


SHORT REPORT

Open Access

# Separating the signal from the noise in metagenomic cell-free DNA sequencing



Philip Burnham<sup>1</sup>, Nardhy Gomez-Lopez<sup>2,3,4</sup>, Michael Heyang<sup>1</sup>, Alexandre Pellan Cheng<sup>1</sup>, Joan Sesing Lenz<sup>1</sup>, Darshana M. Dadhania<sup>5</sup>, John Richard Lee<sup>5</sup>, Manikkam Suthanthiran<sup>5</sup>, Roberto Romero<sup>2,6,7,8,9,10</sup> and Iwijn De Vlaminc<sup>1\*</sup> 

## Abstract

**Background:** Cell-free DNA (cfDNA) in blood, urine, and other biofluids provides a unique window into human health. A proportion of cfDNA is derived from bacteria and viruses, creating opportunities for the diagnosis of infection via metagenomic sequencing. The total biomass of microbial-derived cfDNA in clinical isolates is low, which makes metagenomic cfDNA sequencing susceptible to contamination and alignment noise.

**Results:** Here, we report low biomass background correction (LBBC), a bioinformatics noise filtering tool informed by the uniformity of the coverage of microbial genomes and the batch variation in the absolute abundance of microbial cfDNA. We demonstrate that LBBC leads to a dramatic reduction in false positive rate while minimally affecting the true positive rate for a cfDNA test to screen for urinary tract infection. We next performed high-throughput sequencing of cfDNA in amniotic fluid collected from term uncomplicated pregnancies or those complicated with clinical chorioamnionitis with and without intra-amniotic infection.

**Conclusions:** The data provide unique insight into the properties of fetal and maternal cfDNA in amniotic fluid, demonstrate the utility of cfDNA to screen for intra-amniotic infection, support the view that the amniotic fluid is sterile during normal pregnancy, and reveal cases of intra-amniotic inflammation without infection at term.

**Keywords:** Cell-free DNA, Metagenomics, Biomarkers, Infectious disease, Prenatal health

## Background

Metagenomic sequencing of cell-free DNA (cfDNA) offers a highly sensitive approach to screen for pathogens in clinical samples [1–4]. The sensitivity of metagenomic sequencing of cfDNA in plasma can be boosted by the implementation of library preparations optimized to recover short, degraded microbial cfDNA [5], or by strategies that selectively enrich microbial DNA or deplete host DNA [6–8]. A major remaining challenge is the relatively poor specificity of cfDNA metagenomic sequencing, which is limited by alignment noise, annotation errors in reference genomes, and environmental contamination [9].

Here, we report low biomass background correction (LBBC), a tool to filter background contamination and noise in cfDNA metagenomic sequencing datasets. We have

applied LBBC to two independent datasets. We first re-analyzed a dataset from a previous study that investigated the utility of urinary cfDNA as an analyte to monitor urinary tract infection (UTI) [2]. Next, we generated a new dataset of cfDNA in amniotic fluid collected from uncomplicated pregnancies or those complicated with clinical chorioamnionitis at term, a common heterogeneous condition that can occur in the presence or absence of intra-amniotic infection [10]. We report a first, detailed study of the properties of cfDNA in amniotic fluid. For both datasets, detailed microbiologic workups, including results from conventional bacterial culture and/or PCR, were available to benchmark the LBBC workflow. We demonstrate that LBBC greatly improves the specificity of cfDNA metagenomic sequencing, while minimally affecting its sensitivity.

## Results

To extract sequence information from cfDNA isolates, we used a single-stranded DNA library preparation that

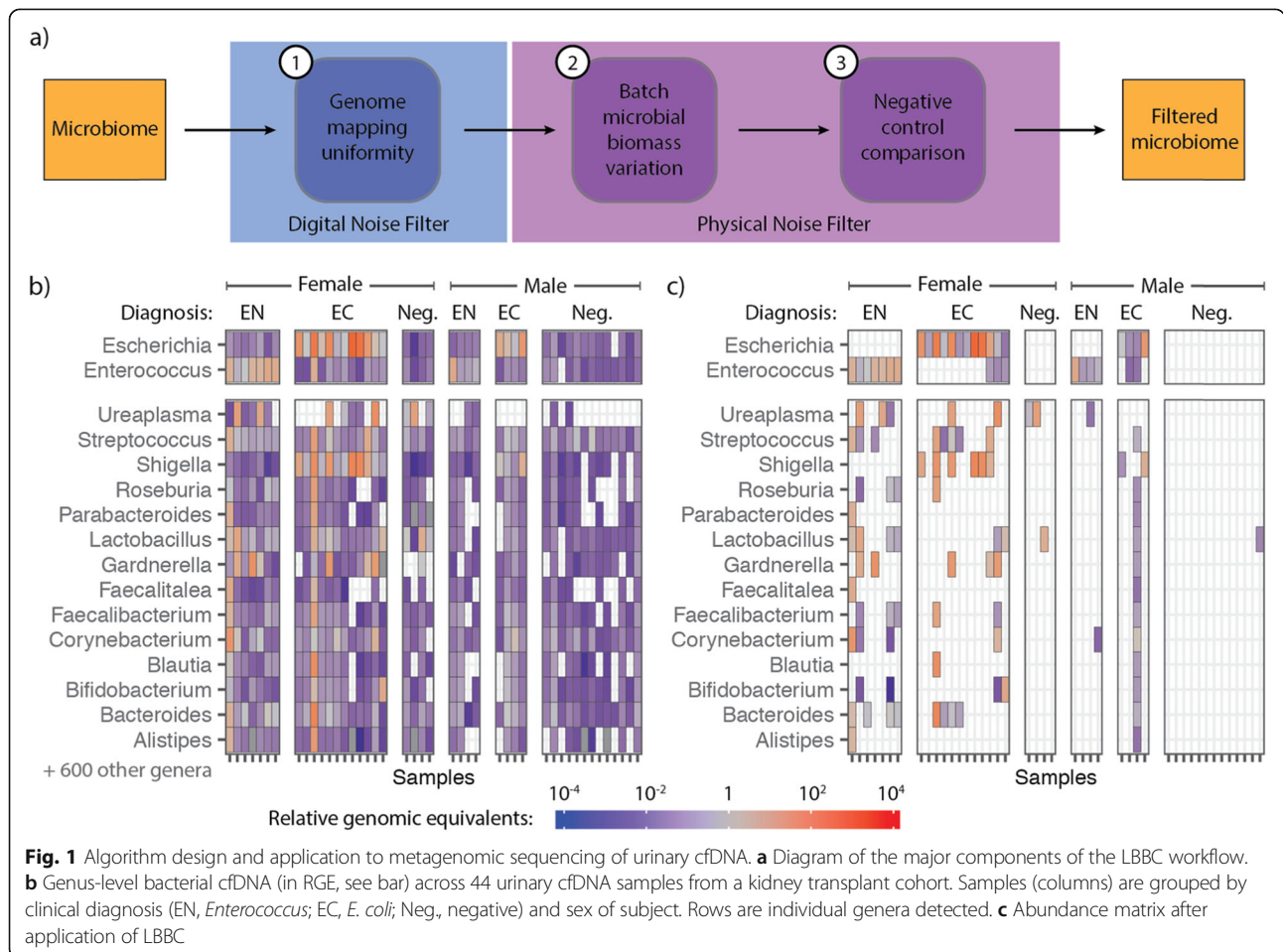
\* Correspondence: [vlaminc@cornell.edu](mailto:vlaminc@cornell.edu)

