost saving was £478.3 million, of which £65.8 million were from directly incurred savings (antibiotics, consumables, etc.) and £412.5 million from opportunity cost savings due to re-allocation of hospital beds and healthcare professionals. The USA model indicates that the bacterial HAI care baseline costs are around $18.3 billion. WGS surveillance costs $169.2 million, and resulted in a net saving of ca.$3.2 billion, while preventing 169 260 HAIs and 4862 deaths. From a 'return on investment' perspective, the model predicts a return to the hospitals of £7.83 per £1 invested in diagnostic WGS in the UK, and US$18.74 per $1 in the USA. Sensitivity analyses show that substantial savings are retained when inputs to the model are varied within a wide range of upper and lower limits. Modelling a proactive WGS system addressing HAI pathogens shows significant improvement in morbidity and mortality while simultaneously achieving substantial savings to healthcare facilities that more than offset the cost of implementing diagnostic genomics surveillance.

Impact Statement

This article estimates the impact of effective whole genome sequencing-based surveillance for tracking and intervening in bacterial nosocomial outbreaks of the 11 most common healtcare-associated infection (HAI) species in both England and the USA. The projected outcome would be to reduce the bacterial morbidity and mortality of HAI in hospitals while simultaneously reducing the cost of patient care and increasing the wider cost savings of England and the USA by £478.3 million and $3.2 billion respectively, with more efficient use of hospital resources.

Confidential. Copyright and IP owned by Genpax.

Limitations of cgMLST in current practice

Previously, scalability challenges of high-resolution analyses using SNVs, or SNV with other differences (e.g. indels, recombination), meant that detailed analysis for outbreak detection and investigation either had to address only small numbers of isolates, be periodically consolidated in major centres, or had to be preceded by an initial low-resolution but practically more deliverable step. Commonly this is a form of Sequence Typing (MLST, cgMSLT/cgST, or wgMLST). After this, strains of the same type or with a certain range of differences are selected for more detailed analysis of some kind. cgST has recognized limitations, as do the minimum spanning trees generated from it. But it has higher resolution than MLST, and can group strains into smaller and more analyzable groups for subsequent investigation. Although this can still present a scaling challenge to compare strains beyond limited time-frames or geographies, especially for the most clinically important and common clones.

Because cgST correlates well with other phylogenomic information in population-scale studies when compared to findings using more detailed analyses, it has been assumed that it will perform reliably and sensitively in selecting the strains that are potentially parts of outbreaks and transmission-linked clusters. However, the nature and scale of differences spanning populations does not necessarily reflect performance in distinguishing more closely related strains, with a method that is reported to generate slightly different results when using different assemblers, less than 40x coverage, and varied addressed alleles, even when reanalyzing the same sequencing files (e.g. Abdel-Glil *et al*. J. Clin Micro, 2022). However, because cgST when used in clinical and epidemiology studies is used as a pre-selection step, this assumption is not normally tested because ungrouped strains are not compared in detail.

The resources within IDEM perform analysis in detail throughout. Further, because it is a natural-reference free solution, the results remain fully comparable because they are not divided into groups that have been analyzed using different references, or references with different degrees of divergence from analyzed isolates. This enables the performance of cgST in the identification of putative outbreak members and transmission-linked isolates to be assessed.

Such an analysis has been performed using and comparing findings to the results of two large published studies of *Campylobacter* species. One of *C. jejuni*, the other of *C. coli*. This genus illustrates several challenges for WGS analysis, because it undergoes relatively frequent recombination resulting in a panmictic population structure. This means that defining genes and alleles may be difficult using the search methods used in cgST (BLAST), and that it is difficult (perhaps impossible) to establish a consistent set of good quality reference genomes. The analysis methods used within IDEM are natural reference free, and overcome these barriers. Thereby enabling a test of the performance of cgST in these species. The findings have potential implications not only for analyses performed solely using cgST, but also for all pipelines that use it as a preliminary step.

