SCIENTIFIC REPORTS

OPEN

Microbial metagenome of urinary tract infection

Ahmed Moustafa[1], Weizhong Li[1,2], Harinder Singh[3], Kelvin J. Moncera[2], Manolito G. Torralba[2], Yanbao Yu[3], Oriol Manuel[4], William Biggs[1], J. Craig Venter[1,2,3], Karen E. Nelson[1,2,3], Rembert Pieper[1,3] & Amalio Telenti[2]

Received: 17 August 2017
Accepted: 26 February 2018
Published online: 12 March 2018

Urine culture and microscopy techniques are used to profile the bacterial species present in urinary tract infections. To gain insight into the urinary flora, we analyzed clinical laboratory features and the microbial metagenome of 121 clean-catch urine samples. 16S rDNA gene signatures were successfully obtained for 116 participants, while metagenome sequencing data was successfully generated for samples from 49 participants. Although 16S rDNA sequencing was more sensitive, metagenome sequencing allowed for a more comprehensive and unbiased representation of the microbial flora, including eukarya and viral pathogens, and of bacterial virulence factors. Urine samples positive by metagenome sequencing contained a plethora of bacterial (median 41 genera/sample), eukarya (median 2 species/sample) and viral sequences (median 3 viruses/sample). Genomic analyses suggested cases of infection with potential pathogens that are often missed during routine urine culture due to species specific growth requirements. While conventional microbiological methods are inadequate to identify a large diversity of microbial species that are present in urine, genomic approaches appear to more comprehensively and quantitatively describe the urinary microbiome.

Urinary tract infections (UTIs) occur in a high proportion of the population and are a significant health economic burden[1]. The criteria for diagnosis includes multiple clinical parameters and laboratory tests[2], and the clinical suspicion of a UTI frequently triggers the prescription of broad spectrum antibiotics, with or without confirmation of the infecting organisms. The most common organism in uncomplicated UTIs is *Escherichia coli* followed by a number of gram-positive cocci and other Enterobacteriaceae[3]. Other organisms, including difficult-to-culture prokaryotes, eukaryotes such as *Candida albicans* and viruses, are involved in UTIs or other manifestations of genitourinary tract infection such as urethritis and sexually transmitted diseases. Because the care of UTI is streamlined, it is only after treatment failure that molecular tests and additional non-molecular investigations are launched.

Conventional microbiological methods are inadequate to fully determine the diversity of bacteria that are present in urine[4]. Next generation sequencing techniques create the possibility of investigating the microbial metagenome associated with infection and inflammation of the urinary tract. Metaproteomic methods have enabled a deeper characterization of the inflammatory response towards uropathogens in cases of UTI and asymptomatic bacteriuria[5,6]. Sixteen studies have characterized the urinary microbiome by 16S rRNA sequencing in adults. A cumulative number of 603 subjects were investigated across the various studies (UTI, n = 50; other urinary manifestations, n = 219; sexually transmitted diseases, n = 20; renal transplant samples n = 60; urine in bacterial vaginosis, n = 109; healthy, n = 145)[4,7–21]. However, a complete view of the microbiome, including eukarya and viruses, as well as an unbiased characterization of abundance and the identification of virulence factors as presented in this study can only be achieved by comprehensive microbiome analyses using metagenome sequencing. A study of the urine metagenome (35 samples) was published in 2014 by Hasman and colleagues[22]. Sequencing directly from the urine using Ion Torrent technology enabled bacterial identification in polymicrobial samples and the identification of putative pathogenic strains in some culture-negative samples.

The aims of this study were to discover new microbial and viral components in clinical urine specimens using metagenomics sequencing, and to examine the question of whether the microbial compositions of urine specimens justifies the description of a urinary microbial metagenome. The metagenomics component allowed for the exploration of organisms and their abundances from all microbial kingdoms and allowed us to investigate

[1]Human Longevity Inc., San Diego, CA, 92121, USA. [2]J. Craig Venter Institute, La Jolla, CA, 92037, USA. [3]J. Craig Venter Institute, Rockville, MD, 20850, USA. [4]Centre Hospitalier Universitaire Vaudois, 1011, Lausanne, Switzerland. Ahmed Moustafa and Weizhong Li contributed equally to this work. Correspondence and requests for materials should be addressed to R.P. (email: rpieper@jcvi.org) or A.T. (email: atelenti@jcvi.org)

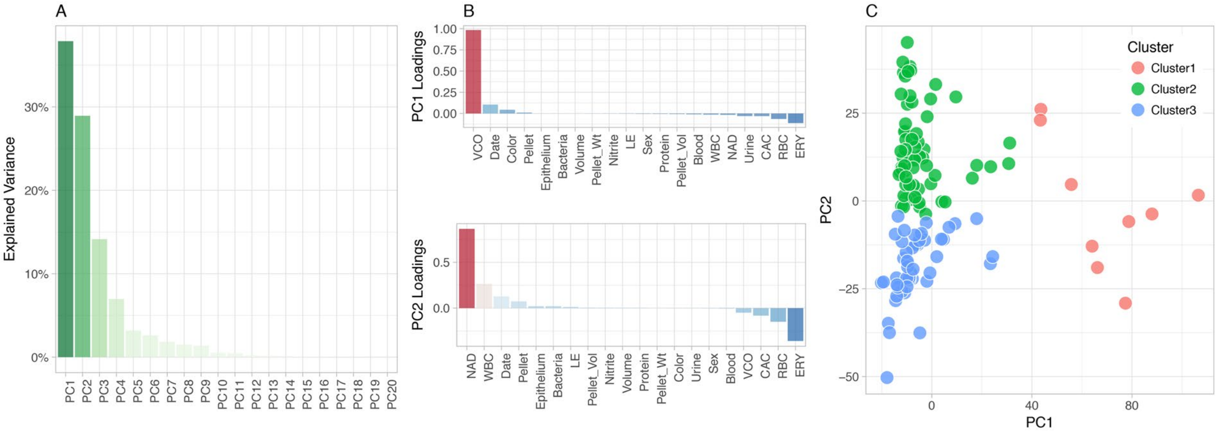

Figure 1. Definition of clinical and laboratory groups. The study used an unbiased approach to the classification of specimens using 20 parameters from the laboratory analysis of urine. (**A**) Explained variance from PCA; the first two PCs were retained for downstream clustering analyses. (**B**) Contributing factors (loadings) to the first two PCs. Note that the directionality of the loadings reflect enrichment independently of sign and direction. (**C**) Clustering of samples is based on the method of partitioning around medoids (pam).

	Cluster 1		Cluster 2		Cluster 3		Total	
Microbiome	Count	Percent	Count	Percent	Count	Percent	Count	Percent
MG + 16S	5	56%	35	56%	7	14%	47	39%
MG	0	0%	1	2%	1	2%	2	2%
16S	4	44%	27	43%	38	78%	69	57%
None	0	0%	0	0%	3	6%	3	2%
Total	9		63		49		121	

Table 1. Microbiome Sequencing performance rate. Clusters are defined on the basis of clinical and laboratory metadata. Cluster 1 is interpreted as reflecting contamination, Cluster 2 is most consistent with urinary infection, Cluster 2 is of unclear nature. MG: metagenome sequencing, 16S: 16S rDNA sequencing.

the distribution of known virulence genes across the various study groups. Overall, the study reveals patterns of peri-urethral colonization and vaginal contamination of urine samples and of different profiles of what can be considered active infection. The study also contributes to the identification of difficult-to-culture and potential novel pathogens and addresses the presence of various human viruses and eukarya that are important in genitourinary medicine.

Results

Clinical laboratory data representation.

To support an unbiased assessment of the clinical nature of the specimens, we approached the urine sample laboratory and microbiology data using dimensionality reduction, and clustering analysis. A listing of clinical data is provided in Tables S1 and S2. The PCA representation of the clinical laboratory data is presented in Fig. 1. The PCA analysis showed that the first two components (PC1, PC2) explained 65% of the variance in the clinical laboratory dataset. PC1 was driven by the vaginal contamination score (VCO), PC2 was contributed primarily by neutrophil activation and degranulation score (NAD), and secondarily by the erythrocyte and vascular injury score (ERY) and the presence of red blood cells (RBC) and leukocytes (WBC) (Fig. 1B). The partitioning around medoids clustering resulted in three Clusters, with 9 individuals in Cluster #1, 63 individuals in Cluster #2, and 49 individuals in Cluster #3 (Fig. 1C). Clinical metadata were compared between these three clusters. NAD, CAC (Complement activity and coagulation score), WBC, VCO and ERY showed most significant difference ($p < 0.001$, Kruskal-Wallis rank sum test, Table S2).