<sup>1</sup>Meinig School of Biomedical Engineering, Cornell University, Ithaca, NY, USA  
Full list of author information is available at the end of the article



improves the recovery of microbial cfDNA relative to host cfDNA by up to 70-fold for cfDNA in plasma [5]. We quantified microbial cfDNA by alignment of sequences to microbial reference genomes [11, 12] (see the “Methods” section). We identified two classes of noise, which we addressed using a bioinformatics workflow that implements both novel and previously described filtering approaches [13, 14] (Fig. 1a). The first type of noise can be classified as “digital crosstalk” and stems from errors in alignment and contaminant sequences that are present in microbial reference genomes, including human-related sequences or sequences from other microbes. Digital crosstalk affects distinct segments of a microbial genome and gives rise to inhomogeneous coverage of the reference genome. We computed the coefficient of variation in the per-base genome coverage for all identified species (CV, computed as the standard deviation in genome coverage divided by the mean coverage) and removed taxa for which the CV differed greatly from the CV determined for a uniformly sampled genome of the same size (see the “Methods” section), because this indicated that a significant number of sequences assigned to the genome are due to digital crosstalk.

A second class of noise is due to physical contamination of the sample with environmental DNA present at the time of collection and in reagents used for DNA isolation and sequencing library preparation [13]. We reasoned that the total biomass of environmental DNA would be consistent for samples prepared in the same batch. LBBC filters environmental contaminants by performing batch variation analysis on the absolute abundance of microbial DNA quantified with high accuracy. The core elements of LBBC can be implemented using any metagenomics abundance estimation algorithm which makes use of sequence alignment to full microbial genomes. In our analysis, we estimate the genomic abundance of each species using a maximum likelihood model implemented in GRAMMy [12] (see the “Methods” section). GRAMMy helps ameliorate the impact of closely related genomes [12]. From the relative abundance of species, we compute the absolute number of molecules in a dataset corresponding to a specific species, considering differences in genome sizes for all identified microbes. The total biomass of microbial DNA is then estimated as the proportion of sequencing reads derived from a species, multiplied by the measured biomass



**Fig. 1** Algorithm design and application to metagenomic sequencing of urinary cfDNA. **a** Diagram of the major components of the LBBC workflow. **b** Genus-level bacterial cfDNA (in RGE, see bar) across 44 urinary cfDNA samples from a kidney transplant cohort. Samples (columns) are grouped by clinical diagnosis (EN, *Enterococcus*; EC, *E. coli*; Neg., negative) and sex of subject. Rows are individual genera detected. **c** Abundance matrix after application of LBBC

inputted in the library preparation reaction. Recent approaches have identified environmental contaminants by (i) looking for batch-by-batch covariation in the relative abundance of microbes measured by metagenomic sequencing or (ii) examining the (inverse) correlation between biomass of the sample and the relative abundance of microbial DNA in the sample [13, 14]. These studies have shown the dramatic effect of environmental contamination in low biomass settings. LBBC effectively combines these two prior approaches into one. Using this analysis applied to the metagenomic cfDNA datasets described below, we estimate that the total biomass of environmental, contaminant DNA can exceed 100 pg (range of 0 to 230.4 pg). This is a small amount of DNA (<1% of sequencing reads) that nonetheless can significantly impact the interpretation of metagenomic sequencing results. We further incorporated a known-template, negative control in the library preparation procedures to identify any remaining contaminant sequences. The use of a negative control is recommended for metagenomics studies [9] and was implemented in our previous work [2, 15]. Here, we compared the microbial abundance detected in samples to those in controls to set a baseline for environmental contamination. This analysis indicated that, on average, only 46% of physical contaminant species determined by LBBC are removed using comparison to a negative control alone, supporting the need for the additional filters implemented in LBBC.

We evaluated and optimized LBBC using a dataset available from a recently published study that assessed the utility of urinary cfDNA for the monitoring of bacterial infection of the urinary tract [2]. We analyzed 44 cfDNA datasets from male and female kidney recipients. These included 16 datasets from subjects with *E. coli* UTI, 11 datasets from subjects with *Enterococcus* UTI, and 17 datasets from subjects without UTI, as determined by conventional urine culture performed on the same day. Prior to application of the LBBC algorithm, the ratio of sequences assigned as non-host vs host (paired host reads relative to sequences assigned to microbial taxa) was  $4.4 \times 10^{-1} \pm 1.68$  in this dataset. We detected 616 bacterial genera across all 44 samples (Fig. 1b; RGE >  $10^{-6}$ ), many of which were atypical in the urinary tract, including *Hermiimonas* and *Methylobacterium*, albeit at very low abundance.

We defined two parameters for threshold-based filtering; these are (1) the maximum difference in the observed CV and that of a uniformly sequenced taxon for the same sequencing depth and genome size,  $\Delta CV_{\max}$ , and (2) the minimum allowable within-batch variation,  $\sigma_{\min}^2$ . A third, fixed parameter was used to remove species identified in the negative controls (threshold 10-fold the observed representation in the negative controls). We optimized these parameters based on following metric:

$$BC_{\text{score}} = k_{\text{TP}}(\text{TP}) + k_{\text{TN}}(\text{TN}) + k_{\text{FP}}(\text{FP}) + k_{\text{FN}}(\text{FN}) + k_U(U),$$

where {TP, TN, FP, FN} is the number of true positives, true negatives, false positives, and false negatives, respectively,  $U$  is the total number of identified taxa for which an orthogonal measurement was not performed, and the coefficients  $k$  for these values represent weights to optimize the filtering parameters. Here, we chose  $\{k_{\text{TP}}, k_{\text{TN}}, k_{\text{FP}}, k_{\text{FN}}, k_U\} = \{4, 2, -1, -2, -0.2\}$  and used nonlinear minimization by gradient descent on the variable  $BC_{\text{score}}$  to determine an optimal set of threshold parameters:  $\{\Delta CV_{\max}, \sigma_{\min}^2\} = \{2.00, 3.16 \text{ pg}^2\}$ .

