Biotia

Clinical-Grade Metagenomics in Urinary Tract Infections: Improving Performance of Next-Generation Sequencing Assays Using Internal Controls and Machine Learning

Mara Couto-Rodriguez[1], David C Danko[1], Xavier Jirau Serrano[1], Taylor Paisie[1], John Papciak[1], Eszter Szollosi[1], Christopher E. Mason[1,2-5], Caitlin Otto[1], Niamh B. O'Hara[1,6], Dorottya Nagy-Szakal[1,6]

Biotia Inc., New York, NY, USA
Tri-Institutional Computational Biology & Medicine Program, Weill Cornell Medicine of Cornell University, New York, NY, USA
The HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY, USA
The WorldQuant Initiative for Quantitative Prediction, Weill Cornell Medicine, New York, NY, USA
The Feil Family Brain and Mind Research Institute, Weill Cornell Medicine, New York, NY, USA
SUNY Downstate Health Sciences University, The Department Cell Biology/College of Medicine, New York, NY, USA

SUNY Downstate Health Sciences University
Weill Cornell Medicine
CORNELL TECH HOME OF THE JACOBS TECHNION-CORNELL INSTITUTE

Contact us!
Mara Couto-Rodriguez
couto-rodriguez@biotia.io

Dorottya Nagy-Szakal
MD PhD
nagy-szakal@biotia.io
biotia.io

ABSTRACT

Objectives

Shotgun sequencing-based metagenomics is a useful approach to profiling microbiomes in environmental and patient samples. In a clinical setting, metagenomic techniques have the advantage of identifying organisms, which cannot be readily cultured or confirmed by other techniques. We have developed a clinical-grade, streamlined metagenomics-based pipeline that includes regulatory compliant method considerations, such as an internal control followed by a machine-based learning (ML) process to identify pathogens in urine samples.

Methods

We built an optimized novel end-to-end NGS assay pipeline that harnesses pathogen-specific genome data to detect bacterial species. We processed de-identified clinical urine specimens, collected from patients symptomatic for urinary tract infection (UTI). This workflow includes an IPC, QIACube-MDx extraction, library preparation and Illumina NextSeq 550 sequencing and a novel interpretable ML based analytic approach, Biotia-DX. Clinical culture results and qPCR were used as a baseline for the assay to train the ML model and to establish accuracy relative to the clinical standard of care.

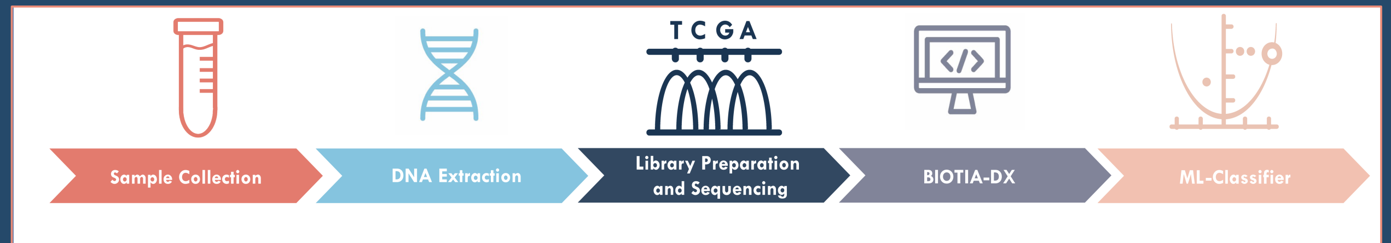
Findings

We clinically validated over 40 key uropathogens and conducted clinical studies of specificity, intra/inter reproducibility, accuracy in urine specimens (n=300), and limit of detection in *E. coli, K. pneumoniae*, *P. mirabilis*, *S. aureus*, *E. faecalis* and *Candida*. Additionally, the implementation of an internal control coupled with our Biotia-DX software provides an accurate (F1 score 95.9%) and highly sensitive clinical grade diagnostic tool.