Limitations of sequence typing for isolate inclusion in outbreak investigations

Poster No. P2151
Date: 27/04/2024

K. Jackowska [1], D. Frampton [1], J.F. Peden [1], N.J. Saunders [1]. [1]Genpax - London (United Kingdom)

Email via: research@genpax.co
Tel: +41 44 586 86 06

Introduction

- Campylobacteriosis, the leading cause of gastrointestinal disease in the EU, is primarily caused by two species: *Campylobacter jejuni* and *Campylobacter coli*.
- Correct detection of outbreaks is key to infection control.
- MLST and cgMLST are commonly used to identify clusters for further analysis.
- Here, we assess their performance against an openly scalable solution comparing strains at SNV resolution to determine their sensitivity and specificity for likely transmission-linked strains, using data from two recent publications. (Hsu *et al.,* Harrison *et al.*)

Methods

- In total, 3762 readsets were processed through the Genpax IDEM platform.
- All pairwise SNP distances for 844 *C. jejuni* and 2918 *C. coli* isolates were determined using the platform.
- Allele difference groups, as well as cgST and MLST were defined and derived for each sample in the two publications. (Hsu *et al.,* Harrison *et al.*)
- The results were combined and compared.

Key

AD0	Allele difference of 0
AD5	Allele difference of 5
AD10	Allele difference of 10
AD25	Allele difference of 25
AD100	Allele difference of 100
AD200	Allele difference of 200
cgST	Core genome sequence type
MLST	Multi-locus sequence typing

Results - Overview

- Of the pairs which were linked by IDEM at a distance of ≤10 SNV, 80/318 (25%) of *C. jejuni* and 1374/2920 (47%) of *C. coli* isolate pairs were not clustered by cgMLST at an allelic difference of 10 (AD10); representing thresholds used for outbreak detection.
- Using a higher threshold of AD25, a portion (21/318 (7%)) of the ≤10 SNP *C. jejuni* isolates remained ungrouped.
- Strikingly, 4/39 (10%) *C. jejuni* and 47/117 (40%) *C. coli* 0-SNV isolates are also not grouped at AD10.
- Examination of reported sequence types (ST) showed that 11/318 (3%) of *C. jejuni* and 200/2920 (7%) *C. coli* sample pairs were identified as different sequence types (STs) while differing by ≤ 10 SNVs.
- In contrast, samples belonging to the same ST had a median pairwise SNV distance of 568 (IQR: 75-1961) for *C. jejuni* and 95 (IQR: 66-153) for *C. coli* respectively.

Campylobacter jejuni results

Figure 1. Distribution of cgST/AD/MLST pairs at each SNV distance threshold for *C. jejuni*

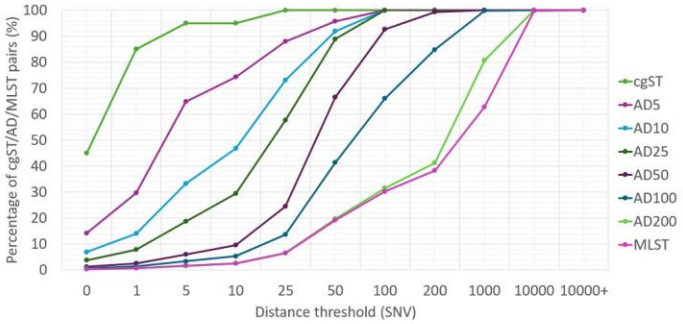
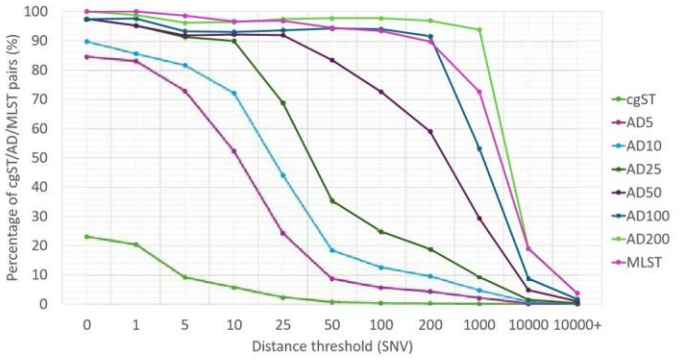


Figure 2. Percentage of pairs at each SNV distance threshold that have the same cgST/AD/MLST classification for *C. jejuni*



Campylobacter coli results

Figure 3. Distribution of AD/MLAST pairs at each SNV distance threshold for *C. coli*