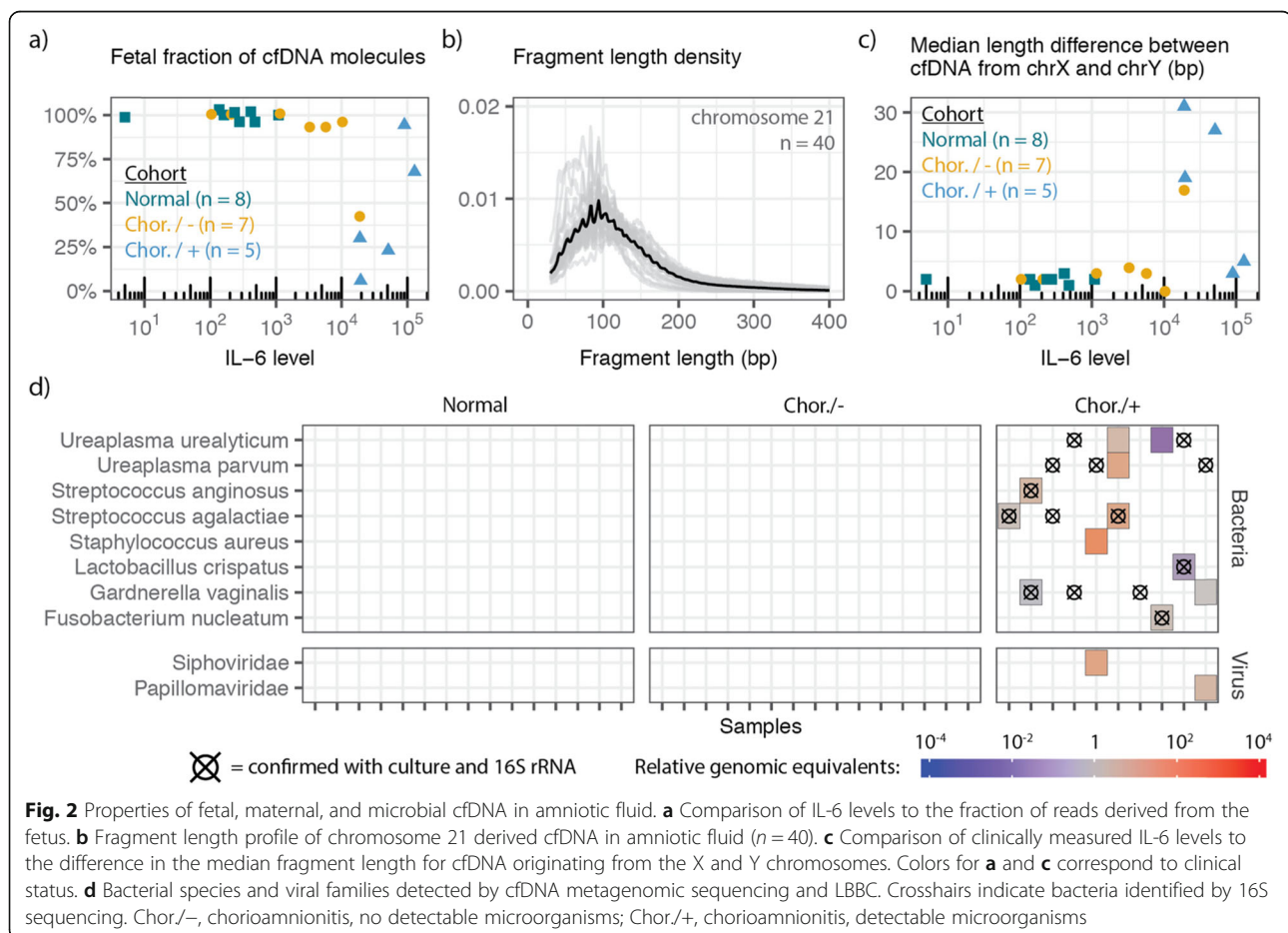
Applying LBBC with these parameters to urinary cfDNA microbiome profiles led to a diagnostic sensitivity of 100% and specificity of 91.8%, when analyzed against results from conventional urine culture. We computed a confusion matrix (see the “Methods” section) and determined the accuracy of the test to be 0.886 (no information rate, NIR = 0.386,  $p < 10^{-10}$ ). Without LBBC, the test achieved a sensitivity of 100% but a specificity of 3.3%, and an accuracy of 0.000 (as most samples have both *E. coli* and *Enterococcus*). Applying a simple filter that excludes taxa with relative abundance below a pre-defined threshold (RGE > 0.1) led to an accuracy of 0.864 (sensitivity of 81.5%, specificity of 96.7%); however, such filtering does not remove sources of physical or digital noise at high abundance and may remove pathogens present at low abundance. After applying LBBC, we observed far fewer bacterial genera outside of *Escherichia* and *Enterococcus* in samples from patients diagnosed with UTI (Fig. 1c). LBBC did not remove bacteria that are known to be commensal in the female genitourinary tract, including species from the genera *Gardnerella* and *Ureaplasma* [16]. For male subjects without UTI, we detected a single *Lactobacillus* species among all subjects, consistent with the view that the male urinary tract is sterile in the absence of infection. For patients with UTI, the urinary microbiomes were less diverse in males compared with females, as previously reported [17]. These examples illustrate that LBBC conserves key relationships between pathogenic and non-pathogenic bacteria.