From these data, we established a preliminary definition of Cluster #1 as likely representing urine from non-infected individuals, while Clusters #2 and #3 are consistent with separate manifestations of infectious and inflammatory processes of the urinary tract. The performance of 16S rDNA and metagenome sequencing across clinical laboratory clusters is presented in Table 1.

16S rDNA sequencing.

16S rDNA sequencing was successful for 116 (96%) samples (Table 1) with an average of 39,288 paired end high quality reads (2 × 300 bp) per sample (Table S3). The median (range) number of genera identified per individual was 38 (6–220). The median (range) number of genera varied across clinical Clusters: 51 (16–106) for Cluster 1, 32 (7–172) for Cluster 2, and 60 (6–220) for Cluster 3. Analysis of the normalized abundance of the classified bacterial genera across the clinical groups (Fig. 2) confirmed that proteobacteria were the predominant phylum in cluster 2 - the Cluster that represents infection, with prominent identification of *Citrobacter sp*., *Enterobacter sp*., and *Escherichia sp*. Clusters 1 and 3 were more diverse in composition (Fig. 2).

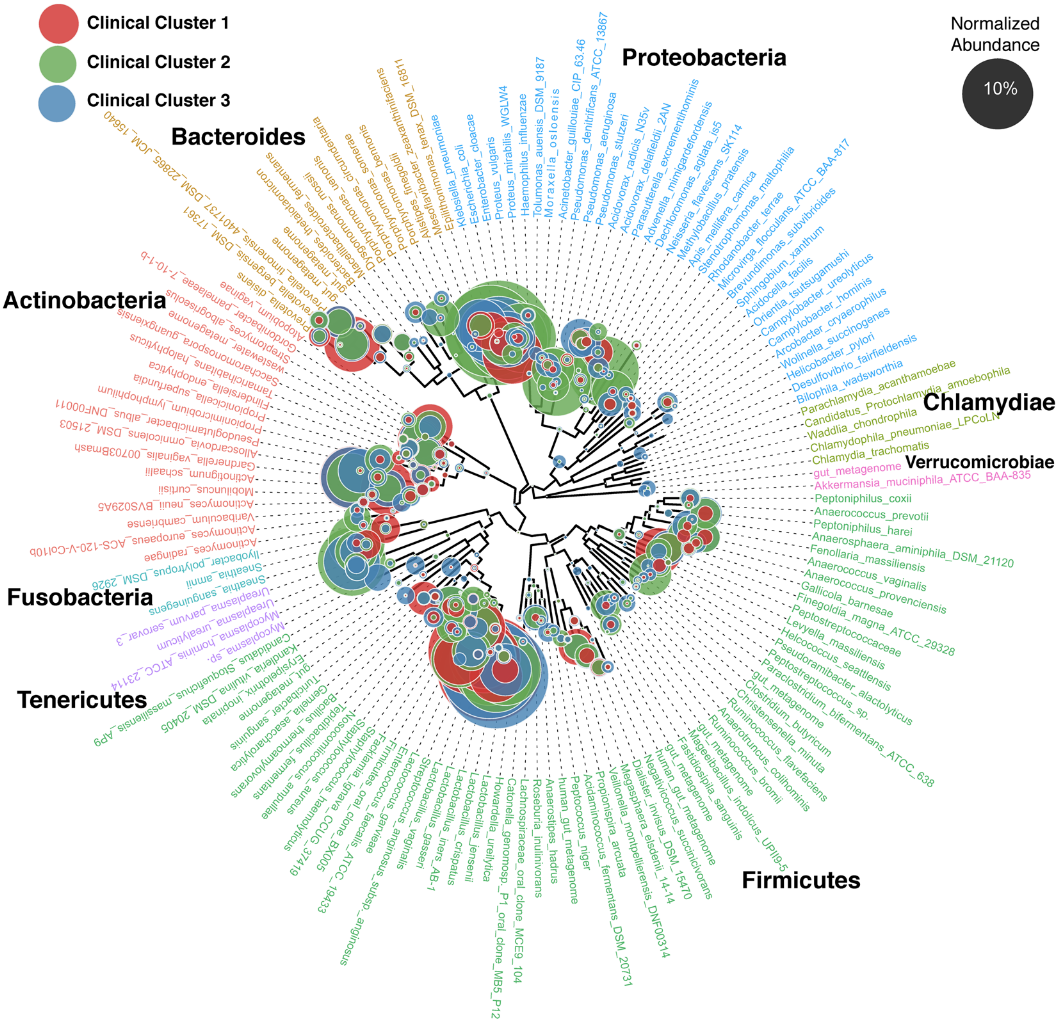

Figure 2. Normalized abundance of bacterial genera across the clinical groups using 16S rDNA. 116 samples were successfully analysed by 16S rDNA sequencing and grouped according to the clinical laboratory clusters. Proteobacteria were the predominant phylum in Cluster 2 - the cluster that represents infection, with prominent identification of *Citrobacter*, *Enterobacter*, *Escherichia*. Clusters 1 and 3 were more diverse in composition (Fig. S1).

Cluster 1 had greater abundance of *Actinotignum*, *Aerococcus*, *Atopobium*, *Facklamia*, *Gardnerella*, *Lactobacillus*, *Megasphaera*, *Oligella*, *Prevotella*, and *Streptococcus* species. Cluster 3 had greater abundance of *Acidovorax*, *Alloscardovia*, *Epilithonimonas*, *Lachnospira*, *Peptostreptococcus*, *Pseudomonas*, *Rhodanobacter*, *Riemerella*, *Sphingobium* and *Ureaplasma* (Fig. S1).

Metagenome sequencing. Shotgun metagenome analysis was successfully performed on 49 samples highlighting that, although samples with limited microbial content may amplify via 16S rDNA, insufficient reads or failed sequencing will occur if starting DNA material is limiting. However, metagenomics data will reflect quantitatively more accurate analyses compared to 16S rDNA data. Metagenomic sequencing generated 26.6 million paired end high quality reads (2 × 125 pb) per sample on average (Table S4). After removing reads that from human host, which range from 1.3% to 99.9%, on average, 4.26 million paired end high quality reads per sample were used for metagenomic analysis (Table S4). In the samples that were successfully investigated by microbial whole genome sequencing (WGS), the average composition of the reads per kingdom was 94.6% Bacteria, 0.05% Eukarya, 0.0027% Viruses, and 0.0001% Archaea (Fig. 3). The archaeal component was discarded from subsequent analyses. We also observed a significant proportion, 4.9%, of unmapped non-human sequence reads. The largest microbial content was observed in Cluster 1, the lowest in Cluster 3.

Bacteria. The median (range) number of bacterial species – genera - identified per individual was 41 (27–49). The median (range) number of species across clinical Clusters was 44 (29–48) for Cluster 1, 41 (28–49) for Cluster 2, and 38 (28–47) for Cluster 3. Figure 4 depicts the read counts for genera across clinical groups, as well as the highest genome coverage of strains within each genus. Genomes of 27 strains in 9 genera were recovered with >90% genome coverage. In 33 genera, there were 411 strains whose genomes were recovered with >50%.