Conclusion

Urine has historically presented a challenge for diagnostics via culturing, with a high rate of culture-negative results (~30% on average). We improved the clinical utility of an NGS urine assay by leveraging an IPC and ML software. This decreased the rate of false positive species called in a sample relative to other NGS techniques and allows for greater sensitivity and taxonomic specificity. This assay may be especially useful for low colony-count or negative-culture samples to diagnose and guide patient treatment.

STUDY DESIGN



Sample Collection → DNA Extraction → Library Preparation and Sequencing → BIOTIA-DX → ML-Classifier

Sample collection and extraction De-identified left-over urine specimens were collected and processed under the IRB numbered Pro00038083 (Advarra). Midstream clean-catch urine specimens were preserved in UTT. Genomic DNA was isolated from clinical specimens and spike ins using a QIAcube-MDx extraction and were quantified with Qubit-Flex.

Culture Clinical isolates and reference strains used in this validation study were cultured in Blood Agar at 37C.

BIOTIA-ID Urine NGS Assay Metagenomic libraries were prepared using Illumina DNA Prep Library preparation kit. Libraries were quality checked for size and concentration using Tapestation 4200 and Qubit-Flex, respectively. Libraries were pooled in 24-plex reactions and sequenced on an Illumina NextSeq 550 platform using a NextSeq 500/550 Mid Output kit (Illumina, San Diego, CA) set to 150bp single-end reads with i5 and i7 indexes.

BIOTIA-DX The BIOTIA-DX pipeline included removal of low quality reads and human reads. The remaining reads were psuedo-aligned to a large database of microbial genomes in a coarse classification step. Organisms identified from coarse classification were filtered for identification quality and the remaining candidates were sent to a fine classification step. Reads were aligned to curated pangenomes for each organism and summary statistics were generated. These statistics were fed into a machine learning classifier which assigned a confidence score for whether the organism was present or absent.

HIGH QUALITY CLINICAL-GRADE METAGENOMICS

We have designed a clinical grade urine metagenomics assay that incorporates various controls to ensure validity, sensitivity, accuracy and performance for diagnosis of urinary tract infection. Our assay contains the following Quality Checks (QC):

- *Positive Control (PC)* – External control containing two yeasts, three Gram negative, and five Gram positive bacteria for validating reagent integrity and assay performance.
- *Negative Extraction Control (NEC)* – Negative urine matrix used to evaluate extraction reagent performance and cross contamination.
- *Internal Positive Control (IPC)* – Spike in control to assess clinical specimen integrity and performance to minimize false negative results due to inhibition.
- *No Template Control (NTC)* – Negative control for monitoring reagent purity and library preparation cross contamination.