We next applied LBBC to the analysis of cfDNA in amniotic fluid. Circulating cfDNA in maternal plasma has emerged as a highly valuable analyte for the screening of aneuploidy in pregnancy [18], but no studies have examined the properties of cfDNA in amniotic fluid. No studies have furthermore assessed the utility of amniotic fluid cfDNA as an analyte to monitor clinical chorioamnionitis, the most common diagnosis related to infection made in labor and delivery units worldwide [19]. Traditionally, it was thought that clinical chorioamnionitis was due to microbial invasion of the amniotic cavity (i.e., intra-amniotic infection), which elicits a maternal inflammatory response characterized by maternal fever,

uterine tenderness, tachycardia, and leukocytosis as well as fetal tachycardia and a foul-smelling amniotic fluid [20, 21]. However, recent studies in which amniocentesis has been used to characterize the microbiologic state of the amniotic cavity and the inflammatory response [amniotic fluid interleukin (IL)-6 > 2.6 ng/ml [22]] show that only 60% of patients with the diagnosis of clinical chorioamnionitis have proven infection using culture or molecular microbiologic techniques [10]. The remainder of the patients has clinical chorioamnionitis in the presence of intra-amniotic inflammation (i.e., sterile intra-amniotic inflammation) or without neither intra-amniotic inflammation nor microorganisms in the amniotic cavity [10]. Therefore, the emergent picture is that clinical chorioamnionitis at term is a heterogeneous syndrome, which requires further study to optimize maternal and neonatal outcomes [23]. We analyzed 40 amniotic cfDNA isolates collected from the following study groups of women: (1) with clinical chorioamnionitis and detectable microorganisms ( $n = 10$ ), (2) with clinical chorioamnionitis without detectable microorganisms ( $n = 15$ ), and 93 without clinical chorioamnionitis (i.e., normal full-term pregnancies) ( $n = 15$ ). Microorganisms were detected by cultivation and

broad-range PCR coupled with electrospray ionization mass spectrometry or PCR/ESI-MS (see the “Methods” section). Data from several independent clinical assays were available, including levels of interleukin 6 (IL-6), white and red blood cell counts, and glucose levels (see the “Methods” section).

We obtained  $77.7 \pm 31.8$  million paired-end reads per sample, yielding a per-base human genome coverage of  $1.90 \times \pm 0.88 \times$ . The data provide unique insight into the properties of amniotic fluid cfDNA. For women carrying a male fetus, we used the coverage of the Y chromosome relative to autosomes to estimate the fetal fraction of cfDNA in amniotic fluid (see the “Methods” section). The fetal fraction ranged from 6.0 to 100% and was strongly anticorrelated with inflammatory markers such as IL-6 [24, 25] (Spearman’s rho of  $-0.763$ ,  $p = 1.34 \times 10^{-4}$ ,  $n = 20$ ; Fig. 2a). We attribute this observation to the recruitment of immune cells to the amniotic cavity during infection [26, 27]. We next used paired-end read mapping to determine the fragment length profiles of cfDNA in amniotic fluid (Fig. 2b). We found that amniotic fluid cfDNA was highly fragmented (median length 108 bp) and lacked the canonical peak at 167 bp typically





observed in the fragmentation profile of plasma cfDNA [18, 28]. To determine size differences between fetal and maternal cfDNA in amniotic fluid, we computed the median fragment length for molecules derived from the X and Y chromosomes in cfDNA from male pregnancy samples. We hypothesized that if all cfDNA in a sample originated from the male fetus, the median fragment lengths for the X- and Y-aligned DNA would be equivalent, and, conversely, in samples with a large fraction of cfDNA originating from the mother, a length discrepancy may arise. Using this approach, we found that fetal-derived cfDNA was shorter than maternal-derived cfDNA (up to 31 bp shorter; Fig. 2c). Previous reports have similarly noted that fetal cfDNA in urine and plasma is shorter than maternal cfDNA [29, 30].

We next examined the utility of LBBC for the diagnosis of clinical chorioamnionitis. Prior to the application of the LBBC algorithm, the ratio of sequences assigned as non-host vs host (paired host reads relative to sequences assigned to microbial taxa) was  $1.08 \times 10^{-2} \pm 4.76 \times 10^{-2}$  in this dataset. After applying LBBC with a relaxed batch variation minimum to account for species-level analysis ( $\sigma_{\min}^2 = 1 \text{ pg}^2$ ), no bacteria were detected in the normal pregnancy group (Fig. 2d), in line with recent studies that point to a sterile amniotic cavity and placenta in the absence of infection [31, 32]. The cfDNA sequencing assay detected only 6 of the 14 bacterial genera identified by bacterial culture or PCR/ESI-MS, and was unable to identify a fungal pathogen, *Candida albicans*, detected by PCR/ESI-MS (see the “Methods” section). We asked if these false negatives were due to LBBC filtering. Relaxation of the filtering thresholds revealed that *Ureaplasma* was removed in four samples by the batch variation filter; other false negatives were not due to LBBC filtering. Interestingly, in all cases of chorioamnionitis without detectable microorganisms, no bacterium was identified (Fig. 2d), in line with previous evidence showing that chorioamnionitis and intra-amniotic inflammation can occur in the absence of microbial invasion of the amniotic cavity [10]. Last, in two samples, we identified a high burden of viral DNA, including papillomavirus in one sample and bacteriophage in another (Fig. 2d), demonstrating the utility of cfDNA paired with LBBC to detect viruses in the amniotic fluid.

## Discussion

cfDNA metagenomic sequencing is emerging as a powerful approach to screen for infection [3]. The technique has inherent high sensitivity, but lower specificity. Here, we described LBBC, a simple computational workflow to filter background contamination and noise in cfDNA metagenomic sequencing datasets. LBBC analyzes batch effects, the uniformity of the genome coverage and the relationship between microbial abundance

and total biomass of the sample to identify and filter noise contributions. Though batch effects can arise at any step, we found some steps are more prone to contamination and hence batch effects, in particular the cfDNA extraction batch, while others had very little effect, e.g., the sequencing instrument. Other possible batch effects include the date of processing (influencing reagent batch) and location where samples were prepared (e.g., in a clean room or in a lab environment with multiple experiments being performed); the sources of contamination in metagenomic sequencing and batch effects have been reviewed recently [9, 33].

The three filtering steps implemented in LLBC are appropriate for the analysis of any low-biomass sample, not limited to cfDNA isolates, and can be readily implemented, in a modular fashion, provided that (1) the total DNA biomass going into the sample preparation is measured and recorded, (2) batch information is available, and (3) the microbial abundance is determined by a sequence-based alignment method [12]. This last point is of importance, because of the several popular algorithms for metagenomic sequence classification, including Metaphlan, which relies on reduced reference genomes [34]. Such approaches preclude the ability to estimate sequencing coverage uniformity across the genome, required for the CV filter that is part of LBBC [12]. To our knowledge, LBBC is the first filtering scheme to analyze sequencing coverage heterogeneity across thousands of microbial genomes and filter results based on the coefficient of variation in sequence coverage.