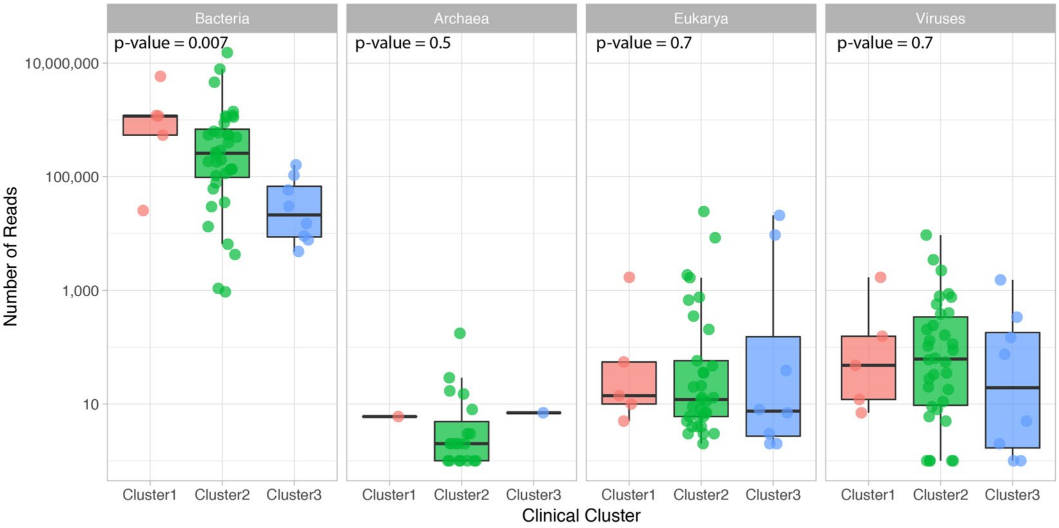

Figure 3. Metagenome sequencing mapped reads per sample. 49 samples were successfully sequenced and grouped according to the clinical laboratory clusters. Each point represents a sample. The thick line in the boxplot represents the median number of reads for the cluster.

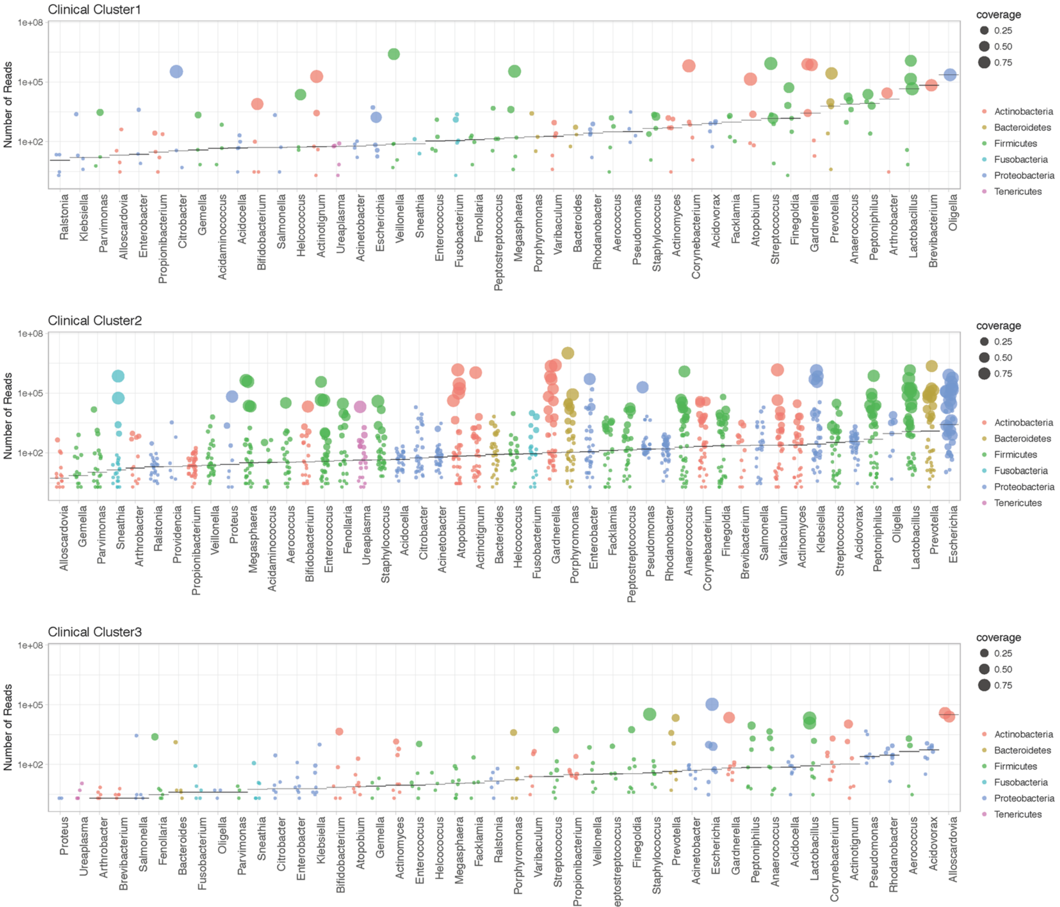

Figure 4. Ranking of bacterial genera by counts from metagenome sequencing across clinical laboratory clusters. Shown are bacteria observed with at least 1% of total reads in a sample. Analyses reflect results from 49 samples that were successfully sequenced and grouped according to the clinical laboratory clusters. Each point represents a genus in a sample. The horizontal represents the median number of reads for the genus.

Clinical and laboratory Cluster	**Majority species (≥10E5 reads)**	**Number of samples**	**Culture results**
Cluster 1	*Actinotignum* + *Citrobacter*	1	Negative
	Atopobium + *Gardnerella* + *Lactobacillus* + *Megasphaera* + *Streptococcus* + *Veillonella*	1	Negative
	Corynebacterium + *Lactobacillus* + *Oligella*	1	Negative
	Gardnerella + *Prevotella*	1	Negative
Cluster 2	*Escherichia*	5	***Escherichia*** (in 4 of 5 samples)
	Klebsiella	3	Negative
	Lactobacillus	2	Negative
	Actinotignum	1	Mixed
	Anaerococcus + *Peptoniphilus* + *Porphyromonas* + *Varibaculum*	1	Mixed
	Atopobium + *Enterobacter* + *Escherichia*	1	***Escherichia***
	Atopobium + *Escherichia* + *Gardnerella* + *Lactobacillus* + *Prevotella*	1	***Escherichia***
	Atopobium + *Gardnerella* + *Lactobacillus* + *Megasphaera*	1	***Staphylococcus***
	Atopobium + *Gardnerella* + *Megasphaera* + *Prevotella* + *Sneathia*	1	Negative
	Enterobacter	1	***Enterobacter***
	Enterococcus	1	Mixed
	Escherichia + *Lactobacillus*	1	Negative
	Gardnerella	1	Negative
	Gardnerella + *Prevotella*	1	Negative
	Klebsiella + *Prevotella*	1	***Klebsiella***
	Pseudomonas + *Lactobacillus*	1	Mixed
Cluster 3	*Escherichia*	1	Negative

Table 2. Relationship between metagenome sequencing and routine culture.

This analysis indicates that proteobacteria are the predominant phylum in Cluster 2 - the cluster that represents infection, including classic uropathogens such as *Escherichia*, *Klebsiella*, *Pseudomonas*, *Enterobacter*, *Citrobacter*, as well as species with unclear or unknown role in infection, such as *Acidovorax*, *Rhodanobacter*, and *Oligella* (Fig. S2). Cluster 1 had greater abundance of *Actinomyces*, *Anaerococcus*, *Atopobium*, *Facklamia*, *Finegoldia*, *Gardnerella*, *Lactobacillus*, *Megasphaera*, *Peptoniphilus*, *Staphylococcus*, and *Streptococcus* (Fig. S2). Given the depletion in total number of reads in Cluster 3, we could not identify any uniquely enriched genus.

We specifically chose to represent the metagenome data as read counts, a surrogate of absolute abundance, because the process does not involve amplification and thus, relative abundance may misrepresent actual content of microbiota. However, we compared the relative abundance as estimated by 16S rDNA with the absolute read number from metagenome sequencing to assess the degree of correlation. The correlation was high ($R^2 = 0.88$), however, there were some discrepancies where relevant organisms appeared better identified by WGS than by 16S rDNA sequencing (eg. *Gardnerella*, Fig. S3). The presence of *Gardnerella vaginalis* in urine has been also recognized through metaproteomic approaches[5].

We also explored the nature of samples in Cluster 2 that were negative by WGS – despite the expectation that samples in this group would be indicative of infection. For this, we inspected differences in 16S rDNA read counts for 27 samples in Cluster 2 that were negative in WGS compared with 35 samples in Cluster 2 that were positive in WGS. We did not identify differences in median 16S rDNA bacterial read counts across these two sets, nor a significant difference in pattern of bacterial abundance. Therefore, it remains unclear what the true nature of those Cluster 2 samples is: inflammatory reactions, traumatic (for example, passage of a kidney stone), low grade infection, or technical limits to WGS that limit sensitivity.

