

PROFILING INFECTIOUS DISEASE VIA SINGLE-CELL AND SINGLE-MOLECULE  
SEQUENCING

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Philip Smith Burnham

May 2019

© 2019 Philip Smith Burnham

PROFILING INFECTIOUS DISEASE VIA SINGLE-CELL AND SINGLE-MOLECULE  
SEQUENCING

Philip Smith Burnham, Ph. D.

Cornell University 2019

The global burden of infectious disease has declined in recent decades. Yet, patients who are immunocompromised and individuals in resource-limited settings remain at high risk of infection. In this dissertation, I will present several next generation sequencing assays that we have created that enable new ways to monitor and study infectious diseases. I will present two classes of technologies that target two different analytes: (1) cell-free DNA (cfDNA) in biological fluids and (2) viral transcripts within single cells. We have developed a library preparation assay that is sensitive to ultrashort cfDNA, which captures information about the pathogen and host. We applied this cfDNA sequencing assay to a large number of urine samples collected from patients with viral and bacterial urinary tract infections. Our findings indicate cfDNA sequencing can accurately detect a broad range of uropathogens and describe functional information about the infectious agent and host. We have also developed a complementary analytical pipeline to reduce false-positive identifications and background contamination. We have recently applied this pipeline in the monitoring of infectious diseases that are endemic in low-income countries. Using DNA sequencing, we proved that genome replication dynamics can be observed during MTB infections and that an abundance of enteric bacteria is present in the plasma of children suffering environmental enteropathy. In the second part of the dissertation, I will introduce a new high-throughput single-cell RNA sequencing tool that combines enrichment measurements of targeted RNA sequences with unbiased profiling of the polyadenylated transcriptome across thousands of single cells in the same biological sample. We applied this technique to simultaneously characterize the non-A-tailed transcripts of a segmented dsRNA viruses

and the transcriptome of the infected cells. In addition, we applied the technology to simultaneously determine the natively paired, variable region heavy and light chain amplicons and the transcriptome of B lymphocytes. In summary, we have created new tools to capture and sequence nucleic acids in biological fluids and single cells, thereby providing novel ways to understand pathogens as well as host immunity and damage.

## BIOGRAPHICAL SKETCH

Phil was born in Philadelphia, PA in 1991. He was inspired to become a scientist by his mother. He fell in love with physics and mathematics in high school (Seneca High School, Tabernacle, NJ) and undergraduate studies (Villanova University, Villanova, PA). He began his graduate work at Cornell University to obtain a doctorate in physics in 2013. Following the start of the West Africa Ebola outbreak, he was moved to pursue the development of novel methods to understand and diagnose microbial pathogens. In the future he hopes to become a leader in understanding cellular identity within the context of environmental and community influences, particularly as it relates to infection.

This manuscript is dedicated to my family.

## ACKNOWLEDGMENTS

My research progress over the past several years would not have been possible without the guidance and leadership of Iwijn De Vlaminck. Iwijn's vision to become a leader in diagnostics and genomics first drew me to join his lab. It has been a privilege to have helped in starting the lab and to have watched it grow and evolve since 2015.

My family has always encouraged my interests, pursuits, passions, and general love of science. Foremost I thank my wife, Melanie Burnham, who has constantly supported my career as a scientist. Even when research is daunting, she motivates me to achieve at the highest level. I thank my mother and Adhoedha, Elizabeth and Eric Jolly, for their support, guidance, and dedication to education and equity. My mom has encouraged me to pursue work that will leave a positive impact on the world. My Adhoedha has helped me recognize that the science is not just the act of discovery but those who are performing it. I thank my grandmother, Carole Robertson, who has always been my biggest fan. She reminds me of the importance of dreams. I thank my brother, Michael Burnham, and sister, Victoria Jolly. They have always pushed me to be a better person. I thank my father and stepmother, Phil and Diane Burnham; my aunt and uncle, Peggy Anne and Manny D'Alessio, and my cousins; and my in-laws, Todd, Betty, and Kelsey Berger. Lastly, I thank my grandfather, Henry Freda, for all he did to shape my appreciation for life.

I would not have made it through my degree without the support of Hao Shi, Andre Frankenthal, Ti-yen Lan, and Archishman Raju. They have shown true friendship when life or research have been overwhelming and have become family over the past six years. They have made me think critically about both scientific and social problems.

The members of the De Vlaminck lab have been instrumental in the success of my projects. I would especially like to thank: Alexandre Pellan Cheng for his friendship and shared interest in cell-free DNA and diagnostics; Mridusmita Saikia, for her guidance in the lab and for sharing her joy in experimental biology; Joan Sesing Lenz, for her support in experimental aspects; and my mentees, Sara Keshavjee,

Michael Heyang, and Fanny Chen, for their drive and compassionate pursuits. One of the greatest parts of my research career so far has been in training and working with Sara, Michael, and Fanny. Meleana Hinchman and John Parker have been instrumental in matters related to virology. Peter Schweitzer and his team at the Cornell Genomics Center have been a tremendous resource in helping me develop sequencing related technologies over the past four years.

I thank the wonderful people who were part of my tenure as a graduate resident fellow at Hans Bethe House, in particular Karen and John Smeda, Erica Ostermann, Scott MacDonald, Julia Thom-Levy, Elizabeth Feeney, Xine Yao, Elizabeth Wayne, Nancy and Charlie Trautmann, and David Miller. They helped me build a home in Ithaca, for which I will always be thankful.

My undergraduate mentor Dr. Georgia C. Papaefthymiou was the first person who gave me the opportunity to perform research. She has championed my work since 2011, and I am fortunate to have had her mentorship.

I have been blessed with an amazing and supportive committee of physicists to guide my doctoral studies. Michelle Wang helped me tremendously to secure funding as a first year student, allowing me the freedom to pursue my areas of interest. Itai Cohen has shown me the importance of, and art to, scientific communication. Chris Myers has asked great questions of my work, which formed the second half of my doctorate and opened up new avenues to study viral infections in single cells.

There are many others.

## TABLE OF CONTENTS

BIOGRAPHICAL SKETCH .....	v
ACKNOWLEDGMENTS .....	vii
LIST OF FIGURES .....	xi
LIST OF TABLES .....	xviii
LIST OF ABBREVIATIONS.....	xix
Introduction.....	1
PART I: Cell-free DNA sequencing to monitor infection .....	4
Chapter 1: Physiological and historical origins of cell-free DNA .....	5
Chapter 2: Single-stranded DNA library preparation uncovers the origin and diversity of ultrashort cell-free DNA in plasma .....	10
1.2.0    A single-stranded ligation strategy to enrich ultrashort cfDNA .....	11
1.2.1    Mitochondrial cfDNA in plasma measured by digital PCR.....	12
1.2.2    ssDNA library preparation and fragmentation profiles.....	14
1.2.3    Improved recovery of mitochondrial and microbial cfDNA.....	16
1.2.4    Donor-specific cfDNA from the mitochondria and autosomes .....	19
1.2.5    Methods and sampling cohort description .....	22
1.2.6    Ultrashort cell-free DNA provides a new perspective of origin and pathology.....	25
Chapter 3: Urinary cell-free DNA is a versatile analyte for monitoring infections of the urinary tract .	28
1.3.0    cfDNA sequencing to inform infections of the urinary tract .....	29
1.3.1    Biophysical properties of urinary cfDNA.....	30
1.3.2    Screening the infectome for multiple pathogenic agents .....	32
1.3.3    Profiling the urinary microbiome.....	34
1.3.4    Broad screening for viruses via cfDNA.....	35
1.3.5    Quantifying bacterial growth rates.....	39
1.3.6    Antimicrobial resistome profiling.....	41
1.3.7    Measuring the host response to infection.....	42
1.3.8    Identification of the tissue-of-origin of urinary cfDNA.....	45
1.3.9    Detailed methods and sample selection .....	47
1.3.10   Urinary cfDNA accurately predicts infections and host cell damage .....	55
Chapter 4: Pathogen screening and microbiome profiling from low-biomass isolates of cell-free DNA .....	57
1.4.0    Factors contributing to false-positive microbial identification .....	58
1.4.1    The low-biomass background correction pipeline .....	60

1.4.2	Training correction parameters using clinically informed interpretation.....	65
1.4.3	Background corrected cfDNA sequencing confirms chorioamnionitis in sterile wombs ...	67
1.4.4	Detection of microbial cfDNA in peritoneal dialysis effluent .....	69
Chapter 5: Applications of cfDNA sequencing for global health .....		74
1.5.0	cfDNA sequencing to study infectious disease in global health .....	75
1.5.1	Analysis of MTB genome replication dynamics.....	76
1.5.2	Enteric bacterial cfDNA detected in plasma during environmental enteropathy.....	81
Chapter 6: Perspective in cell-free DNA diagnostics .....		85
PART II: Single-cell sequencing in infected cell populations – understanding the innate and adaptive immune response.....		89
Chapter 1: Introduction to virus-inclusive single-cell RNA sequencing .....		90
Chapter 2: Simultaneous multiplexed amplicon sequencing and transcriptome profiling in single cells .....		98
2.2.0	Targeted amplicon sequencing in single cells to fully describe cell heterogeneity .....	99
2.2.1	DART-seq primer bead synthesis .....	100
2.2.2	DART-seq reveals heterogeneity of cellular phenotypes and viral genotypes .....	104
2.2.3	DART-seq allows high-throughput paired repertoire sequencing of B lymphocytes .....	107
2.2.4	Detailed methods of DART-seq assay .....	110
2.2.5	Conclusions from DART-seq proof-of-principle experiments .....	114
Chapter 3: Virus-inclusive scRNA-seq to understand enteric viral infections .....		115
2.3.0	Motivations to expand studies to greater pathogenicity and complexity .....	116
2.3.1	Rotavirus infections in monkey fibroblasts.....	117
2.3.2	Single-cell sequencing of complex cellular communities.....	122
2.3.3	Future experiments: rotavirus infections in organoid systems.....	126
Chapter 4: Afterwards on virus-inclusive single-cell sequencing.....		128
Conclusions.....		131
REFERENCES .....		137

LIST OF FIGURES

**Figure 1.1.0.1** cfDNA properties strongly indicate nucleosomal origin. (a) Histogram of fragment length distribution for a plasma cfDNA sample prepared using standard, commercial library preparation. (b) Fourier transform was performed on distribution in (a) between 60 and 170 bp, peak indicates most prominent periodicity. (c) A dinucleotide heatmap (colored by mean A/T frequency) is shown for cfDNA fragments of varying length, relative to central position. (A/T) dinucleotide combinations are scored with a value of 1, other combinations are scored with value 0, heatmap reflects mean score. Top: Cross-section of heatmap shown above for 167 bp. In (a) and (c) dashed line corresponds to most abundant fragment length, 167 bp. {1} ..... 7

**Figure 1.1.0.2** Patterns of cfDNA fragmentation observed by sequencing depth along reference genome. Sequencing coverage of human aligned cfDNA originating from plasma across a small section of the genome for 140-180 bp fragments. Plasma cfDNA prepared for sequencing using standard, commercial library preparation. Positions illustrated are from human chromosome 12. {2} ..... 8

**Figure 1.2.1.1** Digital PCR measurements reveal an abundance of ultrashort cfDNA originating from the mitochondria. Number of genomic copies (per milliliter of plasma) is shown as a function of amplicon length for cfDNA originating from the nucleus (black, five amplicons) and mitochondria (green, nine amplicons). {3} ..... 13

**Figure 1.2.2.1** Schematic overview of single-stranded library preparation for cfDNA. (i) cfDNA from biofluid is extracted and is highly heterogeneous with respect to degradation patterns. (ii) cfDNA is denatured at 95 °C. (iii) single-stranded cfDNA is ligated to biotinylated adapters and bound to streptavidin-coated magnetic beads. (iv) Bound ssDNA is end-repaired, extended, and adapted with double-stranded sequencing primers. (v) Sequencing libraries are PCR amplified and indexed. Final libraries represent degradation diversity present in original set. {4} ..... 15

**Figure 1.2.2.2** Single-stranded library preparation reveals an abundance of ultrashort cfDNA molecules. Density plot of the fragment sizes of nuclear genomic cfDNA measured after ssDNA (blue) and dsDNA (red) library preparation. The inset shows the sample-to-sample variability (n = 40), as well as the difference in GC content for short (< 100 bp) and long (> 100 bp) fragments. Significance of  $p \ll 10^{-5}$ , Mann-Whitney U Test. Boxplot description given in section 1.3.9. {5} ..... 16

**Figure 1.2.3.1** ssDNA ligation based library preparation reveals enrichment for microbial and mitochondrial cell-free DNA. Fragment length distribution for matched samples using ssDNA ligation and traditional library preparation techniques is compared, as normalized to nuclear cfDNA molecules, for (a) mitochondrial-aligned and (b) microbial-aligned fragments. {6} ..... 17

**Figure 1.2.3.2** ssDNA library preparation captures more bacterial cfDNA with higher diversity than standard methods. (a) Single-stranded ligation method has significantly more microbial cell-free DNA than the traditional method (double-stranded ligation). (b) The abundance of microbial genomes is compared by ligation method, for matched species within matched samples. Species are colored by superkingdom (Non-fungi eukaryotes are grey in color). {7} ..... 18

**Figure 1.2.4.1** Single-stranded library preparation captures donor fraction information from autosomal cfDNA in matched samples. The donor fraction from autosomes was calculated and compared for both library preparation methods. Error bars represent standard error in measurement in error calculated from sampling error. Red line indicates exact correspondence. {8} ..... 19

**Figure 1.2.4.2** Donor-specific cfDNA originating from the mitochondria reveals graft damage. (a) Schematic representation of analysis workflow used to discriminate donor and recipient specific mt-cfDNA. Examples of an ambiguous assignment and a fragment assigned to the donor are shown. (b) Fraction of donor-specific mt-cfDNA as function of time post-transplant for five double lung

transplant patients (25 samples, samples with fewer than 20 informative fragments excluded); Inset: The fraction of donor-specific mitochondrial and nuclear genomic DNA for the same samples is compared (corr. = 0.463, Pearson, $p = 0.020$ ). {9} .....	20
<b>Figure 1.2.4.3</b> Smoothed (distribution five nearest-neighbor running mean) of donor mt-cfDNA (red) is compared to that of recipient mt-cfDNA (blue). Inset: median fragment size for the donor mt-cfDNA (red line) compared to the fragment size of 10,000 subsets sampled from the recipient mt-cfDNA length set (median depicted by black line). {10} .....	22
<b>Figure 1.2.6.1</b> Fragment length distributions indicate the heterogeneity of degradation in various biological fluids. Fragment length distributions are shown for cfDNA originating from plasma (n = 40), urine (n = 40), and peritoneal dialysis effluent (n = 40). Samples were prepared using ssDNA library preparation. Figure assembled by Fanny Chen. {11} .....	26
<b>Figure 1.3.1.1</b> Shotgun sequencing assay and biophysical properties of urinary cfDNA. (a) Study design included patient samples, in health and infection, from one day to six years after transplant. Fragment length distributions and Fast Fourier Transform (FFT, 60 – 140 bp, inset) for cfDNA from organisms with nucleosomal packaging (b) and without such packaging (c). {12} .....	31
<b>Figure 1.3.2.1</b> Diagnostic comparison of microbial cfDNA sequencing in UTI to gold standard technologies. (a) The relative genomic abundance is shown for BK polyomavirus for patients who tested positive or negative for polyomavirus nephropathy via renal biopsy, and for those untested. (b) Bacterial species were ranked according to relative genomic abundance within each sample for patients with culture-diagnosed UTI. The ranked position of the species was compared to that identified by standard urine culture. (c) Receiver-operator characteristics for samples originating from samples with <i>Enterococcus</i> , <i>E. coli</i> , and <i>P. aeruginosa</i> UTI. {13} .....	32
<b>Figure 1.3.3.1</b> Bacterial diversity and abundance measurements in urinary cfDNA for patient samples separated by donor sex, recipient sex, and collection methods. Metrics calculated at genus taxonomic level. {14} .....	35
<b>Figure 1.3.4.1</b> Urinary cfDNA profiles the human virome. Viral cfDNA was detected in 66 samples of the 141 samples. cfDNA reveals frequent occurrence of viruses that are potentially clinically relevant (left panel); red crosses identify samples belonging to subjects who developed an infection of the corresponding viral agent. Right-most panel shows boxplots of the viral cfDNA abundance across all samples. Color of points and boxplots by viral family. {15} .....	36
<b>Figure 1.3.4.2</b> Phylogenetic reconstruction of BK polyomavirus VP1 consensus sequences assembled from urinary cfDNA. BEAST software was used to find the most likely tree given a constant population over time. Nodes represent points of genetic separation. Leaves are annotated with sample identifier and positioned at time of sampling. Colors correspond to BK polyomavirus spread assessed in renal biopsy samples (blue = low, magenta = moderate, orange = diffuse). {16} .....	37
<b>Figure 1.3.5.1</b> Estimating bacterial population growth rates from urinary cfDNA. (a) Normalized bacterial genome coverage for four representative bacterial species. The coverage was binned in 1 kbp tiles and normalized. Each panel represents a single sample, with the exception of <i>C. acnes</i> (*) for which the coverage was aggregated across 99 samples (solid line is a LOESS filter smoothing curve, span=0.70). The non-uniform genome coverage for <i>E. coli</i> and <i>K. pneumoniae</i> , with an overrepresentation of sequences at the origin of replication, is a result of bi-directional replication from a single origin of replication. The initial and final 5% of the genome is removed for display. (b) Box plots of growth rates for species in 14 genera grouped by patient groups (at least 2500 alignments, 41 samples, see Methods for definition of pre/post-UTI). Each point indicates a bacterial species in a sample. Triangles indicate culture-confirmed bacteria by genus. Boxplot features are described in Section 1.3.9. {17} .....	39

**Figure 1.3.6.1** cfDNA-based antimicrobial resistome profiling reveals vancomycin resistant enterococcus. For 42 samples from subjects with clinically confirmed UTI, AR gene profiling reveals the presence of genes conferring resistance to various antimicrobial classes. These data are organized in three sample groups: samples from subjects with vancomycin-resistant *Enterococcus* (Resistant), samples from subjects with vancomycin-susceptible *Enterococcus* (Susceptible), and samples from subjects for which vancomycin resistance testing was not performed. Blue highlight indicates the AR class in which vancomycin resides. Black squares indicate at least one alignment. More than one alignment is indicated by red shading. {18} ..... 41

**Figure 1.3.7.1** The host response to infection can be quantified using urinary cfDNA. (a) Proportion of donor-specific cfDNA in urine of subjects that are BKVN positive per kidney allograft biopsy (BKVN) in urine collected in the first 5 days after transplant surgery (Early), urine collected from subjects that are bacterial UTI negative per culture in the first month following transplantation (No UTI), samples collected before or after bacterial UTI (Pre-/Post-UTI), and samples collected at the time of bacterial UTI diagnosis (UTI). The single outliers in Pre-/Post-UTI and UTI groups correspond to the same patient, who suffered an acute rejection episode in the months prior. Low donor fractions in the Pre-/Post-UTI and UTI groups are likely due to increased immune cell, i.e., WBC, presence in the urinary tract; subjects with higher WBC counts have lower donor fractions (inset, red color indicates pyuria). (b) Absolute abundance of donor cfDNA in the urine of subjects not diagnosed with infection in the first month post-transplant (red line is a LOESS filter smoothing curve, span = 1). Dotted lines connect samples from the same patient. (c) Genome coverage at the transcription start site, binned by the gene expression level across all samples in the study. TSS = transcription start site. FPKM = fragments per kilobase of transcript per million mapped reads, an RNA-seq measure of gene expression. {19} ..... 43

**Figure 1.3.8.1** Bisulfite-treated cfDNA sequencing indicates tissue of origin of cfDNA ensemble. (a) Measured donor fraction is compared to the kidney fraction for bisulfite-treated urinary cfDNA samples. Solid line indicates exact match. (b) The mean tissue proportion for each pathological group in the study is shown. (c) Tissue fraction is multiplied by total cfDNA extracted in urine to determine amount of cfDNA originating from organ. We compare Kidney (left) and White Blood Cell (WBC, right) cfDNA levels (ng / mL urine) across four pathological groups. {20} ..... 46

**Figure 1.4.1.1** A pipeline for sparse metagenomic background correction with simulated cases of false identification. (a) Overview of pipeline. An initial file containing the abundance of microbes across samples is compared to external information (blue boxes) regarding alignment statistics, batch information, measured biomass, and taxa in negative control samples. (b-d) Simulated examples to identify contaminant through different means. (b) Sequencing coverage across three artificial microorganisms was simulated in the case of high and homogenous coverage (yellow), high and inhomogeneous coverage (blue), and low coverage (green). At right, the coefficient of variation (CV) was calculated for a uniformly sampled, uniformly sequenced at the same depth of sequencing for each organism and compared to the calculated CV was compared to the theoretical CV. Dotted line shows 1:1 correspondence. (c) Simulated cases for a microbe as a contaminant (orange) and true identification (blue) show the theoretical relationship between measured biomass and proportion of all sampled microbes. (d) Simulated cases of bacterial abundance for four microbes. Orange shows low variability within batches, so is likely a contaminant. {21} ..... 61