## Conclusions

We have described LBBC, a bioinformatics noise filtering tool informed by the uniformity of the coverage of microbial genomes and the batch variation in the absolute abundance of microbial cfDNA. We applied LBBC to a recently published urinary cfDNA dataset. Comparison to clinical testing showed that LBBC greatly improves the specificity of metagenomic cfDNA sequencing while minimally affecting the sensitivity of the assay (Fig. 1). We next applied LBBC to a novel dataset of cfDNA from the amniotic fluid of subjects with and without clinical chorioamnionitis. This dataset allowed us to characterize the properties of maternal and fetal DNA in the amniotic sac for the first time (Fig. 2). While LBBC greatly reduces the noise in metagenomic sequencing, some technical challenges, inherent to metagenomic read assignments, remain. For example, some reads, originating from a source microbe, can incorrectly align to taxa with a highly similar genome; LBBC reduces the frequency of erroneous read assignments, but it does not completely remove these reads.

The application of LBBC to a new dataset of cfDNA in amniotic fluid revealed a bacteria-free environment in healthy full-term pregnancies and in a subset of patients

with clinical chorioamnionitis and intra-amniotic inflammation as well as in the presence of pathogenic bacteria in many cases of clinical chorioamnionitis with intra-amniotic infection and inflammation. In addition, few microbial taxa were identified in cases of chorioamnionitis with no detectable bacteria via culture or PCR/ESI-MS. In summary, metagenomic cfDNA sequencing, complemented with a background reduction workflow, enables identification of potential pathogens in clinical samples with both high sensitivity and specificity.

## Methods

### Sample description—urinary cfDNA

Forty-four sample datasets were selected from a recent study [2]. Urine samples were collected under an Institution Review Board protocol that was approved at Weill Cornell Medicine. All subjects provided written informed consent. Datasets were selected from the study from one of two groups: (1) UTI—those corresponding to a same-day positive urine culture ( $>10,000$  CFU/mL) indicating monomicrobial *E. coli*, *Enterococcus faecium*, or *Enterococcus faecalis* UTI. A single sample from the original study [2] (GU14) was excluded due to the high likelihood that it was *R. ornithinolytica* infection incorrectly diagnosed as an *E. coli* UTI. (2) No UTI—samples from patients with same-day negative standard urine culture and no microorganisms detected at earlier or later dates. Sample metadata is included in Additional file 1.

### Sample description—amniotic fluid cfDNA

Forty samples were collected from a cohort of subjects with full-term pregnancy, which were uncomplicated ( $n = 15$ ), or burdened with clinical chorioamnionitis with detectable microorganisms ( $n = 10$ ) or clinical chorioamnionitis without detectable microorganisms ( $n = 15$ ). Amniotic fluid samples were obtained by transabdominal amniocentesis performed for evaluation of the microbial and inflammatory status of the amniotic cavity in patients with clinical chorioamnionitis, whereas women approaching term underwent an amniocentesis for assessment of fetal lung maturity. Twenty of the 40 samples were from mothers pregnant with male fetus. Clinical chorioamnionitis was diagnosed by the presence of maternal fever (temperature  $>37.8^{\circ}\text{C}$ ) accompanied by two or more of the following criteria: (1) uterine tenderness, (2) foul-smelling amniotic fluid, (3) fetal tachycardia (heart rate  $>160$  beats/min), (4) maternal tachycardia (heart rate  $>100$  beats/min), and (5) maternal leukocytosis (leukocyte count  $>15,000$  cells/ $\text{mm}^3$ ) [20, 24]. Amniotic fluid samples were transported to the clinical laboratory in a sterile capped syringe and cultured for aerobic and anaerobic bacteria, including genital Mycoplasmas. The clinical tests also included the determination of amniotic fluid white blood cell (WBC) count [35], glucose concentration [36], and Gram stain [37]. Microbial invasion

of the amniotic cavity was defined as a positive amniotic fluid culture and/or polymerase chain reaction with electrospray ionization mass spectrometry (PCR/ESI-MS) (Ibis<sup>®</sup> Technology—Pathogen, Carlsbad, CA, USA) test result [38]. Intra-amniotic inflammation was defined as an amniotic fluid IL-6 concentration  $>2.6$  ng/mL [22]. Sample metadata is included in Additional file 1.

### cfDNA extraction and library preparation

Amniotic fluid samples were thawed from  $-80^{\circ}\text{C}$  and centrifuged at  $1500\times g$  for 5 min. The top 175  $\mu\text{L}$  of supernatant was removed and placed in a 1.5-mL tube with 825  $\mu\text{L}$  of  $1\times$  PBS and pipette mixed. The amniotic fluid was diluted to 1 mL in PBS, and cfDNA was isolated using the “Urine Supernatant 1 mL” protocol of the QiaAmp circulating nucleic acid extraction kit. Total cfDNA was eluted into 30  $\mu\text{L}$  of the elution buffer. The DNA concentration was determined using the Qubit 3.0 Fluorometer (dsDNA HS Qubit). Libraries of extracted amniotic fluid cfDNA were prepared using a single-stranded DNA library preparation method. For this study, sample batches were not continuous between the cfDNA extraction, library preparation, and sequencing steps due to sample processing constraints. LBBC can address batch effects at any stage but will perform best if samples are maintained in the same batch throughout sample processing.

### cfDNA sequencing

Paired-end DNA sequencing was performed on Illumina NextSeq 500 ( $2\times 75$  bp) at Cornell University or Illumina HiSeq ( $2\times 100$  bp) at Michigan State University. Paired-end fastq files were trimmed to 75 bp, and samples processed on both NextSeq and HiSeq platforms were concatenated into a single file for each sample.

### Fetal fraction determination

Adapter-trimmed reads were aligned to the UCSC hg19 build using bwa mem [39]. Duplicates, low-quality reads, and reads with secondary sequence alignments were removed. Aligned bam files were processed in 500 bp windows using the R package HMMcopy (version 1) [40]. We determined the coverage exclusively in these regions with high mappability scores to extrapolate the coverage of the whole chromosome. The fetal fraction was determined as  $2Y/A$  for subjects who were known to be pregnant with male fetuses, where  $Y$  and  $A$  are the inferred sequencing coverage of the Y chromosome and autosomes, respectively. To confirm the accuracy of the measurement, we ran the algorithm on samples from subjects with female fetuses, which we would expect to have a zero fetal fraction. We determined very few misalignments to the Y chromosome (median 2.6%,  $n = 20$ ).