It was also important to assess the correspondence of WGS and routine culture used in the clinics. A total of 23 samples in Cluster 2 presented dominant flora (*post hoc* defined as $>10^5$ reads). For those, we observed eight samples with consistent WGS and culture results, 1 with a discrepant growth, and 4 reported as mixed flora in culture (Table 2). There was no reported culture growth for four samples in Cluster 1, and only one sample in Cluster 3 despite the observation of dominant flora in sequencing. Two samples, one in Cluster 1 and one in Cluster 2 contained high number of reads of *Actinotignum* sp. This facultative anaerobic gram-positive rod (in particular, *A. schaali*) has been claimed to be part of the urinary microbiota of healthy individuals while also responsible for UTIs, particularly in elderly men and young children[23]. Use of matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry (MALDI-TOF MS) supports the better identification of this organism[24].

The metagenome approach permits the identification of virulence genes in the bacterial pool. Searching for virulence factors against VFDB[25], we observed enrichment of specific factors, in particular in Cluster 2 (Fig. 5). While the identification of virulence genes does not necessarily inform on potential for expression and pathogenicity, it serves to illustrate the differences in output of WGS versus 16S rDNA sequencing.

Eukarya. The median (range) number of species identified per individual was 2 (1–8). The median (range) number of species was 2 (1–8) for cluster 1, 2 (1–6) for cluster 2, and 2 (1–3) for cluster 3. Nine species were

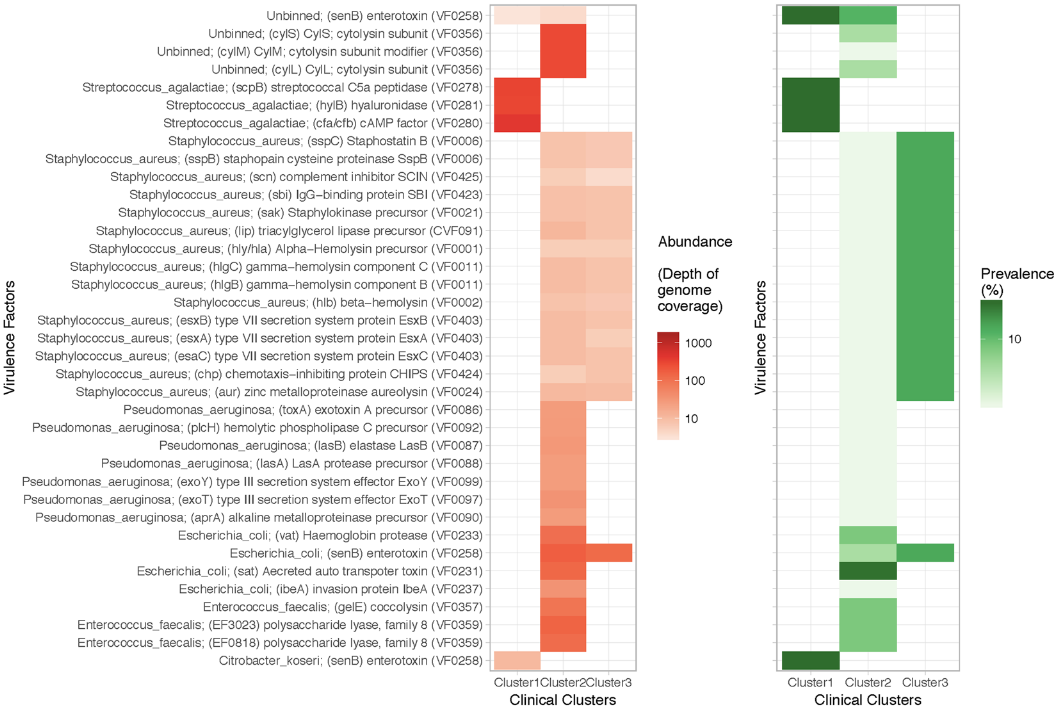

Figure 5. Virulence factors across clinical laboratory clusters. Metagenome sequencing data was used to search for open reading frames (ORFs) compared against the database VFDB[25] to identify virulence factor genes with over 95% sequence identity. Listed are the factors identified in the dataset, grouped by taxonomic binning, with the VFDB accession number in parenthesis. The left panel shows enrichment in the abundance of ORFs across clusters. Here, the abundance is the depth of coverage of the genome where the ORFs were predicted. The right panel shows prevalence of samples that contain organisms carrying the corresponding virulence factor in each cluster.

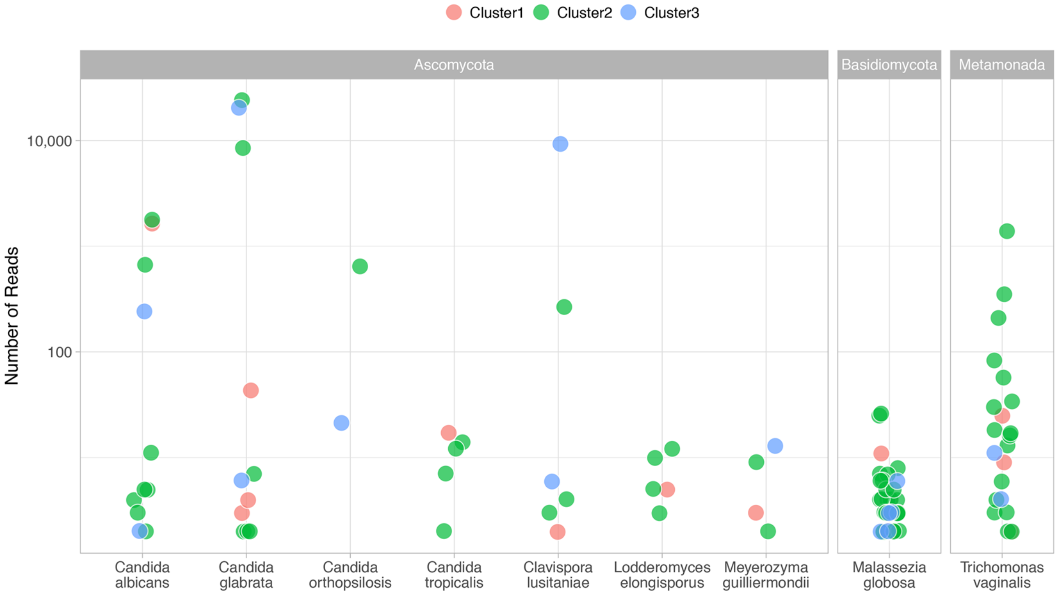

Figure 6. Eukarya read counts across clinical laboratory clusters. Shown are eukarya observed with at least 10 sequence reads in a sample. Analyses reflect results from 49 samples that were successfully sequenced and grouped according to the clinical laboratory clusters. Each point represents a species in a sample.

identified (minimum 10 reads per sample): eight fungal species (*Candida albicans*, *C. glabrata*, *C. orthopsilosis* and *C. tropicalis*, *Clavispora lusitaniae*, *Lodderomyces elongisporus*, *Meyerozyma guilliermondii and Malassezia globosa*) and a metamonada (*Trichomonas vaginalis*). Figure 6 depicts the read counts for genera across clinical

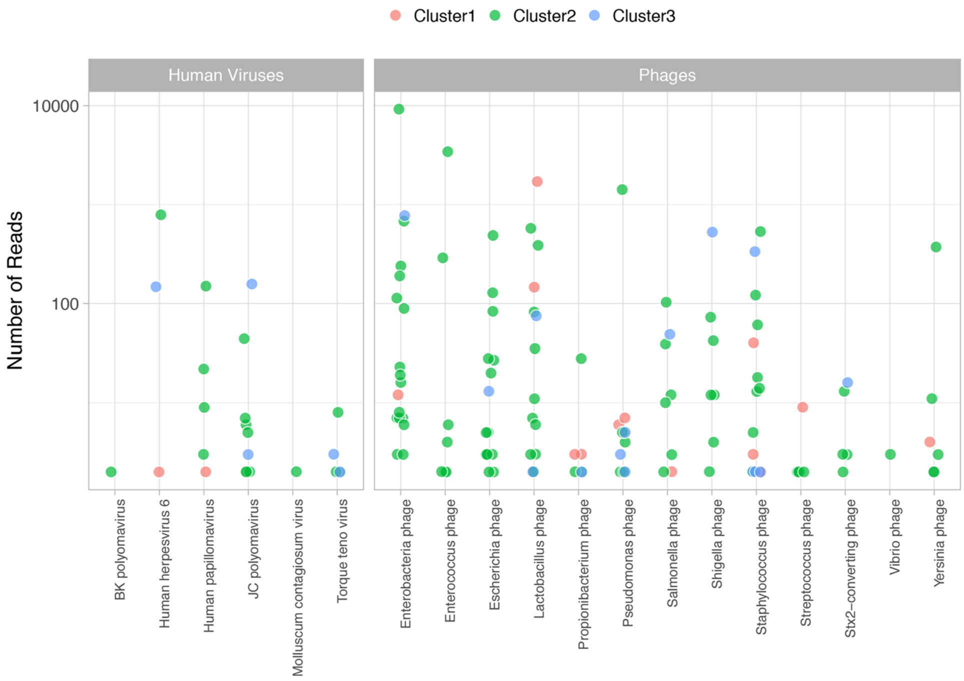

Figure 7. Viral read counts across clinical laboratory clusters. Shown are viruses observed with at least one sequence read in a sample. Analyses reflect results from 49 samples that were successfully sequenced and grouped according to the clinical laboratory clusters. Each point represents a virus in a sample.

