Highly Accurate and Reliable Next-Generation Sequencing-Based Urine Assay Overperforms Conventional Diagnostics

Mara Couto-Rodriguez[1], David C. Danko[1], Heather L. Wells[1], Xavier Jirau Serrano[1], John Papciak[1], Gabor Fidler[1], P. Ford Combs[1], Adam Nagyhazy-Horvath[1], Eszter Szollosi[1], Patrik Blik[1], Taylor Paisie[1], Christopher Mason[1-5], Caitlin Otto[1], Niamh O'Hara[1,6], Dorottya Nagy-Szakal[1,6]

[1]Biotia Inc., New York, NY, USA
[2]Tri-Institutional Computational Biology & Medicine Program, Weill Cornell Medicine of Cornell University, New York, NY, USA
[3]The HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY, USA
[4]The WorldQuant Initiative for Quantitative Prediction, Weill Cornell Medicine, New York, NY, USA
[5]The Feil Family Brain and Mind Research Institute, Weill Cornell Medicine, New York, NY, USA
[6]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

Antimicrobial resistance is a leading cause of death globally, and by the year 2050, increasing rates of resistance are projected to contribute to 10 million deaths each year. On average, one third of urinary tract infections (UTIs) yield a negative culture result, leading to inappropriate antimicrobial therapy and contributing to increasing rates of antimicrobial resistance. Precision infectious disease diagnostics using next-generation sequencing (NGS)-based metagenomic tools present an opportunity to mitigate this trend by helping to target the use of appropriate antibiotic therapeutics and promote overall antimicrobial stewardship. Here, we have built and validated a clinical-grade sequencing-based pipeline that accurately identifies key pathogens in clinical urine samples using a machine learning classification approach.

De-identified clinical UTI specimens and spike-in samples were processed with a novel end-to-end NGS assay including QIACube-MDx extraction, metagenomic library preparation and Illumina NextSeq 550 sequencing. A bioinformatic pipeline that uses a multiple decision tree-based machine learning approach was developed and trained using data from culture- and qPCR-validated UTI samples. Internal controls and other quality control measures were incorporated into the process to provide rigorous and standardized clinical-grade results.

Clinical validation of the Biotia-ID assay, performed using a unique set of 300 clinical UTI specimens, reached 99% specificity and 97% sensitivity in predicting microbial organisms within each sample. Limit of detection (LoD) was assessed on five of the most prevalent uropathogens demonstrating an overall LoD of <10,000 CFU/mL, which is lower than culture. Specifically, we achieved LoDs of 1,000 CFU/mL for *S. aureus*; 7,500 CFU/mL for *E. coli*, *K. pneumoniae* and *P. mirabilis*; and 10,000 CFU/mL for *E. faecalis*. Apparent false positive and false negative results, where the ML classification differed from the original lab culture result, were additionally tested by qPCR or Sanger sequencing, and 85% of these NGS results were ultimately supported and considered true positives. Overall, these results demonstrate that Biotia-ID is a highly accurate clinical grade diagnostic tool with notable advantages over current culture-based diagnostics.

STUDY DESIGN

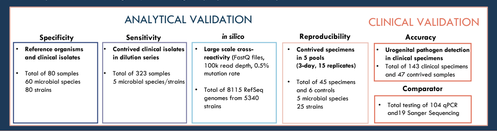
Sample collection and extraction De-identified left-over urine specimens were collected and processed under the IRB numbered Pro00038083 (Advarra). Midstream clean-catch urine specimens were preserved in Urine Transport Tube (UTT). Genomic DNA was isolated from clinical specimens and spike ins using 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 pseudo-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 (Hyun, J.C., et al. 2022) 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.

Comparator testing Genomic DNA from urine specimens were used to perform confirmatory qPCR or Sanger Sequencing tests. A total of 8 qPCR and 7 Sanger (taxa or group specific) assays were used.



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.