### Microbial abundance determination

Fastq files were trimmed (Trimmomatic-0.32 [41]) and aligned to the human genome (UCSC hg19 build) using bowtie2 [42] (in very sensitive mode, version 2.3.5.1). Human-unaligned reads were retrieved and aligned to an annotated NCBI microbial database using BLAST [11] (blastn, NCBI BLAST 2.2.28+). After read alignment, a maximum likelihood estimator, GRAMMy (version 1), was used to adjust the BLAST hits [12]. The adjusted hits to each taxon and respective genome size of each taxon were used to compute the taxon genome coverage. The ratio of each taxon's genomic coverage to that of human chromosome 21 was used to compute the relative genomic abundance of each taxon in each sample.

### Low biomass background correction

The biomass correction method was employed in three steps: (1) BLAST hits were collected for every taxon with ten alignments or more. Genomes were aggregated into 1-kbp bins and the number of alignments within each bin was determined. The coefficient of variation (the standard deviation in alignments per bin divided by the mean number of alignments per bin) was calculated for each taxon in the sample. Given the number of alignments to a specific taxon and the taxon size, we randomly generated reads across the genome to simulate uniform sampling. The CV of this simulated taxon was calculated ( $CV_{sim}$ ). The difference between the CV and  $CV_{sim}$  ( $\Delta CV$ ) was then determined to look at coverage statistic discrepancy. CV and  $\Delta CV$  were calculated for every taxon in every sample in the cohort. Taxa were removed if they exceeded a maximum allowable  $\Delta CV$  value.

(2) The mass of each taxon present in a sample was calculated by calculating the adjusted number of BLAST hits from GRAMMy, dividing by the total number of sequencing reads, and multiplying by the mass of DNA added into library preparation (measured using a Qubit 3.0 Fluorometer). Taxon biomasses were compared across samples extracted or prepared within batches using the “cov” command standard in R. The diagonal of the output matrix reveals the variation within the batch for a given taxon. Taxa with variation below the minimum filtering parameter ( $\sigma^2$ ) were removed from every sample in the batch.

(3) For all of our wet lab procedures, a negative control (dsDNA synthetic oligos of length 25 bp, 40 bp, 55 bp, and 70 bp; each resuspended 0.20  $\mu$ M eluted in TE buffer) was processed alongside samples in batches. Microbial controls were sequenced alongside samples and were designed to take up 1–3% of the sequencing lane (roughly four to 12 million reads). Control samples were processed through the bioinformatics pipeline, and the taxa read proportion was calculated (raw BLAST hits to a taxon divided by total raw sequencing reads). The taxa

read proportion was calculated in samples and compared with that in the controls. Taxa for which the read proportion did not exceed 10-fold higher than the contaminant read proportion were removed. Following processing, the relative genomic abundance (measured in relative genomic equivalents, RGE) was summed for taxa to the species, genus, or family level, depending on desired output.

### Correction optimization

To facilitate the optimization of filtering parameters  $\Delta CV_{max}$  and  $\sigma^2_{min}$ , we created a store based on a linear combination of values related to the true positive, true negative, false positive, and false negative rates. We optimized these parameters based on the following metric:

$$BC_{score} = k_{TP}(TP) + k_{TN}(TN) + k_{FP}(FP) + k_{FN}(FN) + k_U(U),$$

where {TP, TN, FP, FN} is the number of true positives, true negatives, false positives, and false negatives, respectively;  $U$  is the total number of identified taxa for which a secondary method of identification was not performed; and the coefficients  $k$  for these values represent weights to optimize the filtering parameters based on the specifics of the application. Here, we chose  $\{k_{TP}, k_{TN}, k_{FP}, k_{FN}, k_U\} = \{4, 2, -1, -2, -0.25\}$  and used nonlinear minimization by gradient descent to minimize  $(1 - BC_{score})$  to determine an optimal set of threshold parameters.

### Other statistical analyses

All statistical analyses were performed in R. Correlation measurements were performed using Spearman correlations (function `cor.test`). To compute the confusion matrix in analysis of the urinary cfDNA datasets, we constructed four possible observable states for each sample: *Escherichia* positive, *Enterococcus* positive, both *Escherichia* and *Enterococcus* positive, and double negative. Observation of the state was determined with the reduced microbial matrix after filtering. Observed state was compared with standard urine culture as the reference. A  $4 \times 4$  confusion matrix was constructed, and statistics, including the accuracy and no information rate, were determined using the “confusionMatrix” command from the R caret package.

### Versions of software and references

Reads were aligned to human genome build hg19. Non-human reads were aligned to a NCBI reference database (downloaded 2015). The following packages (with versions) were used to build the LBBC package and analyze the data in R (version 3.6.1): caret (6.0-84), data.table (1.12.6), devtools (2.2.1), ggplot2 (3.2.1), ggpubr (0.2.3), ineq (0.2-13), MASS (7.3-51.4), reshape2 (1.4.3), roxygen2 (6.1.1), and taxize (0.9.9).

## Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1186/s40168-020-0793-4>.

**Additional file 1:** Sample metadata presented as two spreadsheets, corresponding to urinary cell-free DNA and amniotic fluid cell-free DNA. Metadata includes relevant clinical metadata and library preparation metadata.

### Abbreviations

cfDNA: Cell-free DNA; Chor: Chorioamnionitis; CV: Coefficient of variation; LBBC: Low biomass background correction; UTI: Urinary tract infection

### Acknowledgements

The authors acknowledge Hao Shi and other members of the De Vlaminck lab for helpful discussions. The authors thank Peter Schweitzer (Cornell Genomics Facility) and Susan Land and Katherine Gurdziel (Applied Genomics Technology Core, Wayne State University) for the assistance with Illumina sequencing assays.

### Authors' contributions

PB, NG-L, DD, JRL, MS, RR, and IDV contributed to the study design. NG-L and RR collected samples new to this study. PB, MH, and JSL performed the experiments. PB, APC, and IDV analyzed the data. PB and IDV wrote the manuscript. All authors provided comments and edits. All authors read and approved the final manuscript.

### Funding

This work was supported by R21AI133331 (to IDV and JRL), R21AI124237 (to IDV), DP2AI138242 (to IDV), K23AI124464 (to JRL), R01AI146165 (to IDV). PB is supported by an NSF GRFP, DGE-1144153. APC is supported by the National Sciences 33 and Engineering Research Council of Canada (401236174) fellowship. This research was supported, in part, by the Perinatology Research Branch, Division of Obstetrics and Maternal-Fetal Medicine, Division of Intramural Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, U.S. Department of Health and Human Services (NICHD/NIH/DHHS); and, in part, with federal funds from NICHD/NIH/DHHS under Contract No. HHSN275201300006C. RR has contributed to this work as part of his official duties as an employee of the United States Federal Government. NG-L was also supported by the Wayne State University Perinatal Initiative in Maternal, Perinatal and Child Health. MS is supported by NIH MERIT Award R37 AI051652.

