Unraveling the impact of host genetics and factors on the urinary microbiome in a young population

Leying Zou[1,†], Zhe Zhang[2,†], Ruijin Guo[2], Xin Tong[2], Junhong Chen[1], Yanmei Ju[2,3], Haorong Lu[4], Huanming Yang[2,5], Jian Wang[2,5], Yang Zong[2], Xun Xu[2], Xin Jin[2], Liang Xiao[2,6], Tao Zhang[1,7*], Xiaomin Liu[2,7*]

1. BGI Research, Wuhan 430074, China;
2. BGI Research, Shenzhen 518083, China
3. College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100049, China;
4. China National Genebank, BGI, Shenzhen 518120, China;
5. James D. Watson Institute of Genome Sciences, Hangzhou 310058, China;
6. Shenzhen Engineering Laboratory of Detection and Intervention of Human Intestinal Microbiome, BGI-Shenzhen, Shenzhen 518083, China;
7. Shenzhen Key Laboratory of Human Commensal Microorganisms and Health Research, BGI Research, Shenzhen 518083, China

[†] These authors contributed equally.

* To whom correspondence should be addressed: X.L., liuxiaomin@genomics.cn and T.Z., tao.zhang@genomics.cn

Abstract

The significance of the urinary microbiome in maintaining health and contributing to disease development is increasingly recognized. However, a comprehensive understanding of this microbiome and its influencing factors remains elusive. Utilizing whole metagenomic and whole-genome sequencing, along with detailed metadata, we characterized the urinary microbiome and its influencing factors in a cohort of 1579 Chinese individuals. Our findings unveil the distinctiveness of the urinary microbiome from other body sites, delineating five unique urotypes. We identified 137 host factors significantly influencing the urinary microbiome, collectively explaining 13.16% of the variance in microbial composition. Notably, gender-related factors, including sex hormones, emerged as key determinants in defining urotype groups and microbial composition, with the urinary microbiome exhibiting strong predictive ability for gender (AUC=0.815). Furthermore, we discovered 43 genome-wide significant associations between host genetic loci and specific urinary bacteria, particularly *Acinetobacter,* linked to 8 host loci ($p < 5x10^{-8}$). These associations were also modulated by gender and sex hormone levels. In summary, our study provides novel insights into the impact of host

genetics and other factors on the urinary microbiome, shedding light on its implications for host health and disease.

Introduction

Mounting evidence has implicated the presence of a microbiome community within the urinary tract, commonly referred to as urinary microbiome or urobiome[1,2]. This discovery has been substantiated through high-throughput 16S rRNA/metagenomic sequencing, as well as enhanced urine culture techniques across multiple studies[3-6]. In recent years, the urinary microbiota has garnered considerable attention owning to its potential implications for urinary tract health and diseases, such as urine incontinence[7,8], urinary tract infection(UTI)[9,10], chronic kidney disease[11], and bladder cancer[12].

Understanding the urinary microbiome and elucidating the factors that influence its composition and diversity are essential for developing effective diagnostic and therapeutic interventions. Previous studies have identified certain host factors, such as sex and age, that may influence composition of the urinary microbiota[13,14]. A recent study in mainly postmenopausal women has suggested a role for host genetics in microbial variations among individuals[15]. However, these studies have predominantly utilized 16S rRNA sequencing that had relatively low resolution and allowed for the identification and quantification of bacterial taxa at the genus level. Furthermore, existing research has primarily focused on elderly or middle-aged individuals, in which the microbiome may be largely influenced by the environmental factors in the aging process. Thus, there is a gap in knowledge concerning urinary microbiota in young individuals, which our study aims to address. Moreover, there is a necessity to explore additional potential host factors, such as blood metabolites, urine factors, lifestyle, etc., as they may also impact the urinary microbiota. Current research in this area remains limited and calls for more in-depth exploration.

Here in this study, leveraging shotgun metagenomics sequencing and comprehensive metadata from 1579 Chinese individuals from the 4D-SZ cohort[16-21], which is primarily comprised of young adults around 30 years old, we conducted a thorough investigation to enhance our understanding of the urinary microbiota and its influencing factors. Initially, we characterized the urinary microbiome by analyzing its microbial composition and comparing it to microbial communities in other body sites. We then systematically investigated its influencing factors among 386 traits, encompassing sex, age, BMI, diets, lifestyles, as well as blood and urinary measurements. Moreover, by performing the so far first urinary Metagenome-genome wide association study (M-GWAS) in a subset of 687 individuals with available whole genome

sequencing data, we endeavor to unravel the intricate interplay between host genetics and the urinary microbiome.

Results

Urinary microbiome: A unique microecology distinct from other body sites

We collected urine samples of 1579 individuals, with 66.12% being females and a mean age of 29 years. Utilizing shotgun metagenomic sequencing, we generated an average of 19.18 ± 7.90 Gb of data per sample (**Supplementary Table 1**; **Supplementary Fig. 1**). These samples are from the 4D-SZ cohort[16-21], with multi-omics data collected including human whole genome, whole metagenomes across multiple body sites, blood metabolites, and detailed questionaries. Since we have metagenomic data from not only the urinary tract but also the other body sites (**Fig. 1a**), including the gut (n=1754), saliva (n=3222), and reproductive tract (n=686), it prompted us to compare the urinary microbiome with the microbial communities inhabiting these distinct body sites. Permutational mutivariate analysis of variance (PERMANOVA) revealed that the urinary microbiome are significantly distinct from the other examined microbiomes. It showed a much lower α-diversity compared to the rich and diverse microbiome harbored in the gut and the oral, with the highest disparity observed in the phylum level with the gut microbiome (R^2=0.455, a higher R^2 values corresponding to a larger difference between the compared sites, p<0.001; **Supplementary Fig. 2**) and in the genus level with the saliva microbiome (R^2 = 0.361, p<0.001; **Fig. 1b**). In comparison to the vaginal microbiome, which is known to characterize a very low α-diversity, the urinary microbiome exhibited the smallest but significant dissimilarity at both phylum (R^2=0.214) and genus level (R^2 = 0.111), with p<0.001 for both (**Supplementary Fig. 2 and Fig. 1b)**, along with a significantly higher α-diversity (**Supplementary Fig. 3)**.

Further analysis was conducted on metagenomic data from 75 females to corroborate the above findings, wherein data from all four body sites were available. Consistent with the broader dataset, the urinary microbiome exhibited the most pronounced differences from the gut microbiome at the phylum level (R^2 = 0.299; **Supplementary Fig. 4**), and from the salivary microbiome at the genus level (R^2 = 0.307; **Fig. 1c**), while displaying relatively smaller disparities with the vaginal microbiome (R^2 = 0.11 at the phylum level and R^2 = 0.051 at the genus level; **Supplementary Fig. 4** and **Supplementary Fig. 1c**). Importantly, these variations remained statistically significant (adjusted p < 0.001).

It is evident that these distinctions arise from variations in genus composition across different body sites (**Supplementary Table 2**). For example, *Lactobacillus* was the most abundant genus found in both urinary (an average relative abundance of 17.6%) and vaginal samples (58.9%). while its average relative abundance in saliva and gut was less than 0.13%. The second and third most prevalent genera in urine were *Variovora*x and *Acinetobacter*, with mean relative abundances of 10.6% and 9.4%, respectively; however, their average relative abundances in the other examined sites were less than 0.03%. *Prevotella* is found to be enriched in the saliva with an average relative abundance of 29.1%, while it is less abundant in the gut at 13.4%, followed by urine at 6.7% and vagina at 3.6%. The distinctive microbial compositions and species' diversities across different human body sites resulted in marked dissimilarities in the microbiota across these regions.

In summary, our findings underscore the uniqueness of urinary microbiota in comparison to other anatomical regions such as the gut, saliva, and vagina. This distinct microbial ecosystem warrants further in-depth investigation.

Characteristics of the urinary microbiome

Since the urinary microbiome represents a distinctive microecology, we further characterized its community composition in detail. First, we examined the top ten most abundant genera and species in the cohort (**Supplementary Fig. 5**). At the genus level, *Lactobacillus* had the highest relative abundance, followed by *Variovorax, Acinetobacter, Prevotella,* and *Sphingobium*. At the species level, the top five bacteria with the highest relative abundances, in descending order, were *Lactobacillus iners, Variovorax sp. PDC80, Sphingobium fluviale, Lactobacillus crispatus*, and *Gardnerella vaginalis*. Notably, we observed high Pearson correlations between the relative abundances of bacteria commonly found in the vagina, such as *Lactobacillus*, *Gardnerella*, *Prevotella*, and *Ureaplasma*, in the urinary and vaginal microbiomes (Pearson's r >0.5).

Second, using the partitioning around medoids (PAM) clustering method (Methods), the urinary microbial communities were clustered into five distinct clusters (referred to as "urotypes" analogous to the "enterotypes"). By applying Linear Discriminant Analysis (LDA) for detecting species discriminating each urotype from the rest and ranking them by LDA score (Methods), we determined the dominant species for each urotype. They are *Gardnerella vaginalis* (accounting for 23.10% of the samples)*, Sphingobium fluviale* (34.11%)*, Lactobacillus iners* (15.76%), *Variovorax sp. PDC80* (16.58%), and *Acinetobacter junii* (10.38%), respectively (**Fig. 1d**). Consistent with the clustering results, each urotype showed a significant enrichment of the respective dominant bacteria (**Fig. 1e**).