groups. Relatively elevated counts were observed for *C. glabrata* and *Clavispora lusitaniae* in four individuals from Clusters 2 and 3. Candida species both colonize and cause invasive disease in the urinary tract[26]. The identification of the lipophilic fungi *Malassezia* is not unexpected as these fungi predominate in most skin sites in healthy adults[27].

Trichomonas vaginalis colonizes the genitourinary tract of men and women. Young women with urinary symptoms in the absence of documented UTI were found more likely to have *Trichomonas vaginalis* compared to those with a documented UTI[28]. Molecular amplification detects *Trichomonas vaginalis* in penile-meatal swabs and urine specimens of men[29]. The identification in the present study of sequence reads in 18 of 38 females (47%) and 4 of 11 (36%) males suggests the common presence of this organism in the genitourinary region – at least in populations in a clinical setting.

Viruses.

The median (range) number of viruses identified per individual was 3 (1–9). The median (range) number of viruses was 3 (2–6) for cluster 1, 3 (1–9) for cluster 2, and 2 (1–7) for cluster 3 (Fig. 7). We identified 13 phages that were generally dominant and associated with the cognate bacteria in the sample.

We identified 6 human viruses consistent with a genitourinary source (human papillomavirus and molluscum contagiosum virus), urinary excretion (BK and JC polyomavirus) or viruses possibly leaked into urine from bleeding and inflammation (Herpesvirus 6 and Anellovirus). As previously reported, excretion of polyomavirus is more commonly observed for JC than BK virus among nonimmunosuppressed individuals[30,31], and excretion increases with immunosuppression[32]. Herpesvirus 6 is rarely excreted in urine[33]. During acute infection, some children with exanthema subitum may present sterile pyuria[34]. However, a likely source of significant number of viral reads in urine may be the sloughing off of cells in individuals with integrated copies of the HHV6 in the host genome – occurring in 0.5 to 1% of the population[35,36].

Gender.

We observed differences in the microbiome content across sex (Fig. S4A). The greatest differences (not significant after multiple testing correction) were greater numbers of sequence reads for *Lactobacillus* and *Prevotella* in women, and of *Enterococcus*, and *Pseudomonas* in men (Fig. S4B).

Discussion

This study provides a detailed view of the microbial metagenomes of urine specimens. The study departs from a classical analysis in that it maximizes a data-driven approach that extracts laboratory metadata features and matches them to metagenomic profiles. It provides an unbiased identification of the flora associated with samples colonized or contaminated with vaginal commensal organisms or local flora, and with samples associated with infection. Colonizing bacteria may be present at the urinary meatus, the distal urethra or along the entire urothelium. Where such bacteria reside cannot be determined from voided urine samples. Our study extends the identity of possible pathogens to include unconventional microorganisms and thus represents a new view of the nature of infection in the genitourinary region and an approach to the question of a normal urinary tract flora. Indeed, the concept of urine not being sterile has been raised in the past[8,17].

We identified 16 reports[4,7–21] in the literature that used 16S rDNA sequencing in the analysis of urinary microbiome. The target of these studies was very diverse, encompassing the study of urine samples from healthy individuals, urinary tract infection, clients of sexual transmitted clinics, and various disorders of the upper urinary tract. These studies identified microbial communities, and characterized the impact of various perturbations, including antibiotherapy. One report[22] used whole genome sequencing of urinary specimens in the setting of infection. The complementarity and properties of either approach are not well studied. To evaluate those questions, we used both 16S rDNA and metagenome sequencing techniques. In a review, Jovel *et al*.[37] concluded that, in the study of other human microbial niche, WGS offers increased resolution, enabling a more specific taxonomic and functional classification of sequences as well as the discovery of new bacterial genes and genomes, and offering a greater potential for identification of strains. A recent paper underscores the existence of sub-populations (subspecies) in the majority of abundant gut prokaryotes – leading to a better functional and ecological understanding of the human gut microbiome[38]. This dimension is not captured by 16S rDNA sequencing.

In our hands, 16S rDNA sequencing provided a greater sensitivity, as it identified bacterial species across the majority of samples and clinical groups. In contrast, less than half of the clean catch urine samples generated sequencing libraries for WGS. The basis for the lower sensitivity rests on the fact that WGS uses limited technical amplification of the nucleic acid content in the sample, thus more closely reflecting the proportionate biomass contributed by microbes in the urinary metagenomes. WGS also provides a unique view of non-prokaryotic content of urine through the identification of eukarya - mainly Candida species, and of human viruses and phages. These differences notwithstanding, both sequencing techniques identify a substantial diversity of microbial species. WGS also provides a representation of virulence factors in the bacterial pool of the individual. Not unexpectedly, the analysis identifies differences in the microbial metagenome across sexes.

The study convincingly identifies high numbers of sequence reads of conventional uropathogens, but also proposes novel bacterial species associated with features of infection. It also challenges the cutoffs used to define infection: generally, 10^5 colony forming units in culture. The quantitative nature of the WGS approach identifies traditional uropathogens at lower quantities in samples with features of infection. It identifies non-cultured/difficult to grow bacteria long discussed as a possible pathogenic organism, for example, *Alloscardovia*[39–42] and *Actinotignum* sp.[23]. *A. schaalii* may be an underestimated cause of UTIs because of its fastidious growth on usual media and difficulties associated with its identification using phenotypic methods[23]. WGS also provides a broader screen compared to the conventional urinary culture. For example, we identified sequence reads of *Ureaplasma* – a potential pathogen that requires dedicated culture systems or molecular testing. It is expected that the approach will identify *Mycoplasma*, *Chlamydia* and other agents associated with sexually transmitted diseases.

Use of WGS also captures viral DNA sequence reads. The identification of viruses in the genitourinary tract is important because of the potential for transmission from local disease (e.g., HSV2, papillomaviruses), or because of the interest in monitoring of shedding (e.g., CMV, BK virus). WGS also identified shedding of common blood viruses such as Anellovirus (Torque teno virus)[43]. There is however limited information on the role of viruses as a cause of UTI[44]. Consistent with the work of Santiago-Rodriguez the al.[45], we observed the abundant presence of phages that match the presence of the cognate bacteria in urine. Metagenomic analyses could thus expand the understanding of viruses as flora of the genitourinary tract.

The present study uses specimens collected for clinical diagnostic purposes, but de-identified and considered medical waste. This limits in-depth understanding of the clinical setting beyond what can be established from the urine laboratory metadata. However, it allows the assessment of the metagenome content on the basis of objective laboratory data, while excluding subjective clinical interpretation. We propose that future studies on the urinary microbiome should use baseline unbiased microbial metagenome analysis to prospectively understand the nature of infection and of treatment response. We speculate that "Cluster 3" may to some extent include urine samples from individuals that received treatment with antibiotics. This cluster has the least amount of sequence reads in WGS, and the presence at low titers of classical uropathogens such as *Pseudomonas aeruginosa* or *Escherichia coli*. Another consideration for the interpretation of Cluster 3 is that we did not use negative extraction controls as they rarely generate the appropriate libraries for sequencing, and thus, cannot formally exclude reagent or environmental contamination. A systematic, prospective use of metagenomic tools may also shed light on the role of unknown and unconventional microorganisms in the urinary tract. Additional aspects that could be approached by urinary metagenomics are the characteristics of the urinary "normal flora" – as it is increasingly observed that the urinary tract may not be sterile. These studies could be performed via suprapubic collection of urine. Overall, the present study underscores that the current understanding of the etiology of UTIs can be improved through the combined used of unbiased clinical laboratory data and microbial metagenome analysis.