**Figure 1.4.2.1** Low-biomass background correction (LBBC) on microbial abundance reveals pathogens, commensals, and sterile biomes. We applied LBBC to the microbial abundance arrays calculated in 44 samples for *Enterococcus* UTI in females and males (Groups 1 and 2,  $n = 7$  and 4, respectively), *E. coli* UTI in females and males (Groups 3 and 4,  $n = 12$  and 4, respectively), and healthy females (Group 5,  $n = 4$ ) and males (Group 6,  $n = 13$ ). The genera detected were then subset

from the microbial abundance matrix without background correction (left of arrow). Relative genome abundance is indicated by color gradient, on left. Green borders indicate positive clinical identification by standard urine culture. {22} .....	66
<b>Figure 1.4.3.1</b> cfDNA sequencing with background correction reveals causative chorioamnionitis pathogen. Species-level abundance is shown across 44 samples in cohort. Abundance value occupy a range from $10^{-2}$ RGE (blue) to $10^3$ RGE (red). Samples are divided by clinical pathology and species are divided by superkingdom. Black crosshairs indicate clinical confirmation by culture and 16S sequencing. {23} .....	68
<b>Figure 1.4.4.1</b> Microbial cfDNA sequencing in PD effluent reveals pathogens and broad detection of skin flora. Relative genomic abundance is shown across samples (n = 55) for five genera tested using bacterial culture. Samples are grouped by clinical presentation of culture-positive peritonitis (Peritonitis/Positive), culture-negative peritonitis (Peritonitis/Negative), and no peritonitis. X = taxa identification removed by LBBC algorithm. ▲ = clinical confirmation of taxa in patient. Colored by log scale of relative genomic abundance. {24} .....	70
<b>Figure 1.4.4.2</b> Microbial cfDNA sequencing in PD effluent is able to detect causative pathogens at low abundance and weeks after treatment. Relative genomic abundance of the clinically determined pathogen is shown for culture-positive peritonitis patients. Inset: Days 0 to 4 post-peritonitis are shown. Dotted line is the absolute limit of detection, $10^{-3}$ RGE. Red points are both detected and remain in abundance profiles after filtering. Blue points are detected but are filtered out through LBBC algorithm. Green points were not detected through microbial cfDNA sequencing. Lines connect samples from same patient. {25} .....	72
<b>Figure 1.5.1.1</b> Whole genome sequencing of MTB reveals effects of antibiotics. (a) MTB cultures were treated with one of several drugs (Untreated, ethambutol (ETH), moxifloxacin (MXF), rifampicin (RIF), isoniazid (IZD)). The percentage of replicating bacteria was determined using the replication score, and the concentration of genomic DNA (gDNA, normalized for volume and optical density) and supernatant DNA (superDNA) were calculated for time 12 hours before and up to 24 hours after dosing. Dotted line indicates time of drug inoculation. (b) The sequencing coverage (normalized to mean coverage) across the MTB genome (binned to 10 kbp) is shown for untreated and ETH-treated MTB prior to dosing and 24 hours after dosing. Solid and dashed vertical lines represent positions of replication origin and terminus, respectively. (c) A model of ETH effects on MTB in culture is shown. When ETH is inoculated into culture, MTB cell walls do not grow. Concurrently, genomes duplicate until there are two copies. (d) Confocal fluorescent imaging of a MTB after 24 hours of dosing. Bacteria were stained with DAPI. {26} .....	77
<b>Figure 1.5.1.2</b> WGS of rabbit caseum shows presence of MTB, but lack of active replication. Samples collected from days 0, 5, 7, 14, and 21 post-inoculation are shown. Relative genomic abundance was calculated by dividing MTB genome coverage by the sequencing coverage of rabbit genome. Mean coverage shown for 10 kbp genome bins. Solid and dashed vertical lines represent positions of replication origin and terminus, respectively. {27} .....	79
<b>Figure 1.5.1.3</b> Comparison of <i>Mycobacterium</i> (genus) and MTB cfDNA in plasma samples from patients with sputum diagnosis reveals no difference in relative genomic abundance. For 61 patients, relative genomic abundance for <i>Mycobacterium</i> genus and <i>Mycobacterium tuberculosis</i> species was calculated and compared by clinical diagnosis. Samples registering $10^{-5}$ RGE showed no cfDNA. {28} .....	80
<b>Figure 1.5.2.1</b> cfDNA sequencing of plasma microbiome in pediatric patients with EE reveals presence of gut flora. Genus-level identifications are presented for the bacteria identified across all samples in the EE cohort after sparse metagenomic background correction. Size of each point is relative	

to log <sub>10</sub> measurement of relative genomic abundance. Samples organized by sugar ratio measurement (top). {29}	82
<b>Figure 1.5.2.2</b> Plasma virome in patient-samples with EE indicate diversity and breadth of viral infections. (a) UMAP clustering was performed on the family-level relative genomic abundance of viruses identified across samples in EE cohort. Ellipse indicates cluster with a high amount of human-tropic viruses. (b) Relative genomic abundance of viruses aggregated by host tropism (Bacteria, top; Eukaryote bottom). Color gradient (grey to red) corresponds to sugar ratio (low to high). {30}	83
<b>Figure 2.1.0.1</b> Clustering of infected cells reflects dynamics of cell populations during EHV1. (a) t-SNE representation of 13,156 PBMCs from mock (blue) and EHV1-infected (yellow) horses. Cell subtypes, listed beside clusters, were determined by marker genes after k-nearest neighbor clustering. Black diamonds represent cells with detectable amounts of EHV1 transcripts; size of diamond corresponds to relative viral transcript abundance. (b) Percentage of cells from mock (blue) and EHV1-infected (yellow) horses making up each cluster (numbers corresponding to groups in (a)). {31}	92
<b>Figure 2.1.0.2</b> Representation of viral genetic reassortment occurring when two genetically distinct viruses infect and replicate within a cell. Progeny viruses have a likelihood of encapsidating gene segments from both parent viruses, allowing for the formation of novel viral genotypes. {32}	95
<b>Figure 2.2.1.1</b> DART-seq is an easily implemented and multiplex technology for single cell studies. Drop-seq beads (with oligos containing a poly(dT) tail) are combined with a diverse mixture of toehold molecules including the targeting primer and a splint oligo with a polyA overhang. A DNA ligase is added to the suspension, which binds the toehold oligos to the Drop-seq beads. A light heat treatment (65 °C) is used to denature splint oligos which are subsequently washed away. The procedure retains all oligo information present on the original bead. {33}	100
<b>Figure 2.2.1.2</b> Qubit fluorometer measurements can resolve the number of fluorescent oligos bound to DART-seq beads. (a) To evaluate DART-seq ligation efficiency we designed an assay able to indirectly measure probe binding. (1) Toehold molecules are added to Drop-seq beads via the DART-seq conversion protocol. (2) Fluorescent oligos with sequence complementary to DART-seq probes are added to DART-seq beads. (3) The fluorescence of suspensions of 2000 beads are measured via Qubit 3.0 fluorometer in the 647 nm channel. (b) For four different toehold probes we compared the fluorescence (in A.U., arbitrary units) as a response to the number of fluorescent probes added to the Qubit measurement. {34}	101
<b>Figure 2.2.1.3</b> Conversion of Drop-seq beads to DART-seq beads is efficient and uniform, and DART-seq beads enrich targeted RNAs. (a) DART-seq beads were created from a mixture of four probes and complementary fluorescent oligos were bound to beads and measured by Qubit fluorometer. Dotted line represents 100% conversion efficiency. Inset: diagram of fluorescent oligos bound to beads. (b) DART-seq beads with bound fluorescent oligos were imaged using a fluorescence microscope and the average pixel intensity across 741 beads was determined. (c) Enrichment of targeted RNAs with respect to <i>Gapdh</i> , as measured by qPCR on cDNA from bulk RNA samples, is shown for various concentrations of probes added (10 <sup>6</sup> to 10 <sup>12</sup> probes per bead). (d) DART-seq was used to capture two viral mRNAs at seven loci and qPCR was performed at positions above the plot and compared to <i>Gapdh</i> . {35}	102
<b>Figure 2.2.2.1</b> DART-seq reveals heterogeneity in viral genotypes and host response to infection. (a) Experimental design. MOI, multiplicity of infection. b, Schematic of DART-seq designs (design-1, red bars; design-2, blue bars). c, Comparison of sequence coverage (normalized to host UMI detected × 10 <sup>6</sup> ) of the ten reovirus gene segments (columns) for three library preparations (rows). An A5 pentanucleotide sequence part of segment M3 is shown (arrow). Dotted lines, DART-seq target positions. d, Per-base coverage upstream (5' end) of ten custom primers of DART-seq design-1 (light red; average shown in dark red), and mean coverage achieved with Drop-seq (yellow). e, Per-base	

coverage of S2 gene segment achieved with DART-seq design-2 (bottom; dashed lines indicate custom primer positions) and Drop-seq (top). f, Frequency and pattern of base mutations (top); histogram of nucleotide ratios for positions with reference nucleotide G detected in single cells (bottom). g, Clustering analysis for variable gene expression of reovirus-infected L cells (DART-seq design-1; yellow/purple indicates higher/lower expression). Similar clustering was observed in all three experiments with infected cells. h, Relative abundance of viral transcripts in L cell clusters (P values determined by two-tailed Wilcoxon rank-sum test). Lower and upper hinges correspond to 25th and 75th percentiles, respectively. Lower/upper whisker corresponds to smallest/largest value within 150% of the interquartile range from the nearest hinge (cluster 1, n = 411; cluster 2, n = 397; cluster 3, n = 50; cluster 4, n = 69). i, Fraction of cells in metaclusters for four experiments depicted in panel a with assay type and infection status (+ or -) indicated. {36} ..... 105

**Figure 2.2.3.1** DART-seq measures paired heavy and light chain B cell transcripts at single-cell resolution. (a) DART-seq custom primer design targeting the constant region of human heavy and light isotypes. (b) cDNA copies of immunoglobulin (Ig) transcripts relative to *GAPDH* as a function of the number of custom primers included in the ligation reaction (left panel, LC- $\lambda$  + V primers; right panel, IgG + V primers; 62,500 cells, 12,000 beads, bulk assay). Points are mean of two replicate measurements; bars indicate minimum and maximum. (c) Percentage of B cells for which heavy and/or light chain transcripts were detected as a function of the UMI count per cell. Cells were binned by the number of UMI detected (bin width 200 UMI, 0–2,400 UMI per cell, bins with fewer than 20 cells omitted, 26–2,396 cells per bin). Distributions were fit with a sigmoid curve (Methods). (d) Drop-seq and DART-seq assays of human PBMCs. Experiments were performed on two distinct PBMC samples (n = 2). Representative t-SNE for one DART-seq assay shown here (4,997 single cells). Cells are colored on the basis of heavy and/or light chain transcript detection. € Bar graph of isotype distribution for CD27+ B cells and B cells for which CD27 was not detected. (f) CDR3L and CDR3H length distribution. n = 818 B cells. (g) Paired heavy (IGHV) and light (IGKV and IGLV) variable chain usage in B cells; n = 164 single cells. {37} ..... 108

**Figure 2.3.1.1** Targeted sequencing is not necessary for all viral transcripts. (a) The number of A(5) repeats per kilonucleotide (knt) is shown for three dsRNA, segmented viruses. (b) Mean normalized sequencing coverage across eleven rotavirus segments was calculated based on experimental results. {38} ..... 117

**Figure 2.3.1.2** Rotavirus infected fibroblast quality analysis in single cells sequencing. (a-d) For Mock, Low MOI (MOI 0.1), and High MOI (MOI 5.0) single-cell sequencing analysis revealed the number of (a) UMI per cell, (b) genes per cell, (c) percentage of transcripts from the mitochondria, and (d) the percentage of transcripts from rotavirus. (e) Sequencing coverage across the rotavirus genome segments (ordered largest to smallest) 25 cells with the highest percentage of viral reads. Traces of each cell are in grey and mean of the traces is shown in black. Scale is square-root transformed. {39} ..... 119

**Figure 2.3.1.3** Viral gene transcription is altered through infection progression. Viral mRNAs are averaged in groups of thirty cells after ordering by the percent of transcripts of viral origin in each cell. The fractional abundance of each transcript in the cell bin (30 cells) is shown as a stacked barplot. Line indicates the fraction of reads originating from the virus in each cell bin. Dotted line represents the cutoff of uninfected cells. Bar color represents viral gene segment (annotated on right). {40} ..... 121

**Figure 2.3.2.1** Dissociation of organoids leaves cellular aggregates but represents all cell types. (a) 10x phase contrast image of dissociated enteroid on Fuchs-Rosenthal hemocytometer. (b-c) The diversity of cell types is observed by morphological features, as shown at 40x by the presence of (b) goblet cells and (c) remnants of crypts. Images taken on Zeiss Axio Observer Z1 under phase contrast. {41} ..... 123

**Figure 2.3.2.2** Differential gene expression and clustering analysis of low quality cells from enteroids. (a) The percentage of transcripts originating from the mitochondria is compared to the total number of genes expressed. Dotted lines represent cutoffs for filtering. Density plot overlaid on cells (colored by experiment). (b,c) t-SNE dimensional reduction of cell-cycle regressed gene expression. (b) Colors indicate Mock and Infected datasets. (c) Colors indicate increased expression of *Sis*, *Lyz1*, *Chga*, and *Tff3*. Intensity of color corresponds to intensity of gene expression. (1) Enterocyte cluster. (2) enteroendocrine + goblet cell cluster. (3) Paneth cell + goblet cell cluster. {42} ..... 124

## LIST OF TABLES

<b>Table 1.6.0.1</b> Summary of datasets acquired using cfDNA sequencing with ssDNA library preparation. cfDNA sequencing datasets acquired in the De Vlaminck lab, according to: disease, interest, host organism, fluid from which biological fluid was extracted, country of sample origin, infectious organism superkingdom, whether cfDNA from the organism was general detectable, and the number of samples processed. ....	86
--	----

## LIST OF ABBREVIATIONS

AR = antimicrobial resistance

AUC = area under the curve

BKV = BK polyomavirus

BKVN = BK polyomavirus nephropathy

bp = basepair (unit)

BT = bisulfite treatment (bisulfite treated)

CDR3 = complementarity-determining region 3

cfDNA = cell-free DNA

cfRNA = cell-free RNA

CFU = colony forming units

CMV = cytomegalovirus

CoNS = coagulase-negative *Staphylococcus*

DART-seq = droplet-assisted RNA targeting by single-cell sequencing

(d)dPCR = (droplet) digital PCR

DMR = differentially methylated region

dsDNA = double-stranded DNA

EE = environmental enteropathy

EHV1 = equid herpesvirus 1

ETH = ethambutol (antimicrobial drug)

FFT = Fast Fourier Transform

HPV = human papillomavirus

Ig/IG = immunoglobulin

IZD = isoniazid (antimicrobial drug)

LBBC = low-biomass background correction

MOI = multiplicity of infection

MTB = *Mycobacterium tuberculosis*

mt-cfDNA = mitochondrial cell-free DNA

mt-DNA = mitochondrial DNA

MXF = moxifloxacin (antimicrobial drug)

nt = nucleotide (unit)

PABP = poly-A binding protein

PBMC = peripheral blood mononuclear cell

PCR = polymerase chain reaction

PD = peritoneal dialysis

RdRp = RNA-dependent RNA polymerase

RGE = relative genomic equivalent

RIF = rifampicin (antimicrobial drug)

ROC = receiver operating characteristic

scRNA-seq = single-cell RNA sequencing

SNP = single nucleotide polymorphism

ssDNA = single-stranded DNA

STH = soil-transmitted helminth

t-SNE = t-stochastic neighborhood embedding

Tx = transplant

UMAP = uniform manifold approximation and projection

UMI = unique molecular identifier

UTI = urinary tract infection

V/D/J/C = variable / diversity / constant / joining (in context of immune repertoire)

VRE = vancomycin resistant *Enterococcus*

WBC = white blood cell

WGS = whole genome sequencing

## Introduction

In October 2014, I attended an evening lecture given by Alfonso Torres, a former Associate Dean at Cornell University College for Veterinary Medicine, where he described West African Ebolavirus outbreak. Alfonso's presentation expressed deep concern about the epidemic, but also described the heroism of veterinary and medical assistants who were performing science in the field. Underlying his talk was the success analysis using molecular epidemiology – the use of molecular detection to infer the spatiotemporal dynamics of epidemics. I was absorbed by analysis of the viral genetics over time and geographical space, and became interested in using genome sequencing to surveil pathogens. Around the same time, I was fortunate to discover a lab where I could combine my joy in performing benchtop research and my interest in applied mathematics with Iwijn De Vlaminck. My work in the De Vlaminck lab has focused on two separate aspects of genomics as applied to understanding infectious disease. These areas of study will make up the major two parts of this dissertation.

In Part I, I aim to describe the field of cell-free diagnostics as it was when I started my research in 2015. Prior to that time, few studies have analyzed the presence of non-host cell-free DNA (cfDNA) in bodily fluids. Work by De Vlaminck et al. determined the presence of an abundant virome observable in the plasma cell-free DNA of lung and heart transplant recipients [1–3]. The sequencing data from this work was brought into the newly-formed De Vlaminck lab and led to an investigation on using donor and recipient identified cell-free DNA molecules to predict cases of rejection from mitochondrial cfDNA [4]. As we discovered through means detailed in this dissertation, these molecules are much smaller than cfDNA fragments derived from chromosomes, owing mostly to the lack of protection from histone octamers and large, persistent transcription factors.

We designed a method with the ability to capture and sequence short fragments of mitochondrial cell-free DNA. The technique also captured cell-free DNA fragments from other non-nucleosome bound sources (Part I, Chapter 2). We realized the potential of this new assay to identify viral and bacterial pathogens alongside host cell-free DNA, and expanded earlier studies of plasma cell-free DNA in lung

transplant recipients [4] to include cell-free DNA extracted from urine and peritoneal dialysis effluent in kidney transplant recipients [5] (Part I, Chapter 3). Furthermore, these studies have enabled us to develop analytic tools to overcome shortcomings inherent to cfDNA sequencing, such as background contamination and cell-free DNA originating from microbial sources with annotated reference genomes (Part I, Chapter 4). We also show that exploratory datasets derived from plasma cell-free DNA from individuals in developing nations has yielded insights into cell-free DNA diagnostics as a tool for molecular epidemiology genomes (Part I, Chapter 5). Recently published work from Karius, Inc. have shown the potential in using microbial cell-free DNA sequencing as a rapid diagnostic tool, able to return reports of identified pathogens in 24-36 hours back to the consumer [6]. I believe our advances, coupled with the progress made in the private sector, will lead to microbial cell-free DNA sequencing as a clinically useful diagnostic in the coming years.

Part II of this thesis focuses on the insights gained by analyzing transcriptomes of single cells infected with viruses. Single-cell transcriptomics and genomics is a nascent field, with the first sequencing experiments performed and published in 2009. Since then the ability to simultaneously and affordably (< \$0.10 per cell) prepare sequencing libraries for tens of thousands of cells has been achieved [7]. These methods, however, are often unable to capture non-polyadenylated or non-5'-end capped transcripts, the signatures of eukaryotic messenger RNAs. We identified the need to enrich for transcripts that current gold-standard technologies fail to capture, while still using the framework that these technologies have advanced. To meet our goal, we created a cheap and multiplexable technology to capture dozens of targeted amplicons alongside polyadenylated messenger RNAs, allowing, in principle, the sequencing of any RNA species, regardless of origin or property [8] (Part II, Chapter 2).

I will describe the validation of this new technique, as we applied it to infections with mammalian orthoreovirus in cell culture and provide evidence of its use in other areas analysis of host-pathogen interaction (such as immunology). We have recently expanded this work to include other viral systems, in particular, orthoreovirus in complex intestinal organoids and rotavirus in cell culture (Part II, Chapter 3).

Through analysis at the single-cell level, interesting phenomena can be observed during the infection cycle that otherwise would not be observed in bulk RNA studies. In summary, the proof of our new method and follow-up studies provide a new tool to understand innate and adaptive immunity during infection.

While these two areas of study, single-cell ‘omics and cell-free nucleic acid diagnostics, are somewhat disparate, they overlap in both molecular biology techniques used to generate libraries for sequencing and the bioinformatics principles used to organize and analyze the sequencing data. Moreover, they reflect a similarity in my own curiosity and motivations when we are pursuing novel genomics technologies, namely, answering the following questions: “what information are we missing?” and “what elegant methods are available to capture it?”. Recent works from some labs around the world are utilizing single-cell ‘omics technologies to determine novel biomarkers of disease, accessible in the form of nucleic acids circulating through the body. The last portion of the dissertation will summarize my work and describe new technologies to diagnose infections using single-cell and single-molecule sequencing technologies.