While the dominant species characterized the highest LDA score for each urotype, there were more species identified, especially in clusters 1, 2, and 5 **(Supplementary Fig. 6**; **Supplementary Table 3**). For cluster 1 (*Gardnerella vaginalis* urotype), the second-ranked species *Lactobacillus crispatus* followed the dominant bacterium *Gardnerella vaginalis* so close in LDA score, just like their close relative abundances with an average of 3.91% versus 3.92%. The individuals characterizing this urotype also harbored significantly higher levels of species belonging to the genus *Prevotella*, including *Prevotella bivia* (3.67%), *Prevotella timonensis* (3.43%), *Prevotella disiens* (3.08%), *Prevotella colorans* (2.84%), and *Prevotella corporis* (2.70%). For cluster 2 (*Sphingobium fluviale* urotype), species such as *Sphingomonas ursincola* (3.76%), *Novosphingobium sp AAP93* (3.62%), *Erythrobacter donghaensis* (3.59%) and *Acinetobacter johnsonli* (3.56%) were also highly represented. For cluster 3 (*Lactobacillus iners* urotype), *Lactobacillus jensenii* ranked second with an LDA score of 2.7 and belonged to the same genus *Lactobacillus* as the dominant species. In cluster 4 (*Variovorax sp PDC80* urotype), *Bradyrhizobium sp CCH1 B1* also showed a high proportion with an LDA score of 3.76. Finally, in cluster 5 (*Acinetobacter junii* urotype), *Herbaspirillum aquaticum* (3.50%), *Cutibacterium acnes* (3.32%), and *Moraxella osloensis* (3.31%) were also most abundant and significantly enriched compared to the other clusters.

Gender-related factors determined urotypes and urinary microbial communities

To understand the factors influencing urotypes, we calculated their correlations with host factors. Notably, gender-related factors emerged as the primary determinants of these urotypes. Menstrual cycle, serum testosterone, serum estradiol, serum creatinine, proportion of urine, menstrual duration, menarche age, hemoglobin, and gender were significantly associated with all five urotypes (**Fig. 2a**, **Supplementary Table 4**). Remarkably, gender distinctly differentiated the urotype groups. Females exhibited higher relative abundances of *Gardnerella vaginalis* (urotype 1) and *Lactobacillus iners* (urotype 3), aligning their urinary microbiome predominantly with these two urotypes. In contrast, males demonstrated higher abundances of *Sphingobium fluviale* (urotype 2), *Variovorax sp. PDC80* (urotype 4), and *Acinetobacter junii* (urotype 5), characterizing their urinary microbial communities with these three respective urotypes. Additionally, the gender-biased urotypes displayed significant correlations with gender-dependent metadata. For instance, female-prevalent urotypes 1 and 3 were correlated with lower levels of serum testosterone, urine specific gravity, and muscle mass; while male-prevalent urotypes 4 and 5 were correlated with higher basal metabolic rate. Besides, some blood amino acids and vitamins were

also related to specific urotypes. For example, homocysteine, proline, and alpha-aminoadipic acid were positively correlated with urotype 4, while glutamic acid, vitamin A, sarcosine, beta-alanine, and hemoglobin all showed positive correlations with urotype 2.

In addition to individual urotypes, we further delved into host factors that contributed to the overall composition of the urinary microbiome and identified a total of 137 significant influencing factors (p.adjust < 0.05; **Supplementary Table 5**). Serum testosterone, gender, urine specific gravity, muscle mass, basal metabolic rate, serum creatine and estradiol were among the top influencing factors associated with β-diversity of the urinary microbiome (**Fig. 2b**). The 137 significant covariates collectively explained 13.16% of the variance of microbial composition (**Fig. 2c**), which encompassed 104 blood factors (an explained variance of 11.03%), four urinary factors (3.87%), gender (3.45%), ten anthropometric measurements (3.3%), six health-related factors (1.84%), seven living habits (1.19%) and five dietary factors (0.84%). Serum hormones had a substantial impact on the urinary microbiome, with serum testosterone (explained R^2 = 4.13%) and estradiol (2.4%) being the first and third most influential ones, respectively. Other metabolites like serum creatinine (2.45%) and uric acid (1.94%) ranked second and sixth. Amino acids in the blood, such as leucine (1.82%), isoleucine (1.69%), valine (1.53%), and glutamic acid (1.35%), also exhibited significant explanatory powers, as did vitamin A (1.38%) among the vitamins. The collective influence of urinary factors ranked as the second significant influence among the seven categories, with urine specific gravity demonstrating the highest explanatory power (3.22%). Several anthropometric factors had explanatory powers exceeding 1%, including muscle mass (2.62%), basal metabolic rate (2.53%), weight (1.91%), height (1.83%), systolic pressure (1.03%), and BMI (0.76%). Within the health-related factors, the test result of the free total prostate specific antigen had the most significant impact on the urinary microbiome (1.14%). In comparison, dietary factors had a comparatively lower effect, with the highest influencing factor being fat meat favorites, which had an explanatory power of only 0.25% (**Supplementary Table 6).**

Motivated by the insights that gender and gender-related factors are primary influencers of the urinary microbiome, we proceeded to investigate differences in the microbial community between genders in detail. We first observed significantly higher α-diversity, as measured by the Shannon index, in the urinary microbiota of females than in males ($p < 8.14 \times 10^{-12}$, **Fig. 3a**). The microbiota composition also significantly differed between males and females (R^2 = 0.03, $p < 1 \times 10^{-4,}$ **Fig. 3b**). We next sought to identify sex-differentiated microbes using Linear Discriminant Analysis Effect Size (LEfSe) (**Fig. 3c**). Females were characterized by higher abundances of genera *Lactobacillus*, *Prevotella, Fannyhessea, Serratia*, and *Citrobacter*. However, males exhibited

more microbiota belonging to *Acinetobacter, Bradyrhizobium, Herbaspirillum, Moraxella*, *Cutibacterium*, and other genera. These results were in line with the Spearman correlations between bacterial profiles and serum testosterone (**Supplementary Fig. 7**), suggesting the important effects of sex and sex hormones on shaping the urinary microbiota.

Given the striking gender-related difference in the microbial community, we employed a random-forest-based machine learning approach to assess the predictive capability of urinary microbiota for gender classification. The predictive accuracy demonstrated by the urinary microbiota in determining gender was notably high, with an area under the curve (AUC) value of approximately 0.815 (**Fig. 3d**). Notably, the representative species of urotype including *Lactobacillus iners*, *Acinetobacter junii*, and *Variovorax sp. PDC80,* were among the top 10 species contributed most significantly to gender differentiation (**Supplementary Table 7,** T-test p-values < 1.8×10^{-3}), which echoes with the gender-dependent distribution of the urotypes. These findings highlight the intricate relationship between microbial composition and host biological characteristics, suggesting that the urinary microbiome's distinct makeup is reflective of and influenced by the host's biological sex.

Host genetics' influences on the urinary microbiota

Out of a total of 1579 individuals, 687 had host whole genome sequencing data available, which enabled us to investigate the influence of the host genome on the urinary microbiome. To achieve this, we conducted a metagenome-genome-wide association study (M-GWAS) on 106 urinary taxa present in over 10% individuals (**Supplementary Table 8**). After adjusting for covariates such as gender, age, BMI, sequencing read counts, and the top ten host Principal Components (PCs), we identified 43 genome-wide significant associations involving 27 host genetic loci and 10 urinary taxa ($p < 5 \times 10^{-8}$, **Fig. 4a**, **Supplementary Table 9**).

Of all microbial taxa examined at the genus and species level, genus *Acinetobacter* was associated with most host genetic loci (n =8; **Fig. 4b**). The strongest association was observed for SNP rs139799338, located in the intronic region of the *PKHD1L1* gene (p=4.82×10^{-12}), negatively associated with the relative abundance of the *Acinetobacter* genus (**Fig. 4a**). *PKHD1L1* has been found to be significantly elevated in the urine of clear cell renal cell carcinoma patients[22]. It was also reported to primarily express in both infant and adult human kidneys and is associated with autosomal recessive polycystic kidney disease[23]. The SNP rs139799338 demonstrated associations with various metabolic traits, including the tumor marker carbohydrate antigen 199 (p=0.006), glycine (p=0.01), vitamin A (p=0.01), beta-alanine (p=0.03), and cobalt (p=0.047), as revealed by the genome-wide

308 association study (GWAS) analysis of host genetic variants and metabolic
309 traits in this cohort (**Supplementary Fig. 8, Supplementary Table 10**).
310
311 The second most significant correlation identified was with rs4241281, located
312 in the intergenic region of genes *CAVIN2* and *TMEFF2*, which was found to
313 have a negative association with the genus *Acinetobacter* (p=9.13x10^{-12}, **Fig.**
314 **4a**, **Supplementary Table 9**). This genetic locus was also correlated with the
315 relative abundance of genus *Bosea*, ranking as the third strongest
316 association. Additionally, it was also linked to the abundance of species
317 *Acinetobacter johnsonii* belonging to the genus *Acinetobacter* (p=1.75x10^{-8}).
318 *CAVIN2* encodes Caveolae Associated Protein 2 in the kidney, and its
319 removal leads to caveolae loss, while overexpression results in caveolae
320 deformation and membrane tubulation[24]. *TMEFF2* has been identified as an
321 epigenetic biomarker for bladder cancer[25], showing significant associations
322 with increasing tumor grade and stage. Notably, SNP rs4241281 in the
323 *CAVIN2* - *TMEFF2* was also correlated with several blood and urinary factors,
324 including concentrations of mercury (p=0.007), alpha amino adipic acid
325 (p=0.015), albumin (p=0.035), body fat mass and rate(p=0.034~0.017), and
326 urine PH (p=0.047**, Supplementary Table 10**). Moreover, rs185971980
327 located in the intergenic region of *FAM170A* and *PRR16*, associated with
328 *Acinetobacter johnsonii* (p=4.46x10^{-10}), was correlated with levels of mean
329 hemoglobin content, mean erythrocyte hemoglobin concentration, average
330 RBC volume and blood iron. Rs117738684 located in the intergenic region of
331 *TMED4* and *OGDH,* associated with *Acinetobacter johnsonii* (p=3.44x10^{-9}),
332 was correlated with levels of blood testosterone, fasting sugar, methionine,
333 and albumin. *LINC01692*, linked to species *Novosphingobium sp. AAP93*,
334 exhibited a positive correlation with urine pH (β=0.087, p=9.3x10^{-4}). *SENP2*,
335 related to the genus *Acinetobacter*, negatively affected urine specific gravity
336 (β=-0.087, p=1.8x10^{-4}, **Supplementary Table 10**). These results suggested
337 host genes may influence the urinary microbiota by regulating the blood
338 metabolites and urinary environment.
339
340 Gender and sex hormones have demonstrated significant impacts on the
341 urinary microbiome, prompting an exploration of gender-gene interactions
342 using the top two loci associated with specific taxa (**Fig. 4c-f**). An intriguing
343 finding emerged where males with elevated testosterone levels exhibited
344 increased abundance of of *Acinetobacter* compared to females, irrespective of
345 genotype (AA or GA/GG). Interestingly, females with genotype AA exhibited a
346 higher abundance of *Acinetobacter* compared to those with genotype GA/GG,
347 while no discernible differences were observed in males with varying
348 genotypes (**Fig. 4c**). Similarly, an interactive effect between rs139799338 and
349 serum testosterone levels on *Acinetobacter* was observed, where individuals
350 with genotype GA/GG and higher serum testosterone level exhibited higher
351 relative abundance of *Acinetobacter* (AA: R=0.28, p=2.7x10^{-13}, GA/GG:

R=0.65, *p*=0.011, **Fig. 4d**). The gender-gene interaction was also evident for another SNP, rs4241281, linked to *Acinetobacter*, where males exhibited greater average relative abundance than females, particularly with the AA allele influencing *Acinetobacter* in the presence of serum testosterone (R=0.28, $p=3.7\times10^{-13}$, **Fig. 4e-f**).

Potential causal relationships between urinary microbiome and diseases

Finally, to explore the potential causal relationships between urinary microbes and diseases by leveraging microbiome-associated host genetic variations, we performed a two-sample Mendelian randomization (MR) analysis. After adjusting for 42 diseases from the BBJ cohort, we identified 7 suggestive causal relationships ($p<=1.19\times10^{-3}=0.05/42$, **Supplementary Table 11, Supplementary Fig. 9**). The strongest MR evidence suggested a link between type 2 diabetes (T2D) and a decreased abundance of *Streptococcus* SGB38301 (β=-0.076, $p=2.83\times10^{-4}$). *Paracoccus haematequi* was causally correlated with a decreased risk of T2D (β=-0.013, $p = 5.12\times10^{-4}$), while the males-enriched *Variovorax sp. PDC80* linked to an increased T2D risk (β=0.017, $p = 8.89\times10^{-4}$). Genus *Herbaspirillum* potentially increased the risk of atopic dermatitis (β=0.064, $p = 4.41\times10^{-4}$). *Bosea robiniae* potentially increased the risk of prostate cancer (β=0.058, $p = 4.65\times10^{-4}$). Keloid was associated with a lower *Bosea* (β = -0.061, $p = 7.64\times10^{-4}$). These inferred relationships underscored a potential connection between urinary microbes and autoimmune diseases, warranting validation in future studies.

Discussion

This study represents the initial profiling of the urinary microbiome in comparison to well-studied human microbiomes from diverse body sites such as the gut, oral cavity, and reproductive tract. While previous studies have also compared the microbial composition across various body sites[26], they relied on databases from different populations and generally involved smaller sample sizes. In contrast, our research provides a more cohesive and comprehensive analysis by utilizing a uniform cohort.

The urinary microbiome was found to possess unique characteristics compared to other sites, evident in the composition of microbial species, as well as α and β diversity differences. The urinary microbiome exhibited significant dissimilarity from other sites, as demonstrated by PCoA analysis in both the entire sample set and the subset of 75 individuals with data from different sites. The difference in microbial communities across body sites was primarily driven by specific taxa. For instance, the gut microbiome was predominantly composed of the genera *Bacteroides* and *Prevotella*, whereas the saliva microbiome showed a gradient of abundance among *Prevotella, Neisseria*, and *Haemophilus*. The urinary microbiome was predominantly

composed of genera such as *Lactobacillus*, *Variovorax*, and *Acinetobacter*, while the vaginal microbiome exhibited more abundant *Lactobacillus* and *Gardnerella*. This highlights the significant role that specific microbial taxa play in characterizing the microbial landscape of different body sites, indicating distinct ecological niches and potentially different functional roles within the human body. Despite the differences, urine and vagina harbored several predominant bacteria in common, supporting the relatively closer relatedness between them[27] compared to other physically distal sites.

We identified five urotypes and their respective dominant bacteria. A previous study extracted microbiota from encrustations on ureteral stents in patients without urinary tract infections and conducted urotype analysis[28]. Their findings, similar to ours, identified types dominated by microbes such as *Lactobacillus*, *Gardnerella vaginalis*, *Achromobacter*, *Corynebacterium*, and *Streptococcus* (**Supplymentary Table 3**). However, they identified extra bacteria like Actinomycetales, Enterobacterales, and *Staphylococcus* which were less abundant in this cohort. These differences probably have stemmed from the distinct sources of the analyzed microbiome. Their study focused on bacteria from ureteral stents, on which nosocomial bacteria may form biofilm. They employed the 16S rRNA sequencing method, which differs from our metagenomic sequencing approach.

Although some studies indicated the urinary microbiome may be impacted by sex[29-32], they were performed in small samples and mainly based on the 16S rRNA data, resulting in limited sex-specific bacteria identification. In this study, we discovered extensive sex differences in the urinary microbiota which could well predict sex (AUC=0.815). We also observed a strong link between the urinary microbiome and sex hormone levels as well as some other sex-specific factors. These findings not only enrich our understanding of the gender-specific susceptibilities to certain urinary tract conditions or diseases but also open avenues for utilizing microbial profiles in gender-specific medical research and potential diagnostic applications.

Our study also marks the first detailed exploration of the impact of blood metabolites on the urinary microbiome. Serum hormones, particularly serum testosterone and serum estradiol, emerged as significant influencers of the urinary microbiome. Other metabolites, such as serum creatinine and serum uric acid, along with blood amino acids like leucine, isoleucine, valine, glutamic acid, and vitamin A, were also found to play crucial roles, highlighting the intricate relationship between various serum components and the composition of the urinary microbiome. Serum creatinine has been regarded as a biomarker of renal function for almost a century[33] and is widely utilized in clinical diagnostics. In our study, serum creatinine showed significant explanatory power for the β-diversity of urinary microbiota and different

urotypes, suggesting that serum creatinine levels might influence kidney function through its effects on the composition of urinary microbiota. The potential causal relationship warrants further investigation. This link emphasizes the intricate connections between metabolic biomarkers and microbiome composition, highlighting the importance of a holistic approach to understanding renal health and disease.

Limited studies investigated the effect of host genome on the urinary microbiome, except a previous study that reported considerable genetic contribution and identified heritable *Escherichia* (A = 0.165) and *Lactobacillus iners* (A = 0.177) by analyzing midstream urine samples from 1,600 older females in the TwinsUK cohort[15]. This study represents a pioneering effort in conducting a genome-wide association analysis of the urinary microbiome along with the host genome. We have successfully identified 27 significant genomic loci associated with 10 urinary microbes, demonstrating the influence of genetic factors on the composition of the urinary microbiome. Genus *Acinetobacter* and its species *Acinetobacter johnsonii* showed the most link to host genetic loci. In addition to the urinary microbiome, *Acinetobacter johnsonii* exhibited a genetic attribute to the skin microbiome[34]. In this GWAS, we also observed an association between genus Acinetobacter and the *MARCO* gene (β=-0.183, p=3.67x10^{-6}), which encoded macrophage receptor with collagenous structure and correlated with stress urinary incontinence (replication p= 0.003) in a GWAS of European women[35]. Moreover, we delved deeper into exploring the impact of interactions between significant loci genotypes, gender, and serum testosterone levels on specific urinary microbes. The insights gained from these analyses offer valuable information on the complex interplay between host genetics, hormonal levels, and microbial abundance.

Methods

Study subjects

All the participants involved in this research belonged to the '4D-SZ' cohort, which encompassed the collection of blood, gut, saliva, vaginal, and urine samples, along with comprehensive metadata, for a multi-omics investigation as previously reported[16-21]. In this study, 1579 urine samples from the cohort were collected for whole metagenomic sequencing (**Supplementary Table 1**). The detailed urine sampling, sequencing, and quality control design were described below. For comparison, we also included metagenomic samples from the gut (n=1754), saliva (n=3222), and reproductive tract (n=686), all from the 4D-SZ cohort. Feces were collected with an MGIEasy kit, and stool DNA was extracted in accordance with the MetaHIT protocol as described previously[36]. For the salivary sample, we employed a 2× concentration of a

stabilizing reagent kit and collected 2 mL of saliva. The DNA extraction from saliva samples was conducted using the MagPure Stool DNA KF Kit B (no. MD5115-02B). For reproductive tract samples, the extraction of genomic DNA from was performed in accordance with the established protocol[37](Hao L,2020). The DNA concentrations from blood, stool, oral, and reproductive tract samples were estimated by Qubit (Invitrogen), and the library preparation and sequencing were conducted in accordance with the procedures described in previous studies[16-21].

All recruitment and study procedures involving human subjects were approved by the Institutional Review Boards (IRB) at BGI-Shenzhen. All participants provided written informed consent at enrolment.