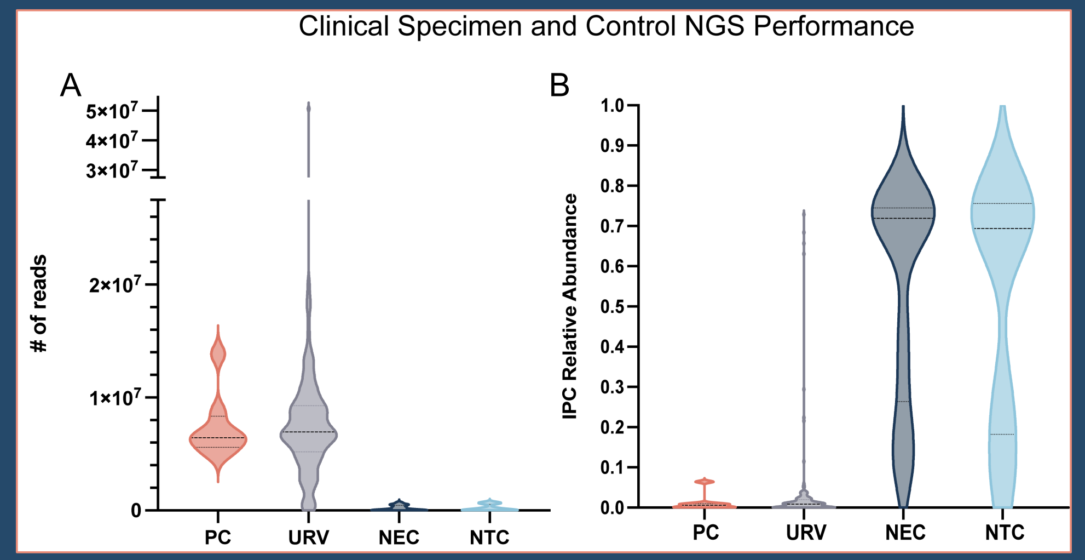


Figure 1. Clinical specimen and control assay performance based on total number of microbial reads obtained (A) and the detection of IPC in clinical specimens and controls tested (B). As expected, the PC and clinical specimens generated equivalent of microbial reads, 7.4 and 7.6 million reads, respectively, while the NEC and NTC generated < 200k microbial reads. The IPC was detected in all specimens and controls, thus validating our process and the integrity of samples tested.

CLINICAL VALIDATION

ACCURACY VERIFICATION IN CLINICAL SPECIMENS

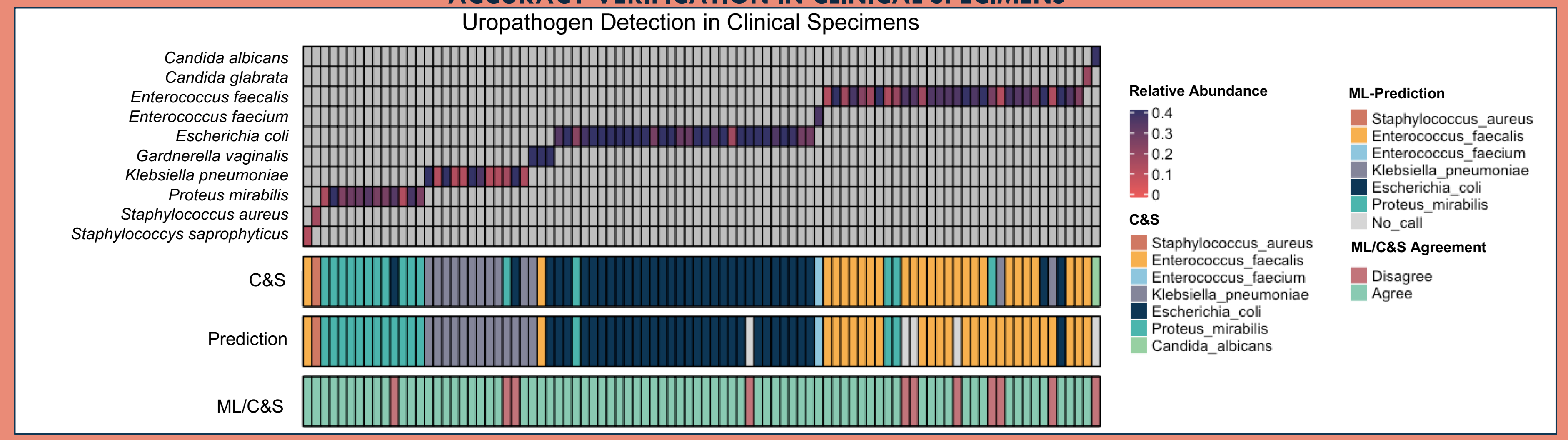


Figure 2. Biotia-ID accurately detects and classifies uropathogens in clinical specimens. As part of the accuracy component of the clinical validation we will test at least 30 UTI specimens for *E. coli, K. pneumoniae, P. mirabilis, S. aureus* and *E. faecalis as* defined by the gold-standard. Heatmap depicts the relative abundance of top organism detected, C&S diagnosis, ML-classifier prediction and ML/C&S agreement for clinical specimens tested (n=92).