Methods

Study participants and urinalysis.

A total of 121 human urine specimens were collected by the Pathology and Clinical Microbiology Laboratory of the Shady Grove Adventist Hospital (SGAH) in Rockville, Maryland. Details on the set of specimens and the urinalysis methods performed were described previously[5]. A total of 92 samples were collected from women, and 29 from men. The study was exempted from review by Institutional Review Boards of the J. Craig Venter Institute (JCVI) and SGAH because the specimens were collected for diagnostic purposes and considered medical waste prior to use for the study. All experiments were performed in accordance with relevant guidelines and regulations. The urine samples left over after clinical urinalysis were de-identified prior to transfer to JCVI. Clinical laboratory records included gender and the results of urinalysis tests, such as presence of bacterial cells, red blood cells, leukocytes, epithelial cells and casts (assessed by phase contrast microscopy), nitrite concentration (associated with bacterial nitrate reduction) and leukocyte esterase activities (derived from the activity of white blood cell proteases and esterases released into urine).

Sample processing. Urine specimens (5 to 30 ml) were stored at 4 °C for up to 6 h after collection and centrifuged at 3,000 × g for 15 minutes at 10 °C. Given that microbiome results are prominently expressed in the log scale throughout the work, we consider the maximal 0.5 log effect of the different volumes of urine as a small component of the variance. Urinary pellets were washed twice with a 10-fold volume of PBS and frozen at −80 °C until used for proteomic analyses as reported[5] or for microbiome and metagenomics analyses. On the day of DNA extraction, 300 µl of TES buffer (20 mM Tris-Cl, pH 8.0, 2 mM EDTA, and 1.2% Triton X-100) was added to a 5 to 25 µl urinary pellet sample. The sample was vortexed, incubated at 75 °C for 10 min and cooled to room temperature. The suspension was supplemented with 60 µl chicken egg lysozyme (200 µg/ml), and 5 µl Linker RNase A, gently mixed and incubated for 60 min at 37 °C. After addition of 100 µl 10% SDS and 42 µl Proteinase K (20 mg/ml), bacterial lysis was allowed to proceed overnight at 55 °C. The DNA was extracted by adding an equal volume of phenol: chloroform: isoamylalcohol (25:24:1; pH 6.6), followed by vortexing and centrifuging at 13,000 RPM for 20 min. The aqueous phase was removed and transferred to a sterile microcentrifuge tube. The residual sample was then re-extracted by repeating the previous step. The aqueous phase was re-extracted with an equal volume of chloroform: isoamylalcohol (24:1) and centrifuged at 13,000 RPM for 15 min. The aqueous phase was transferred to a sterile microcentrifuge tube and 3 M sodium acetate (pH 5.2) was added at a 10% volume. The DNA was precipitated by adding an equal volume of ice cold isopropanol followed by incubation at −80 °C for 30 minutes. Samples were then centrifuged at 16,100 × g for 10 min and the supernatant was removed. The pellet was washed with 80% ethanol and centrifuged again. After air drying, the DNA pellet was resuspended in Tris EDTA buffer in preparation for sequencing.

Special phenotypic tests. Previous work using the same samples focused on the integrated evaluation of urinalysis and proteomic data to diagnose UTI and inflammatory conditions in the genitourinary tract[5]. Specifically, proteomics tools were used to calculate three scores: NAD (neutrophil activation and degranulation), ERY (erythrocyte score) and VCO (vaginal contamination score); see below. The experimental shotgun proteomic methods were based on tryptic peptide analysis via nano-liquid chromatography tandem mass spectrometry (LC-MS/MS) with the high resolution high accuracy Q-Exactive mass spectrometer (V1.4, Thermo Electron) followed by computational searches of a database comprised of the combined sequences of the human proteome and 21 proteomes of microbial species known to colonize the human genitourinary tract[5]. Semi-quantitative proteomic data were obtained counting the peptide-spectral matches for a given proteins using the Proteome Discoverer^{TM} software analysis tool (Thermo Electron) at a 1% peptide and protein false discovery rates. Quantitative analyses for the performance of phenotypic tests utilized the MaxQuant software tool[46]. The iBAQ protein values were computed for 35 proteins highly expressed in activated neutrophils, 32 proteins highly expressed in erythrocytes and five proteins highly expressed in squamous epithelial cells (cornifelin, cornulin, galactin-7, serpin B3, and mucin 5B) compared to the abundance of the urine-specific protein uromodulin. Summed iBAQ values then permitted the calculation of scores, which were termed the NAD score for neutrophil contents, the ERY score for red blood cell contents and the VCO score for squamous epithelial contents[5]. Specifically, the vaginal contamination score is based on the quantification of the VCO proteins defined in a previous publication[5]. VCO markers are strongly expressed in vagina/cervix and/or are strongly associated with stratified squamous epithelium and are not expressed in the urinary tract.

Sequencing and analysis of 16S rDNA genes. DNA extracted from urine samples was amplified using primers that targeted the V1-V3 regions of the 16S rDNA gene[47]. These primers included the i5 and i7 adaptor sequences for Illumina MiSeq and unique 8 bp indices incorporated into both primers such that each sample received its own unique barcode pair. The method of incorporating the adaptors and index sequences into primers at the PCR stage provided minimal loss of sequence data when compared to previous methods that would ligate adaptors to every amplicon after amplification. This method also allowed the generation of all sequence reads in the same 5′-3′ orientation. Using approximately 100 ng of extracted DNA, amplicons were generated with Platinum Taq polymerase (Life Technologies, CA) using the following cycling conditions: 95 °C for 5 min for an initial denaturing step followed by 95 °C for 30 sec, 55 °C for 30 sec, 72 °C for 30 sec for a total of 35 cycles followed by a final extension step of 72 °C for 7 min then stored at 4 °C. Once the PCR for each sample was completed, the amplicons were purified using the QIAquick PCR purification kit (Qiagen Valencia, CA), quantified fluorometrically using SYBR Gold Nucleic Acid Gel Stain (ThermoFisher Scientific), normalized, and pooled in preparation for bridge amplification followed by Illumina MiSeq sequencing using V3 chemistry dual index 2 × 300 bp format (Roche, Branford, CT) following the manufacturer's protocol.

Phylogenetic classification. 16S rDNA amplicons were quality control using Infernal[48]. Only sequences identified as bacterial 16S using Infernal were considered for downstream steps. Bacterial 16S sequences were searched against SILVA (release 128)[49] using blastn[50] to initially determine the species found in the samples to include the corresponding SILVA reference sequences in a reference phylogenetic tree. Identified reference sequences were aligned using MAFFT[51] with the G-INS-i settings for global homology. A maximum likelihood reference tree was inferred under the general time-reversible model with gamma-distributed rate heterogeneity ($GTR + \Gamma$) using FastTree[52]. The 16S reads were mapped onto the reference tree using pplacer[53] with the default settings. The number of sequences assigned to each node on the reference tree was normalized to the total number of sequences from the corresponding samples. The normalized abundances of the mapped reads were visualized using ggtree[54].

Metagenome sequencing. Nextera XT libraries were prepared manually following the manufacturer's protocol (Illumina). Briefly, samples were normalized to 0.2 ng/µl DNA material per library using a Quant-iT picogreen assay system (Life Technologies) on an AF2200 plate reader (Eppendorf), then fragmented and tagged via tagmentation. Amplification was performed by Veriti 96 well PCR (Applied Biosystems) followed by AMPure XP bead cleanup (Beckman Coulter). Fragment size was measured using Labchip GX Touch high-sensitivity. For cluster generation and next generation sequencing, samples were normalized to 1 nM, pooled, and diluted to 8 pM. The paired-end cluster kit V4 was used and cluster generation was performed on an Illumina cBot, with pooled samples in all 8 lanes. Sequencing was performed on an Illumina HiSeq. 2500 using SBS kit V4 chemistry. Median cluster densities (K mm2) were 908.5 for Nextera XT.