Published works referred to in this dissertation:

Burnham P, Kim MS, Agbor-Enoh S, Luikart H, Valantine HA, Khush KK, De Vlaminc I. Single-stranded DNA library preparation uncovers the origin and diversity of ultrashort cell-free DNA in plasma. *Scientific Reports*. 2016. (Part I, Chapter 2).

Burnham P, Dadhania D, Heyang M, Chen F, Westblade L, Suthanthiran M, Lee JR, De Vlaminc I. Urinary cell-free DNA is a versatile analyte for monitoring infections of the urinary tract. *Nature Comm*. 2018. (Part I, Chapter 3).

Saikia M\*, Burnham P\*, Keshavjee SH, Wang MFZ, Heyang M, Moral-Lopez P, Hinchman MM, Danko CG, Parker JSL, De Vlaminc I. Simultaneous multiplexed amplicon sequencing and transcriptome profiling in single cells. *Nature Methods*. 2019. (Part II, Chapter 2).

\* indicates equal contribution.

## PART I: Cell-free DNA sequencing to monitor infection

“Once the cells in a biological machine stop working, it can never be started again. It goes into a cascade of decay, falling toward disorder and randomness...”

— Richard Preston, *The Hot Zone: The Terrifying True Story of the Origins of the Ebola Virus*

## Chapter 1: Physiological and historical origins of cell-free DNA

Insights into the origin of cell-free DNA (cfDNA) and its association with disease began before there was true knowledge of the medium itself. The observation of cell-free DNA predated the publications of the double helix structure of DNA by Watson, Crick, and Franklin in 1952-1953. Mandel and Metais first published their discovery in 1948, recognizing the presence of nucleic acids in plasma extracted from the blood of individuals with various medical conditions [9]. Earlier work had described the presence of nucleic acids in blood, and researchers had optimized the techniques used to extract DNA from biological fluid [10]. However, Mandel and Metais recognized that these works likely misrepresented discoveries of cfDNA and instead captured whole genomic DNA from white blood cells (WBCs). By extracting plasma, and thus depleting the biological fluid of blood cells, they showed the presence of significant amounts of nucleic acids in the plasma – roughly 4.7 ng per mL of plasma – across ten healthy patients.

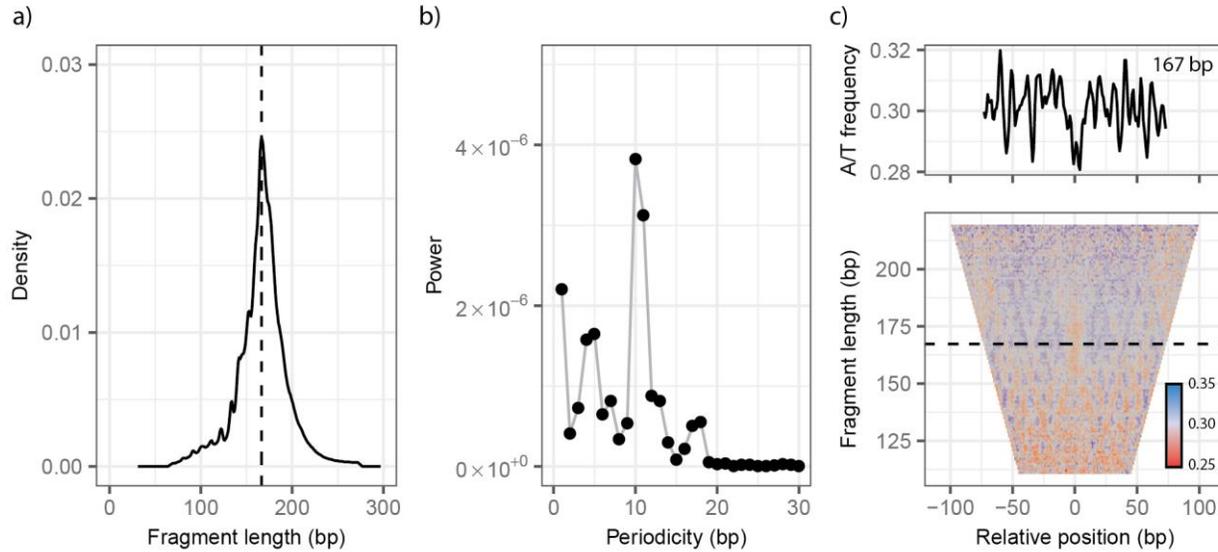
Mandel and Metais also extracted cell-free nucleic acids for fifteen patients with known pathologies or conditions (including pregnancy, autoimmune disorders, and tuberculosis). While the sample size for each of these conditions was small (n of 1 or 2), their importance and foresight cannot be overstated. In the case of a 23-year-old, seven-month pregnant female, the detected amount of nucleic acid in plasma was six standard deviations above the mean of non-pregnant patients.

Half a century later, Lo et al. followed up on Mandel and Metais' observation of elevated cell-free DNA in pregnancy, seeking to determine the origin of the cfDNA molecules in plasma [11]. The researchers isolated cfDNA from serum, plasma, and whole blood from women who were at least twelve weeks pregnant. A polymerase chain reaction (PCR) assay was designed to amplify a fragment of DNA from the Y chromosome [11]. It follows that such an amplicon should only be observed in women pregnant with male fetuses, and would be indicative of cfDNA originating from the fetus. For 24 of 30 samples with male fetuses, Y chromosome cfDNA was detected in plasma (in 21 of the 30 samples for serum). Importantly, the signal was not observed in the plasma of women pregnant with female fetuses.

These observations of fetal cfDNA in maternal plasma preempted the understanding that vascularized tissues within the body can release cfDNA into surrounding blood vessels. Furthermore, the work by Lo et al. effectively moved cfDNA research from a simple observation into a biomarker by which we can monitor disease. Diagnostic tests have been created to detect trisomy 21 and other genetic abnormalities early on in pregnancy [12]. These tests are now frequently used in clinics to determine fetal abnormalities. Later work has shown that cfDNA could be used to inform oncogenesis and as a predictive biomarker of transplant rejection [2, 3, 13, 14]. Both of these diseases often require tissue biopsy for diagnostic confirmation – procedures that are expensive, invasive, and often lack the ability to reflect tissue heterogeneity. The sequencing of cell-free DNA to determine such maladies permits a fast and noninvasive test that avoids some of the technical challenges of standard biopsies. For example, cfDNA from cerebrospinal fluid has indicated the polygenetic composition of a glioma [15]. This identification using standard methods would have otherwise required open brain surgery to remove, dissociate, and sequence the whole tumor.

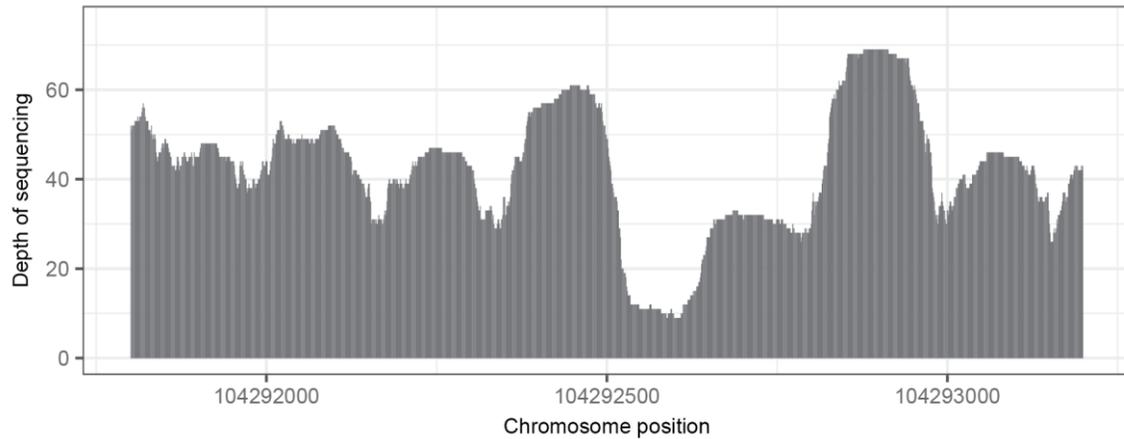
Since 1947, our understanding of the properties of cfDNA has increased along with our appreciation of its utility. It is now understood that cfDNA enters the circulatory system through a variety of mechanisms including apoptosis, necrosis, pyroptosis, and, in some cases, active release [16]. In blood circulation, as was first speculated, dying neutrophils and leukocytes release the majority of cell-free DNA in the blood of healthy individuals. However, other basic studies in the field have found that certain diseases lead to higher amounts of cell-free DNA into the bloodstream from inflammation or injury. For example, individuals suffering traumatic injury and systemic inflammation have shown high amounts of cell-free DNA in the days following disease onset [17, 18]. Once in the circulation, cfDNA is degraded via circulation exonucleases and or immune cells. Naked DNA (i.e. that lacking bound proteins) is degraded at a much faster rate than fragments occupied by proteins [19].

The captured fragments of cell-free DNA statistically represent those with longer lifetimes in the biological fluid in which they are released. Since naked cell-free DNA is degraded by environmental



**Figure 1.1.0.1** cfDNA properties strongly indicate nucleosomal origin. (a) Histogram of fragment length distribution for a plasma cfDNA sample prepared using standard, commercial library preparation. (b) Fourier transform was performed on distribution in (a) between 60 and 170 bp, peak indicates most prominent periodicity. (c) A dinucleotide heatmap (colored by mean A/T frequency) is shown for cfDNA fragments of varying length, relative to central position. (A/T) dinucleotide combinations are scored with a value of 1, other combinations are scored with value 0, heatmap reflects mean score. Top: Cross-section of heatmap shown above for 167 bp. In (a) and (c) dashed line corresponds to most abundant fragment length, 167 bp. {1}

nucleases, those fragments bound to proteins are often protected from degradation. This stabilization leads to, on average, a higher abundance of protein-bound cfDNA in datasets. Evidence of this protection is immediately observable when plotting a histogram of the fragmented lengths of sequenced cfDNA. If only naked DNA were released into the environment and nonspecifically degraded, one would expect an exponentially decaying distribution of fragment lengths. Instead, we observe a strong peak at ~167 bp for cfDNA fragment lengths originating in blood plasma (Fig. 1.1.0.1a). The presence of this peak is explained by the presence of nucleosomes bound to nuclear eukaryotic DNA. Nucleosomes consist of an octamer of four histone proteins to which 145 bp of DNA are wrapped, and an H1 linker histone is then added between nucleosomes [12]. Moreover, the distribution of cfDNA fragment lengths below 167 bp show periodic local maxima roughly every 10 bp (Fig. 1.1.0.1b, Fast Fourier Transform of distribution from Fig. 1.1.0.1a between 60 and 170 bp after background removal). This pattern suggests that the local enrichment is based



**Figure 1.1.0.2** Patterns of cfDNA fragmentation observed by sequencing depth along reference genome. Sequencing coverage of human aligned cfDNA originating from plasma across a small section of the genome for 140-180 bp fragments. Plasma cfDNA prepared for sequencing using standard, commercial library preparation. Positions illustrated are from human chromosome 12. {2}

on protection from the interaction between DNA and the nucleosome core particle with partial degradation at the edges of the molecules [20].

Proteins bound to DNA often do so in a specific manner, determined by the physical properties of nucleotides that act as a signal [21]. This results in the observation of sequence motifs [22]. If one selects human aligned sequences from plasma cfDNA at a particular fragment length, a bias to A/T becomes apparent at positions  $0 \pm 10n$  from the center of the molecule, where  $n$  is an integer value (Fig. 1.1.0.1c, top). We can visualize this bias by plotting the dinucleotide frequency across all molecules of various lengths (Fig. 1.1.0.1c, bottom). Histones have a strong affinity to the minor groove of particular DNA motifs, particularly AAA/TTT and AAT/TTA [20, 22]. Again, this observation supports the reflection of intrinsic helicity of nucleosome-bound DNA (10.4 bp to complete an internal rotation of the DNA double helix). The observations of partial protection and corresponding AT-rich sequences at  $\sim 10.4$  bp intervals for molecules below 167 bp suggest that many sequences are derived from nucleosomes.

cfDNA fragments can also be directly mapped to the human genome to observe the nucleosomal origin of cell-free DNA. This observation requires a high depth of sequencing (i.e. many reads covering a random genomic position, on average), but is striking. Figure 1.0.1.2 shows the coverage for 140-180 bp, blood plasma cell-free DNA). By observing a genomic region which is known to have strong nucleosome

signals (Fig. 1.1.1.2, Chr12: 104,291,750-104,293,250), we can observe the presence of peaks separated by 210 bp, corresponding well to the expected distance between nucleosomes [23]. Conversely, in regions of the genome that are expected to be less compacted for enzymatic accessibility (e.g. a transcription start site of a housekeeping gene), one can see a gap in the nucleosomal occupancy, and a high ordering of nucleosomes surrounding the transcription start site [24].

Because of the strong nucleosomal signals present in cfDNA sequencing data, it could be assumed that other non-nucleosomal sources are of limited scope from their low signal. In fact, cell-free DNA has been observed from non-nucleosomal sources, including from the host and microbial organisms living within the host [1, 23, 25]. In Part I of this dissertation, I will describe the development of a novel library preparation protocol that has allowed researchers to capture and analyze non-nucleosomal cfDNA, which tends to be very short (< 100 bp), and is not captured by traditional commercial methods. Utilizing a ssDNA library preparation of cfDNA has allowed us to expand cfDNA sequencing into underrepresented biological fluids in cfDNA studies, including urine, peritoneal dialysis effluent, and amniotic fluid. Most importantly, we have used this novel library preparation to evaluate cfDNA sequencing as an analytical pipeline to determine viral and bacterial infections. I will show the efficacy of cfDNA as a comprehensive analyte to monitor infectious disease.

## Chapter 2: Single-stranded DNA library preparation uncovers the origin and diversity of ultrashort cell-free DNA in plasma

“Circulating cell-free DNA (cfDNA) is emerging as a powerful monitoring tool in cancer, pregnancy and organ transplantation. Nucleosomal DNA, the predominant form of plasma cfDNA, can be adapted for sequencing via ligation of double-stranded DNA (dsDNA) adapters. dsDNA library preparations, however, are insensitive to ultrashort, degraded cfDNA. Drawing inspiration from advances in paleogenomics, we have applied a single-stranded DNA (ssDNA) library preparation method to sequencing of cfDNA in the plasma of lung transplant recipients (40 samples, six patients). We found that ssDNA library preparation yields a greater portion of sub-100 bp nuclear genomic cfDNA, and an increased relative abundance of mitochondrial and microbial cfDNA. The higher yield of microbial sequences from this method increases the sensitivity of cfDNA-based monitoring for infections following transplantation. We detail the fragmentation pattern of mitochondrial, nuclear genomic and microbial cfDNA over a broad fragment length range. We report the observation of donor-specific mitochondrial cfDNA in the circulation of lung transplant recipients. A ssDNA library preparation method provides a more informative window into understudied forms of cfDNA, including mitochondrial and microbial cfDNA and short nuclear genomic cfDNA, while retaining information provided by standard dsDNA library preparation methods.”

Chapter adapted from [4]:

Burnham P, Kim MS, Agbor-Enoh S, Luikart H, Valantine HA, Khush KK, De Vlaminck I. Single-stranded DNA library preparation uncovers the origin and diversity of ultrashort cell-free DNA in plasma. *Scientific Reports*. 2016.

### 1.2.0 A single-stranded ligation strategy to enrich ultrashort cfDNA

Cell-free DNA exists in circulation in many shapes and forms, including as fragments of the nuclear genome, the mitochondrial genome and microbial genomes [25]. The predominant type of cfDNA is derived from the nuclear genome and has a fragment size centered around 167 bp, approximately the length of a segment of DNA wound around a histone octamer [26, 27]. These nucleosomal fragments of cfDNA are readily accessible for sequencing using standard library preparation methods that are based on ligation of dsDNA sequencing adapters. The most commonly used implementations of this method rely on multiple bead-based size-selective steps that eliminate unwanted adapter-dimer products.

If the host-aligned (i.e. human-aligned) cell-free DNA is separated by chromosome, cell-free DNA originating from the mitochondria may be observed. The evolution of eukaryotic cells and the eventual production of multicellular life rested on integration of a bacterium inside of another cell [28]. Due to the bacterial origin of the mitochondrial organelle, these genomes lack nucleosomal packaging and are more rapidly degraded when removed from the cell [29]. We hypothesized that current library preparation techniques removed many cfDNA molecules that were non-nucleosomal in origin, including mitochondrial cfDNA, in an effort to eliminate adapter-dimer products from sequencing libraries. Adapter-dimer products are often the same size or larger than degraded cfDNA, making it difficult to separate the two using size selection. These methods, although relevant to a wide range of applications, are not sensitive to the full diversity of circulating cfDNA [30]; in particular shorter fragments, highly degraded fragments, nicked dsDNA and single-stranded fragments of DNA in circulation remain undetected.

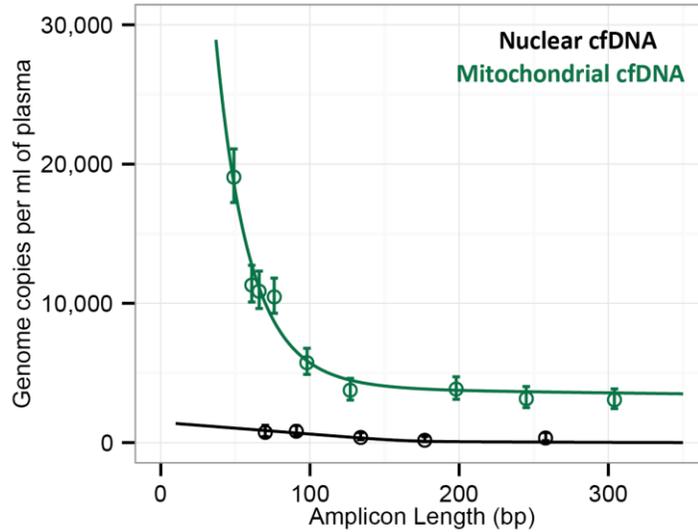
We confirmed an abundance of short mitochondrial cell-free DNA molecules via digital PCR (dPCR) assays and subsequently sought to establish a library preparation protocol sensitive to short cfDNA. A parallel exists with genomic analyses of ancient DNA samples, where the target DNA is often in low abundance and less than 50 bp in length. To address size and abundance issues, researchers introduced a sequencing library preparation technique based on the binding of single-stranded DNA molecules to magnetic beads for enzymatic extension and cleanup [31]. The technique was demonstrated to produce high

quality, low contamination sequencing reads across a Denisovan individual from antiquity [32]. This single-stranded ligation is, in principle, sensitive to the full diversity of cfDNA in the circulation, including ultrashort (< 100 bp) dsDNA, ssDNA and dsDNA with nicks in both strands.

We adapted this protocol and applied it to plasma cfDNA originating from patients receiving bilateral lung transplants. To evaluate the performance of the ssDNA library preparation method, we directly compared data of fragment types, lengths and abundance to results from conventional library preparations performed on the same plasma DNA extracts [1, 3]. Transplant recipients are subject to immunosuppressive therapies that reduce the risk of rejection, but increase their susceptibility to opportunistic infections. Analyses of microbial cfDNA in plasma are therefore particularly relevant in the context of transplantation. Here, we examined the yield of microbial cfDNA that results from the ssDNA and conventional library preparations. Donor-specific nuclear genomic cfDNA is present in the circulation of solid-organ transplant patients and is a marker of transplant rejection [2, 33]. In this study, we used a single-stranded library preparation to study the properties of donor and recipient specific cfDNA across a wide length range. To test whether donor-specific mitochondrial cfDNA can be found in the circulation of transplant recipients, we directly compared data of cfDNA to reference sequences of amplified mitochondrial genomes obtained from pre-transplant samples.

### 1.2.1 Mitochondrial cfDNA in plasma measured by digital PCR

This study was prompted by a retrospective analysis of sequencing data of cfDNA in plasma of transplant recipients available from a previous study [3], which revealed a fractional abundance of mitochondrial cfDNA of  $2 \times 10^{-3}$  %, which is in line with a recent observation [25], but is low considering that there are 50-4,000 mitochondrial genomes per cell [34]. We used digital PCR (dPCR) assays with varying amplicon length (49-304 bp) to assess the abundance of mitochondrial cfDNA prior to library preparation and compared this to the abundance of nuclear genomic cfDNA (Fig. 1.2.1.1). Whereas qPCR



**Figure 1.2.1.1** Digital PCR measurements reveal an abundance of ultrashort cfDNA originating from the mitochondria. Number of genomic copies (per milliliter of plasma) is shown as a function of amplicon length for cfDNA originating from the nucleus (black, five amplicons) and mitochondria (green, nine amplicons). [3]

requires a dilution series of a template oligo for each primer set in order to determine the concentration of a template, digital PCR provides a reference-free technique in which tens of thousands of PCR assays are carried out simultaneously within femtoliter-sized droplets [35]. It is expected that each droplet will contain zero, one, or countably few template molecules. The template in each droplet can then be amplified and a fluorescent dye (such as SYBR green) is bound to DNA to discern template-positive and negative droplets. The Poisson model is applied to determine the initial concentration of the template based on the fractional amount of template positive droplets [35].