VERIFICATION

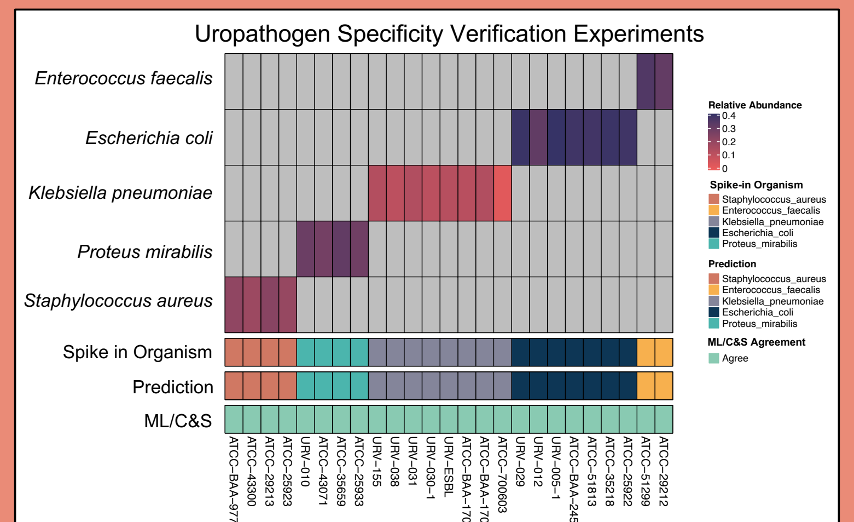
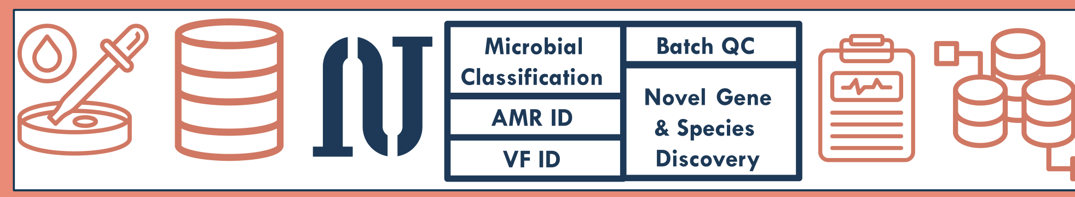


Figure 4. A minimum of five reference and/or clinical isolate strains of *E. coli, K. pneumoniae, P. mirabilis, S. aureus* and *E. faecalis* were spiked into negative urine matrix and validated in the verification run. 100% of the spike-ins samples generated the correct classification with Biotia-DX.

BIOTIA-Dx DATABASE AND PIPELINE



Microbial Classification
AMR ID
VF ID
Batch QC
Novel Gene & Species Discovery

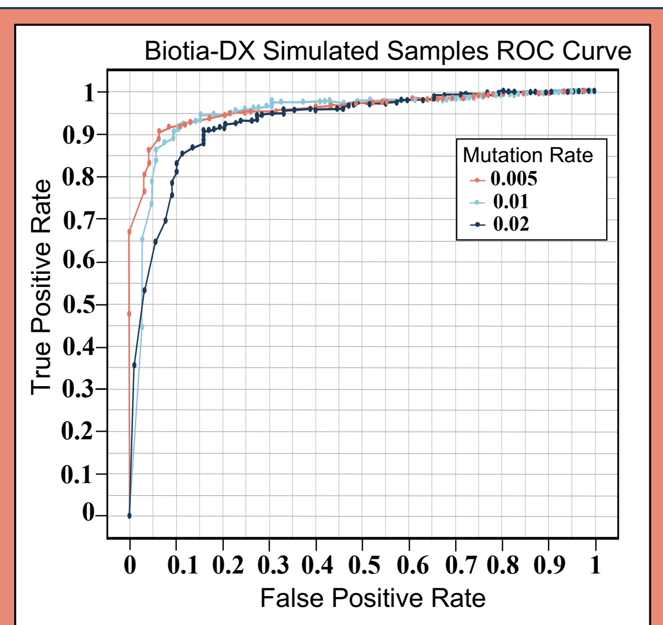


Figure 3. Our proprietary database consists of 27,161 2,279 bacterial species, 238 fungal species, and 11,251 viral strains. Diagnostic accuracy of simulated samples with 0.5%, 1% and 2% mutation rates showed a sensitivity and specificity above 92%.