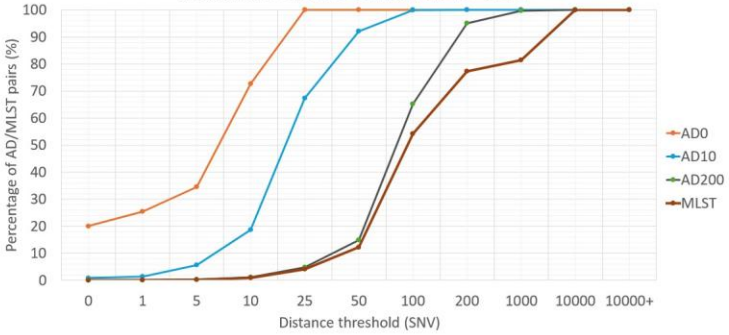
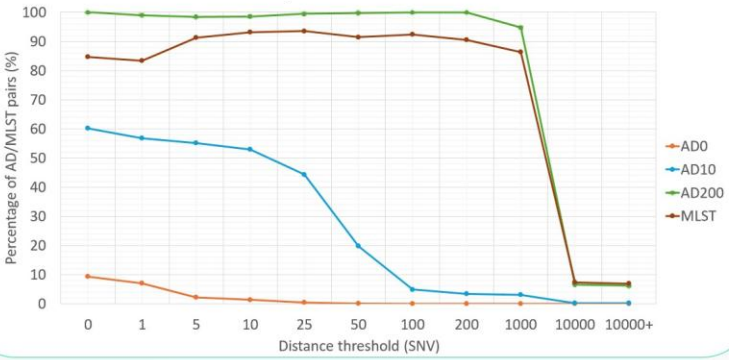


Figure 4. Percentage of pairs at each SNV distance threshold that have the same AD/MLST classification for *C. coli*



Conclusions

- Both MLST and cgMLST under- and over-predict sample linkage in both published studies (generate both false negative and false positive results).
- Across the two studies, *211 pairs of isolates ≤10 SNV distance identified by IDEM do not share the same sequence type* (ST) and **1466 pairs do not share the same cgMLST cluster at AD10**, which would normally exclude them from being identified as potential members of outbreaks and subsequent more detailed comparisons and analysis.
- In addition, of the 6% of isolate pairs identified as the same ST (300959/4514920), only 1% (3039) fall within a distance of 10 SNVs. The majority (99%) of same ST pairs have a greater SNV distance than would normally be considered to indicate outbreak / transmission connection.
- This performance indicates that Sequence Typing (MLST or cgMLST/cgST) is not an optimal first stage analysis for the detection and investigation of transmission-linked strains and outbreaks in these species.
- This may reflect the combined effects of issues inherent to the underlying methodology in the context of the highly recombining and panmictic nature of the genus. Similar analyses seem warranted in other species.

Declaration

This research was entirely funded by Genpax.

Genpax is a bioinformatics company founded in 2020 seeking to develop novel solutions that overcome the limitations of established analysis strategies to maximize the usefulness of bacterial genome sequences in infection prevention and control.

References

1. Hsu C-H, Harrison L, Mukherjee S, Strain E, McDermott P, Zhang Q, Zhao S. Core Genome Multilocus Sequence Typing for Food Animal Source Attribution of Human *Campylobacter jejuni* Infections. Pathogens. 2020; 9(7):532. https://doi.org/10.3390/pathogens9070532
2. Harrison L, Mukherjee S, Hsu C-H, et al. Core genome MLST for source attribution of *Campylobacter coli*. Frontiers. June 21, 2021. Accessed March 19, 2024. https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2021.703890/full.

Check out our website:

Copyright and IP owned by Genpax

Breaking Through the Challenges of *C. difficile* Surveillance

The analysis of *C. difficile* is known to be challenging, with the species encompassing a number of groups of related strains, some of which are further apart than the ANI that would usually indicate separate species. However, despite this evolutionary separation, there is still genetic exchange and recombination that can occur within this species group. This poses substantial challenges to genomics analysis for proactive surveillance and infection prevention and control; partly because results of MLST, cgMLST, and SNP-based analysis will give conflicting and misleading results in relation to the relatedness and connections between strains. For example, highly related strains that are part of outbreaks can appear to be unrelated. Also, it is not a species for which a good set of neutral references can exist for detailed analysis, and the use of different references impedes the ability to see strain relationships. The Genpax IDEM analysis overcomes these issues, and permits open-ended high quality comparison and connections of all strains.