Urine metagenomic sampling, sequencing, and quality control

Midstream voided urine was self-collected using a 40-ml urine cup by 1579 individuals from the 4D-SZ cohort. From the urine cup, 5ml urine was flowed into a sterile tube containing 5ml BGI stabilizing reagent[38] for the preservation of metagenome at room temperature and then stored at -80°C within the day before subsequent metagenomic analysis.

The samples were thawed and transferred into a 15mL centrifuge tube. Then 1mL TE buffer (1M Tris-HCL & 0.1M EDTA) for crystal dissolution[39] and 200uL 50mg/mL lysozyme were added before centrifuge at 8000rpm for 10min. Then DNA extraction was proceeded with MagPure Stool DNA KF Kit B (MD5115, Magen)[40].

DNA extraction and library preparation were performed along with blank negative controls and mock community controls. Specifically, ZymoBIOMICS® Microbial Community Standard (D6300) and ZymoBIOMICS® Microbial Community DNA Standard (D6305) were 1x10^4 times diluted using saline to simulate the low microbial load in urine samples and divided into aliquots (1ml per tube) for storage at -80°C before using as extraction and library preparation controls respectively. During extraction, blank negative control and extraction control (D6300 diluted) were added for each 96-well plate and processed with the same procedure as for the urine samples. Similarly, separate blank negative control and library preparation control (D6305 diluted) were also added for each 96-well plate during library preparation and processed with the same procedure as for the urine samples. Notably, none of the negative controls for both extraction and library preparation passed the library preparation stage for PCR concentration lower than the minimum criteria of 1ng/uL. In total, 33 extraction mock controls and 43 library mock controls generated metagenomic sequencing data.

Filtering and Removal of Possible Contaminants

As low biomass microbiome tends to be more easily influenced by

contaminants and cross-talk, multiple contamination control steps were further applied to the sequencing data.

Step 1. For all the samples in the entire cohort, species were considered as possible contaminants and collectively removed if a species meets one of the following criteria:

A. Relative abundance significantly correlated with PCR concentration (Spearman test, p-value <0.05)

B. Decontam score <0.3 calculated by 'decontam' package[41] based on relative abundance and PCR concentration or occurrence in <4 samples

C. Detected in the 76 mock controls but other than the 10 expected species comprising the mock community with > $1x10^{-4}$ mean relative abundance

D. Reads count less than 1000

Step 2. For all the samples in each specific 96-well plate, remaining species other than the 10 expected species detected in the respective mock control with >$1x10^{-4}$ relative abundance were considered as contaminants specific to that 96-well plate and were removed.

Step 3. Filtering out non-bacterial microbes and retaining only bacteria taxa and renormalization

Constructing the urinary metagenomic taxonomic profile

Taxonomy assignment and functional prediction were performed using MetaPhlAn4[42] version 4.0.6 with default settings. MetaPhlAn 4, an advanced tool for microbial analysis, utilizes around 5.1 million unique clade-specific marker genes derived from approximately 1.01 million microbial genomes (236,600 reference genomes and 771,500 metagenomic assembled genomes), covering 26,970 species-level genome bins (21,978 kSGBs and 4,992 uSGBs). Compared to its predecessor, MetaPhlAn 3, MetaPhlAn 4 identifies a broader and more detailed range of urinary microbiota. Utilizing this tool, we obtained a raw microbial taxonomic dataset composed of a total of 1,063 taxa (11 phyla, 32 classes, 60 orders, 115 families, 248 genera, and 597 species).

Calculating α-, β- diversity, PCOA and urotypes

The microbial diversity metrics, including α-diversity (Shannon and Simpson indices) and β-diversity (Bray–Curtis dissimilarities), were calculated based on species-level abundance data using the 'diversity' and 'vegdist' functions within the R package 'vegan', respectively. Subsequently, principal coordinates analysis (PCoA) was conducted on the computed beta-diversity dissimilarities utilizing the 'capscale' function from the 'vegan' package.

571 For sample clustering based on relative species abundances, the JSD
572 distance metric and the Partitioning Around Medoids (PAM) clustering
573 algorithm were employed. The optimal number of clusters was determined
574 using the Calinski-Harabasz (CH) Index[43] (**Supplementary Fig. 10**).
575 Following this, the Linear Discriminant Analysis Effect Size (LEfSe)[44], a
576 nonparametric cross-sample test utilizing Linear Discriminant Analysis (LDA)
577 for test statistic computation, supported by traditional univariate tests for
578 feature selection, was utilized. The microbial taxa were ranked based on their
579 LDA values in descending order, and the predominant bacteria within each
580 urotype were identified as the species with the highest LDA value.
581
582 **Host factors explained the variance of the urinary microbiome**
583 We investigated the correlations between urinary microbiome compositions
584 and 420 host variables encompassing anthropometric measurements, blood
585 and urine parameters, among others. To identify variables significantly linked
586 to β-diversity, we employed Bray-Curtis distance-based redundancy analysis
587 (dbRDA) and determined the variance explained by these factors using the
588 'capscale' function in the vegan package. Significance of each response
589 variable was verified through analysis of variance (ANOVA) for the dbRDA
590 using the 'anova.cca()' function in the vegan package. Only variables showing
591 significant associations (Benjamini-Hochberg FDR < 0.05) with β-diversity in
592 univariable models were included in the multivariable analysis. These 137
593 significant factors were categorized into distinct groups including urinary
594 parameters, anthropometric measures, health-related indicators, lifestyle
595 habits, and dietary elements. To assess the combined explanatory power of
596 these variable groups on urinary microbiome compositions, we performed a
597 variation partitioning analysis using the vegan package. The 'adj.r.squared'
598 value was calculated using the RsquareAdj option to quantify the proportion of
599 variance explained by the different categories.
600
601 **High-depth whole-genome sequencing for the blood samples**
602 Among the 1,579 individuals, 687 with blood samples were sequenced to a
603 mean of 30× for the whole genome. For blood samples, buffy coat was
604 isolated and DNA was extracted using HiPure Blood DNA Mini Kit (Magen,
605 catalog no. D3111) according to the manufacturer's protocol. The library
606 construction and sequencing methods are consistent with those described in
607 previous articles[16].
608 For genomic analysis, whole-genome reads were aligned to the
609 GRCh38/hg38 human genome using BWA[45] (v0.7.15), with low-quality reads
610 filtered out. Alignments were processed into BAM files using Samtools[46]
611 (v0.1.18) and PCR duplicates were marked with Picardtools (v1.62). The
612 Genome Analysis Toolkit (GATK[47], v3.8) facilitated SNP and INDEL
613 identification and recalibration, referencing dbSNP (v150). Variant discovery
614 and processing were performed with GATK's HaplotypeCaller and subsequent

615 tools. The final dataset, comprising 7.2 million SNPs/INDELs(MAF ≥ 1%),
616 were remained for M-GWAS analyses.
617 For variant inclusion, we set stringent thresholds: (i) a minimum average
618 depth of over 8x; (ii) Hardy-Weinberg equilibrium (HWE) P-value >10−5; and
619 (iii) a genotype calling rate >98%. Regarding sample criteria, we required: (i)
620 mean sequencing depth >20 ×; (ii) a variant call rate >98%; (iii) absence of
621 population stratification, confirmed through principal components analysis
622 (PCA) conducted using PLINK[48](version 1.07); and (iv) the exclusion of
623 related individuals, determined by calculating pairwise identity by descent
624 (IBD, with a Pi-hat threshold of 0.1875) in PLINK. Following these rigorous
625 quality control measures, 687 individuals were left for subsequent analysis.
626
627
628 **Predictive analysis of gender using urinary microbiota**
629 We employed the Random Forest algorithm to explore the potential of using
630 the average relative abundance of urinary microbiota for gender prediction.
631 This approach was chosen due to its efficacy in handling complex biological
632 data and its robustness in feature selection, making it particularly suitable for
633 high-dimensional datasets like microbiome compositions. We first divided our
634 dataset into 70% training and 30% testing sets, ensuring a balanced
635 representation of male and female samples. The Random Forest model was
636 then trained on the training set with gender as the response variable and the
637 average relative abundance of different urinary microbes as predictor
638 variables. We used 500 trees in our model to ensure adequate learning and
639 generalization capability. The model's performance was evaluated using the
640 Area Under the Receiver Operating Characteristic (ROC) Curve (AUC), which
641 provides a comprehensive measure of model accuracy across different
642 classification thresholds. Furthermore, we conducted t-tests to identify the top
643 10 differentially abundant microbes between genders.
644
645 **Association analysis for microbial taxa**
646 Considering the power of GWAS tests, we refined our approach to focus on
647 microbial taxa exhibiting occurrence rates exceeding 10% (present in at least
648 157 individuals) and an average relative abundance above 1×10^{-4}. After
649 filtering, the represented genera of these microbial taxa covered 99.73% of
650 the whole community in the cohort. Subsequently, a total of 106 microbial
651 genera and species were included for association analyses.
652 We examined the relationships between host genetics and the urine
653 microbiome using two different statistical models: a linear model based on the
654 relative abundance and a logistic model based on the presence/absence
655 (P/A) of microbial features. For those taxa present in over 50% individuals, we
656 used a linear model and transformed the relative abundance data using a
657 natural logarithm. Subsequently, we calculated the residuals by employing the
658 'lm' function in R: (log10(Microbe abundance) _ age + sex + BMI +

sequencing read counts + top ten PCs). These residuals were then extracted using the residuals() function from the R stats package and applied in a univariate linear model for association analysis with genotypes. For the microbial taxa observed in more than 10% but less than 90% of individuals, we converted the data into presence/absence patterns to avoid issues with zero inflation. By dichotomizing these data, we could treat the bacterial abundance as a binary trait for logistic regression analysis. This analysis was adjusted for the same set of covariates as mentioned earlier. This dual approach allowed us to comprehensively assess the influence of host genetics on different aspects of the urine microbiome.