SPECIFICITY

Gram Negative Enterobacteriaceae	Gram Negative non Enterobacteriaceae	Other Staphylococcus species	Fungi
Escherichia coli	*Acinetobacter baumannii*	*Streptococcus agalactiae* (Group B)	*Candida albicans*
Klebsiella pneumoniae	*Acinetobacter lwoffii*	Mitis Group Streptococci	*Candida auris*
Klebsiella oxytoca	*Pseudomonas aeruginosa* *Pseudomonas fluorescens*	Anginosus Group Streptococci	*Candida tropicalis*
Proteus mirabilis	*Stenotrophomonas maltophilia*	*Aerococcus urinae*	*Candida parapsilopsis*
Proteus vulgaris	**Anaerobic Bacteria**	**Other bacteria**	*Candida glabrata*
Morganella morganii	*Prevotella spp.*	*Chlamydia trachomatis*	*Cryptococcus neoformans*
Citrobacter koseri *Citrobacter freundii*	*Anaerococcus spp.**	*Gardnerella vaginalis*	*Saccharomyces cerevisiae*
Enterobacter aerogenes	**Gram Positive**	*Mycoplasma genitalium*	*Rhodoturula mucilaginosa*
Enterobacter cloacae	*Enterococcus faecalis*	*Mycoplasma hominis*	**Viruses**
Providencia stuartii *Providencia rettgeri**	*Enterococcus faecium*	*Neisseria gonorrhoeae*	HSV-1*
Serratia marescensens	*Staphylococcus aureus*	*Treponema pallidum**	HSV-2*
*Raoultella ornithinolytica**	*Staphylococcus epidermidis*	*Ureaplasma spp.*	HPV*
Shigella flexneri	*Staphylococcus lugdunensis*	**Protozoa**	CMV*
Salmonella Typhimurium	*Staphylococcus saprophyticus*	*Trichomonas vaginalis**	Adenovirus*

Table 1. Reference organisms and clinical isolates tested and validated on the specificity component of the Biotia-ID clinical validation. Organisms genetically related, organisms that can be isolated from urine specimens and/or organisms that produce similar symptomology or illness were tested by spiking microbial cells into negative urine matrix. (*) Indicates organisms yet to be tested with Biotia-ID assay.

BIOTIA-Dx ACCURACY IN CLINICAL SPECIMENS

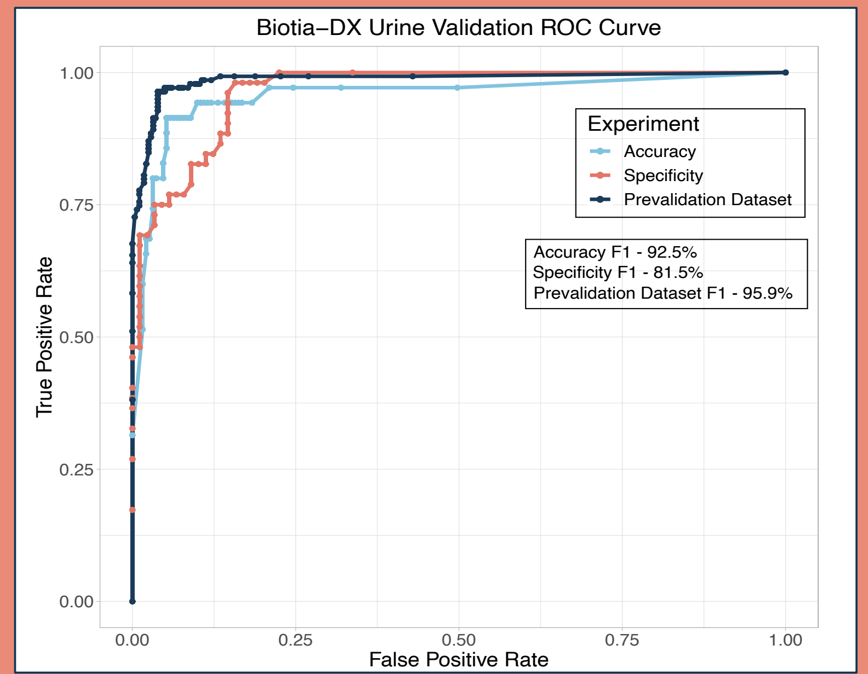
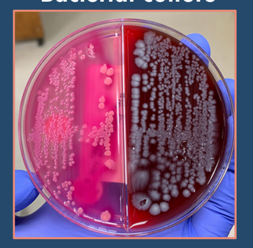