Precision outbreak surveillance of *Clostridioides difficile* through natural reference-free WGS SNP-resolution analysis defining limitations of cgMLST in current practice

A. Pivnyuk, J. Peden, and N. Saunders; Genpax SA, London, United Kingdom

Clostridioides difficile presents substantial challenges in genomic strain identification and surveillance due to its highly diverse strains, for which there are no universally applicable references. We addressed this challenge using a natural reference-free whole genome sequencing (WGS) approach, achieving a high-resolution SNP detection rate that consistently differentiated strains separated by only 0-2 SNP differences, crucial for recent outbreak identification and transmission inference. This scalable solution has been validated against four major studies, encompassing 543 samples, which employed multiple methods to confirm the presence and type of *C.* difficile, including surveillance of outbreaks and performance analyzing prevalent RT types 014, 027, and 078. By employing the Genpax IDEM pipeline, our analysis successfully reproduced findings from these foundational studies, verifying ancestral strains and hybrid /recombining strains. These hybrids, 95% identical to known assemblies but with over 1000 SNPs in the remaining sequence, demonstrate the unique capabilities of IDEM to address such complex events necessary to robustly address this species. The tool also accurately detected the pathogenicity genes *tcdA* and *tcdB*, providing insights into their structure, even when non-functional. With a mean coverage of over 85% of the *C. difficile* genome across the 543 samples, our approach surpasses the resolution of other methods, in a way that is maintained when addressing large numbers of diverse strains. In conclusion, our natural reference-free WGS SNP-resolution analysis offers a potent tool for improving the detection and understanding of *C. difficile* outbreaks, showcasing the potential to transform public health responses to this persistent pathogen.

References:

- Knight, D.R. et al. (2023). European Journal of Clinical Microbiology & Infectious Diseases, 42(5), 607–619.
- Miles, J.A., et al. (2021). Microbial Genomics, 7(6).
- Kuenzil, A.B., et al. (2020). Journal of Hospital Infection, 106(2), 240–245.
- Knight, D.R., et al. (2017). Frontiers in Microbiology, 7, 2138.

Confidential. Copyright and IP owned by Genpax.

From Outbreaks to Prevention: IDEM Enables Rolling Surveillance

Listeria monocytogenes often causes prolonged and dispersed outbreaks with small numbers of differentiating differences from which to infer sources and transmission. The ability of the IDEM natural reference-free analysis to connect strains from multiple centres and studies has been presented previously (*vida supra*). The performance within the detail of outbreak analysis is equally important, to enable rapid detection and effective responses to outbreaks. Also, the precision of the IDEM analysis should enable more rapid and definitive decision making when using information from fewer strains and responses earlier in outbreak situations. To assess this, a reanalysis of an intensively studied and published outbreak is presented.

Outbreak of *Listeria monocytogenes* in enoki mushrooms from South Korea: Comparison of reference-independent SNP-resolution genome comparison tool with existing retrospective methods

G. R. Russell, L. Montemayor, G. Linsmith, D. Frampton, J. Peden, N. Saunders; Genpax, London, United Kingdom

Intro: Enoki mushrooms from a manufacturer in South Korea have been identified as the cause of an outbreak spanning 5 years (2016-2020) and multiple countries (US, Canada, France, and Australia). Identification of these listeriosis cases within nationally agreed wg/ cgMLST thresholds prompted extensive investigations by public health authorities to determine the possible source. *L. monocytogenes* is linked to high hospitalization and mortality, and its distribution is linked to food distribution networks, thus requiring multinational cooperation. *L. monocytogenes* has a population structure with diverse clonal lineages usually analyzed with a single reference strain. This prevents optimal SNP analysis for strains in all lineages. **Objective:** To assess the Genpax IDEM platform in comparison with an existing method in a previously investigated L. *monocytogenes* outbreak. Methods: 72 publicly deposited sequences published by Pereira et al. 2023 were re-analyzed using the reference-independent SNP-level resolution Genpax IDEM platform, and the resultant relationships determined and compared with those from the original study. **Results:** Dendrogram structures show largely congruent clustering of samples with few key differences, demonstrating the ability of the automated IDEM resource to generate similar results to previous carefully curated analysis. We were able to validate SNP differences identified in samples pairs that were previously identified as being 0 SNPs; as well as validating 0 distance sample pairs common to both sets of analysis. We demonstrated a more plausible transmission path for this outbreak within one of the identified clades, with a South Korean enoki mushroom sample basal to the clade. A feature of the Genpax IDEM pipeline is that it allows the direction of transmission to be ascertained, even for small numbers of isolates. **Conclusion:** This outbreak shows that multinational supply chains have many challenges associated with them when trying to ensure safe practices as a single supplier can cause outbreaks in multiple countries. The ability of the Genpax IDEM platform to compare samples to those previously analyzed is of great importance for *L. monocytogenes,* as it allows rolling and preventative surveillance consistent with current best standard methods to look retrospectively at an outbreak and removes the requirement for reference-choice. This level of resolution corroborates earlier results of the Genpax IDEM platform, including a *S. aureus* ring trial where 0 SNP pairs consistent with replicate groups were found.