We searched the oral microbiome-related SNPs in the summary statistics data from this cohort as previously reported to examine their associations with host traits.

Additionally, we conducted GWAS association analyses similar to the steps described above for α, β diversity, and urinary clusters, however, no genome-wide significant signals were observed (**Supplementary Tables 12 & 13**). We also performed gene-environment interaction analyses on the top two significant loci, chr8:109442613 and chr2:191850055. These analyses, focusing on interactions with gender and serum testosterone levels, were executed using the --interaction parameter in PLINK.

Two-sample MR analysis

To explore the causal relationships between the urinary microbiome and various diseases, we accessed summary statistics data of 42 diseases from Biobank Japan[49,50]. The two-sample bidirectional MR analysis was performed by applying the GSMR method[51]. Genetic variants with $P < 1 \times 10^{-5}$ and LD r^2 <0.1 were also selected as instrumental variables for phenotypes in the Japan Biobank study.

Data availability

The metagenomic sequencing data after removing host reads have been deposited to the CNGB Nucleotide Sequence Archive (CNSA: https://db.cngb.org/cnsa; accession number CNP0003862). The summary statistics that support the findings of this study including the associations between host genetics and urinary microbiome are available from https://db.cngb.org/search/project/CNP0005626/. According to the Human Genetic Resources Administration of China regulation and the institutional review board of BGI-Shenzhen related to protecting individual privacy, sequencing data are controlled-access and are available via application on request.

Acknowledgments

We sincerely thank the support provided by China National GeneBank. We thank all the volunteers for their time and for self-collecting the oral samples using our kit. This work was supported by the National Natural Science Foundation of China (No. 32200548).

Author contributions

X.L. and T.Z. conceived and organized this study. J.W. initiated the overall health project. X.X., X.J., L.X., Y.H., and Y.Z. contributed to organization of the cohort, the sample collection and questionnaire collection. H.L led the DNA extraction and sequencing. L.Z. and Z.Z. processed the metagenome data. X.L and X.T. processed the whole genome data. L.Z. and X.L performed the metagenome-genome-wide association analyses. L.Z., Z.Z., and X.L wrote the raw manuscript. All authors contributed to data and texts in this manuscript.

Declaration of interests

The authors declare no competing interest.

References

1 Wolfe, A. J. & Brubaker, L. Urobiome updates: advances in urinary microbiome research. *Nat Rev Urol* **16**, 73-74 (2019). https://doi.org:10.1038/s41585-018-0127-5

2 Whiteside, S. A., Razvi, H., Dave, S., Reid, G. & Burton, J. P. The microbiome of the urinary tract—a role beyond infection. *Nature Reviews Urology* **12**, 81-90 (2015). https://doi.org:10.1038/nrurol.2014.361

3 Pearce, M. M. *et al.* The female urinary microbiome: a comparison of women with and without urgency urinary incontinence. *mBio* **5**, e01283-01214 (2014). https://doi.org:10.1128/mBio.01283-14

4 Ishihara, T. *et al.* Usefulness of next-generation DNA sequencing for the diagnosis of urinary tract infection. *Drug Discoveries & Therapeutics* **14**, 42-49 (2020). https://doi.org:10.5582/ddt.2020.01000

5 Price, T. K. *et al.* The Clinical Urine Culture: Enhanced Techniques Improve Detection of Clinically Relevant Microorganisms. *J Clin Microbiol* **54**, 1216-1222 (2016). https://doi.org:10.1128/jcm.00044-16

6 Thomas-White, K. *et al.* Culturing of female bladder bacteria reveals an interconnected urogenital microbiota. *Nature Communications* **9**, 1557 (2018). https://doi.org:10.1038/s41467-018-03968-5

7 Komesu, Y. M. *et al.* The urinary microbiome in women with mixed urinary incontinence compared to similarly aged controls. *International Urogynecology Journal* **29**, 1785-1795 (2018). https://doi.org:10.1007/s00192-018-3683-6

8 Govender, Y., Gabriel, I., Minassian, V. & Fichorova, R. The Current Evidence on the Association Between the Urinary Microbiome and Urinary Incontinence in Women. *Frontiers in Cellular and Infection Microbiology* **9** (2019).

https://doi.org:10.3389/fcimb.2019.00133

9 Moustafa, A. *et al.* Microbial metagenome of urinary tract infection. *Scientific Reports* **8** (2018). https://doi.org:10.1038/s41598-018-22660-8

10 Vaughan, M. H. *et al.* The Urinary Microbiome in Postmenopausal Women with Recurrent Urinary Tract Infections. *Journal of Urology* **206**, 1222-1231 (2021). https://doi.org:doi:10.1097/JU.0000000000001940

11 Kramer, H. *et al.* Diversity of the midstream urine microbiome in adults with chronic kidney disease. *International Urology and Nephrology* **50**, 1123-1130 (2018). https://doi.org:10.1007/s11255-018-1860-7

12 Wu, P. *et al.* Profiling the Urinary Microbiota in Male Patients With Bladder Cancer in China. *Front Cell Infect Microbiol* **8**, 167 (2018). https://doi.org:10.3389/fcimb.2018.00167

13 Qin, J. *et al.* Characterization of the Genitourinary Microbiome of 1,165 Middle-Aged and Elderly Healthy Individuals. *Front Microbiol* **12**, 673969 (2021). https://doi.org:10.3389/fmicb.2021.673969

14 Ammitzbøll, N. *et al.* Pre- and postmenopausal women have different core urinary microbiota. *Scientific Reports* **11**, 2212 (2021). https://doi.org:10.1038/s41598-021-81790-8

15 Adebayo, A. S. *et al.* The Urinary Tract Microbiome in Older Women Exhibits Host Genetic and Environmental Influences. *Cell Host Microbe* **28**, 298-305.e293 (2020). https://doi.org:10.1016/j.chom.2020.06.022

16 Liu, X. *et al.* A genome-wide association study for gut metagenome in Chinese adults illuminates complex diseases. *Cell Discovery* **7**, 9 (2021). https://doi.org:10.1038/s41421-020-00239-w

17 Liu, X. *et al.* Metagenome-genome-wide association studies reveal human genetic impact on the oral microbiome. *Cell Discovery* **7**, 117 (2021). https://doi.org:10.1038/s41421-021-00356-0

18 Liu, X. *et al.* Mendelian randomization analyses support causal relationships between blood metabolites and the gut microbiome. *Nature genetics* **54**, 52-61 (2022). https://doi.org:10.1038/s41588-021-00968-y

19 Jie, Z. *et al.* A transomic cohort as a reference point for promoting a healthy human gut microbiome. *Medicine in Microecology* **8**, 100039 (2021). https://doi.org:https://doi.org/10.1016/j.medmic.2021.100039

20 Zhu, J. *et al.* Over 50,000 Metagenomically Assembled Draft Genomes for the Human Oral Microbiome Reveal New Taxa. *Genomics, Proteomics & Bioinformatics* (2021). https://doi.org:https://doi.org/10.1016/j.gpb.2021.05.001

21 Chen, C. *et al.* Cervicovaginal microbiome dynamics after taking oral probiotics. *Journal of Genetics and Genomics* **48**, 716-726 (2021). https://doi.org:https://doi.org/10.1016/j.jgg.2021.03.019

22 Yang, Y. *et al.* Excavation of diagnostic biomarkers and construction of prognostic model for clear cell renal cell carcinoma based on urine proteomics. *Frontiers in Oncology* **13** (2023). https://doi.org:10.3389/fonc.2023.1170567

23 Onuchic, L. F. *et al.* PKHD1, the Polycystic Kidney and Hepatic Disease 1 Gene, Encodes a Novel Large Protein Containing Multiple Immunoglobulin-Like Plexin-

791 Transcription–Factor Domains and Parallel Beta-Helix 1 Repeats. *The American*
792 *Journal of Human Genetics* **70**, 1305-1317 (2002).
793 https://doi.org:https://doi.org/10.1086/340448
794 24 Hansen, C. G., Bright, N. A., Howard, G. & Nichols, B. J. SDPR induces membrane
795 curvature and functions in the formation of caveolae. *Nat Cell Biol* **11**, 807-814
796 (2009). https://doi.org:10.1038/ncb1887
797 25 Costa, V. L. *et al.* Three epigenetic biomarkers, GDF15, TMEFF2, and VIM,
798 accurately predict bladder cancer from DNA-based analyses of urine samples. *Clin*
799 *Cancer Res* **16**, 5842-5851 (2010). https://doi.org:10.1158/1078-0432.Ccr-10-1312
800 26 Adebayo, A. S. *et al.* The Urinary Tract Microbiome in Older Women Exhibits Host
801 Genetic and Environmental Influences. *Cell Host & Microbe* **28**, 298-305.e293 (2020).
802 https://doi.org:https://doi.org/10.1016/j.chom.2020.06.022
803 27 Komesu, Y. M. *et al.* Defining the relationship between vaginal and urinary
804 microbiomes. *American Journal of Obstetrics and Gynecology* **222**, 154.e151-
805 154.e110 (2020). https://doi.org:https://doi.org/10.1016/j.ajog.2019.08.011
806 28 Buhmann, M. T. *et al.* Encrustations on ureteral stents from patients without urinary
807 tract infection reveal distinct urotypes and a low bacterial load. *Microbiome* **7**, 60
808 (2019). https://doi.org:10.1186/s40168-019-0674-x
809 29 Siddiqui, H., Nederbragt, A. J., Lagesen, K., Jeansson, S. L. & Jakobsen, K. S.
810 Assessing diversity of the female urine microbiota by high throughput sequencing of
811 16S rDNA amplicons. *BMC Microbiol.* **11**, 244 (2011). https://doi.org:10.1186/1471-
812 2180-11-244
813 30 Fouts, D. E. *et al.* Integrated next-generation sequencing of 16S rDNA and
814 metaproteomics differentiate the healthy urine microbiome from asymptomatic
815 bacteriuria in neuropathic bladder associated with spinal cord injury. *J Transl Med* **10**,
816 174 (2012). https://doi.org:10.1186/1479-5876-10-174
817 31 Lewis, D. A. *et al.* The human urinary microbiome; bacterial DNA in voided urine of
818 asymptomatic adults. *Front Cell Infect Microbiol* **3**, 41 (2013).
819 https://doi.org:10.3389/fcimb.2013.00041
820 32 Gottschick, C. *et al.* The urinary microbiota of men and women and its changes in
821 women during bacterial vaginosis and antibiotic treatment. *Microbiome* **5**, 99 (2017).
822 https://doi.org:10.1186/s40168-017-0305-3
823 33 Bjornsson, T. D. Use of Serum Creatinine Concentrations to Determine Renal
824 Function1. *Clinical Pharmacokinetics* **4**, 200-222 (1979).
825 https://doi.org:10.2165/00003088-197904030-00003
826 34 Moitinho-Silva, L. *et al.* Host genetic factors related to innate immunity, environmental
827 sensing and cellular functions are associated with human skin microbiota. *Nat*
828 *Commun* **13**, 6204 (2022). https://doi.org:10.1038/s41467-022-33906-5
829 35 Cartwright, R. *et al.* Genome-Wide Association Study Identifies Two Novel Loci
830 Associated with Female Stress and Urgency Urinary Incontinence. *Journal of Urology*
831 **206**, 679-687 (2021). https://doi.org:doi:10.1097/JU.0000000000001822
832 36 Qin, J. J. *et al.* A metagenome-wide association study of gut microbiota in type 2
833 diabetes. *Nature* **490**, 55-60 (2012). https://doi.org:10.1038/nature11450
834 37 Lilan, H. & Lilan, H. DNA extraction for human microbe samples. *protocols.io* (2020).