The experimental design with variable amplicon lengths provided information about the underlying fragment length distribution [26]. The genomic abundance of cfDNA, as measured by PCR, is expected to decrease monotonically with amplicon length, with a gradient that is a function of the underlying fragment length distribution. It can then be calculated that the relative fraction of target molecules of length  $x$ , detected in a PCR assay with amplicon of length  $L$ :

$$f(x, L) = \frac{x - L + 1}{x} : L \leq x .$$

And the abundance of a template measured via the digital PCR assay then corresponds to:

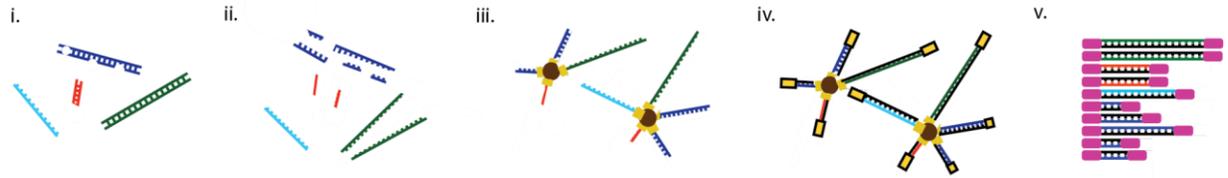
$$A_{PCR}(L) = \sum_{k=L}^{\infty} G(k)f(k, L),$$

where  $G(k)$  is the underlying fragment length distribution (expressed as a density). When fitting the observed dPCR abundance with the model it is apparent the nuclear cell-free DNA fragments can be modeled as a Gaussian (mean = 165 bp, s.d. = 20 bp) and the mitochondrial fragments modeled as an exponential decay.

These experiments revealed that mitochondrial cfDNA is more fragmented than nuclear genomic cfDNA, but present in much greater abundance in plasma (56-fold greater representation, genome equivalents). The exponential shape of the curve indicates that further enrichment could be achieved below the minimum amplicon length used in this study. However, the ability to amplify, sequence, and map molecules shorter than 40 bp is technically challenging. The consequence of the short fragment size of mtDNA is that conventional dsDNA library preparation protocols, which require multiple bead-based size-selective steps to eliminate unwanted adapter-dimer products, are relatively insensitive to mitochondrial sequences. We generalized this finding to hypothesize that all cfDNA of non-nucleosomal origin is likely undersampled by gold-standard preparation and sequencing approaches.

### 1.2.2 ssDNA library preparation and fragmentation profiles

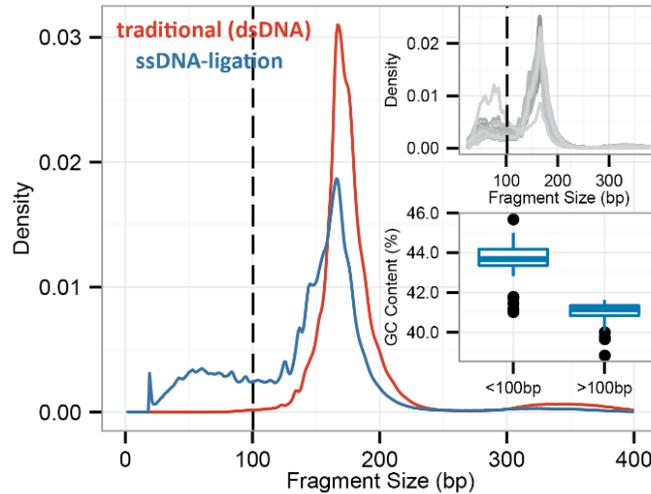
We implemented a ssDNA library preparation protocol first described by Meyer et al. that does not require size-selective steps that eliminate shorter fragments [36] (Fig. 1.2.2.1 and described in Section 1.2.5). We used paired-end sequencing to determine the fragment lengths of nuclear, mitochondrial, and microbial cfDNA (see Section 1.2.5). We found that cfDNA shorter than 100 bp becomes more accessible for sequencing following ssDNA library preparation. The lower limit of efficient capture, as shown by the local maxima of the short fragment cfDNA, for the ssDNA library preparation, was 40-60 bp (for all



**Figure 1.2.2.1** Schematic overview of single-stranded library preparation for cfDNA. (i) cfDNA from biofluid is extracted and is highly heterogeneous with respect to degradation patterns. (ii) cfDNA is denatured at 95 °C. (iii) single-stranded cfDNA is ligated to biotinylated adapters and bound to streptavidin-coated magnetic beads. (iv) Bound ssDNA is end-repaired, extended, and adapted with double-stranded sequencing primers. (v) Sequencing libraries are PCR amplified and indexed. Final libraries represent degradation diversity present in original set. {4}

subclasses), pointing to a limit set by the DNA isolation method. Ancient DNA molecules 30 bp in length were sequenced using a similar preparation, so it is possible to obtain shorted fragments depending on isolation method [32].

The peak in the length profile at 160-167 bp for cfDNA fragments assigned to the nuclear genome (Fig. 1.2.2.2) is a consequence of the protection of these molecules from degradation by nucleases in the blood through tight association with histones. This property has been reported in previous studies and is observed for both the ssDNA and dsDNA library preparation protocols [37]. A second peak at shorter lengths (< 100 bp) is unique to the libraries prepared by single-stranded ligation [2]. The relative proportion of nuclear genomic DNA shorter than 100 bp made up a substantial proportion of nuclear cfDNA (20.54%  $\pm$  11.51%). We partitioned cfDNA prepared via ssDNA ligation into two groups, those with length under and over 100 bp, to examine distinguishing features. The GC content between the groups differed significantly (Fig. 1.2.2.2, inset;  $p \ll 10^{-5}$ , Mann-Whitney U Test); the GC content of the super-100 bp group was 40.9%, while that of the sub-100 bp group was 43.5%. These observations confirm previous observations of high AT density for nucleosome associated cfDNA molecules. Furthermore, the data indicates that a considerable amount of cfDNA originating from the nucleus is not nucleosome protected and, thus, subject to degradation by nucleases in the blood. Previous reports suggest that fetal- and tumor-derived cfDNA are shorter than cfDNA derived from maternal [38] and normal [25] tissue, respectively.

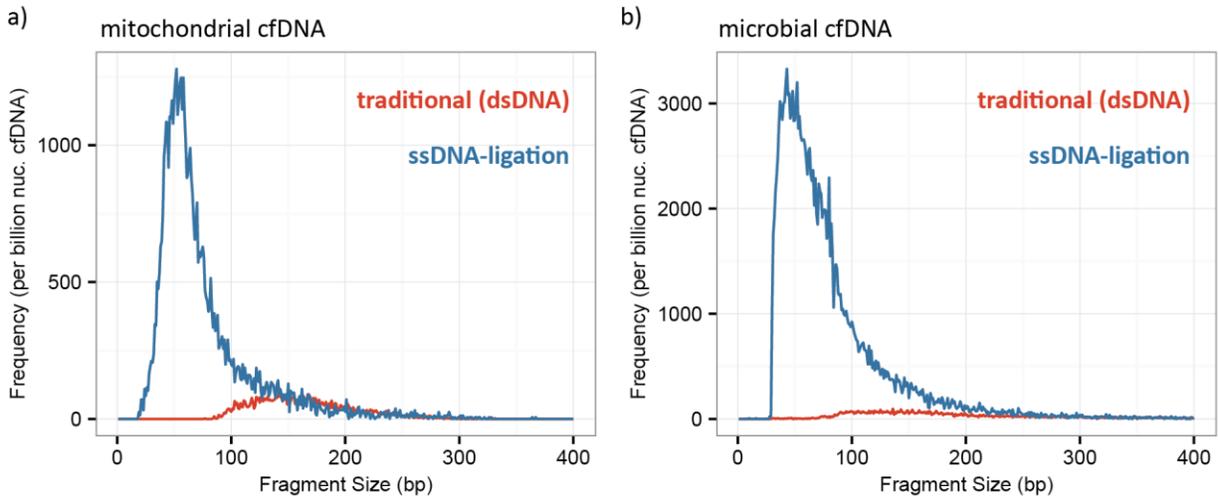


**Figure 1.2.2.2** Single-stranded library preparation reveals an abundance of ultrashort cfDNA molecules. Density plot of the fragment sizes of nuclear genomic cfDNA measured after ssDNA (blue) and dsDNA (red) library preparation. The inset shows the sample-to-sample variability ( $n = 40$ ), as well as the difference in GC content for short ( $< 100$  bp) and long ( $> 100$  bp) fragments. Significance of  $p \ll 10^{-5}$ , Mann-Whitney U Test. Boxplot description given in section 1.3.9. {5}

The sensitivity of the ssDNA library preparation protocol to molecules over a wider length range is therefore a feature that will be useful for applications in prenatal testing and tumor monitoring.

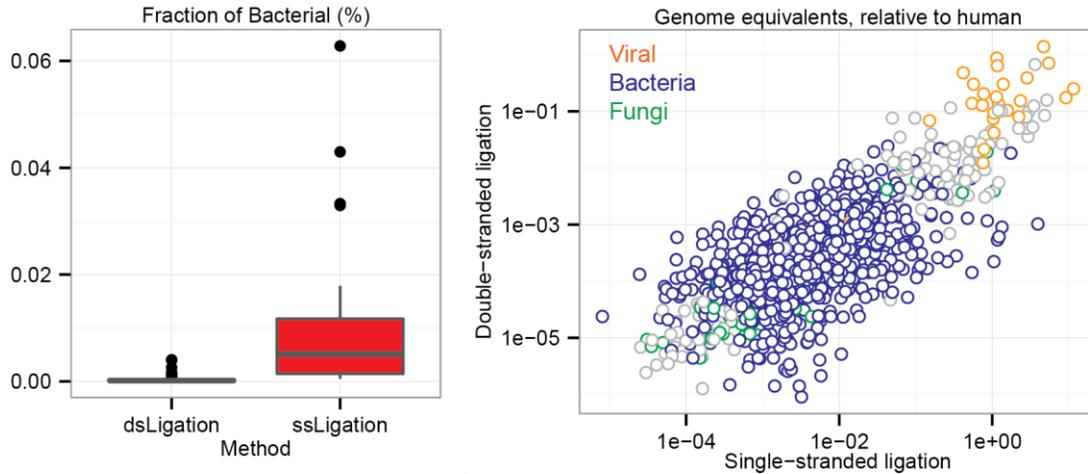
### 1.2.3 Improved recovery of mitochondrial and microbial cfDNA

We next examined the coverage of mitochondrial and microbial genomes relative to the nuclear genome for the dsDNA and ssDNA library preparations. While conventional library preparation resulted in detection of only a few molecules of mitochondrial and microbial cfDNA with length shorter than 100bp, the use of a ssDNA library preparation revealed an abundance of such molecules with lengths between 40 and 100 bp. We found that the ssDNA library preparation gives rise to an increase in the relative number of mitochondrial sequences in the datasets (10.7 fold mean increase,  $p \ll 10^{-5}$ , Mann-Whitney U test) and an increase in the relative coverage of the mitochondrial genome (7.22 fold mean increase,  $p \ll 10^{-5}$ , Mann-Whitney U test; Figure 1.2.3.2). This observation is consistent with the greater sensitivity of the ssDNA library preparation to short fragment DNA described above.



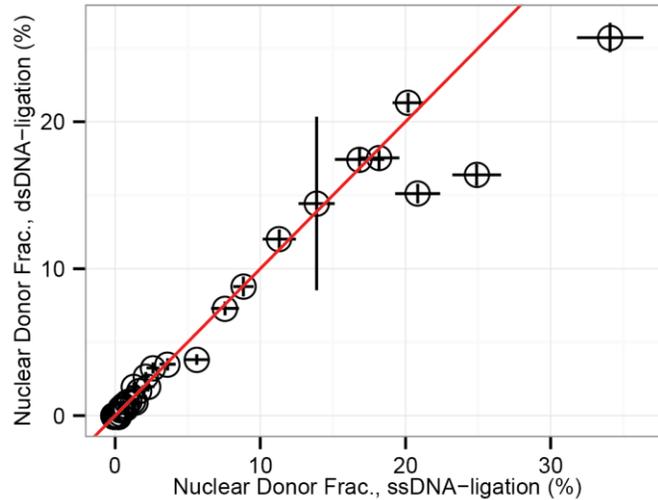
**Figure 1.2.3.1** ssDNA ligation based library preparation reveals enrichment for microbial and mitochondrial cell-free DNA. Fragment length distribution for matched samples using ssDNA ligation and traditional library preparation techniques is compared, as normalized to nuclear cfDNA molecules, for (a) mitochondrial-aligned and (b) microbial-aligned fragments. {6}

A direct comparison of the fraction of sequencing sets made of microbial reads also showed much higher abundance in the instances of using ssDNA library preparation, compared to standard techniques (Fig. 1.2.3.2b). To study the efficiency of recovery of microbial cfDNA, we estimated the genome coverage of each microbe detected across all samples relative to the coverage of the human genome (see Section 1.2.5). We compared the relative genomic coverage of strains or subspecies detected by both methods in matched samples ( $n = 36$ ). We examined over 1,100 direct comparisons and found a significant correlation ( $\text{corr.} = 0.6373$ , Spearman,  $p \ll 10^{-5}$ ) in the relative genomic abundance as measured following ssDNA and dsDNA library preparation (Fig. 1.2.3.3b); the range in the ratio was  $0.277x - 3950x$ , indicative that for most species the ssDNA method led to more efficient detection ( $p \ll 10^{-5}$ ; Mann-Whitney U Test). Importantly, library preparation by ssDNA ligation gave rise to a mean 71-fold increase in the relative genomic coverage of microbial species (74-fold for bacteria, which made up 89% of the sampled species comparisons; Fig. 1.2.3.3a-b). Consistent with the greater recovery efficiency of the ssDNA protocol, we find that most of the species detected in the dsDNA library preparation assays were also detected following ssDNA library preparation (95% species recovery, 934/984). 55% of all species detected were uniquely observed in the ssDNA library preparation assays.



**Figure 1.2.3.2** ssDNA library preparation captures more bacterial cfDNA with higher diversity than standard methods. (a) Single-stranded ligation method has significantly more microbial cell-free DNA than the traditional method (double-stranded ligation). (b) The abundance of microbial genomes is compared by ligation method, for matched species within matched samples. Species are colored by superkingdom (Non-fungi eukaryotes are grey in color). {7}

The greater efficiency in recovery of microbial cfDNA is in line with the greater sensitivity of the ssDNA library preparation protocol to ultrashort cfDNA. This feature offers the potential to profile the bacterial and viral components of the microbiome in plasma both more broadly, in terms of the number of microbes accessible for testing, as well as more deeply. Such an increase in microbial sequencing depth permits infectious disease diagnostics based on sequencing of cfDNA with increased precision and at lower cost. The ssDNA library preparation detected viral fragments with clinical relevance in transplantation, including polyomaviruses (BK polyomavirus, one sample, and Merkel cell polyomavirus, two samples) and single-stranded DNA, transfusion-associated viruses (torque teno virus, 18 samples; SEN virus, 19 samples); the impact of these viruses on the outcome of solid organ transplant patients has been investigated previously [1]. Current methods to detect infections are predominantly limited to testing one pathogen at a time. Metagenomics approaches have the potential to broadly screen for all known pathogens (with a DNA genome) in a single test [1, 39, 40]. Blood can be collected non-invasively and the majority of tissues in the body are connected to the blood circulation, making cfDNA an attractive sample type for such approach. A number of caveats remain; for example, it may be difficult to inform about an infection by organisms that

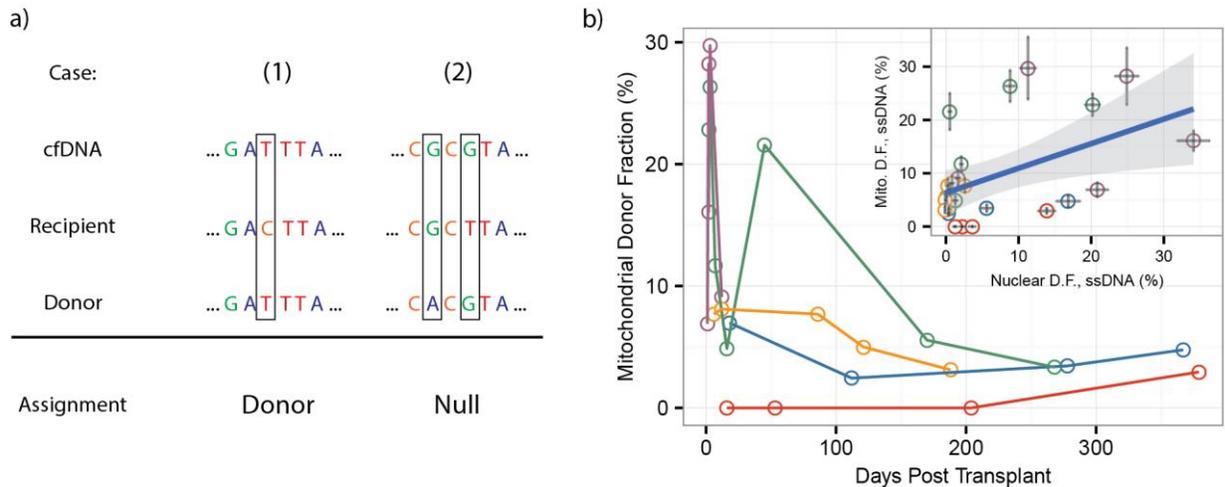


**Figure 1.2.4.1** Single-stranded library preparation captures donor fraction information from autosomal cfDNA in matched samples. The donor fraction from autosomes was calculated and compared for both library preparation methods. Error bars represent standard error in measurement in error calculated from sampling error. Red line indicates exact correspondence. {8}

are part of the normal flora in certain body sites, but are pathogenic in others. cfDNA may be of limited use in such cases as it lacks body-site specificity.

#### 1.2.4 Donor-specific cfDNA from the mitochondria and autosomes

Donor-specific cfDNA is present in the circulation of organ transplant recipients [33] and recent studies have shown that the proportion of donor-specific cfDNA is predictive of acute rejection in heart and lung transplantation [2, 3]. We compared the fractional abundance of donor-specific cfDNA in the lung transplant samples measured following dsDNA and ssDNA library preparation (36 matched samples, six patients, Fig. 1.2.4.1). We found an excellent agreement between matched measurements (corr. = 0.980, Pearson,  $p \ll 10^{-5}$ ). Here, sequences were assigned to the donor and recipient based on genotypic information (single-nucleotide polymorphisms, SNPs) obtained from pre-transplantation whole blood samples [3]. Importantly, we discovered that the number of donor-aligned sequences in patients does not differ significantly, regardless of the increase in microbial reads. The ability to capture high numbers of reads allowed for the maintenance of high precision measurements of the donor fraction (as determined by



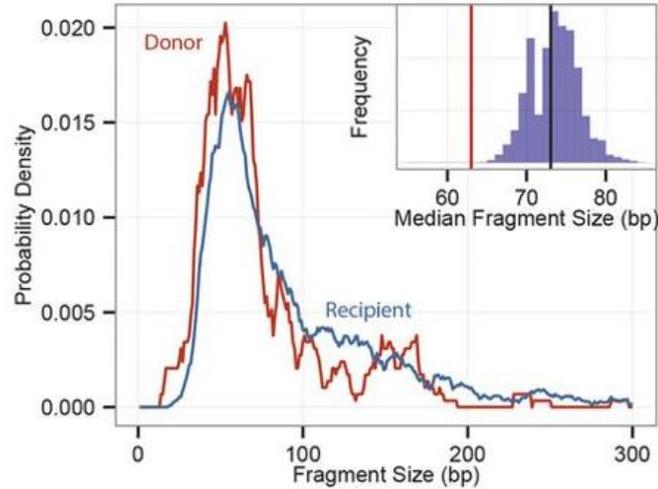
**Figure 1.2.4.2** Donor-specific cfDNA originating from the mitochondria reveals graft damage. (a) Schematic representation of analysis workflow used to discriminate donor and recipient specific mt-cfDNA. Examples of an ambiguous assignment and a fragment assigned to the donor are shown. (b) Fraction of donor-specific mt-cfDNA as function of time post-transplant for five double lung transplant patients (25 samples, samples with fewer than 20 informative fragments excluded); Inset: The fraction of donor-specific mitochondrial and nuclear genomic DNA for the same samples is compared (corr. = 0.463, Pearson,  $p = 0.020$ ). {9}

sampling error). In one sample, we observed preparation via ssDNA ligation increased precision by an order of magnitude (Fig. 1.2.4.1). One patient suffered from a severe rejection event at month 12 post-transplant. The fraction of donor-specific cfDNA measured for this patient using both library preparation methods was elevated, coinciding with the biopsy-proven rejection event.