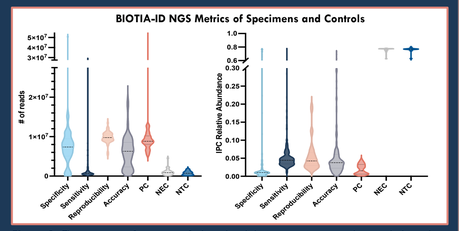


Figure 1. The assay performance of clinical validation specimens and controls based on total number of microbial reads obtained (A) and the detection of IPC in clinical specimens and controls tested (B). As expected, the PC, the clinical and contrived specimens generated similar microbial read depth, while the NEC and NTC had little microbial reads. The IPC was detected in all specimens and controls at expected range (1-5% abundance), thus validating our process and the integrity of samples tested. IPC reads represented the majority of the reads detected in NEC and NTC controls.

SPECIFICITY STUDIES

Gram Negative Enterobacteriaceae	Gram Negative non Enterobacteriaceae	Gram Positive	Fungi
Escherichia coli	*Acinetobacter baumanii*	*Streptococcus agalactiae*	*Candida albicans*
Klebsiella pneumoniae	*Acinetobacter lwoffii*	Mitis Group Streptococci	*Candida auris*
Klebsiella oxytoca	*Pseudomonas aeruginosa*	*Aerococcus urinae*	*Candida glabrata*
Klebsiella varicicola	*Pseudomonas fluorescens*	Other bacteria	*Candida tropicalis*
Proteus mirabilis	*Stenotrophomonas maltophilia*	*Chlamydia trachomatis**	*Candida parapsilosis*
Proteus vulgaris	Anaerobic bacteria	*Gardnerella vaginalis*	*Candida krusei*
Morganella morganii	*Prevotella* spp.	*Mycoplasma genitalium**	*Candida dubliniensis*
Citrobacter koseri	*Anaerococcus* spp	*Mycoplasma hominis**	*Candida lusitaniae*
Citrobacter freundii	Genital Pathogens	*Neisseria gonorrhoeae**	*Candida kefyr*
Enterobacter cloacae	*Enterococcus faecalis*	*Ureaplasma* spp*	*Meyerozyma guilliermondii*
Providencia stuartii	*Enterococcus faecium*	Viruses	*Cryptococcus neoformans**
Providencia rettgeri			
Serratia marcescens	*Staphylococcus aureus*	HSV-1*	*Rhodotorula mucilaginosa**
Rosultella ornithinolytica	*Staphylococcus epidermidis*	HSV-2*	*Saccharomyces cerevisiae**
*Shigella flexneri**	*Staphylococcus lugdunensis*	HPV*	Parasites
*Salmonella Typhimurium**	*Staphylococcus saprophyticus*	CMV*	*Trichomonas vaginalis**

Table 1. Reference organisms and clinical isolates tested and validated on the specificity component of the BIOTIA-ID clinical validation. Whole organisms of the key urogenital pathogens detected by the assay, genetically related organisms, and other organisms that can be present in urine specimens were tested by spiking microbial cells at 50,000 CFU/mL into negative urine matrix. A total of 65 microbial species (bacteria, fungi, viruses and parasites) and 114 strains were evaluated. This study yielded an overall sensitivity and specificity of 100% and 99.93%, respectively. (*) indicates organisms tested with BIOTIA-ID assay but not reported as a key urogenital pathogen. We also conducted an extensive *in silico* analysis to further evaluate the performance and specificity of the assay. A total of 8,266 of simulated samples were processed and tested for key organisms. An overall 99.95% sensitivity and specificity was obtained for simulated samples.