38 Han, M. *et al.* A novel affordable reagent for room temperature storage and transport of fecal samples for metagenomic analyses. *Microbiome* **6**, 43 (2018). https://doi.org:10.1186/s40168-018-0429-0

39 Ackerman, A. L. *et al.* Optimization of DNA extraction from human urinary samples for mycobiome community profiling. *PLoS One* **14**, 26 (2019). https://doi.org:10.1371/journal.pone.0210306

40 Yang, F. M. *et al.* Assessment of fecal DNA extraction protocols for metagenomic studies. *GigaScience* **9**, 12 (2020). https://doi.org:10.1093/gigascience/giaa071

41 Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A. & Callahan, B. J. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome* **6**, 226 (2018). https://doi.org:10.1186/s40168-018-0605-2

42 Blanco-Miguez, A. *et al.* Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat. Biotechnol.* **41**, 1633-+ (2023). https://doi.org:10.1038/s41587-023-01688-w

43 Caliński, T. & Harabasz, J. A dendrite method for cluster analysis. *Communications in Statistics* **3**, 1-27 (1974). https://doi.org:10.1080/03610927408827101

44 Segata, N. *et al.* Metagenomic biomarker discovery and explanation. *Genome Biology* **12**, R60 (2011). https://doi.org:10.1186/gb-2011-12-6-r60

45 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009). https://doi.org:10.1093/bioinformatics/btp324

46 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009). https://doi.org:10.1093/bioinformatics/btp352

47 McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297-1303 (2010). https://doi.org:10.1101/gr.107524.110

48 Purcell, S. *et al.* PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559-575 (2007). https://doi.org:10.1086/519795

49 Ishigaki, K. *et al.* Large-scale genome-wide association study in a Japanese population identifies novel susceptibility loci across different diseases. *Nature Genet.* **52**, 669-+ (2020). https://doi.org:10.1038/s41588-020-0640-3

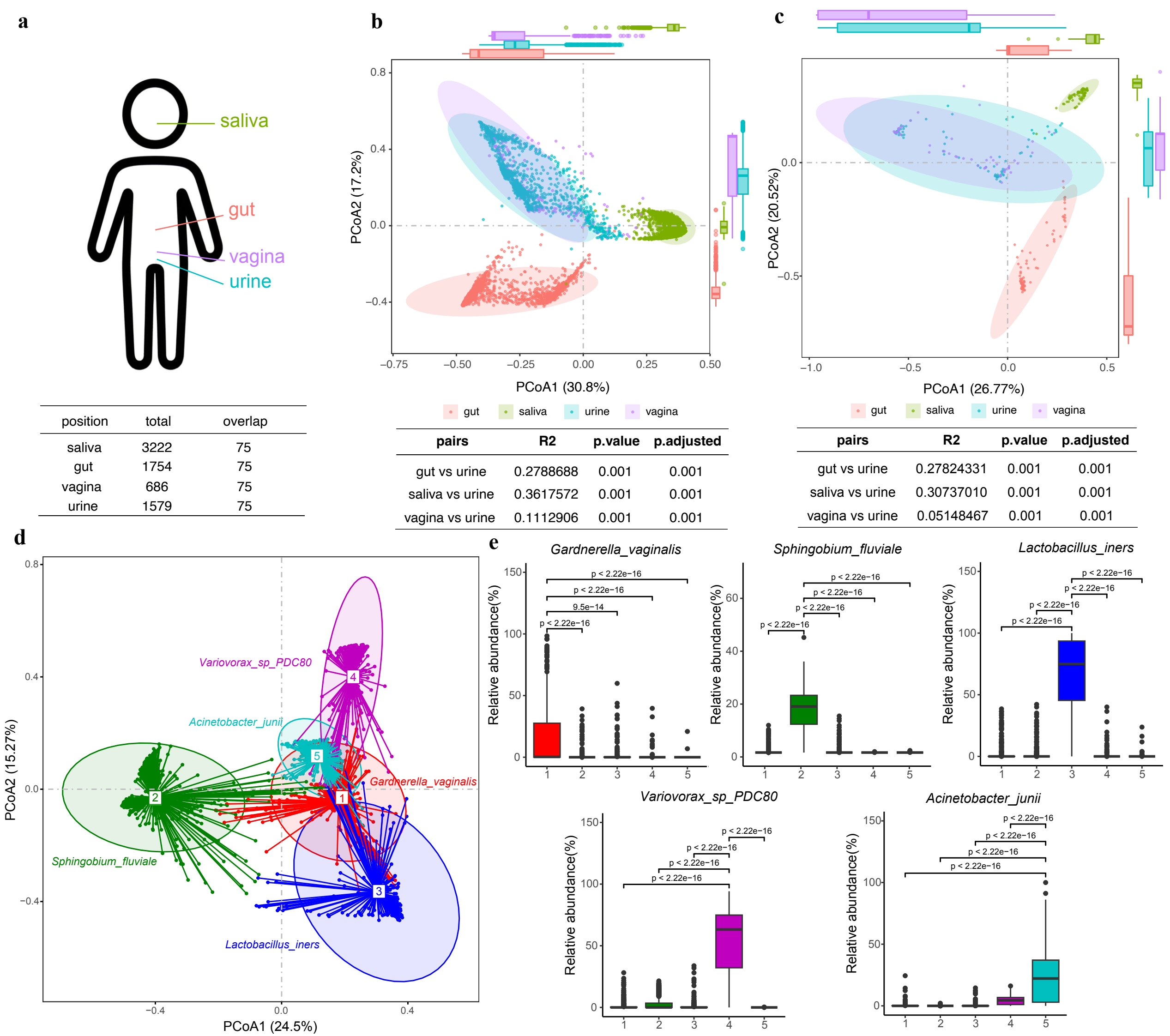
50 Kanai, M. *et al.* Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nature Genet.* **50**, 390-+ (2018). https://doi.org:10.1038/s41588-018-0047-6

51 Zhu, Z. *et al.* Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nature Communications* **9**, 224 (2018). https://doi.org:10.1038/s41467-017-02317-2

878 **Figure captions**
879
880
881 **Figure 1. Comparative analysis of microbial communities across human**
882 **body sites and characterizing the urinary microbiome.**
883 **a**) An illustrative diagram presenting the total and shared sample counts of
884 microbial communities from four body sites: saliva, gut, urine, and the
885 reproductive tract. **b**) A scatterplot illustrating the genus-level PCoA
886 distribution of the total microbial samples from saliva, gut, urine, and
887 reproductive tract, along with PERMANOVA pairwise difference values. **c**) A
888 scatterplot showing the genus-level PCoA distribution of 75 shared samples
889 among saliva, gut, urine, and reproductive tract, with PERMANOVA pairwise
890 difference values. **d**) A scatterplot reflecting the species-level PCoA
891 distribution of urinary microbes, the urotypes classification, and the dominant
892 bacteria present in each urotype. **e**) Boxplots displaying the average relative
893 abundance of the five dominant bacteria among individuals with five different
894 urotypes, along with t-test differential analysis.
895
896 **Figure 2. Impact of host phenotypes on urinary microbiota.**
897 This scatterplot presents the distribution of Pearson correlation estimate
898 values between host phenotypes and five urotypes, highlighting the top 20
899 significant associations for each urotype, with colors corresponding to the
900 urine classification results in Figure 1d. Each phenotype is ranked from high
901 to low according to the average of Pearson correlation estimate value for each
902 urotype. **b**) A bar chart showing the top 40 significant host phenotypes sorted
903 by their explained variance (R^2 values) in the β-diversity of urinary microbiota.
904 **c**) A categorized bar chart displaying the explained variance of significant host
905 phenotypic factors for the urinary microbiota β-diversity, including: 137 total
906 host factors, 104 blood factors, 4 urinary factors, gender, 10 anthropometric
907 factors, 6 health factors, 7 lifestyle factors, and 5 diet-related factors.
908
909 **Figure 3. The urinary microbiome differed between genders and**
910 **effectively distinguished gender.**
911 **a)** Box plot of α-diversity (Shannon index) of the urinary microbiome between
912 males and females. **b)** β-diversity of the urinary microbiome using principal
913 coordinate analysis, as colored by males (cyan) and females (red). The first
914 and second axes, PCoA1 and PCoA2, explaining the highest variance, are
915 shown. Dashed ellipses represent the 95% confidence level of permutational
916 multivariate analysis of variance (PERMANOVA) test. Explained variance (R^2)
917 and p-value were also been showed. **c**) The sex differential urinary
918 microbiota, as sorted by LDA scores in the LEfSe analysis**.** **d**) A ROC curve
919 demonstrating the predictive power of the urinary microbiota for gender using
920 a random forest model. The analysis comprises 520 male samples and 1044
921 female samples.