Because of the high copy number of mtDNA in cells and the relatively high genetic diversity between two unrelated individuals [41, 42], mtDNA is often used in forensic analyses [43] and in studies of population genetics [44]. The same attributes make mitochondrial cfDNA a promising candidate marker of post-transplant graft injury. We asked whether donor-specific mitochondrial cfDNA can be detected in the plasma of transplant recipients. We built mitochondrial reference sequences to assign mitochondrial cfDNA to the transplant donor or recipient. To this end, DNA was extracted from whole blood samples collected from the donor and the recipient prior to the transplant procedure. Mitochondrial DNA was selectively amplified and sequenced. One million sequences led to a per-base coverage greater than 100-fold (genome size 16.5 kb), sufficient to determine subject-specific mitochondrial variants. Based on the

reference sequences, we compiled lists of SNPs that are unique to either the donor or recipient (Fig. 1.2.4.2a, see Methods). On average, 152 informative SNPs were found per donor-recipient pair, leading to a SNP every 114 bp. For samples prepared via ssDNA ligation,  $8.7\% \pm 3.4\%$  of the mitochondrial sequences were informative, and 9.5% of the informative SNPs were assigned to the donor. Donor- and recipient-specific sequences spanned the entire mitochondrial genome.

To the best of our knowledge, this is the first direct observation of graft-derived mitochondrial DNA in the circulation of transplant recipients. We computed the fractional abundance of donor-specific mt-cfDNA as the number of donor-specific mt-cfDNA molecules divided by the total number of informative mt-cfDNA molecules. We studied the variability and time dependence of the levels of donor-specific mt-cfDNA. We observed a higher-than-average fraction of donor-derived mt-cfDNA in the month following transplantation (Fig. 1.4.2.2b). Such an observation echoes elevated levels of donor-derived cfDNA in heart and lung transplant recipients during the first few weeks post-transplant, in the absence of acute rejection [2, 3]. The fraction of donor-specific mt-cfDNA was only modestly correlated (corr. = 0.480, Pearson,  $p = 0.0152$ ) with the fraction of nuclear genomic DNA. Samples for which there were less than 20 informative mitochondrial fragments were removed (11 of 36 samples). Deeper sequencing and an analysis of a greater set of samples will be needed to investigate the relationship between acute rejection and the release of mitochondrial DNA from the graft. It is also likely that the mitochondrial donor fraction does not correspond exactly with the nuclear donor fraction. Cell-free DNA is the collection of all tissues contributing to the nucleic acid abundance in a biological fluid, and each cell contributes a similar (if not exact) amount of nuclear genomic DNA upon death. However, distinct cell types hold various amounts of mitochondria (50-40,000) each with five to ten genome copies, depending on conditions and replication potential [45]. As a result, without knowledge of mitochondrial abundance per cell-type, donor fractions may be confounded due to the underlying physiology.



**Figure 1.2.4.3** Smoothed (distribution five nearest-neighbor running mean) of donor mt-cfDNA (red) is compared to that of recipient mt-cfDNA (blue). Inset: median fragment size for the donor mt-cfDNA (red line) compared to the fragment size of 10,000 subsets sampled from the recipient mt-cfDNA length set (median depicted by black line). {10}

Previous studies have found differences in fragment lengths for fetal and maternal cfDNA [38], tumor and somatic cfDNA [30] and hematopoietic and non-hematopoietic cfDNA [25]. Here, we compared the length of mitochondrial fragments derived from the graft ( $n = 265$ ) to those specific to the recipient ( $n = 1855$ , 40 samples). We generated 10,000 random subsamples of the total collection of recipient-specific fragments (subsampled to the total number of donor fragments detected,  $n = 265$ ). We next computed the median lengths for the random subsamples and compared to the median length of donor fragments (inset Fig. 1.2.4.3). We found that donor sequences were slightly shorter ( $-9 \pm 3$  bp) than recipient-specific mitochondrial sequences. This shortening in fragment length may be indicative of differences in the mechanisms of release, or differences in processes of degradation, of donor and recipient mt-cfDNA.

### 1.2.5 Methods and sampling cohort description

ssDNA sequencing libraries were prepared from cfDNA purified from plasma using a spin column approach (Qiagen QiaAMP nucleic acid extraction kit, 1 mL plasma). We followed a paleogenomics-based protocol with the following exceptions: (1) uracil excision steps using endonuclease VIII were not performed, (2) the amount of CircLigase II enzyme in the protocol was reduced from 4  $\mu$ L to 0.8  $\mu$ L and

amounts of  $\text{MnCl}_2$  and CircLigase II buffer were halved, (3) editing of oligos including extension primer CL9, had an addition N\*N\*N\*N overhang on the 5' end (described by Karlsson et al. [46]) to prevent formation of adapter-dimers. Briefly, cfDNA libraries were denatured at 95 °C for one minute. ssDNA libraries were dephosphorylated and biotinylated adapters were ligated via single-stranded DNA ligation (CircLigase II enzyme). Primer extension was performed on streptavidin functionalized magnetic beads (T4 DNA polymerase) and a second set of adapters was ligated by double-stranded DNA ligation (T4 DNA ligase).

Molecules complementary to the original cfDNA molecules were denatured from the beads at 95 °C and isolated in 25  $\mu\text{L}$  of elution buffer. 1  $\mu\text{L}$  of eluted library was aliquoted and qPCR was performed to determine the optimum number of indexing PCR cycles. Finally, the full library was PCR amplified. A positive control (1  $\mu\text{L}$  of 500  $\mu\text{M}$ , synthetic ssDNA) and a negative control were included with each batch of samples. The efficiency of ligation of cfDNA fragments to biotinylated probes and ligation of double stranded adapters to primer-extended products was estimated using quantitative PCR [36]. On average  $0.8 \times 10^9$  unique ssDNA molecules ( $0.02\text{--}8.2 \times 10^9$ ) were ligated and PCR amplified (8 to 15 cycles). Adjusting the extension sequence primer with a 4-N overhang on the 5' end (as in Karlsson et al. [46]), limited the occurrence of adapter dimers to, on average, one in 1,700 sequences. Libraries were sequenced on the Illumina MiSeq or HiSeq platform ( $2 \times 75$  bp). These settings resulted in an average of  $5.7 \pm 1.4$  million paired-end reads per sample, leading to an average human genome coverage of  $0.23x \pm 0.06x$ .

Forty samples of cfDNA extracted from plasma of six double-lung transplant recipients were analyzed in this study. Results from ssDNA library preparation were compared against sequence data obtained for the same samples following conventional dsDNA library preparation where available (36 matched samples,  $18.8 \pm 9.1$  million paired-end reads per sample). The analysis of matched samples enabled us to assess the effect of different library preparations on measurements of cfDNA.

Mitochondrial consensus sequences were established for every transplant donor and recipient. DNA was extracted from whole blood samples (Qiagen DNeasy Blood & Tissue kit) collected pre-transplant. Mitochondrial DNA was selectively amplified (Qiagen REPLI-g Mitochondrial DNA Kit), and sheared to 300 bp (Covaris). Libraries were prepared for sequencing using the NEBNext Ultra library preparation, characterized (Advanced Analytical Fragment Analyzer and dPCR) and sequenced ( $2 \times 250$  bp, Illumina MiSeq). One million sequences led to a per-base coverage greater than 100-fold (genome size 16.5 kb), sufficient to determine subject-specific mitochondrial variants. Fastq files were trimmed (Trimmomatic [47], LEADING:25 TRAILING:25 SLIDINGWINDOW:4:30 MINLEN:15) and aligned against the human reference genome [GenBank:GCA\_000001305.2] using BWA-mem [48]. Sequences that mapped to the mitochondrial reference sequence (edited from GenBank:NC\_012920) were extracted. A BCF file of SNPs was created and a FASTA consensus sequence was determined. A list of informative SNPs was created.

Nuclear genomic sequences were assigned to the donor or recipient using methods previously described [2, 14]. Briefly, sequences were assigned to the donor or recipient based on SNP genotyping information (IlluminaHumanOmni2.5–8 or HumanOmni1 whole genome arrays) obtained from pre-transplant whole blood samples. Mitochondrial sequences were assigned to the donor and recipient as follows: Raw sequencing datasets were trimmed (Trimmomatic [47], LEADING:20 TRAILING:20 SLIDINGWINDOW:4:20 MINLEN:25), and low quality reads were filtered (FASTX toolkit [49], -q 21 -p 50) and aligned (BWA-mem [48]) to the human reference genome [GenBank:GCA\_000001305.2], with changes made to the mitochondrial genome to account for the circular structure. Sequences that mapped to the mitochondrial reference [GenBank:NC\_012920] were collected and SNPs were listed using SAMtools [50]. Sequences were assigned to the donor or recipient through comparison to the list of informative SNPs compiled for the donor-recipient pair.

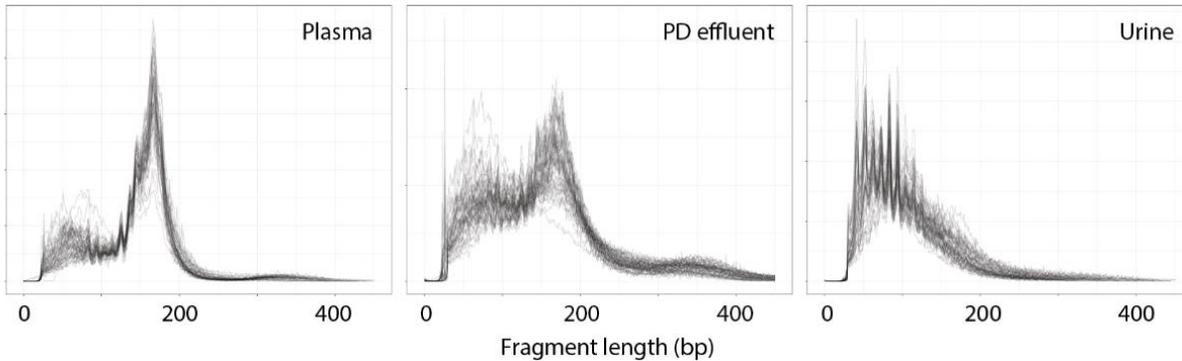
The analysis workflow used to quantify non-human cfDNA sequences is described in detail [2, 3, 14]. Briefly, Low-quality bases and Illumina specific sequences were trimmed (Trimmomatic 0.32

[47]), and read pairs were merged using FLASH 1.2.7 [51]. Reads were aligned (Bowtie 2.1.0 [52]; very sensitive mode) against the human reference (UCSC hg19; <https://genome.ucsc.edu/>). Unaligned reads were extracted and BLASTed (2.2.28+) against a NCBI database [53]. Alignments were required to have an identity of at least 90% across 90% of the bases of the query. A relative genomic abundance of species was determined using GRAMMy [54]. Supplementary material pertinent to this data can be found at <https://www.nature.com/articles/srep27859#supplementary-information>.

#### 1.2.6 Ultrashort cell-free DNA provides a new perspective of origin and pathology

In this work, we have demonstrated that a ssDNA library preparation is sensitive to cfDNA of a broad range of types and lengths. Few studies have focused on ultra-short cfDNA (with lengths shorter than 100 bp) or cfDNA that is not derived from the nuclear genome, including mitochondrial and microbial derived cfDNA. Our present work indicates that these relatively overlooked forms of cfDNA provide a unique window into physiology.

We applied a ssDNA library preparation to the analysis of cfDNA in the plasma of lung transplant recipients. We report the first observation of graft-derived mitochondrial DNA in the plasma of these organ transplant recipients. Donor-derived mitochondrial cfDNA has not been investigated as a marker of acute rejection in solid-organ transplantation, but offers several advantages: (1) the mitochondrial genome is small and relatively straightforward to deeply characterize via sequencing, (2) the mitochondrial genome contains a great number of variants that enables differentiation of donor and recipient sequences, and (3) with thousands of copies of mitochondrial DNA present in every cell, mitochondrial cfDNA is abundant in plasma. Mitochondrial DNA has conserved similarities to bacterial DNA and contains inflammatogenic unmethylated CpG motifs [55]. It is therefore not surprising that mitochondrial DNA was identified as a powerful damage associated molecular pattern – an endogenous molecule that can activate innate immunity when released during cellular injury [56]. It is conceivable that the release of mitochondrial DNA that accompanies graft injury promotes many of the harmful immunologic responses observed in solid-organ transplantation. The results presented here provide the first window into this relationship.



**Figure 1.2.6.1** Fragment length distributions indicate the heterogeneity of degradation in various biological fluids. Fragment length distributions are shown for cfDNA originating from plasma (n = 40), urine (n = 40), and peritoneal dialysis effluent (n = 40). Samples were prepared using ssDNA library preparation. Figure assembled by Fanny Chen. { 11 }

Microbial cfDNA is present in the circulation and is the product of microbial degradation across the body or originates from microorganisms that infect the blood or translocate to the blood [57]. We found that the ssDNA library preparation is more effective at recovering bacterial and viral cfDNA, as compared to a dsDNA library preparation method. We furthermore found that the fragmentation profiles of microbial and mitochondrial DNA in plasma are highly similar, suggesting that they are exposed to similar degradation processes. These observations enable measurement of the bacterial and viral microbiome in plasma with greater sensitivity and at a reduced cost.

Previous studies of the molecular size of nuclear genomic cfDNA have provided insight into the origin and nature of these molecules [58]. Many studies have noted that the predominant fragment size of cfDNA is consistent with the size of DNA wrapped around a single histone octamer [57]. Distinct length profiles are observed for cfDNA depending on their cellular origin with hematopoietically-derived DNA being longer than that of nonhematopoietically-derived [38]. Here, we found that the sequencing library preparation method can have a significant effect on length profile measurements. We report the fragmentation profile of nuclear genomic cfDNA in plasma over a broad range of lengths, and we conclude, in agreement with a recent report [23], that a considerable fraction of nuclear genomic cfDNA is non-nucleosomal and subject to degradation by nucleases, in much the same way that we described for mitochondrial cfDNA.

In its current implementation, the ssDNA library preparation requires more hands-on time compared to standard protocol (~13 hours versus ~6 hours, for 12 samples) at a similar cost per sample (\$35-\$40). This work focused on the cfDNA in plasma, but the methods described herein will further be relevant for genomic measurements of cfDNA in urine. We performed follow-up studies in a number of biological fluids relevant to clinical studies, including urine, amniotic fluid, and peritoneal dialysis effluent. The median fragment length of cfDNA originating in these fluids can vary considerably (Fig. 1.2.6.1). While plasma seems to have the largest fragments, on average, urinary cfDNA is highly fragmented, with relatively few fragments indicating the full nucleosome wrapping length of 167 bp (Fig. 1.2.6.1). It is, thus, necessary for library preparation assays to enrich ultrashort molecules for future discovery of noninvasive biomarkers. The widespread interest in circulating cfDNA as a marker of disease, warrants further investigation into the properties, types and origins of cfDNA and motivates further advances in genomic measurement techniques.

### Chapter 3: Urinary cell-free DNA is a versatile analyte for monitoring infections of the urinary tract

“Urinary tract infections are one of the most common infections in humans. Here we tested the utility of urinary cell-free DNA (cfDNA) to comprehensively monitor host and pathogen dynamics in bacterial and viral urinary tract infections. We isolated cfDNA from 141 urine samples from a cohort of 82 kidney transplant recipients and performed next-generation sequencing. We found that urinary cfDNA is highly informative about bacterial and viral composition of the microbiome, antimicrobial susceptibility, bacterial growth dynamics, kidney allograft injury, and host response to infection. These different layers of information are accessible from a single assay and individually agree with corresponding clinical tests based on quantitative PCR, conventional bacterial culture, and urinalysis. In addition, cfDNA reveals the frequent occurrence of pathologies that remain undiagnosed with conventional diagnostic protocols. Our work identifies urinary cfDNA as a highly versatile analyte to monitor infections of the urinary tract.”

Chapter adapted from [5]:

Burnham P, Dadhania D, Heyang M, Chen F, Westblade L, Suthanthiran M, Lee JR, De Vlaminck I. Urinary cell-free DNA is a versatile analyte for monitoring infections of the urinary tract. *Nature Comm.* 2018.

### 1.3.0 cfDNA sequencing to inform infections of the urinary tract

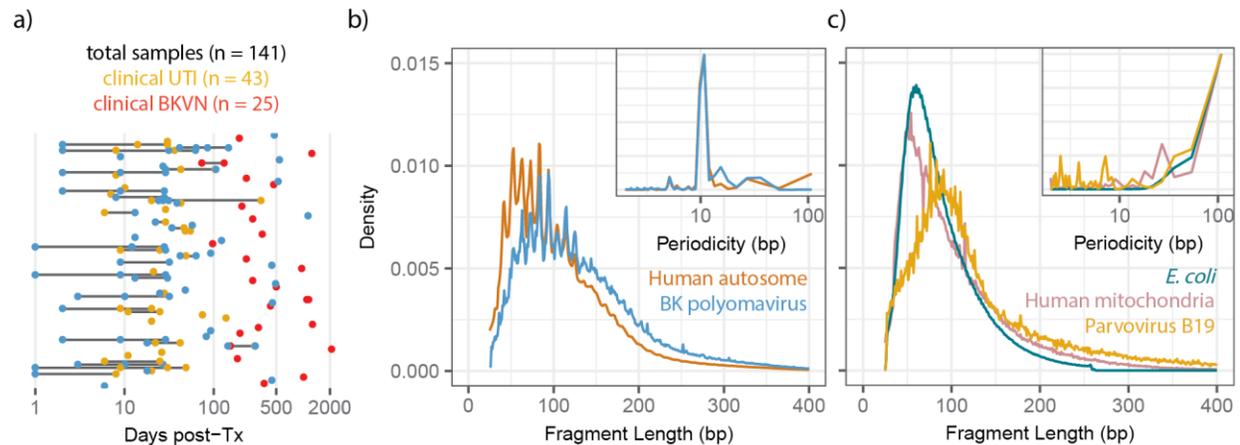
Urinary tract infection (UTI) is one of the most common medical problems in the general population [59]. Among kidney transplant recipients, UTIs occur at an alarmingly high rate [60]. Bacterial UTI affects approximately 20% of kidney transplant recipients in the first year after transplantation [61] and at least 50% in the first three years after transplantation [62]. In addition, complications due to viral infection often occur. Up to 8% of kidney transplant recipients suffer nephropathy from BK polyomavirus (BKV) infection in the first three years after transplantation [63, 64]. Other viruses that commonly cause complications in kidney transplantation include adenovirus, JC polyomavirus, cytomegalovirus (CMV), and parvovirus [65]. The current gold standard for diagnosis of bacterial UTI is *in vitro* urine culture [66]. Although improved culture methods are being investigated [67, 68], bacterial culture protocols implemented in clinical practice remain limited to the detection of relatively few cultivable organisms [67]. Furthermore, urinalysis is often required in conjunction with culture to make treatment decisions [66]. A large number of ultrashort cfDNA fragments are present in plasma and urine [11, 27, 38, 69]. These molecules are the debris of the genomes of dead cells and offer opportunities for precision diagnostics based on ‘omics principles, with applications in pregnancy, cancer and solid-organ transplantation [2, 11–13].

Here, we investigate the utility of urinary cfDNA to comprehensively monitor host and pathogen interactions that arise in the setting of viral and bacterial infections of the urinary tract. Using shotgun DNA sequencing, we assay cfDNA isolated from 141 urine samples collected from a cohort of 82 kidney transplant recipients, including recipients diagnosed with bacterial UTI and BKV nephropathy (BKVN). We recently developed a single-stranded DNA (ssDNA) library preparation, optimized for the analysis of short, highly fragmented DNA [4, 32, 36], and were able to sequence cfDNA isolated from relatively small volumes of urine supernatant (1 mL or less) with this method. We find that urinary cfDNA sequencing agrees in the vast majority of cases with conventional diagnostic testing, while also uncovering frequent occurrence of bacteria and viruses that remain undetected in conventional diagnostic protocols. We further investigated cfDNA-based analytic methods that go beyond microbial identification and provide a deeper

understanding of the infectious process. We show that rate of bacterial population growth can be estimated from an analysis of the bacterial genome structure and that this measurement can inform diagnosis of UTI. We further mined cfDNA for antimicrobial resistance (AR) genes and show that AR gene profiling can be used to evaluate AR. We observe that the relative proportion of kidney donor-specific cfDNA correlates with graft tissue injury in the setting of viral infection, and host immune cell activation in the setting of bacterial infection. As a follow-up study, we chose a subset of cfDNA patient samples from the cohort, treated them with sodium bisulfite, performed ssDNA library preparation [4], and used analysis of differentially methylated regions (DMRs) from methylated cytosines. These measurements indicate the tissue-of-origin of the cfDNA ensemble, which we used to show high abundance of kidney-derived and leukocyte-derived cfDNA in cases of BKVN and UTI, respectively. Collectively, our study supports the use of shotgun DNA sequencing of urinary cfDNA as a comprehensive tool for monitoring patient health and studying host-pathogen interactions.