ANALYTICAL SENSITIVITY

Analyte	Detection Limit (CFU/mL)	CFU/mL									
		10	$1x10^3$	$1x10^4$	$5x10^3$	$7.5\ x10^3$	$1x10^4$	$1.25x10^4$	$2.5x10^4$	$5x10^4$	$1x10^5$	$1x10^6$
Staphylococcus aureus	1,000	0/6	0/6	6/6	-	-	6/6	-	6/6	-	6/6	6/6
Escherichia coli	5,000	0/3	0/3	3/9	5/6	6/6	6/6	6/6	6/6	6/6	6/6	6/6
Klebsiella pneumoniae	7,500	0/3	0/3	0/6	4/6	6/6	6/6	6/6	6/6	6/6	6/6	6/6
Proteus mirabilis	7,500	0/6	0/6	3/12	3/6	6/6	6/6	6/6	6/6	6/6	6/6	6/6
Enterococcus faecalis	10,000	0/3	0/3	3/6	0/6	4/6	6/6	6/6	6/6	6/6	6/6	6/6

Table 2. Analytical sensitivity was assessed in five of the most prevalent urogenital pathogens, by spiking reference whole organisms into negative urine matrix. The LoD was reproducibly verified and determined based on a 100% positivity rate resulting in an overall LoD of <10,000 CFU/mL. Specifically, the LoDs determined with this assay were 1,000 CFU/mL for *S. aureus*; 7,500 CFU/mL for *E. coli*, *K. pneumoniae* and *P. mirabilis*; and 10,000 CFU/mL for *E. faecalis*.

ACCURACY IN CLINICAL SPECIMENS

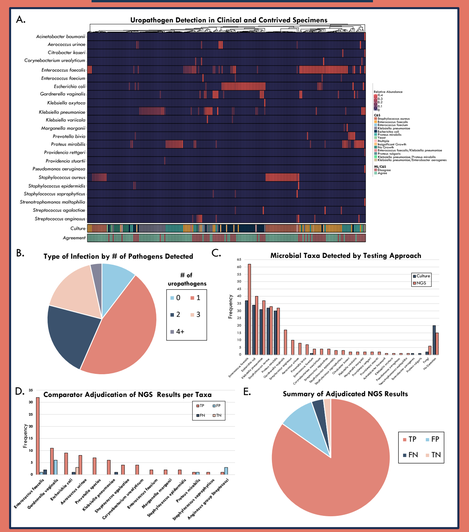


Figure 2. BIOTIA-ID accurately detects and classifies urogenital pathogens in clinical (n=143) and contrived (n=47) specimens in the accuracy studies. (A) Heatmap depicts the relative abundance of taxa detected by NGS and is annotated with C&S diagnosis, and agreement between BIOTIA-DX ML/C&S. (B) Infection type distribution based on # of pathogens detected with NGS assay. (C) Distribution of microbial taxa detected with C&S and NGS. (D) Apparent false positive and false negative results from 15 analytes , where BIOTIA-ID differed from the original C&S result (36% of samples), were tested by qPCR or Sanger sequencing (n=123), and (E) 87% of these NGS results were found to be correct. This accuracy study showed how NGS outperforms conventional diagnostics with an overall 96.77% sensitivity and 99.57% specificity.

INTER/INTRA REPRODUCIBILITY

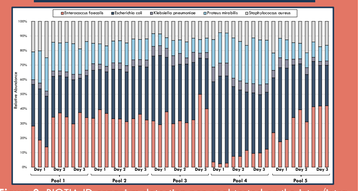


Figure 3. BIOTIA-ID organism detection per pool tested on the inter/intra reproducibility studies. Five different pools containing one clinical strain of *E. coli*, *E. faecalis*, *K. pneumoniae*, *P. mirabilis* and *S. aureus* were tested. Each pool had a different strain of the five species tested in this study (altogether 25 different strains tested). Each pool was tested daily for three consecutive days in three technical replicates for a total of 15 replicates per pool, 45 replicates per analyte and 6 negative samples. All analytes were reproducibly and correctly detected in each sample tested, obtaining a sensitivity and specificity of 100%.

BIOTIA-DX ML PERFORMANCE

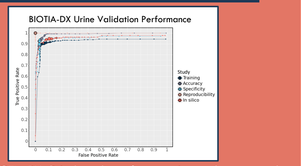


Figure 4. BIOTIA-DX clinical grade performance in samples processed on the training, *in silico*, accuracy, specificity and reproducibility studies.