Figure 4. Influences of host genetics and their interactions with gender/sex hormones on urinary microbiota.

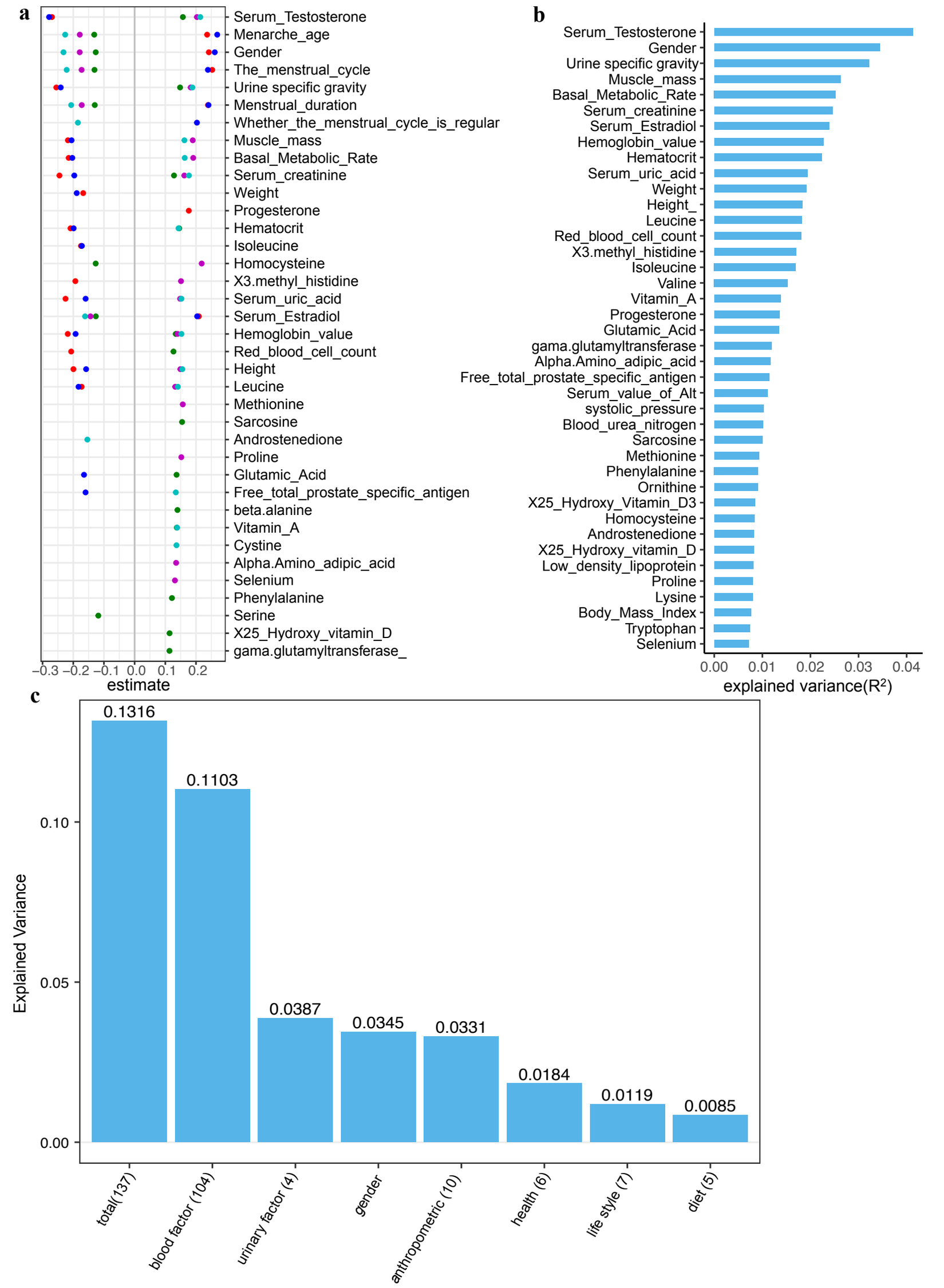
a) A Manhattan plot showing the genetic variants associated with urine microbial taxa (n=106), with genome-wide significance denoted by horizontal black lines ($p = 5 \times 10^{-8}$) and gene names annotated for 27 significant loci. **b**) A bar chart displaying the urinary microbes associated with the significant 27 loci and the number of associations. **c**) Boxplots illustrating the average relative abundance of genus *Acinetobacter* among individuals with different genotypes (AA vs GA/GG) at the top locus rs139799338, as well as stratified by male and female samples. **d**) A scatterplot depicting the association between serum testosterone and the genus *Acinetobacter* relative abundance among individuals with different genotypes (AA vs GA/GG) at the top locus rs139799338. Spearman correlation's R and p values were also listed. **e**) Boxplots showing the average relative abundance of genus *Acinetobacter* among individuals with different genotypes (AA vs GA/GG) at the top locus rs4241281, as well as stratified by male and female samples. **f**) A scatterplot illustrating the association between serum testosterone and the genus *Acinetobacter* relative abundance among individuals with different genotypes (AA vs GA/GG) at the top locus rs4241281. Spearman correlation's R and p values were also listed.