### 1.3.1 Biophysical properties of urinary cfDNA

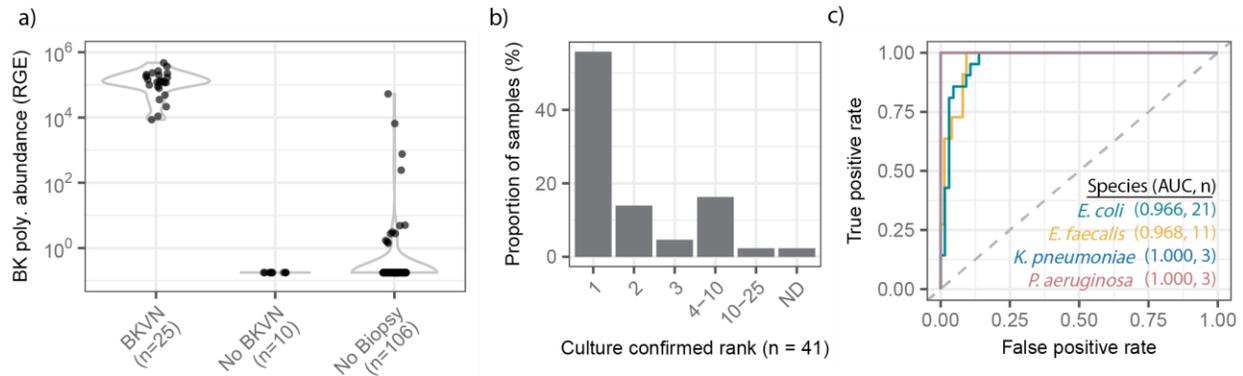
Urinary cfDNA is composed of human chromosomal, mitochondrial, and microbial cfDNA released from host cells and microbes in the urinary tract, and of plasma-derived cfDNA that passes from blood into urine [70]. Urine can be collected non-invasively in large volumes, and therefore represents an attractive target for diagnostic assays. Compared to plasma cfDNA, relatively few studies have examined the properties and diagnostic potential of urinary cfDNA. The urinary environment degrades nucleic acids more rapidly than plasma resulting in fewer DNA fragments that are also shorter [71]. Consequently, sequence analyses of urinary cfDNA have to date required relatively large (> 10 mL) volumes of urine [38,



**Figure 1.3.1.1** Shotgun sequencing assay and biophysical properties of urinary cfDNA. (a) Study design included patient samples, in health and infection, from one day to six years after transplant. Fragment length distributions and Fast Fourier Transform (FFT, 60 – 140 bp, inset) for cfDNA from organisms with nucleosomal packaging (b) and without such packaging (c). [12]

72]. Here, we applied a ssDNA library preparation technique that employs ssDNA adapters and bead ligation to create diverse sequencing libraries that capture short, highly degraded cfDNA [4, 36] (Fig. 1.2.2.1). We find that single-stranded library preparation enables sequence analyses of urinary cfDNA from just one milliliter of urine supernatant. We tested 141 urine samples collected from 82 kidney transplant recipients, including subjects diagnosed with bacterial UTI and BKVN (overview of post-transplant dates and categories depicted in Fig. 1.3.1.1a). We obtained  $43.5 \pm 17.3$  million paired-end reads per sample, yielding a per-base human genome coverage of  $0.49x \pm 0.24x$ . Many fragments derived from microbiota; for example, for subjects diagnosed with bacterial UTI, bacterial cfDNA accounted for up to 34.7% of the raw sequencing reads and in cases of BKVN, BKV cfDNA accounted for up to 10.3% of raw sequencing reads. To account for technical variability and sources of environmental contamination during extraction and library preparation, a known-template control sample was included in every sample batch and sequenced (see Section 1.3.9).

We analyzed the fragment length profiles of urinary cfDNA at single nucleotide resolution using paired-end read mapping [48]. This analysis confirmed previous observations of the highly fragmented nature of urinary cfDNA compared to plasma cfDNA [72] (Fig. 1.3.1.1b). We observed a 10.4 bp periodicity in the fragment length profile of chromosomal cfDNA (Fourier analysis, Fig. 1.3.1.1b, inset),



**Figure 1.3.2.1** Diagnostic comparison of microbial cfDNA sequencing in UTI to gold standard technologies. (a) The relative genomic abundance is shown for BK polyomavirus for patients who tested positive or negative for polyomavirus nephropathy via renal biopsy, and for those untested. (b) Bacterial species were ranked according to relative genomic abundance within each sample for patients with culture-diagnosed UTI. The ranked position of the species was compared to that identified by standard urine culture. (c) Receiver-operator characteristics for samples originating from samples with *Enterococcus*, *E. coli*, and *P. aeruginosa* UTI. {13}

consistent with the periodicity of DNA-histone contacts in nucleosomes [73]. BKV is known to hijack histones of infected host cells, and to form mini-chromosomes after infection [74]. The periodicity in the fragment length profiles of BKV cfDNA in urine reflect this biology (Fig. 1.3.1.1b). We did not observe a similar nucleosomal footprint for bacterial and mitochondrial cfDNA, or cfDNA arising from parvovirus B19, which is expected given the non-nucleosomal compaction of the genomes that contribute these cfDNA types (Fig. 1.3.1.1c).

### 1.3.2 Screening the infectome for multiple pathogenic agents

We assessed the presence of cfDNA from bacterial and viral pathogens reported by conventional diagnostic assays. We used bioinformatics approaches to estimate the relative genomic representation of different species (see Section 1.3.9) [1, 54]. To directly compare the microbial abundance across samples, we computed the representation of microbial genome copies relative to human genomes copies, and expressed this quantity in relative genome equivalents (RGE).

We detected a very high load of BKV cfDNA in all 25 samples collected from 23 subjects diagnosed with BKVN by needle biopsy (mean  $1.49 \pm 1.08 \times 10^5$  RGE, Fig. 1.3.2.1a), but not in 10 samples from 10 subjects that were BKVN negative per biopsy (all below detection limit). In these 35 biopsy-associated samples, the BKV cfDNA abundance correlated with a matched urine cell pellet BKV VP1 mRNA copy measurement that we previously validated as a noninvasive marker for BKVN (corr. = 0.74, Spearman,  $p = 3.48 \times 10^{-7}$ ) [75, 76].

We quantified bacterial urinary cfDNA in 43 urine samples from 31 subjects who had a corresponding positive culture from urine obtained on the same day. For 41 of the 43 positive urine specimens, a bacterial organism was reported to the species level by conventional culture. In 40 of these 41 samples, sequencing of urinary cfDNA detected the clinically reported organisms to the species level (Fig. 1.3.2.1b). For a single sample, urinary cfDNA did not match with the bacterial culture: *Raoultella ornithinolytica* was isolated in culture, but not detected by cfDNA sequencing (see Methods for a detailed discussion of this discordant readout). For two of the 43 clinically positive samples, the suspected etiologic agent was identified to the genus level (*Staphylococcus*, reported as coagulase-negative *Staphylococcus* species [CoNS], and *Streptococcus*, reported as viridans group streptococci) by culture. In both these cases the reported organism was detected as the most prevalent within the sample. From the 43 cultures, we examined six with polymicrobial bacterial infection (defined as two individual bacterial taxon detected at the genus or species level). For five out of six of these cases, we observed both species among the 10 most abundant species by cfDNA sequencing. In one sample, the secondary bacterial agent, CoNS (< 10,000 colony forming units [cfu]/mL), was not detected.

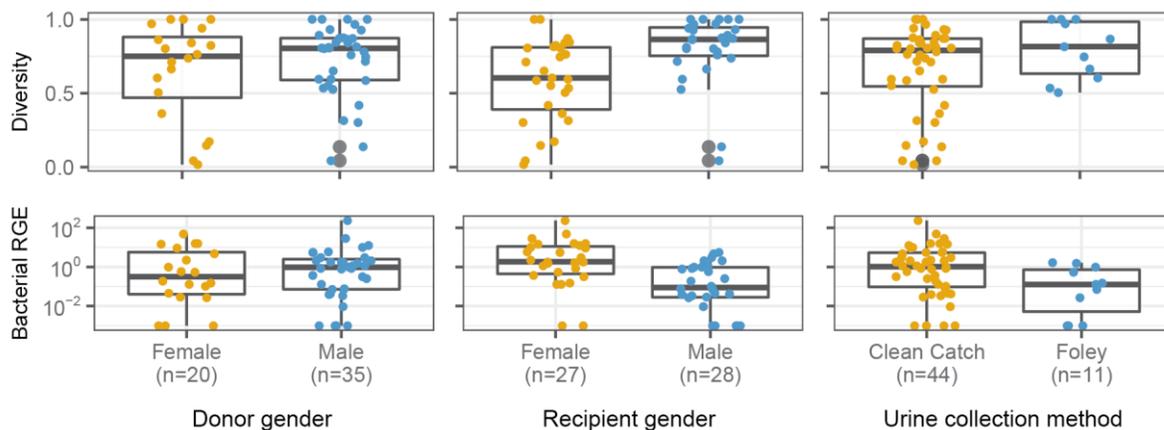
To further assess the performance of urinary cfDNA for microbial identification, we compared the relative genomic abundance (in RGE) of bacterial cfDNA for subjects diagnosed with bacterial infection (49 bacterial isolates identified from 43 clean-catch midstream cultures), to the relative genomic abundance (in RGE) measured for 43 negative clean-catch midstream urine cultures, (Fig. 1.3.2.1c). We found agreement between urinary cfDNA and culture based isolation of *Enterococcus faecalis* (number of

matched positive cultures, n = 11, Area Under the Curve [AUC] = 0.97, 95% Confidence Interval [CI] = 0.935-1), *Enterococcus faecium* (n = 2, AUC = 0.98, CI = 0.976-1), *Escherichia coli* (n = 21, AUC = 0.97, CI = 0.93-1), *Klebsiella pneumoniae* (n = 3, AUC = 1.00, CI = 1), *Pseudomonas aeruginosa* (n = 3, AUC = 1.00, CI = 1), CoNS (n = 4, AUC = 0.78, CI = 0.46 - 1). Here, receiver operating characteristic (ROC) analysis was performed for bacterial species where there was at least one positive culture of the same organism available (n > 1).

In only 60% of examined samples (26/43 UTI cases) was the organism identified in culture the most prevalent organism in the sample as measured by cfDNA (Fig. 1.3.2.1b). Whereas bacterial culture is skewed towards species that are readily isolated on routine bacteriological media employed for urine culture, cfDNA sequence analyses permit the identification of a broader spectrum of bacterial species. To evaluate this concept further, we assayed two samples collected from one of the subjects included in the analysis above diagnosed with *Haemophilus influenzae* bacteriuria. *H. influenzae* is an uncommon uropathogen that does not routinely grow on media employed for conventional urine culture (tryptic soy agar with sheep blood and MacConkey agar) [77]. Repeated urine cultures for this patient were negative, but given a urinalysis suggestive of a UTI and given that the patient developed *H. influenzae* bacteremia, the original urine specimen collected at presentation was re-plated onto chocolate agar, upon which *H. influenzae* was isolated. We observed *H. influenzae* cfDNA in the sample taken at the time of presentation and a sample taken four days after presentation (0.037 RGE and 0.41 RGE, respectively). This case supports the utility of urinary cfDNA to identify infections where conventional culture fails.

### 1.3.3 Profiling the urinary microbiome

The urinary tract was regarded as sterile but recent studies have revealed that it may harbor microbiota [68, 78, 79]. We examined the composition of the urinary microbiome by urinary cfDNA profiling (absence of UTI, n = 43), or collected within the first three days post-transplant (n = 12). We find

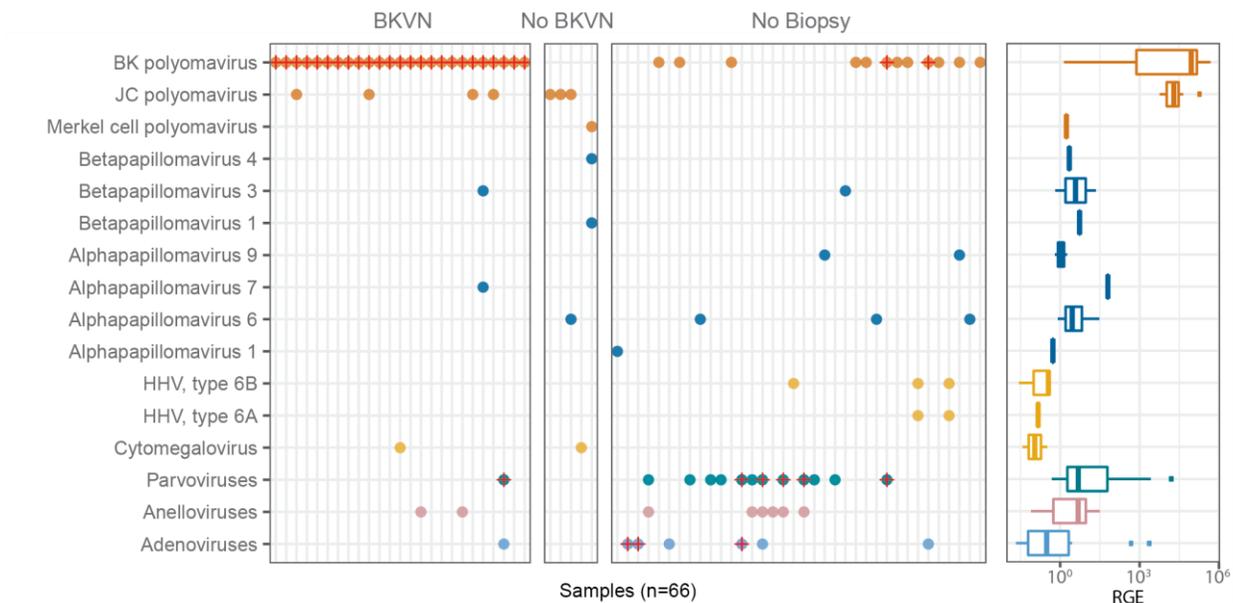


**Figure 1.3.3.1** Bacterial diversity and abundance measurements in urinary cfDNA for patient samples separated by donor sex, recipient sex, and collection methods. Metrics calculated at genus taxonomic level. {14}

that the species-level abundance and the species-level diversity of the bacteriome are a function of the transplant recipient gender but not the donor gender (Fig. 1.3.3.1). On average, we observed two to three orders of magnitude more cfDNA from *Gardnerella* (6125x), *Ureaplasma* (1686x), and *Lactobacillus* (321x) across female transplant recipients who did not have a UTI at time of sampling compared to male recipients who did not have UTI; these bacterial genera have been characterized as microbial components of the vaginal microbiome [80]. We examined the relationship between urine collection methods and the abundance and diversity of the bacteriome and find a notably reduced bacterial load for samples collected by indwelling catheter (samples collected within four days after transplant) versus clean catch urine samples. cfDNA may be an ideal tool to study the urinary microbiome, but such future studies need to account for effects of gender and sample collection approaches.

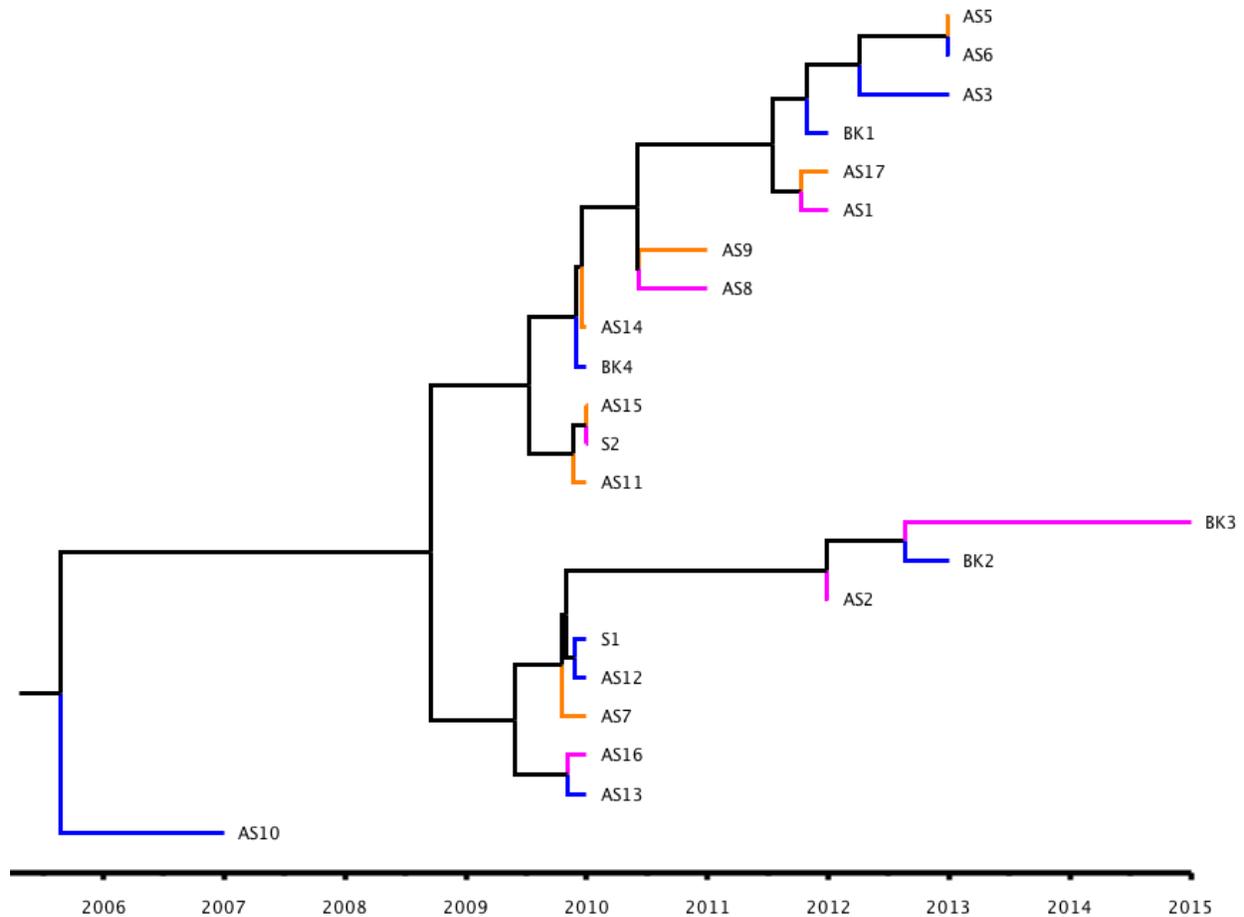
#### 1.3.4 Broad screening for viruses via cfDNA

We next screened for the occurrence of cfDNA derived from viruses. Nearly half of the samples (66/141) had detectable levels of cfDNA derived from eukaryotic viruses that are potentially clinically relevant. Figure 1.3.4.1 highlights the frequent occurrence of JC polyomavirus, parvovirus B19, Merkel cell polyomavirus, CMV, human herpesvirus 6A, human herpesvirus 6B, and various known oncoviruses



**Figure 1.3.4.1** Urinary cfDNA profiles the human virome. Viral cfDNA was detected in 66 samples of the 141 samples. cfDNA reveals frequent occurrence of viruses that are potentially clinically relevant (left panel); red crosses identify samples belonging to subjects who developed an infection of the corresponding viral agent. Right-most panel shows boxplots of the viral cfDNA abundance across all samples. Color of points and boxplots by viral family. {15}

across different patient groups. In several samples, we detected cfDNA from multiple polyomavirus species concurrently (JC polyomavirus or BKV). To shed light on the potential clinical utility of broad screening for viruses via cfDNA, we assayed serial urine from three subjects diagnosed with viral infections that are relatively uncommon in kidney transplant recipients and consequently not routinely screened for in our patient cohorts. In samples from two subjects with clinically diagnosed parvovirus B19 infection, we detected urinary cfDNA from parvovirus B19 eight days prior to the clinical diagnosis in one subject and urinary parvovirus B19 cfDNA 80 days before diagnosis and 25 days after diagnosis in another subject. In the former subject, we observed a high abundance of both BKV ( $3.54 \times 10^4$  RGE) and parvovirus B19 ( $2.48 \times 10^4$  RGE), which correlated with positive results of individual viral-specific PCR tests for BKV and parvovirus B19. For a third patient, we observed a high abundance of human adenovirus B DNA, in samples obtained 15 days before ( $2.52 \times 10^3$  RGE) and nine days after ( $5.08 \times 10^2$  RGE) adenovirus infection diagnosis by clinical urine DNA PCR. These data support the utility of urinary cfDNA sequencing for the detection of both common and uncommon viral agents.