Figure 5. Clinical grade performance of Biotia-Dx in urine clinical specimens. Biotia-Dx is a highly accurate diagnostic tool yielding an F1 score of 95.9% based on the training and testing set of the Pre-validation dataset (n=93), 92.6% for accuracy in urine clinical specimens (n= 92) and 81% for specificity (n=60). Due to limitations of both C&S and NGS, we are performing cross validation experiments with a third comparator assay to address discrepancies found on the pathogen diagnosed.

CLINICAL RELEVANCE OF PRECISION INFECTIOUS DISEASE DIAGNOSTICS

The recent COVID-19 pandemic has catalyzed a radical transformation in the field of infectious disease diagnostics. Common infectious diseases diagnostic tools, such as culture, the gold standard technology, and polymerase chain reaction (PCR), a rapid alternative, have entered public awareness like never before, as the pandemic progresses. However, despite their ubiquity, these limited tools have slowed progress in diagnosing and researching infectious diseases, contributing to skyrocketing drug resistance, as well as 18M diagnostic errors annually, >$100B in losses, extensive patient suffering, and 80K deaths.

Bacterial culture

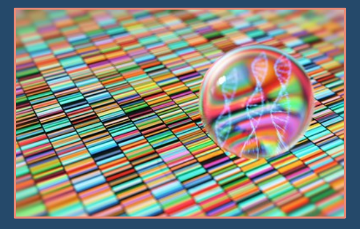


- Older technology
- 33% of UTI samples are culture negative
- Slow diagnostic turnaround
- Limited information