### Availability of data and materials

Raw sequencing has been made available for both the urinary cfDNA datasets (dbGaP accession number phs001564.v2.p1) and amniotic fluid cfDNA datasets (phs001564.v3.p1). LBBC is made available as an R package: <https://github.com/pburnham50/LowBiomassBackgroundCorrection>.

### Ethics approval and consent to participate

The collection of samples was approved by the Institutional Review Boards of the Detroit Medical Center (Detroit, MI, USA), Wayne State University, and the Perinatology Research Branch, an intramural program of the Eunice Kennedy Shriver National Institutes of Health, U.S. Department of Health and Human Services (NICHD/NIH/DHHS). All participating women provided written informed consent prior to the collection of samples.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

### Author details

<sup>1</sup>Meinig School of Biomedical Engineering, Cornell University, Ithaca, NY, USA. <sup>2</sup>Perinatology Research Branch, Division of Obstetrics and Maternal-Fetal Medicine, Division of Intramural Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, U.S. Department of Health and Human Services (NICHD/NIH/DHHS), Bethesda, MD, USA. <sup>3</sup>Department of Biochemistry, Microbiology and

Immunology, Wayne State University School of Medicine, Detroit, MI, USA.

<sup>4</sup>Department of Obstetrics and Gynecology, Wayne State University School of Medicine, Detroit, MI, USA. <sup>5</sup>Department of Transplantation Medicine, New York Presbyterian Hospital–Weill Cornell Medical Center, New York, NY, USA. <sup>6</sup>Center for Molecular Medicine and Genetics, Wayne State University, Detroit, MI, USA. <sup>7</sup>Department of Epidemiology and Biostatistics, College of Human Medicine, East Lansing, MI, USA. <sup>8</sup>Department of Obstetrics and Gynecology, University of Michigan Health System, Ann Arbor, MI, USA. <sup>9</sup>Detroit Medical Center, Detroit, MI, USA. <sup>10</sup>Department of Obstetrics and Gynecology, Florida International University, Miami, Florida, USA.

Received: 14 November 2019 Accepted: 20 January 2020

Published online: 11 February 2020

### References

1. De Vlaminck I, Khush KK, Strehl C, Kohli B, Luikart H, Neff NF, et al. Temporal response of the human virome to immunosuppression and antiviral therapy. *Cell*. 2013;155:1178–87.
2. Burnham P, Dadhania D, Heyang M, Chen F, Westblade LF, Suthanthiran M, et al. Urinary cell-free DNA is a versatile analyte for monitoring infections of the urinary tract. *Nat Commun*. 2018;9:2412. <https://doi.org/10.1038/s41467-018-04745-0>.
3. Blauwkamp TA, Thair S, Rosen MJ, et al. Analytical and clinical validation of a microbial cell-free DNA sequencing test for infectious disease. *Nat Microbiol*. 2019;4:663–74. <https://doi.org/10.1038/s41564-018-0349-6>.
4. De Vlaminck I, Martin L, Kertesz M, Patel K, Kowarsky M, Strehl C, et al. Noninvasive monitoring of infection and rejection after lung transplantation. *Proc Natl Acad Sci U S A*. 2015;112(43):13336–41. <https://doi.org/10.1073/pnas.1517494112>.
5. Burnham P, Kim MS, Agbor-Enoh S, Luikart H, Valentine HA, Khush KK, et al. Single-stranded DNA library preparation uncovers the origin and diversity of ultrashort cell-free DNA in plasma. *Sci Rep*. 2016;6:27859. <https://doi.org/10.1038/srep27859>.
6. Marotz CA, Sanders JG, Zuniga C, Zaramela LS, Knight R, Zengler K. Improving saliva shotgun metagenomics by chemical host DNA depletion. *Microbiome*. 2018;6:42. <https://doi.org/10.1186/s40168-018-0426-3>.
7. Carpenter ML, Buenostro JD, Valdiosera C, Schroeder H, Allentoft ME, Sikora M, et al. Pulling out the 1%: whole-genome capture for the targeted enrichment of ancient DNA sequencing libraries. *Am J Hum Genet*. 2013;93:852–64.
8. Gu W, Crawford ED, O'Donovan BD, Wilson MR, Chow ED, Retalack H, et al. Depletion of abundant sequences by hybridization (DASH): using Cas9 to remove unwanted high-abundance species in sequencing libraries and molecular counting applications. *Genome Biol*. 2016;17:41.
9. Eisenhofer R, Minich JJ, Marotz C, Cooper A, Knight R, Weyrich LS. Contamination in low microbial biomass microbiome studies: issues and recommendations. *Trends Microbiol*. 2019;27:105–17.
10. Romero R, Miranda J, Kusanovic JP, Chaiworapongsa T, Chaemsaitong P, Martinez A, et al. Clinical chorioamnionitis at term I: microbiology of the amniotic cavity using cultivation and molecular techniques. *J Perinat Med*. 2015;43:19–36. <https://doi.org/10.1515/jpm-2014-0249>.
11. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215(3):403–10.
12. Xia LC, Cram JA, Chen T, Fuhrman JA, Sun F. Accurate genome relative abundance estimation based on shotgun metagenomic reads. *PLoS One*. 2011;6:e27992. <https://doi.org/10.1371/journal.pone.0027992>.
13. de Goffau MC, Lager S, Salter SJ, Wagner J, Kronbichler A, Charnock-Jones DS, et al. Recognizing the reagent microbiome. *Nat Microbiol*. 2018;3:851–3.
14. Davis NM, Proctor DM, Holmes SP, Relman DA, Callahan BJ. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome*. 2018;6:226.
15. Cheng AP, Burnham P, Lee JR, Cheng MP, Suthanthiran M, Dadhania D, et al. A cell-free DNA metagenomic sequencing assay that integrates the host injury response to infection. *Proc Natl Acad Sci*. 2019;116:18738 LP–18744. <https://doi.org/10.1073/pnas.1906320116>.
16. Chaban B, Links MG, Jayaprakash TP, Wagner EC, Bourque DK, Lohn Z, et al. Characterization of the vaginal microbiota of healthy Canadian women through the menstrual cycle. *Microbiome*. 2014;2:23.
17. Lewis D, Brown R, Williams J, White P, Jacobson S, Marchesi J, et al. The human urinary microbiome; bacterial DNA in voided urine of asymptomatic adults. *Fron Cell Infect Microbiol*. 2013;3:41.