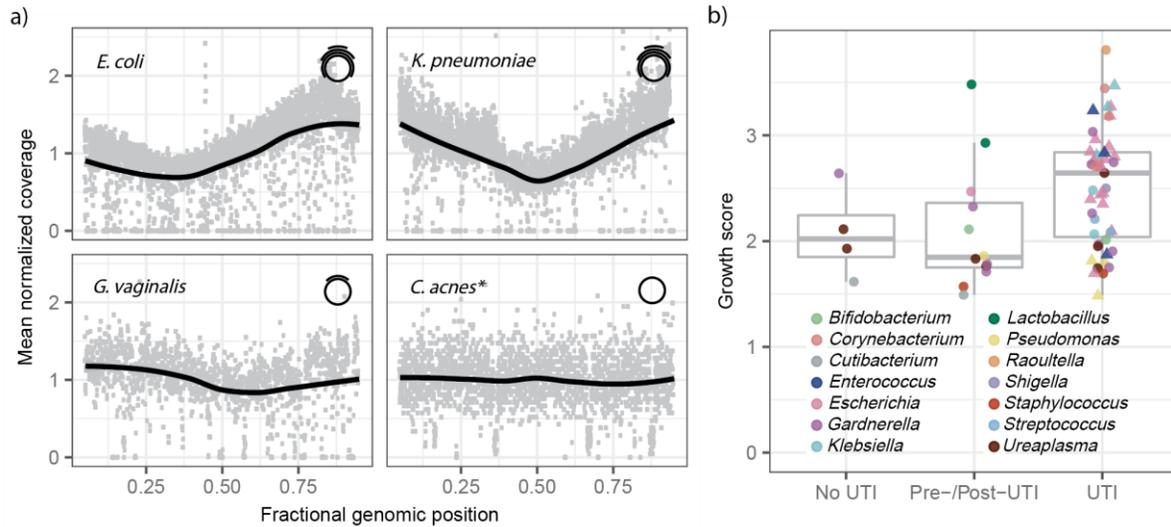


**Figure 1.3.4.2** Phylogenetic reconstruction of BK polyomavirus VP1 consensus sequences assembled from urinary cfDNA. BEAST software was used to find the most likely tree given a constant population over time. Nodes represent points of genetic separation. Leaves are annotated with sample identifier and positioned at time of sampling. Colors correspond to BK polyomavirus spread assessed in renal biopsy samples (blue = low, magenta = moderate, orange = diffuse). { 16}

The high prevalence of cfDNA attributed to viral genomes allows additional, more detailed genetic analysis. Multiple BK polyomavirus genomic subtypes exist in the global population with various antigenic properties [81]. Recently, Morel et al. have demonstrated a method to rapidly subtype BK polyomavirus through algorithmic query of SNPs [82]. To our knowledge, it is unknown if certain BK polyomavirus subtypes are more likely to result in more acute cases of BKVN or higher viral abundance during reactivation. However, it is accepted that the pathogenic BK polyomavirus is derived from the donor, and HLA mismatches result in faster and more severe cases of BK reactivation in renal transplant recipients [83].

The ability to construct viral genomes at single nucleotide resolution allowed us to subtype BK polyomavirus using a SNP-based decision tree [82]. Of the twenty-two patients in this study with BKVN, we identified most had BK polyomavirus subtype I (seven patients with subtype Ia, three patients with subtype Ib-1, eight patients with subtype Ib-2), while few had BK polyomavirus subtypes II ( $n = 2$ ) and III ( $n = 2$ ). We did not observe an association between BK subtype and clinical measurements of pathogenicity, though studies have indicated subtype distribution differs by geographic location [84]. For two patients with follow up time points, we compared VP1 sequences to identify any selective evolution on shorter time scales (~50 days) within a singular patient. In both cases we observed identical VP1 consensus regions.

Our patient sampling method did not allow for us to observe BK subtype changes in kidney recipients over time. However, we were able to reconstruct a phylogenetic tree for the twenty-two patients with clinically diagnosed BKVN. We aligned non-human reads directly to the BK polyomavirus reference genome [48] (NCBI accession NC\_001538) and removed duplicate regions (Samtools removedup) [50]. We then constructed a consensus sequence by calling the most abundant nucleotide at each positive across the genome (Samtools mpileup and vcftools). We restricted the consensus sequence to the VP1 gene segment (positions 1564 to 2652 bp) on the forward strand. BK polyomavirus VP1 protein is the outer capsid protein and is most often used for subtyping. We used sampling dates recorded by clinicians and the 988 bp VP1 consensus regions for all twenty-two samples into BEAST [85], a tool to construct phylogenetic trees using Bayesian evolutionary analysis. We ran 250 million samplings in the Monte Carlo simulator assuming constant population size and heterogeneous rates of evolution among the three positions in each codon. The program produced a highest maximum likelihood tree (Fig. 1.3.4.2) that identified two distinct lineages beginning in 2008 (21 of 22 samples were sequenced after this point). We analyzed the tree with respect to dispersion of polyomavirus particles as seen in needle biopsies. We observed a higher likelihood of diffuse infection for patients with BK polyomavirus in the upper branch of the phylogenetic tree (Fig.1.3.4.2). The probability of six of the seven diffuse cases in the upper branch is low,  $P = 0.118$ , but not significant. By comparing the date of biopsy with the phylogenetic information, we could determine the



**Figure 1.3.5.1** Estimating bacterial population growth rates from urinary cfDNA. (a) Normalized bacterial genome coverage for four representative bacterial species. The coverage was binned in 1 kbp tiles and normalized. Each panel represents a single sample, with the exception of *C. acnes* (\*) for which the coverage was aggregated across 99 samples (solid line is a LOESS filter smoothing curve, span = 0.70). The non-uniform genome coverage for *E. coli* and *K. pneumoniae*, with an overrepresentation of sequences at the origin of replication, is a result of bi-directional replication from a single origin of replication. The initial and final 5% of the genome is removed for display. (b) Box plots of growth rates for species in 14 genera grouped by patient groups (at least 2500 alignments, 41 samples, see Methods for definition of pre/post-UTI). Each point indicates a bacterial species in a sample. Triangles indicate culture-confirmed bacteria by genus. Boxplot features are described in Section 1.3.9. [17]

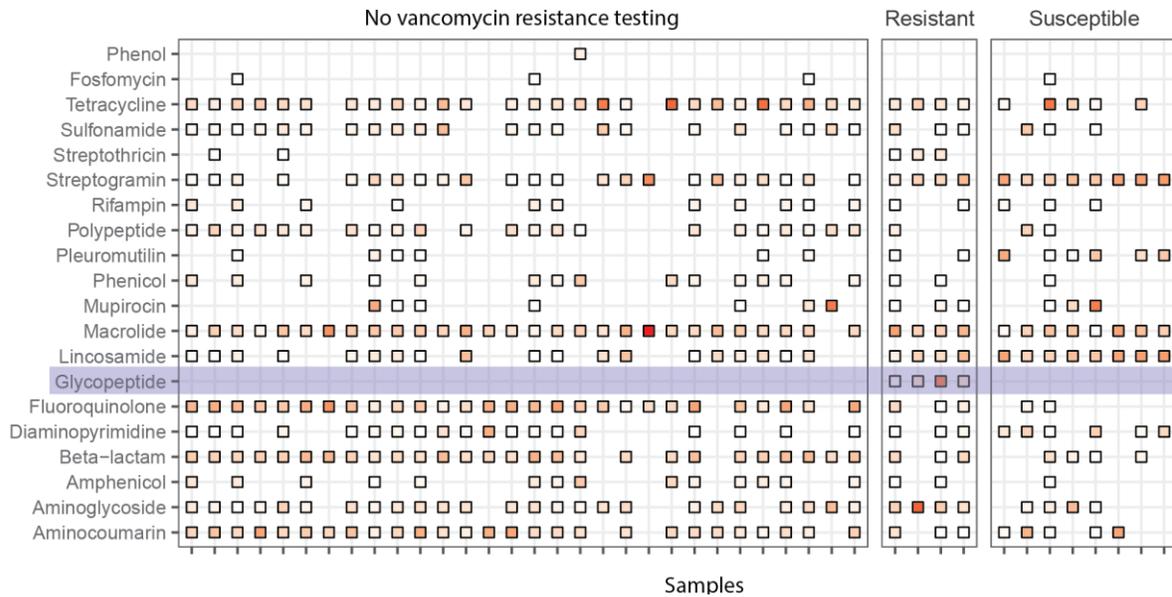
average rate of mutation of polyomavirus among those infected in the cohort as  $7.6 \times 10^{-6}$  substitutions per site per year on the VP1 exon. This is similar to previous reports for BK polyomavirus isolated from transplant recipients [86]. Our analysis indicates the addition of more samples and samples taken longitudinally from the same patients will identify the inter- and intrahost evolutionary patterns.

### 1.3.5 Quantifying bacterial growth rates

Conventional metagenomic sequencing can provide a snapshot of the microbiome, yet does not inform about microbial life cycles or growth dynamics. In a recent study, Korem and colleagues reported that the pattern of metagenomic sequencing read coverage across a microbial genome can be used to quantify microbial genome replication rates for microbes in complex communities [87]. We tested whether this concept can be used to estimate bacterial population growth from measurements of cfDNA. Figure

1.3.5.1a shows the urinary cfDNA sequence coverage for four bacterial species, *E. coli*, *K. pneumoniae*, *Gardnerella vaginalis* and *Cutibacterium acnes*. For two subjects diagnosed with *E. coli* and *K. pneumoniae* UTI (Fig. 1.3.5.1a), the *E. coli* and *K. pneumoniae* genome coverage was non-uniform, with an overrepresentation of sequences at the origin of replication and an underrepresentation of sequences at the replication terminus. The shape of the *E. coli* and *K. pneumoniae* genome coverage is a result of bi-directional replication from a single origin of replication. The skew in genome coverage reflects the bacterial population growth rate, where a stronger skew signals faster population growth [88]. The genome coverage of a common inhabitant and sometimes uropathogenic bacterial species, *G. vaginalis*, exhibited non-uniform genome coverage (Fig. 1.3.5.1a), similar to the *E. coli* and *K. pneumoniae* cases above but less pronounced. *C. acnes* has been recognized as a common skin commensal and contaminant in the setting of molecular assays [89]. The genome coverage for *C. acnes*, was highly uniform, indicative of slow or no growth (aggregate across 99 samples, which had *C. acnes* cfDNA detected, Fig. 1.3.5.1a).

We asked whether this measure of bacterial growth can be used to inform bacterial UTI diagnosis. We calculated an index of replication based on the shape of the sequencing coverage using methods described previously [88]. We used BLAST to identify abundant bacterial strains and then re-aligned all sequences with BWA [48] to a curated list of bacterial species. Samples for which the genome coverage was too sparse were excluded from this analysis (see Section 1.3.9). Figure 1.3.5.1b compares the index of replication for bacteria detected by cfDNA in samples from subjects diagnosed with UTI, to the index of replication for bacteria detected by cfDNA in samples from subjects with negative cultures and in samples collected from subjects prior to UTI development (Pre-UTI Group) or after UTI development (Post-UTI Group). Species categorized in the UTI group had markedly greater growth rates, than those in the no UTI and pre-/post-UTI groups (two-tailed Wilcoxon rank sum test,  $p = 9.0 \times 10^{-3}$ ).



**Figure 1.3.6.1** cfDNA-based antimicrobial resistome profiling reveals vancomycin resistant enterococcus. For 42 samples from subjects with clinically confirmed UTI, AR gene profiling reveals the presence of genes conferring resistance to various antimicrobial classes. These data are organized in three sample groups: samples from subjects with vancomycin-resistant *Enterococcus* (Resistant), samples from subjects with vancomycin-susceptible *Enterococcus* (Susceptible), and samples from subjects for which vancomycin resistance testing was not performed. Blue highlight indicates the AR class in which vancomycin resides. Black squares indicate at least one alignment. More than one alignment is indicated by red shading. {18}

### 1.3.6 Antimicrobial resistome profiling

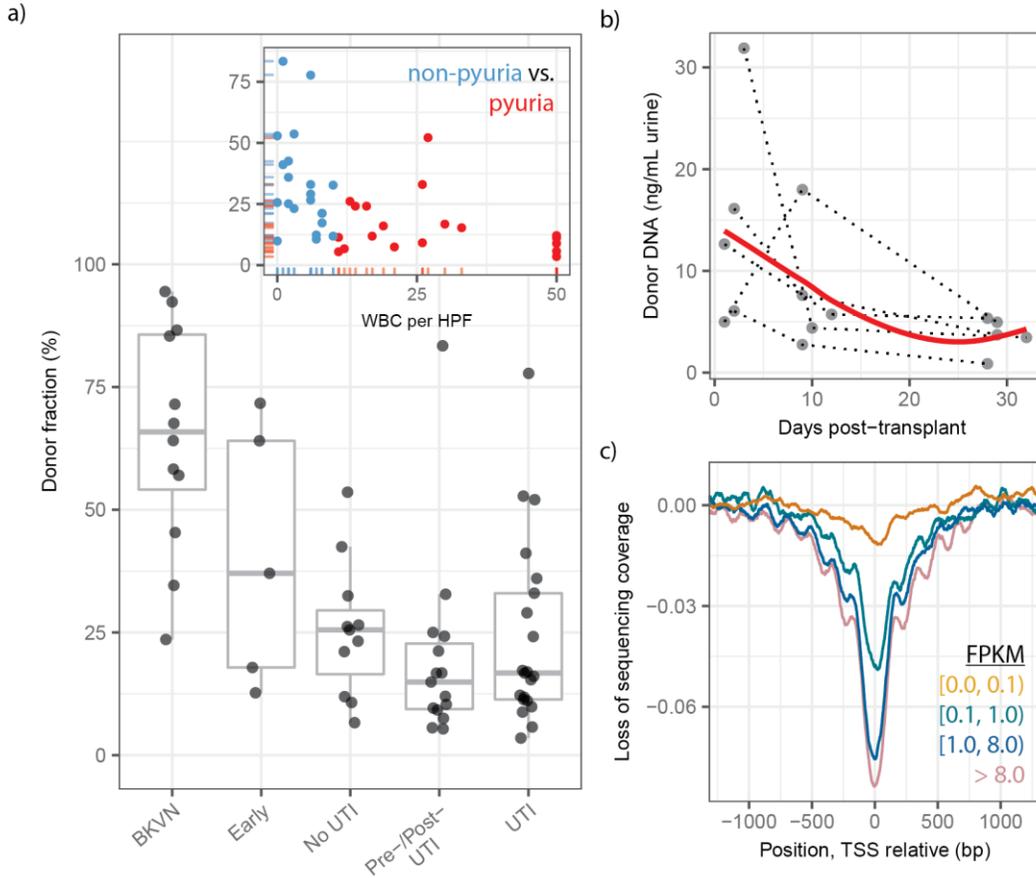
For 42 of 43 samples collected from subjects with clinically confirmed UTIs, we determined the relative abundance of genes conferring resistance to several classes of antimicrobials (a single sample, for which no AR gene fragments were observed, was excluded from this analysis). We used blastp to align non-human sequences against known AR genes and mutations [90]. AR gene sequences were aggregated and called against the non-redundant Comprehensive Antibiotic Resistance Database that indicates the drug resistance conferred by the given gene [90].

We compared the results of phenotypic antimicrobial susceptibility testing (see Section 1.3.9) to the resistance profiles determined by cfDNA sequencing. For most samples, there was a high diversity in alignments with highly abundant resistance classes including resistance to macrolides, aminoglycosides, and beta-lactams (Fig. 1.3.6.1). We studied vancomycin-resistant *Enterococcus* (VRE) infections, which

often lead to complications after transplantation [65], in depth. Resistance to vancomycin was clinically assessed via measurement of the minimum inhibitory concentration value using broth microdilution on the MicroScan WalkAway platform according to the manufacturer's instructions. We detected fragments of genes conferring resistance to the glycopeptide antibiotic class, of which vancomycin is a member, for all VRE positive samples (n = 4). Moreover, for samples with *Enterococcus* that tested as vancomycin susceptible (n = 7), we did not detect fragments of glycopeptide class resistance genes (Fig. 1.3.6.1). These data indicate potential to predict antimicrobial susceptibility from measurements of urinary cfDNA.

### 1.3.7 Measuring the host response to infection

We next examined the host response to viral and bacterial infections. Recent work has identified transplant donor-specific cfDNA in plasma as a marker of graft injury in heart, lung, liver and kidney transplantation [2, 3, 33, 91]. Here, we quantified donor-specific cfDNA in urine for sex-mismatched donor recipient pairs (i.e., male donor, female recipient; female donor, male recipient) by counting cfDNA molecules aligning to the human Y chromosome (Fig. 1.3.7.1a). We observed elevated levels of donor cfDNA in the urine of subjects diagnosed with BKVN (mean proportion of donor DNA 65.1%, n = 12) compared to the urine of subjects who had normal biopsies (no BKVN, mean 51.4%, n = 4) and samples from subjects who did not develop a clinical UTI in the first three months of transplantation (mean 25.5%, n = 11, samples collected within five days after transplant excluded). The release of donor DNA reflects severe cellular and tissue injury in the graft, a hallmark of BKVN. In contrast to subjects with BKVN, subjects diagnosed with bacterial UTI had lower proportions of donor DNA as compared to individuals without bacterial UTI. This is likely explained by an elevated number of recipient immune cells in the urinary tract following immune activation. Indeed, comparison to clinical urinalysis indicates that the donor fraction decreases with increasing white blood cell (WBC) count per high power field (HPF) 400x microscope magnification (inset Fig. 1.3.7.1a; corr. = -0.57, Spearman,  $p = 1.3 \times 10^{-4}$ ). Furthermore, clinical cases of pyuria, defined as greater than ten WBC per HPF [92], had a lower donor fraction than those



**Figure 1.3.7.1** The host response to infection can be quantified using urinary cfDNA. (a) Proportion of donor-specific cfDNA in urine of subjects that are BKVN positive per kidney allograft biopsy (BKVN) in urine collected in the first 5 days after transplant surgery (Early), urine collected from subjects that are bacterial UTI negative per culture in the first month following transplantation (No UTI), samples collected before or after bacterial UTI (Pre-/Post-UTI), and samples collected at the time of bacterial UTI diagnosis (UTI). The single outliers in Pre-/Post-UTI and UTI groups correspond to the same patient, who suffered an acute rejection episode in the months prior. Low donor fractions in the Pre-/Post-UTI and UTI groups are likely due to increased immune cell, i.e., WBC, presence in the urinary tract; subjects with higher WBC counts have lower donor fractions (inset, red color indicates pyuria). (b) Absolute abundance of donor cfDNA in the urine of subjects not diagnosed with infection in the first month post-transplant (red line is a LOESS filter smoothing curve, span = 1). Dotted lines connect samples from the same patient. (c) Genome coverage at the transcription start site, binned by the gene expression level across all samples in the study. TSS = transcription start site. FPKM = fragments per kilobase of transcript per million mapped reads, an RNA-seq measure of gene expression. {19}

without (two-tailed Wilcoxon test,  $p = 8.0 \times 10^{-4}$ ). In addition, we found that the level of donor cfDNA in the first few days after transplant was elevated, consistent with early graft injury. We tracked the relative and absolute abundance of donor-specific urinary cfDNA in the first few days after transplantation for a small

subset of subjects (n = 5). The initial elevated level of donor cfDNA quickly decayed to a lower baseline level (Fig. 1.3.7.1b), in line with previous observations in heart and lung transplantation [2, 3].