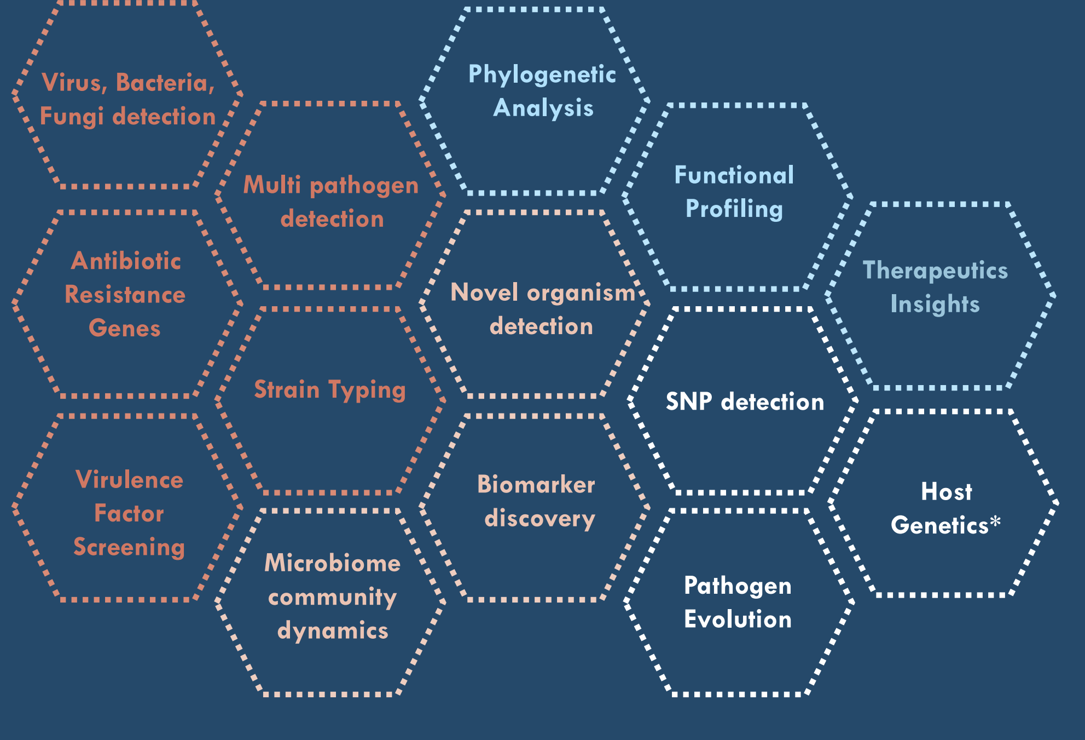
PCR

- Presence/absence of target genes and markers only
- Limited target number
- Intolerant to mutations limiting pathogen discovery

Sequencing

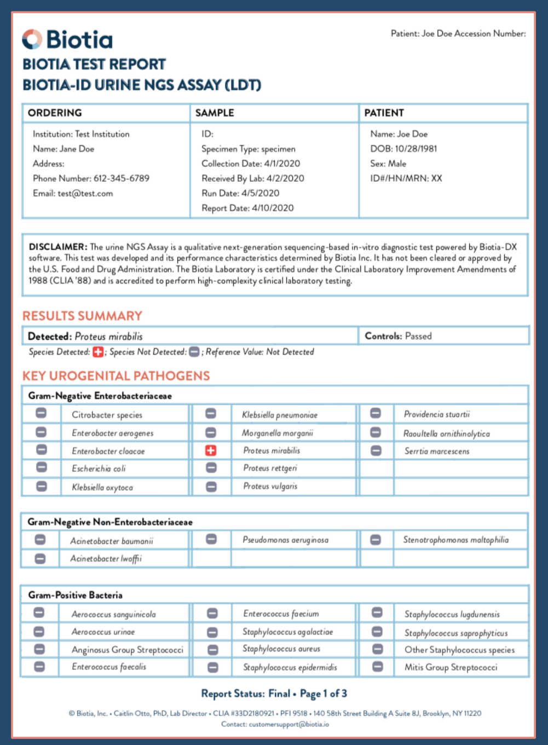


- High diagnostic sensitivity and specificity
- Presence/Absence of target genes and markers
- Robust insight into genetic information (AMR, virulence factors, strain typing)
- Scalable



Virus, Bacteria, Fungi detection
Multi pathogen detection
Phylogenetic Analysis
Functional Profiling
Antibiotic Resistance Genes
Novel organism detection
Therapeutics Insights
Strain Typing
SNP detection
Virulence Factor Screening
Biomarker discovery
Host Genetics*
Microbiome community dynamics
Pathogen Evolution

Next-generation sequencing (NGS) offers the opportunity to identify important species, resistance markers, and pathogen evolution, at a scale unmatched by existing technologies, and can alter clinical care to provide insight into pathogens beyond presence or absence.



Biotia
Patient: Joe Doe Accession Number:

BIOTIA TEST REPORT
BIOTIA-ID URINE NGS ASSAY (LDT)

ORDERING	SAMPLE	PATIENT
Institution: Test Institution Name: Jane Doe Address: Phone Number: 612-345-6789 Email: test@test.com	ID: Specimen Type: specimen Collection Date: 4/1/2020 Received By Lab: 4/2/2020 Run Date: 4/5/2020 Report Date: 4/10/2020	Name: Joe Doe DOB: 10/28/1981 Sex: Male ID#/HN/MRN: XX

DISCLAIMER: The urine NGS Assay is a qualitative next-generation sequencing-based in-vitro diagnostic test powered by Biotia-DX software. This test was developed and its performance characteristics determined by Biotia Inc. It has not been cleared or approved by the U.S. Food and Drug Administration. The Biotia Laboratory is certified under the Clinical Laboratory Improvement Amendments of 1988 (CLIA '88) and is accredited to perform high-complexity clinical laboratory testing.