Taxonomic assignments, microbial abundance, and virulence markers. Sequences were processed using the Human Longevity Inc. microbiome annotation pipeline as described in[55]. Briefly, after trimming adapter sequence, removing low quality bases, excluding reads shorter than 90 nucleotides, removing duplicated reads, reads were aligned to the human reference genome hg38 using BWA[56]. Reads that were mapped to hg38 were excluded from downstream analyses. Non-human reads were mapped to Human Longevity Inc. reference genomes database, which is composed of almost 19,023 NCBI reference genomes of bacteria, archaea, eukarya, and viruses. Successfully mapped reads were taxonomically classified using the Expectation Maximization algorithm[57]. The relative abundance of a reference genome was estimated as the genome coverage divided by the sum of all genome coverages. Non-human reads were assembled using IDBA-UD[58] and ORFs are predicted from assembled scaffolds with Metagene[59]. An assembled scaffold was binned to a species if more than 50% of the reads that mapped to the scaffold were also mapped to the species using BWA. ORFs were compared against VFDB[25] to identify virulence factor genes. An ORF is considered as a virulence gene if (a) it is over 95% identity to a gene in VFDB, and (b) the alignment must cover over 90% of the length of the ORF and over 50% of the gene in VFDB, and (c) the scaffold from which the ORF was predicted must be taxonomically binned to a species that contains the gene in VFDB, and (d) all the assembled scaffolds from that species must cover at least 33% of the genome size.

Dimensionality reduction of clinical laboratory data and clustering. The clinical laboratory metadata matrix was imputed for missing entries using MissForest[60]. Then principal component analysis (PCA) was conducted for a matrix of twenty clinical and sampling meta parameters (collection date, sex, urine appearance, urine volume, urine color, urine blood, hemoglobin presence with urine dipstick, red blood cells (RBCs), vascular injury score (ERY, see definition above), protein presence with urine dipstick, nitrate concentration, number of leukocytes, neutrophil activation and degranulation score (NAD, see definition above), complement system activity and coagulation, leukocytes microscopy, squamous epithelial cells [Epithelium], vagina contamination score [VCO, see definition above], urinary pellet appearance and color, urinary pellet volume and weight) from 121 individuals. The first two components from the PCA analysis, which explained 35% and 30% of the variance, were used to cluster the individuals using the partitioning around medoids (pam) method[61]. The optimal number of clusters was determined to be three using the silhouette method[62]. Microbial taxa were filtered for those with relative abundance ≥1e-4 in at least one individual. Clinical laboratory parameters were compared between clusters and the differences were tested with Kruskal-Wallis rank sum test.

Data resources. The metagenomic sequence data is available at NCBI under BioProject with accession PRJNA385350 https://www.ncbi.nlm.nih.gov/bioproject/385350.

References

1. Flores-Mireles, A. L., Walker, J. N., Caparon, M. & Hultgren, S. J. Urinary tract infections: epidemiology, mechanisms of infection and treatment options. *Nat Rev Microbiol* **13**, 269–284, https://doi.org/10.1038/nrmicro3432 (2015).
2. Schmiemann, G., Kniehl, E., Gebhardt, K., Matejczyk, M. M. & Hummers-Pradier, E. The diagnosis of urinary tract infection: a systematic review. *Dtsch Arztebl Int* **107**, 361–367, https://doi.org/10.3238/arztebl.2010.0361 (2010).
3. Foxman, B. The epidemiology of urinary tract infection. *Nat Rev Urol* **7**, 653–660, https://doi.org/10.1038/nrurol.2010.190 (2010).
4. Lewis, D. A. *et al.* The human urinary microbiome; bacterial DNA in voided urine of asymptomatic adults. *Front Cell Infect Microbiol* **3**, 41, https://doi.org/10.3389/fcimb.2013.00041 (2013).
5. Yu, Y. *et al.* Diagnosing inflammation and infection in the urinary system via proteomics. *J Transl Med* **13**, 111, https://doi.org/10.1186/s12967-015-0475-3 (2015).
6. Yu, Y. *et al.* Similar Neutrophil-Driven Inflammatory and Antibacterial Responses in Elderly Patients with Symptomatic and Asymptomatic Bacteriuria. *Infect Immun* **83**, 4142–4153, https://doi.org/10.1128/IAI.00745-15 (2015).
7. Wolfe, A. J. *et al.* Evidence of uncultivated bacteria in the adult female bladder. *J Clin Microbiol* **50**, 1376–1383, https://doi.org/10.1128/JCM.05852-11 (2012).
8. Hilt, E. E. *et al.* Urine is not sterile: use of enhanced urine culture techniques to detect resident bacterial flora in the adult female bladder. *J Clin Microbiol* **52**, 871–876, https://doi.org/10.1128/JCM.02876-13 (2014).
9. Fricke, W. F., Maddox, C., Song, Y. & Bromberg, J. S. Human microbiota characterization in the course of renal transplantation. *Am J Transplant* **14**, 416–427, https://doi.org/10.1111/ajt.12588 (2014).
10. Nelson, D. E. *et al.* Bacterial communities of the coronal sulcus and distal urethra of adolescent males. *PloS one* **7**, e36298, https://doi.org/10.1371/journal.pone.0036298 (2012).
11. Siddiqui, H., Nederbragt, A. J., Lagesen, K., Jeansson, S. L. & Jakobsen, K. S. Assessing diversity of the female urine microbiota by high throughput sequencing of 16S rDNA amplicons. *BMC Microbiol* **11**, 244, https://doi.org/10.1186/1471-2180-11-244 (2011).
12. Siddiqui, H., Lagesen, K., Nederbragt, A. J., Jeansson, S. L. & Jakobsen, K. S. Alterations of microbiota in urine from women with interstitial cystitis. *BMC Microbiol* **12**, 205, https://doi.org/10.1186/1471-2180-12-205 (2012).
13. Fouts, D. E. *et al.* Integrated next-generation sequencing of 16S rDNA and metaproteomics differentiate the healthy urine microbiome from asymptomatic bacteriuria in neuropathic bladder associated with spinal cord injury. *J Transl Med* **10**, 174, https://doi.org/10.1186/1479-5876-10-174 (2012).
14. Nelson, D. E. *et al.* Characteristic male urine microbiomes associate with asymptomatic sexually transmitted infection. *PloS one* **5**, e14116, https://doi.org/10.1371/journal.pone.0014116 (2010).
15. Dong, Q. *et al.* The microbial communities in male first catch urine are highly similar to those in paired urethral swab specimens. *PloS one* **6**, e19709, https://doi.org/10.1371/journal.pone.0019709 (2011).
16. Willner, D. *et al.* Single clinical isolates from acute uncomplicated urinary tract infections are representative of dominant *in situ* populations. *mBio* **5**, e01064–01013, https://doi.org/10.1128/mBio.01064-13 (2014).