18. Fan HC, Blumenfeld YJ, Chitkara U, Hudgins L, Quake SR. Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. *Proc Natl Acad Sci U S A*. 2008;105:16266–71.
19. Malloy MH. Chorioamnionitis: epidemiology of newborn management and outcome United States 2008. *J Perinatol*. 2014;34:611. <https://doi.org/10.1038/jp.2014.81>.
20. Gibbs RS, Blanco JE, St. Clair PJ, Castaneda YS. Quantitative bacteriology of amniotic fluid from women with clinical intraamniotic infection at term. *J Infect Dis*. 1982;145:1–8. <https://doi.org/10.1093/infdis/145.1.1>.
21. Gibbs RS, Dinsmoor MJ, Newton ER, Ramamurthy RS. A randomized trial of intrapartum versus immediate postpartum treatment of women with intra-amniotic infection. *Obstet Gynecol*. 1988;72:823–8.
22. Yoon BH, Romero R, Bin MJ, Shim S-S, Kim M, Kim G, et al. Clinical significance of intra-amniotic inflammation in patients with preterm labor and intact membranes. *Am J Obstet Gynecol*. 2001;185:1130–6. <https://doi.org/10.1067/mob.2001.117680>.
23. Romero R, Gomez-Lopez N, Kusanovic JP, Pacora P, Panaitescu B, Erez O, et al. Clinical chorioamnionitis at term: new insights into the etiology, microbiology, and the fetal, maternal and amniotic cavity inflammatory responses. *Nogyogy es szuleszeti Tovabbk Szle*. 2018;20:103–12.
24. Romero R, Chaemsaihong P, Korzeniewski SJ, Tarca AL, Bhatti G, Xu Z, et al. Clinical chorioamnionitis at term II: the intra-amniotic inflammatory response. *J Perinat Med*. 2016;44:5–22. <https://doi.org/10.1515/jpm-2015-0045>.
25. Gomez-Lopez N, Romero R, Maymon E, Kusanovic JP, Panaitescu B, Miller D, et al. Clinical chorioamnionitis at term IX: in vivo evidence of intra-amniotic inflammasome activation. *J Perinat Med*. 2019;47:276–87.
26. Gomez-Lopez N, Romero R, Xu Y, Leng Y, Garcia-Flores V, Miller D, et al. Are amniotic fluid neutrophils in women with intraamniotic infection and/or inflammation of fetal or maternal origin? *Am J Obstet Gynecol*. 2017;217:693.e1–693.e16.
27. Gomez-Lopez N, Romero R, Xu Y, Miller D, Leng Y, Panaitescu B, et al. The immunophenotype of amniotic fluid leukocytes in normal and complicated pregnancies. *Am J Reprod Immunol*. 2018;79:e12827.
28. Snyder MW, Kircher M, Hill AJ, Daza RM, Shendure J. Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin. *Cell*. 2016;164:57–68.
29. Tsui NBY, Jiang P, Chow KCK, Su X, Leung TY, Sun H, et al. High resolution size analysis of fetal DNA in the urine of pregnant women by paired-end massively parallel sequencing. *PLoS One*. 2012;7:1–7.
30. Fan HC, Blumenfeld YJ, Chitkara U, Hudgins L, Quake SR. Analysis of the size distributions of fetal and maternal cell-free DNA by paired-end sequencing. *Clin Chem*. 2010;56:1279–86.
31. Leiby JS, McCormick K, Sherrill-Mix S, Clarke EL, Kessler LR, Taylor LJ, et al. Lack of detection of a human placenta microbiome in samples from preterm and term deliveries. *Microbiome*. 2018;6:196.
32. Theis KR, Romero R, Winters AD, Greenberg JM, Gomez-Lopez N, Alhousseini A, et al. Does the human placenta delivered at term have a microbiota? Results of cultivation, quantitative real-time PCR, 16S rRNA gene sequencing, and metagenomics. *Am J Obstet Gynecol*. 2019;220:267.e1–267.e39. <https://doi.org/10.1016/j.jajog.2018.10.018>.
33. Weiss S, Amir A, Hyde ER, Metcalf JL, Song SJ, Knight R. Tracking down the sources of experimental contamination in microbiome studies. *Genome Biol*. 2014;15:564. <https://doi.org/10.1186/s13059-014-0564-2>.
34. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Methods*. 2012;9:811. <https://doi.org/10.1038/nmeth.2066>.
35. Romero R, Quintero R, Nores J, Avila C, Mazor auJMoshe, Hanaoka S, et al. Amniotic fluid white blood cell count: a rapid and simple test to diagnose microbial invasion of the amniotic cavity and predict preterm delivery. *Am J Obstet Gynecol* 1991;165:821–830. [https://doi.org/10.1016/0002-9378\(91\)90423-O](https://doi.org/10.1016/0002-9378(91)90423-O).
36. Romero R, Jimenez C, Lohda AK, Nores J, Hanaoka S, Avila C, et al. Amniotic fluid glucose concentration: a rapid and simple method for the detection of intraamniotic infection in preterm labor. *Am J Obstet Gynecol*. 1990;163:968–74. [https://doi.org/10.1016/0002-9378\(90\)91106-M](https://doi.org/10.1016/0002-9378(90)91106-M).
37. Romero R, Emamian M, Quintero R, Wan M, Hobbins JC, Mazor M, et al. The value and limitations of the Gram stain examination in the diagnosis of intraamniotic infection. *Am J Obstet Gynecol*. 1988;159:114–9. <https://doi.org/10.5555/uri:pii:0002937888905030>.
38. Romero R, Miranda J, Chaiworapongsa T, Chaemsaihong P, Gotsch F, Dong Z, et al. A novel molecular microbiologic technique for the rapid diagnosis of microbial invasion of the amniotic cavity and intra-amniotic infection in preterm labor with intact membranes. *Am J Reprod Immunol*. 2014;71:330–58. <https://doi.org/10.1111/aji.12189>.
39. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25:1754–60. <https://doi.org/10.1093/bioinformatics/btp324>.
40. Ha G, Roth A, Lai D, Bashashati A, Ding J, Goya R, et al. Integrative analysis of genome-wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted pathways in triple-negative breast cancer. *Genome Res*. 2012;22:1995–2007.
41. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30:2114–20.
42. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012;9:357–9.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