RESULTS SUMMARY

Detected: *Proteus mirabilis* Controls: Passed

Species Detected: 🟥; Species Not Detected: ⊖; Reference Value: Not Detected

KEY UROGENITAL PATHOGENS

Gram-Negative Enterobacteriaceae

⊖	Citrobacter species	⊖	*Klebsiella pneumoniae*	⊖	*Providencia stuartii*
⊖	*Enterobacter aerogenes*	⊖	*Morganella morganii*	⊖	*Raoultella ornithinolytica*
⊖	*Enterobacter cloacae*	🟥	*Proteus mirabilis*	⊖	*Serrtia marcescens*
⊖	*Escherichia coli*	⊖	*Proteus rettgeri*		
⊖	*Klebsiella oxytoca*	⊖	*Proteus vulgaris*		

Gram-Negative Non-Enterobacteriaceae

⊖	*Acinetobacter baumanii*	⊖	*Pseudomonas aeruginosa*	⊖	*Stenotrophomonas maltophilia*
⊖	*Acinetobacter lwoffii*				

Gram-Positive Bacteria

⊖	*Aerococcus sanguinicola*	⊖	*Enterococcus faecium*	⊖	*Staphylococcus lugdunensis*
⊖	*Aerococcus urinae*	⊖	*Staphylococcus agalactiae*	⊖	*Staphylococcus saprophyticus*
⊖	Anginosus Group Streptococci	⊖	*Staphylococcus aureus*	⊖	Other Staphylococcus species
⊖	*Enterococcus faecalis*	⊖	*Staphylococcus epidermidis*	⊖	Mitis Group Streptococci

Report Status: Final • Page 1 of 3

© Biotia, Inc. • Caitlin Otto, PhD, Lab Director • CLIA #33D2180921 • PFI 9518 • 140 58th Street Building A Suite 8J, Brooklyn, NY 11220
Contact: customersupport@biotia.io

Infectious disease is the second leading cause of death for cancer patients with about 1/3 of cancer patients dying from an infection, not from cancer. Urinary tract infections (UTIs) are the most common outpatient infection and a persistent problem for cancer patients, however the gold standard technology for UTI diagnosis, culturing, misses about 1/3 of diagnoses and requires many tests to identify the wide range of pathogens involved and characterize drug resistance. Biotia developed and optimized an NGS-based urine assay providing a valuable tool for diagnosis and guided treatment. Such, precision infectious disease diagnoses and management is an urgent need for cancer patients and other high risk patient population.

LIMITATIONS AND FUTURE WORK

We will expand our clinical validation for antimicrobial resistance (including gyrA, parC) and virulence factor detection. Future studies are needed to collect clinical metadata, standard of care and disease management specifics in relation to culture and NGS data in high-risk patient population (cancer, transplant and other immunocompromising conditions) to enable monitoring disease outcome, hospital admission/stay, and sepsis development.

ACKNOWLEDGEMENTS

We thank Zaineb Bello for clinical specimen collection and laboratory operation support; Tamara Goncharuk for sequencing support; Yehudah Gruenstein and LabQ Team for clinical specimen collection; the Abesse Team, Adam Nagyhazy-Horvath, Pierre Davidoff, Cory Mason for IT support, and Talisa Vega-Kline for supporting conference logistics.

The BIOTIA-DX software pipeline work used CromwellOnAzure, the Microsoft Genomics supported implementation of the Broad Institute's Cromwell workflow engine on Azure. The development and validation work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant (ACI-1548562). Specifically, it used the Bridges system, which is supported by NSF award (ACI-1445606), at the Pittsburgh Supercomputing Center (PSC).

DISCLAIMER: MCR, DCD, XJS, TP, JP, ES, CEM, CO, NBO and DNS are employees at Biotia Inc.