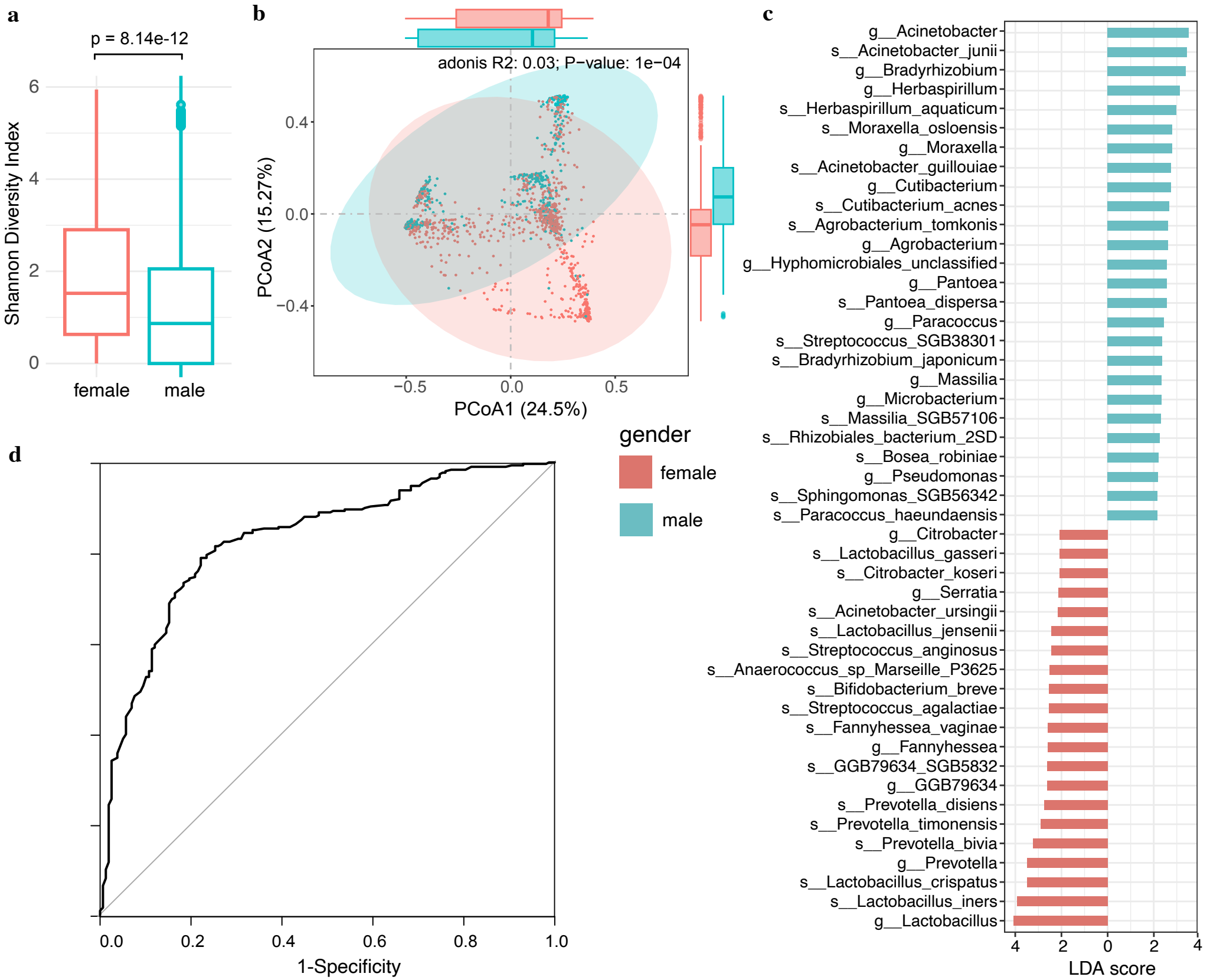


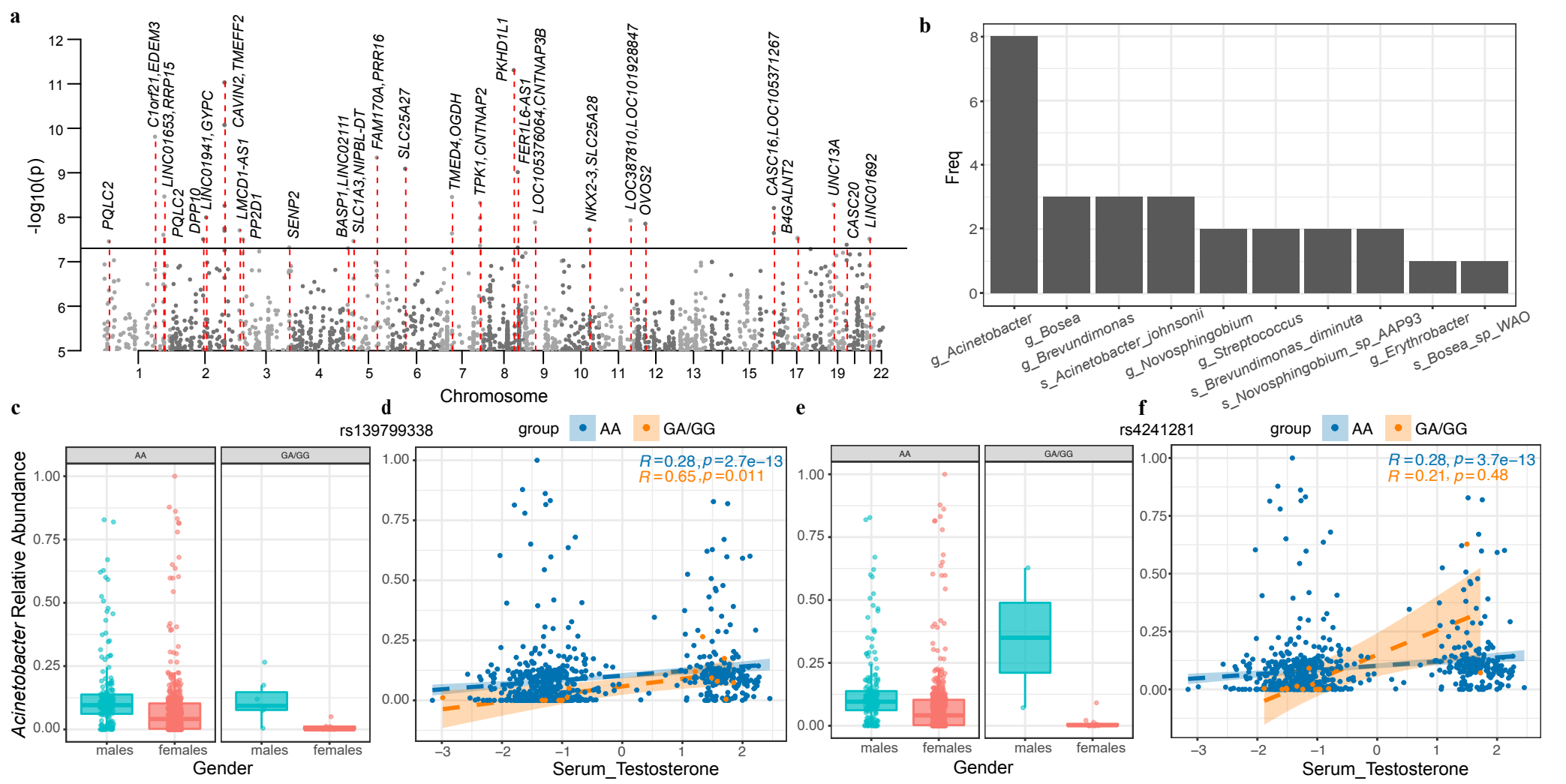
position	total	overlap
saliva	3222	75
gut	1754	75
vagina	686	75
urine	1579	75

pairs	R2	p.value	p.adjusted
gut vs urine	0.2788688	0.001	0.001
saliva vs urine	0.3617572	0.001	0.001
vagina vs urine	0.1112906	0.001	0.001

pairs	R2	p.value	p.adjusted
gut vs urine	0.27824331	0.001	0.001
saliva vs urine	0.30737010	0.001	0.001
vagina vs urine	0.05148467	0.001	0.001


a
Serum_Testosterone
Menarche_age
Gender
The_menstrual_cycle
Urine specific gravity
Menstrual_duration
Whether_the_menstrual_cycle_is_regular
Muscle_mass
Basal_Metabolic_Rate
Serum_creatinine
Weight
Progesterone
Hematocrit
Isoleucine
Homocysteine
X3.methyl_histidine
Serum_uric_acid
Serum_Estradiol
Hemoglobin_value
Red_blood_cell_count
Height
Leucine
Methionine
Sarcosine
Androstenedione
Proline
Glutamic_Acid
Free_total_prostate_specific_antigen
beta.alanine
Vitamin_A
Cystine
Alpha.Amino_adipic_acid
Selenium
Phenylalanine
Serine
X25_Hydroxy_vitamin_D
gama.glutamyltransferase_
−0.3 −0.2 −0.1 0.0 0.1 0.2
estimate
b
Serum_Testosterone
Gender
Urine specific gravity
Muscle_mass
Basal_Metabolic_Rate
Serum_creatinine
Serum_Estradiol
Hemoglobin_value
Hematocrit
Serum_uric_acid
Weight
Height_
Leucine
Red_blood_cell_count
X3.methyl_histidine
Isoleucine
Valine
Vitamin_A
Progesterone
Glutamic_Acid
gama.glutamyltransferase
Alpha.Amino_adipic_acid
Free_total_prostate_specific_antigen
Serum_value_of_Alt
systolic_pressure
Blood_urea_nitrogen
Sarcosine
Methionine
Phenylalanine
Ornithine
X25_Hydroxy_Vitamin_D3
Homocysteine
Androstenedione
X25_Hydroxy_vitamin_D
Low_density_lipoprotein
Proline
Lysine
Body_Mass_Index
Tryptophan
Selenium
0.00 0.01 0.02 0.03 0.04
explained variance(R²)
c
Explained Variance
0.00
0.05
0.10
0.1316
0.1103
0.0387
0.0345
0.0331
0.0184
0.0119
0.0085
total(137)
blood factor (104)
urinary factor (4)
gender
anthropometric (10)
health (6)
life style (7)
diet (5)


a
p = 8.14e-12
Shannon Diversity Index
female
male
b
adonis R2: 0.03; P−value: 1e−04
PCoA2 (15.27%)
PCoA1 (24.5%)
gender
female
male
c
g__Acinetobacter
s__Acinetobacter_junii
g__Bradyrhizobium
g__Herbaspirillum
s__Herbaspirillum_aquaticum
s__Moraxella_osloensis
g__Moraxella
s__Acinetobacter_guillouiae
g__Cutibacterium
s__Cutibacterium_acnes
s__Agrobacterium_tomkonis
g__Agrobacterium
g__Hyphomicrobiales_unclassified
g__Pantoea
s__Pantoea_dispersa
g__Paracoccus
s__Streptococcus_SGB38301
s__Bradyrhizobium_japonicum
g__Massilia
g__Microbacterium
s__Massilia_SGB57106
s__Rhizobiales_bacterium_2SD
s__Bosea_robiniae
g__Pseudomonas
s__Sphingomonas_SGB56342
s__Paracoccus_haeundaensis
g__Citrobacter
s__Lactobacillus_gasseri
s__Citrobacter_koseri
g__Serratia
s__Acinetobacter_ursingii
s__Lactobacillus_jensenii
s__Streptococcus_anginosus
s__Anaerococcus_sp_Marseille_P3625
s__Bifidobacterium_breve
s__Streptococcus_agalactiae
s__Fannyhessea_vaginae
g__Fannyhessea
s__GGB79634_SGB5832
g__GGB79634
s__Prevotella_disiens
s__Prevotella_timonensis
s__Prevotella_bivia
g__Prevotella
s__Lactobacillus_crispatus
s__Lactobacillus_iners
g__Lactobacillus
LDA score
d
1-Specificity


a
-log10(p)
Chromosome
PQLC2
C1orf21,EDEM3
LINC01653,RRP15
PQLC2
DPP10
LINC01941,GYPC
CAVIN2,TMEFF2
LMCD1-AS1
PP2D1
SENP2
BASP1,LINC02111
SLC1A3,NIPBL-DT
FAM170A,PRR16
SLC25A27
TMED4,OGDH
TPK1,CNTNAP2
PKHD1L1
FER1L6-AS1
LOC105376064,CNTNAP3B
NKX2-3,SLC25A28
LOC387810,LOC101928847
OVOS2
CASC16,LOC105371267
B4GALNT2
UNC13A
CASC20
LINC01692
b
Freq
g_Acinetobacter
g_Bosea
g_Brevundimonas
s_Acinetobacter_johnsonii
g_Novosphingobium
g_Streptococcus
s_Brevundimonas_diminuta
s_Novosphingobium_sp_AAP93
g_Erythrobacter
s_Bosea_sp_WAO
c
Acinetobacter Relative Abundance
AA
GA/GG
males
females
Gender
d
rs139799338
group
AA
GA/GG
R=0.28, p=2.7e−13
R=0.65, p=0.011
Serum_Testosterone
e
AA
GA/GG
males
females
Gender
f
rs4241281
group
AA
GA/GG
R=0.28, p=3.7e−13
R=0.21, p=0.48
Serum_Testosterone