BIOTIA-ID VALIDATION PERFORMANCE

Method	TP	TN	FN	FP	Total	Sensitivity	Specificity
Clinical	210	8078	7	35	8330	96.77%	99.57%
Contrived	573	6897	0	29	7501	100.00%	99.58%
in silico	8115	272778	4	147	281044	99.95%	99.95%
Total	8900	287753	11	211	296875	99.88%	99.93%

Table 3. Performance characteristics summary of the BIOTIA-ID Urine NGS Assay validation study which yielded an overall performance of 99.88% sensitivity and 99.93% specificity.

BIOTIA-ID PERFORMANCE CHARACTERISTICS

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.

- ✓ **COMPREHENSIVE** including anaerobes and polymicrobial infections
- ✓ **RAPID** turnaround time of 36-48 hours
- ✓ **ACCURATE AND SENSITIVE** compared to standard of care (culture)
- ✓ **OPTIMIZED** therapy based on microbial and drug resistance profile

7,500+	Organisms in our microbial database
400+	Clinical specimens tested
1000+	Reactions completed for extensive validation
97%, 99%	Sensitivity and specificity

Precision infectious disease diagnostics and UTI patient management is an urgent need for immunocompromised patients at high risk of developing sepsis.

Annually, 11 million people in the United States and 404 million people worldwide are diagnosed with a UTI (Yang et al., 2022). Immunocompromised patients are at higher risk to develop complicated and/or recurrent UTI as well as to progress to urosepsis, which has significantly higher morbidity and mortality as well as a much higher cost of care. In hospitalized patients, UTIs are associated with an attributed mortality rate of 2.3% and an estimated annual cost of $340 to $450 million in the United States alone (Yang et al., 2022). For high-risk patients, it is critical to rapidly identify the urogenital pathogens causing UTIs to provide appropriate treatment and reduce the use of broad-spectrum antibiotics.

Urine culture is the standard of care (SOC) for identification of urogenital pathogens causing UTIs. However, this method has several limitations, including: (1) an inability to identify hard-to-grow microbial organisms, such as anaerobic bacteria; (2) a long diagnostic turnaround time; (3) limitations in identifying rare pathogens or organisms that are not routinely cultured; and (4) false negative results for patients being treated with antibiotics. Long diagnostic TATs or inconclusive results can lead to the treatment of suspected infections with broad-spectrum antibiotics that are often inappropriate and may contribute to increased rates of drug resistance.

BIOTIA-ID is a highly accurate diagnostic assay that can accurately and rapidly identify pathogens in clinical urine specimens from patients with recurrent and complicated UTI. Selected use cases:

- ✓ **Recurrent UTIs**
- ✓ **Complicated UTIs**
- ✓ **Drug Resistant UTIs**
- ✓ **Drug Resistant STIs**
- ✓ **Culture negative symptomatic patients**
- ✓ **Immunocompromised patients**
- ✓ **Women's health**
- ✓ **Interstitial cystitis, Prostatitis**

LIMITATIONS AND FUTURE WORK

We will expand our clinical validation for antimicrobial resistance (including markers for extended spectrum beta-lactamase resistance) 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, Pierre Davidoff, Cory Mason for IT support, and Talisa Vega-Kline for supporting conference logistics.

The BIOTIA-DX software pipeline work used Cromwell OnAzure, 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, HLW, XJS, JP, PFC, GF, ANH, ES, PB, TP, CEM, CO, NBO and DNS are employees at Biotia Inc.

REF: Yang, X., *et al.* (2022). Disease burden and long-term trends of urinary tract infections: A worldwide report. Frontiers in Public Health.
Hyun, J.C., et al. (2022). Comparative pangenomics: analysis of 12 microbial pathogen pangenomes reveals conserved global structures of genetic and functional diversity. BMC genomics.