17. Pearce, M. M. *et al*. The female urinary microbiome: a comparison of women with and without urgency urinary incontinence. *mBio* **5**, e01283–01214, https://doi.org/10.1128/mBio.01283-14 (2014).
18. Karstens, L. *et al*. Does the Urinary Microbiome Play a Role in Urgency Urinary Incontinence and Its Severity? *Front Cell Infect Microbiol* **6**, 78, https://doi.org/10.3389/fcimb.2016.00078 (2016).
19. Groah, S. L. *et al*. Redefining Healthy Urine: A Cross-Sectional Exploratory Metagenomic Study of People With and Without Bladder Dysfunction. *J Urol* **196**, 579–587, https://doi.org/10.1016/j.juro.2016.01.088 (2016).
20. Shrestha, E. *et al*. Profiling the Urinary Microbiome in Men with Positive versus Negative Biopsies for Prostate Cancer. *J Urol* **199**, 161–171, https://doi.org/10.1016/j.juro.2017.08.001 (2018).
21. Gottschick, C. *et al*. The urinary microbiota of men and women and its changes in women during bacterial vaginosis and antibiotic treatment. *Microbiome* **5**, 99, https://doi.org/10.1186/s40168-017-0305-3 (2017).
22. Hasman, H. *et al*. Rapid whole-genome sequencing for detection and characterization of microorganisms directly from clinical samples. *J Clin Microbiol* **52**, 139–146, https://doi.org/10.1128/JCM.02452-13 (2014).
23. Lotte, R., Lotte, L. & Ruimy, R. Actinotignum schaalii (formerly Actinobaculum schaalii): a newly recognized pathogen-review of the literature. *Clin Microbiol Infect* **22**, 28–36, https://doi.org/10.1016/j.cmi.2015.10.038 (2016).
24. Stevens, R. P. & Taylor, P. C. Actinotignum (formerly Actinobaculum) schaalii: a review of MALDI-TOF for identification of clinical isolates, and a proposed method for presumptive phenotypic identification. *Pathology* **48**, 367–371, https://doi.org/10.1016/j.pathol.2016.03.006 (2016).
25. Chen, L., Zheng, D., Liu, B. & Yang, J. & Jin, Q. VFDB 2016: hierarchical and refined dataset for big data analysis–10 years on. *Nucleic acids res* **44**, D694–697, https://doi.org/10.1093/nar/gkv1239 (2016).
26. Kauffman, C. A. Diagnosis and management of fungal urinary tract infection. *Infect Dis Clin North Am* **28**, 61–74, https://doi.org/10.1016/j.idc.2013.09.004 (2014).
27. Jo, J. H. *et al*. Diverse Human Skin Fungal Communities in Children Converge in Adulthood. *J Invest Dermatol* **136**, 2356–2363, https://doi.org/10.1016/j.jid.2016.05.130 (2016).
28. Huppert, J. S. *et al*. Urinary symptoms in adolescent females: STI or UTI? *J Adolesc Health* **40**, 418–424, https://doi.org/10.1016/j.jadohealth.2006.12.010 (2007).
29. Dize, L. *et al*. Comparison of self-obtained penile-meatal swabs to urine for the detection of C. trachomatis, N. gonorrhoeae and T. vaginalis. *Sex Transm Infect* **89**, 305–307, https://doi.org/10.1136/sextrans-2012-050686 (2013).
30. Egli, A. *et al*. Prevalence of polyomavirus BK and JC infection and replication in 400 healthy blood donors. *The Journal of infectious diseases* **199**, 837–846 (2009).
31. Zhong, S. *et al*. Age-related urinary excretion of BK polyomavirus by nonimmunocompromised individuals. *J Clin Microbiol* **45**, 193–198, https://doi.org/10.1128/JCM.01645-06 (2007).
32. Marshall, W. F., Telenti, A., Proper, J., Aksamit, A. J. & Smith, T. F. Survey of urine from transplant recipients for polyomaviruses JC and BK using the polymerase chain reaction. *Mol Cell Probes* **5**, 125–128 (1991).
33. Suga, S., Yoshikawa, T., Kajita, Y., Ozaki, T. & Asano, Y. Prospective study of persistence and excretion of human herpesvirus-6 in patients with exanthem subitum and their parents. *Pediatrics* **102**, 900–904 (1998).
34. Ko, H.-R., Shin, S. M. & Park, S. W. Predicting Factors of Roseola Infantum Infected with Human Herpesvirus 6 from Urinary Tract Infection *Child. Kidney Dis* **20**, 69–73 (2016).
35. Daibata, M., Taguchi, T., Nemoto, Y., Taguchi, H. & Miyoshi, I. Inheritance of chromosomally integrated human herpesvirus 6 DNA. *Blood* **94**, 1545–1549 (1999).
36. Moustafa, A. *et al*. The blood DNA virome in 8,000 humans. *PLoS pathogens* **13**, e1006292, https://doi.org/10.1371/journal.ppat.1006292 (2017).
37. Jovel, J. *et al*. Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics. *Front Microbiol* **7**, 459, https://doi.org/10.3389/fmicb.2016.00459 (2016).
38. Costea, P. I. *et al*. Subspecies in the global human gut microbiome. *Mol Syst Biol* **13**, 960, https://doi.org/10.15252/msb.20177589 (2017).
39. Huys, G. *et al*. Alloscardovia omnicolens gen. nov., sp. nov., from human clinical samples. *Int J Syst Evol Microbiol* **57**, 1442–1446, https://doi.org/10.1099/ijs.0.64812-0 (2007).
40. Mahlen, S. D. & Clarridge, J. E. Site and clinical significance of Alloscardovia omnicolens and Bifidobacterium species isolated in the clinical laboratory. *J Clin Microbiol* **47**, 3289–3293, https://doi.org/10.1128/JCM.00555–09 (2009).
41. Brown, M. K., Forbes, B. A., Stitley, K. & Doern, C. D. Defining the Clinical Significance of Alloscardovia omnincolens in the Urinary Tract. *J Clin Microbiol* **54**, 1552–1556, https://doi.org/10.1128/JCM.03084-15 (2016).
42. Ogawa, Y. *et al*. Bacteremia secondary to Alloscardovia omnicolens urinary tract infection. *J Infect Chemother* **22**, 424–425, https://doi.org/10.1016/j.jiac.2015.12.013 (2016).
43. Rani, A. *et al*. A diverse virome in kidney transplant patients contains multiple viral subtypes with distinct polymorphisms. *Scientific reports* **6**, 33327, https://doi.org/10.1038/srep33327 (2016).
44. Paduch, D. A. Viral lower urinary tract infections. *Curr Urol Rep* **8**, 324–335 (2007).
45. Santiago-Rodriguez, T. M., Ly, M., Bonilla, N. & Pride, D. T. The human urine virome in association with urinary tract infections. *Front Microbiol* **6**, 14, https://doi.org/10.3389/fmicb.2015.00014 (2015).
46. Cox, J. *et al*. A practical guide to the MaxQuant computational platform for SILAC-based quantitative proteomics. *Nat Protoc* **4**, 698–705, https://doi.org/10.1038/nprot.2009.36 (2009).
47. Rajagopala, S. V. *et al*. Gastrointestinal microbial populations can distinguish pediatric and adolescent Acute Lymphoblastic Leukemia (ALL) at the time of disease diagnosis. *BMC Genomics* **17**, 635, https://doi.org/10.1186/s12864-016-2965-y (2016).
48. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935, https://doi.org/10.1093/bioinformatics/btt509 (2013).
49. Quast, C. *et al*. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic acids res* **41**, D590–596, https://doi.org/10.1093/nar/gks1219 (2013).
50. Camacho, C. *et al*. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421, https://doi.org/10.1186/1471-2105-10-421 (2009).
51. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids res* **30**, 3059–3066 (2002).
52. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2–approximately maximum-likelihood trees for large alignments. *PloS one* **5**, e9490, https://doi.org/10.1371/journal.pone.0009490 (2010).
53. Matsen, F. A., Kodner, R. B. & Armbrust, E. V. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* **11**, 538, https://doi.org/10.1186/1471-2105-11-538 (2010).
54. Yu, G., Smith, D. K., Zhu, H. & Lam, T. T.-Y. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Meth Ecol Evol* **8**, 28–36 (2017).
55. Anderson, E. L. *et al*. A robust ambient temperature collection and stabilization strategy: Enabling worldwide functional studies of the human microbiome. *Scientific reports* **6**, 31731, https://doi.org/10.1038/srep31731 (2016).
56. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760, https://doi.org/10.1093/bioinformatics/btp324 (2009).

57. Do, C. B. & Batzoglou, S. What is the expectation maximization algorithm? *Nature biotechnology* **26**, 897–899, https://doi.org/10.1038/nbt1406 (2008).
58. Peng, Y., Leung, H. C., Yiu, S. M. & Chin, F. Y. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420–1428, https://doi.org/10.1093/bioinformatics/bts174 (2012).
59. Noguchi, H., Park, J. & Takagi, T. MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic acids res* **34**, 5623–5630, https://doi.org/10.1093/nar/gkl723 (2006).
60. Stekhoven, D. J. & Buhlmann, P. MissForest–non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28**, 112–118, https://doi.org/10.1093/bioinformatics/btr597 (2012).
61. Reynolds, A. P., Richards, G., de la Iglesia, B. & Rayward-Smith, V. J. Clustering Rules: A Comparison of Partitioning and Hierarchical Clustering Algorithms. *J Math Model Algor* **5**, 475–504 (2006).
62. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* **20**, 53–65 (1987).

Acknowledgements

A.M., W.L., W.B., J.C.V., K.E.N., are employees and or stock holders of Human Longevity Inc. There is no IP associated with this work.

Author Contributions

J.C.V., K.E.N., R.P. and A.T. designed the study, R.P. and A.T. designed the analyses, A.M., H.S., W.L., K.J.M., M.G.T., Y.Y. performed analyses and laboratory work, O.M. and A.T. established clinical validity, W.B. directed sequencing, A.M., O.M., R.P., A.T. wrote the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-22660-8.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.