References:

Confidential. Copyright and IP owned by Genpax.

Pereira E, et al., 2023. Multinational Outbreak of *Listeria monocytogenes* Infections Linked to Enoki Mushrooms Imported from the Republic of Korea 2016-2020. Journal of Food Protection, 86(7),p.100101. Littlefair JC, et al., Calling Zero: A new foundation for diagnostic bacterial genomics. ECCMID, 2022.

Transformative differences to practice enabled by Genpax IDEM

From reactive to proactive WGS for hospital IPC

Current use of WGS-based genomics is largely limited to the investigation of otherwise suspected outbreak-associated isolates. It has a reactive and remediation function, but it is not primarily a detection and prevention resource. In the rare settings where it is currently used proactively, it is typically limited to low resolution analysis (e.g. cgMLST), followed by later detailed analyses of selected strains; or to limited sampling space and time-windows to limit the number of strains compared (normally to not more than 50-100). This is because of multiple factors, and requires local expert teams, all of which are overcome by the Genpax IDEM solution.

Using IDEM information from proactive sequencing of targeted pathogens can **identify transmission-connected infections, outbreak clusters**, and for tracking and source identification, that can be augmented by environmental surveillance. Enabling rapid responses, without depending upon other indicators, to contain high-risk strains within hospital and other healthcare environments. It can also i**dentify sites with more transmissible, virulent, and resistant strains** for targeted containment, preadmission screening, and follow-up interventions.

By detecting otherwise unrecognized connections between infections early, from the second isolate of a strain transmitted or acquired in the hospital, or even the first isolate of a strain identified as high risk, IPC responses can be targeted **to prevent onward transmission** and **reduce the size of outbreaks** and the number of healthcare-associated infections that occur. This will result in smaller outbreaks, fewer outbreaks, more rapid detection and remediation of routes and sources of transmission, and **greater protection of both patients and staff** from emergent pathogens. Thus, protecting patient safety, the biosecurity of hospital environments, and reducing direct and indirect costs of care, that conservative models show cover the costs of WGS-led rapid surveillance many times over.

Due to its **open scalability**, the more surveillance data collected within IDEM, from patients of the environment, **the more informed the IPC team becomes**, and with it their ability to rapidly deliver effective interventions and identify sources. Thereby, **protecting both patients and the hospital** from the ever-increasing biosecurity challenges of more resistant and virulent healthcare adapted strains. The new paradigm being maximal immediate patient prevention, coupled with the creation and maintenance of the safest possible healthcare environment, though new IPC capabilities enabled by proactive pathogen sequencing, analyzed and connected through IDEM.

So what?

- Smaller outbreaks.
- Fewer outbreaks.
- Better infection prevention.
- Improved patient safety.
- Improved hospital reputation and IPC practices.
- Saves money while saving lives.

.

Confidential. Copyright and IP owned by Genpax.

The clinical and monetary value of knowing that you don't have an outbreak

IPC resources are limited and must be used with maximum efficiency to protect the patients and hospital environments from highly transmissible, virulent, and resistant strains. Being able to distinguish outbreaks from non-outbreaks and to know which infection and colonizations are transmission-linked or not is fundamental to infection prevention. The pursuit of connections between patients with infections that are not linked, and the institution of control measures for coincidental but not connected similar infections, consumes limited resources and distracts IPC teams from investigating genuinely linked infections, and confuses those investigations. Meanwhile professional medical practice requires precautions and actions to address possible risks to patients, so recognizing unconnected cases is important.

Analysis in IDEM provides pre-emptive information. A team can see whether a *C. difficile* isolate is part of the hospital associated strains, or one that was imported by the patient; and costly cohorting and outbreak response meetings and actions avoided. When increased incidence of infections are noted, the proactively sequenced isolated information can be used to determine whether there is an outbreak, and which isolates are members of which part of coincident outbreaks. This latter issue is greatest relevance to early stage adopters in which multiple long-standing hospital-associated outbreaks often coexist.

Examples in early users of IDEM have already illustrated situations in which non-outbreaks have been rapidly recognized, allowing IPC teams **not to spend resources inappropriately** and to focus upon other impactful activities. Substantial IPC resources can be wasted investigating non-outbreak strains, where IDEM would enable more correctly targeted and effective responses. And, IDEM consistently identfies multiple transmission-linked clusters of only 2 or 3 isolates that would otherwise not have been recognized at all due to not being alert organisms, and therefore not being on the 'radar'.

So what?

- Cost savings from avoiding unnecessary IPC meetings, interventions pending investigations, and precautionary ward closures for non-outbreaks.
- More effective IPC from better and more detailed information on connected strains and transmission chains
- More effective IPC from reducing time investigating non-outbreaks, freeing time to focus upon real outbreaks and transmission events and other prevention-focused activities.
- Defensible positions with patients demonstrably infected by patient-linked, rather than hospital associated strains

Confidential. Copyright and IP owned by Genpax.

Detailed and timely information communicated directly to the IPC team

IDEM is not genomics only for the bioinformaticians, report-focused epidemiologists, and academics. IDEM is about optimized genomics presented directly to those who can act on it to improve patient and public safety in a way that can be easily understood; which means the Infection Prevention and Control teams, and the patient-facing clinical staff. Batch and QC reports are generated for the sequencing laboratory to ensure that the data generated is of consistently high quality and for any issues to be quickly addressed. An individual sample report is also generated for record keeping. But, the most important report is an interactive, continuously updated, easy to access, interpret, and interrogate resource for front-line clinicians.

The IPC reporting system enables the IPC team, from the nurses and infectious disease physicians, to the Director of Infection Prevention and Control (and equivalents) to access and act upon the information as soon as the sequencing data has been analyzed and integrated. (Typically within 2 hours of a sequencing run being completed, or first thing in the morning if a run has completed during the night.) There are no delays or intermediate interpretations between the frontline patient-facing teams and the results of the sequencing analysis. With minimal induction training, any user with a professional understanding of infection prevention and control has direct access to the usable information, and is empowered to act upon it. No long per-sample multiple page pdf reports from multiple samples to work through, no large tables of information that doesn't impact clinical decision making, no periodic data consolidation. Just simple, focused, clearly communicated information in a format tailored for the people who need it that highlights connected strains and potential outbreak clusters. It also provides access to other findings of relevance to IPC and containment of more hazardous or resistant isolates such as resistance and virulence genes.

So what?

- The information gets directly to the people who need it, not delayed or stuck in the lab x Patient-focused responses are enabled more quickly.
- Fewer and smaller outbreaks .
- Improved patient care and safety.

Confidential. Copyright and IP owned by Genpax.

Making immediate connections in Healthcare Surveillance and Public Health

No more waiting for periodic consolidation of data. All strains with evidence of outbreak- and/or transmission-connections are immediately detected and accessible though IDEM. Within a target turnaround time of under 2 hours from receipt of the FASTQ (the file from the DNA sequencer).

IDEM does not use a Sequence Typing step, thereby avoiding associated errors, avoiding missing some outbreak members, and putting strains into groups that are too large for optimal detailed scalable analyses. The IDEM pipeline is also **natural reference-free**, meaning that all strains are comparable at the same high-resolution regardless of reference sequence availability or quality; and that all strains are comparable and connectable with an optimum high-resolution analysis. Thereby **avoiding missing outbreak members** by separating them prior to detailed comparisons or using difference references. Thus, all sequenced isolates are directly compared in a system that is continuously updated to detect transmission connections and determine relationships in real clinical time.

Once weekly, or other periodic consolidation (as is typically practiced in reference laboratories, and other settings) **is no longer necessary to identify connections**. Previously a slow and

computationally intensive costly process that could not reasonably be performed on a rolling basis for every newly sequenced isolate, this is now integrated into the core analysis process. Thus a delay associated with the final stage of analysis when addressing larger numbers of isolates in detail of 1 to 7 days, or more, is avoided in recognition and response times; and the actionable information can be obtained in close to the time it takes to isolate and sequence the DNA.

So what?

- Surveillance in healthcare can operate over longer time periods, necessary to detect some outbreaks linked to the environment, healthcare workers, or patient re-admissions.
- Critical information for public health responses available more quickly.
- No consolidation delays with public health teams able to respond on a rolling basis, rather than following periodic updates.
- Can be used to connect data between labs, enabling wider connected surveillance and closer to isolation laboratory sequencing.
- Faster responses in pathogen eradication programs, such as for *M. tuberculosis.*
- Connection possible between all historic, surveillance, and clinical samples with ability to look over multiple years and origins.

Confidential. Copyright and IP owned by Genpax.



WWW.GENPAX.CO.UK

Copyright and IP owned by Genpax