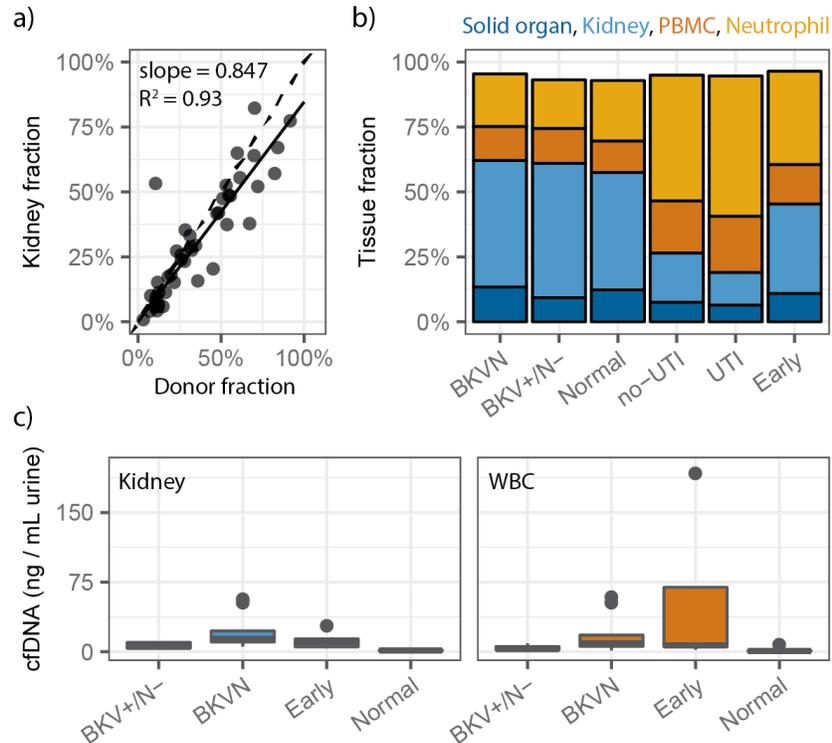
Two studies recently demonstrated that the structure of chromatin in gene promoters is conserved within circulating cfDNA in plasma [23, 24]. Ulz et al. employed whole-genome sequencing of plasma DNA to show that nucleosomal occupancy at transcription start sites results in different read depth coverage patterns for expressed and silent genes [24]. Here, we found that footprints of nucleosomes in gene promoters and transcriptional regulatory elements are conserved within urinary cfDNA (Fig. 1.3.7.1c, aggregation and normalization across all samples), and that the extent of nucleosomal protection is proportional to gene expression. Measurements of nucleosomal depletion can serve as a proxy for increased gene expression and may be used to investigate host-pathogen interactions in more detail.

Mitochondrial cfDNA (mt-cfDNA) in the urine was recently identified as a possible biomarker for hypertensive kidney damage [93]. Furthermore, recent data indicate a role for extracellular mitochondrial DNA as a powerful damage-associated molecular pattern (DAMP) [55]. Elevated levels of mtDNA in plasma have been reported in trauma, sepsis and cancer, and recent studies have identified mtDNA released into the circulation by necrotic cells [94]. For a small subset of subjects diagnosed with BKVN (eight samples from seven subjects), we quantified donor- and recipient-specific mt-cfDNA in urine, using an approach we have previously described in Section 1.2 [4]. We found that the graft is the predominant source of mitochondrial urinary cfDNA in seven of the eight samples (two-tailed Student's t-test,  $p < 10^{-6}$ ; see Section 1.3.9). Molecular techniques to track DAMPs in urine released in the setting of kidney graft injury may provide a non-invasive window into the potential role of these molecules in immune-related complications.

### 1.3.8 Identification of the tissue-of-origin of urinary cfDNA

Metagenomic cfDNA sequencing is only able to determine donor from recipient molecules by analysis of SNPs [2, 3, 95] or sex chromosomes, as shown in Section 1.3.7. It follows that tissue- and cell-level origin of molecules is difficult to determine due to the lack of SNP markers. We have previously mentioned techniques using cfDNA sequencing coverage to infer tissue-of-origin of the collection of cfDNA molecules [23, 24], though these techniques require sequencing at high depth. However, epigenetic marks are unique to DNA originating from particular organs, tissues, and cell lines within an individual organism [96]. One such signature present with tissue-level uniqueness is the presence of cytosine methylation in eukaryotic DNA [97]. 5-methylcytosine is a transcriptional regulator and is most often associated with the suppression of transcription for particular genes [97]. These marks persist on cfDNA molecules, but are not observed using standard sequencing methods. However, when sodium bisulfite is applied to the single-stranded DNA, unmethylated cytosines are deaminated into uracils, and methylated cytosines are unaffected [98]. When standard library preparation is carried out on these molecules, the resulting sequencing libraries carry a thymine in the place of unmethylated cytosines and retain their identity as a cytosine if a methyl group was present originally [98]. By comparing the treated sample to reference genomes, it is possible to determine the location of methylated CpGs at single base resolution [99].

Recent studies have applied the bisulfite treatment (BT) protocol to plasma and urine cfDNA [100, 101]. However, the bisulfite conversion process requires that DNA molecules are single-stranded, thus reducing the efficiency of library preparation using standard methods [102]. We applied our single-stranded library preparation approach to a subset of samples from the kidney transplant cohort ( $n = 51$ ), to test the efficiency of the library preparation process from cfDNA acquired from a milliliter of urine or less. Conversion efficiencies for this process exceeded 90% in samples (as measured at CH dinucleotides), and we were able to sequencing hundreds of millions of molecules per sample, with low duplication rates.



**Figure 1.3.8.1** Bisulfite-treated cfDNA sequencing indicates tissue of origin of cfDNA ensemble. (a) Measured donor fraction is compared to the kidney fraction for bisulfite-treated urinary cfDNA samples. Solid line indicates exact match. (b) The mean tissue proportion for each pathological group in the study is shown. (c) Tissue fraction is multiplied by total cfDNA extracted in urine to determine amount of cfDNA originating from organ. We compare Kidney (left) and White Blood Cell (WBC, right) cfDNA levels (ng / mL urine) across four pathological groups. {20}

A full description of techniques used in BT cfDNA sequencing analysis are outside of the scope of this work. Briefly, a collection of differentially methylated regions (DMRs) from a series of reference databases from whole-genome BT sequencing is filtered to reveal regions of the genome uniquely describing the methylation state of tissues within the reference [96, 99]. The methylation percentage is then calculated across the genome of the cfDNA sample and quadratic programming is used to delineate the contribution of reference tissues to the samples [103]. We applied the sex-mismatched donor fraction measurements described above to the BT cfDNA samples and compared that to the fraction of molecules determined to originate from the kidney (Fig. 1.3.8.1a). The measurements were highly correlated with one another (corr. = 0.879, Spearman,  $p < 10^{-10}$ ,  $n = 44$ ), and agreed with donor fraction determined in matched samples in our previous work. Using linear regression (with y-intercept fixed at 0) we found the tissue

fraction was 84.7% of the measured donor fraction, on average. In two samples we identified a high donor fraction but a low kidney fraction. This patient received a bone marrow transplant from the kidney donor, making it impractical to determine the donor fraction, while the kidney fraction could still be obtained.

Importantly, when breaking the samples down into patient groups, an average profile of all the tissues contributing to the cfDNA in the urinary tract can be observed (Fig. 1.3.8.1b). For instance, in samples with corresponding BK polyomavirus infections (with and without nephropathy), we observed an enriched fraction in cfDNA originating from the kidney compared to other tissue types (~50% from kidney). In samples with clinically-confirmed UTIs, we determined an enrichment in neutrophils and other WBC groups compared to healthy controls, confirming the insights we made in the previous study (Fig. 1.3.8.1b).

To further stratify the patient groups, we multiplied the tissue fractions by the concentration of cfDNA in urine, measured after cfDNA extraction. This measurement revealed the total amount of cfDNA originating from each tissue present in the urine. We observed significant enrichment ( $p < 0.01$ , two-tailed Wilcoxon rank sum test) in the amount of kidney-derived cfDNA between patients with BK polyomavirus nephropathy compared to patients with BK polyomavirus reactivations without nephropathy (BKV+/N-) (Fig. 1.3.8.1c). Similarly, cfDNA originating from leukocytes was enriched in patients with BKVN and BKV+/N-, compared to patients with normal biopsies. These measurements allow for the stratification of samples by pathology with high resolution. We also determined that the use of bisulfite treatment does not alter measurements of the relative abundance of microbial pathogens. Our results indicate that BT cfDNA sequencing may be used to more deeply describe the origin of cfDNA as well as distinguish cases of infection from infectious disease.

### 1.3.9 Detailed methods and sample selection

*Study cohort and sample collection.* One hundred and forty one urine samples were collected from kidney transplant recipients who received care at New York Presbyterian Hospital–Weill Cornell Medical Center.

We assayed urine samples from a total of 82 subjects. We assayed urine samples from a total of 82 subjects. The study was approved by the Weill Cornell Medicine Institutional Review Board (protocols 9402002786, 1207012730, 0710009490). All patients provided written informed consent. Bacterial group: We included 99 urine samples from 34 subjects who developed bacterial UTI diagnosed within the first 12 months of transplantation and 14 subjects who never developed UTI within the first 3 months of transplantation. For the 34 subjects who developed UTIs, we assayed 43 urine samples corresponding to same day positive urine cultures (UTI group); we assayed 15 urine samples from 15 subjects, collected at least 2–16 days (median 7 days) prior to development of the positive urine cultures (Pre-UTI group), and we assayed 12 urine samples from 9 subjects, collected at least 3–26 days (median 9 days) after development of the positive urine cultures (Post-UTI group) (7 of the 9 subjects were treated with antibiotics). We assayed a total of 29 samples collected within 3 months after transplantation from 14 subjects who never developed UTI in the first 3 months of transplantation. Viral group: The study further included 25 samples from 23 subjects who had a corresponding positive diagnosis of BKVN by needle biopsy of the kidney allograft (BKVN-positive group, 1 sample was also associated with a positive bacterial urine culture) and 10 samples from 10 subjects who had a normal protocol biopsy and were negative for BKV (BKVN-negative group). Finally, the study analyzed seven samples from three subjects who developed clinically diagnosed rare viral infections, including parvovirus or adenovirus.

*Conventional bacterial culture, bacterial identification.* Ninety of the 141 samples in the study had a corresponding same day urine culture. Each of these clean-catch midstream culture urine samples was inoculated onto tryptic soy agar with sheep blood (Becton, Dickinson and Company [BD], Franklin Lakes, NJ) and MacConkey agar (BD) using a 1- $\mu$ L inoculation loop and incubated in ambient air at 35 °C. Four urine samples were reported as mixed bacterial flora and were excluded because of lack of further identification and 86 samples were defined as either negative urine culture (n = 43) or positive urine culture (n = 43) for the cfDNA/bacterial correlation analyses and ROC analyses. A positive urine culture was defined as a culture growing an organism identified to at least the genus level (almost all bacterial isolates

were recovered at a colony count  $\geq 10,000$  cfu/mL, while three isolates were recovered at a colony count  $< 10,000$  cfu/mL). A urine culture was defined as negative when either no organism was isolated in culture (35 cultures,  $< 1000$  cfu/mL) or the organism was unidentified to either the genus or species level (i.e., unidentified) and the colony count was  $< 10,000$  cfu/mL (8 cultures). Bacterial isolates were identified using either abbreviated identification algorithms [104] or MicroScan (Beckman Coulter, Inc., West Sacramento, CA) identification panels: Neg ID Type 2 panel for Gram-negative bacteria and Pos Combo 33 panel for Gram-positive bacteria, in conjunction with the WalkAway plus system (Beckman Coulter, Inc.). In two cases, the organism isolated in culture was identified using the MALDI Biotyper CA System (Bruker Daltonics, Inc., Billerica, MA). Testing on the WalkAway plus and MALDI Biotyper CA systems was performed per the manufacturer's instructions.

*Antimicrobial susceptibility testing.* Antimicrobial susceptibility testing was performed using broth microdilution or disk diffusion. Broth microdilution testing was accomplished using MicroScan antimicrobial susceptibility testing panels on the WalkAway plus system according to the manufacturer's instructions. Gram-negative organisms were tested using the Neg MIC 42 panel or the Pos Combo 33 panel for Gram-positive organisms. In a single instance, an *E. faecalis* isolate was assayed with the Pos MIC 34 panel. An isolate of *H. influenzae* was tested using the disk diffusion method as recommended by the Clinical and Laboratory Standards Institute (CLSI) M02-A12 [105]. All antimicrobial susceptibility data were interpreted according to the CLSI M100 document. The M100 version used for interpretation varied depending on the year the isolate was recovered in culture: M100-S25 (2015) [106], M100-S26 (2016) [107], and M100-S27 (2017) [104].

*Conventional viral identification.* Quantitative adenovirus and parvovirus B19 PCR was performed on urine samples in an outside reference laboratory (Viracor Eurofins, Lee's Summit, MO).

*Analysis of discordance against bacterial culture.* In a single sample, urinary cfDNA did not identify the organism reported by conventional culture: *R. ornithinolytica*. The patient had developed an *E. coli* UTI on

postoperative day 6 and was treated initially with aztreonam but switched to cephalexin for a 14-day course. The subject subsequently developed a UTI that conventional bacterial culture revealed to be *R. ornithinolytica* on postoperative day 25. cfDNA analysis on urine samples collected on postoperative days 6 and 25 revealed a high abundance of *E. coli* cfDNA and no evidence of *R. ornithinolytica* cfDNA. Given the discordant results, it is unclear if the second culture growing *R. ornithinolytica* is a recurrence as suggested by the cfDNA analysis or is an infection with a different organism as suggested by the urine culture data.

*Urine collection and supernatant isolation.* Most urine samples were collected via the conventional clean-catch midstream culture method (n = 130). Samples obtained prior to post-transplant day 4 were collected via indwelling catheter (n = 11). Approximately 50 mL of urine was centrifuged at  $3000 \times g$  on the same day for 30 min and the supernatant was stored at  $-80^{\circ}\text{C}$  in 1 or 4 mL aliquots (except for a single *H. influenzae* UTI sample which was centrifuged for cfDNA analysis 5 days after collection). cfDNA was extracted from 1 mL (131 samples) or 4 mL (10 samples) of urine according to the manufacturer's instructions (Qiagen Circulating Nucleic Acid Kit, Qiagen, Valencia, CA).

*Negative control.* To control for environmental and sample-to-sample contamination, a known-template control sample (IDT-DNA synthetic oligo mix, lengths 25, 40, 55, 70 bp; 0.20  $\mu\text{M}$  eluted in TE buffer) was included with every sample batch and sequenced to a fraction of the depth of the cfDNA extracts (~5 million fragments). The number of bacterial and viral reads detected in the controls was quantified for each genus and normalized to the total reads across the controls. This fractional representation was used to filter out genera detected at low level in the clinical samples: any genus for which the fractional representation in the clinical sample was within five standard deviations from the mean measured in the controls was removed. Possible sources of contamination in these experiments include: environmental contamination during sample collection in the clinic, nucleic acid contamination in reagents used for DNA isolation and library preparation, and sample-to-sample contamination due to Illumina index switching [108].

*Library preparation and next-generation sequencing.* Sequencing libraries were prepared using a single-stranded library preparation optimized for the analysis of ultrashort fragment DNA, described in Section 1.2. Libraries were characterized using the AATI fragment analyzer. Samples were pooled and sequenced on the Illumina NextSeq platform (paired-end,  $2 \times 75$  bp). Approximately 45 million paired-end reads were generated per sample.

*Determining the composition of the urinary microbiome.* Low-quality bases and Illumina-specific sequences were trimmed (Trimmomatic-0.32 [47]). Reads from short fragments were merged and a consensus sequence of the overlapping bases were determined using FLASH-1.2.7. Reads were aligned (Bowtie2, very sensitive mode [52]) against the human reference (UCSC hg19). Unaligned reads were extracted, and the non-redundant human genome coverage was calculated (Samtools 0.1.19 rmdup [50]). To derive the urinary microbiome, reads were BLASTed (NCBI BLAST 2.2.28+) to a curated list of bacterial and viral reference genomes [109]. The relative abundance of different species in a sample was estimated based on the BLAST reports using GRAMMy, a software that implements a maximum likelihood algorithm and takes into account the ambiguity of read mapping [1, 3, 54]. The relative abundance of higher level taxa was determined based on the relative abundance at the strain or species level. For positive identification of viruses, we required at least 10 BLAST hits. In addition, due to the high load and genetic similarity of BK and JC polyomaviruses, we implemented a conservative filter for incompleteness and heterogeneity of genome coverage (Gini index  $< 0.8$  with at least 75% of the genome covered) for these two species only.

*ROC analysis.* ROC analyses were performed using the function “roc” in the R package pROC. For each species, we compared the relative genomic abundance of species (in RGE) in urine samples matched to a positive culture to the relative genomic abundance of the same species in culture-negative samples.

*Bacterial growth dynamics.* Bacterial genome replication rates were determined using the approach described by Brown et al. [88]. Briefly, all bacterial strains within a sample were sorted and the GC-skew

was used to identify the origin and terminus of replication (minimum and maximum GC-skew, respectively). Bacterial genomes were binned in 1 kbp tiles. The coverage was smoothed based on a running mean of 100 nearest neighboring tiles. The coverage in each tile was quantified and tiles were sorted by coverage. Linear regression was performed between the origin and terminus of replication after further removing the 5% least and most covered bins. The product of the slope of the regression line and the genome length was defined as the growth rate, a metric applied in previous analyses. We applied the analysis to all bacterial strains with genome lengths  $> 0.5$  Mbp,  $R^2$  linear regression correlation, as previously described,  $> 0.90$ , and Gini coefficient  $< 0.2$ , for which at least 2500 BLAST hits were detected in the sample.

*Nucleosome footprints in gene bodies.* Paired-end reads were aligned using BWA-mem [48]. The sequence read coverage in 2-kbp windows around the transcription start sites of all genes was determined using the SAMtools depth function [24]. A list of transcription start sites organized by transcriptional activity was obtained from Ulz et al. [24]. The depth of coverage was summed across genes with similar transcriptional activity.

*Proportion of donor-specific cfDNA in urine.* The fraction of donor-specific cfDNA in urine and plasma samples was estimated for sex-mismatched, donor–recipient pairs. The donor fraction was determined as:  $\{2Y/A, 1 - (2Y/A)\}$ , depending if the recipient is female or male, respectively.  $Y$  and  $A$  represent the sequencing coverage of the mappability-adjusted Y chromosome and autosomes, respectively. Sequence mappability was determined using HMMcopy [110].

*Mitochondrial donor fraction.* The proportion of donor-specific mt-cfDNA was quantified using methods previously described in Section 1.2 [4]. Briefly, mtDNA was extracted from pre-transplantation whole blood samples, amplified, underwent library preparation, and was sequenced. Processing was separate for donor and recipient samples. Raw sequencing data was trimmed and aligned to the human genome, and mitochondrial sequences were selected. We estimated the presence of each nucleotide at each position using bam-readcount (<https://github.com/genome/bam-readcount>), and a mitochondrial consensus sequence was

determined for the donor and recipient. The consensus sequences were compared and single-nucleotide polymorphisms (SNPs) discriminating the two individuals were identified. Downstream cfDNA sequence data aligning to the mitochondrial were compared to the bases with SNPs that discriminate the donor and recipient. For each SNP position across the mitochondrial genome, we determined the donor fraction by dividing the donor SNP count by the total number of donor and recipient counts at the SNP. We discarded a donor fraction estimation at a point if the depth of sequencing was  $< 50x$ . We determined the mean of the estimated donor fraction at all SNPs to quantify the mitochondrial donor fraction. One sample was removed owing to low depth of sequencing across all SNPs.

*Antimicrobial resistance profiling.* Paired-end nonhuman sequencing reads were merged (FLASH-1.2.7). Subsequently, reads were aligned to a database of protein sequences, in fasta format, of known AR genes (CARD 1.1.5, 2158 genes) using blastx [111] (e-value  $10^{-7}$ , culling limit 8 blastx hits). We implemented a filter post-alignment that eliminated reads with  $< 90\%$  similarity between the query and reference. Hits with the highest identity and overlap length were selected for each read. CARD data includes an ontology for the AR class to which each gene confers resistance, if known [90]. We matched ontology-derived antimicrobial classes to the gene alignment from blastx, giving a measure of each antimicrobial class for each gene hit. If a gene conferred resistance to multiple antimicrobial classes, each class was attributed a hit. The hits were aggregated to provide an antimicrobial susceptibility profile of the sample.

*Bisulfite treatment.* 5  $\mu$ L to 20  $\mu$ L of cfDNA in elution buffer was isolated. We applied a sodium bisulfite treatment protocol to the eluate (Zymo EZ DNA Methylation Kit, Irvine CA). After treatment, samples were eluted in 30  $\mu$ L of buffer. We applied ssDNA library preparation to 15  $\mu$ L in the manner described above and sequenced to a depth of roughly 2x across the human genome (Illumina NextSeq 2x75 bp).

*BT cfDNA sequencing pipeline.* BT cfDNA sequencing libraries were aligned to the human reference genome (hg19) using bwa-meth [112]. Tissue-of-origin measurements were calculated by quadratic

programming (using limSolve package in R) where the reference was based on DMRs of a tissue panel. DMRs were found using metilene [99].

*Statistical analysis.* All statistical analyses were performed using R version 3.3.2. Unless otherwise noted, groups were compared using the nonparametric Mann–Whitney U test. Fourier analyses were performed using the spec.pgram function, part of the standard stats package, in R.

*Boxplots.* Boxes in the boxplots indicate the 25th and 75th percentiles, the band in the box indicates the median, lower whiskers extend from the hinge to the smallest value at most  $1.5 \times$  IQR of the hinge, and higher whiskers extend from the hinge to the highest value at most  $1.5 \times$  IQR of the hinge.

*Data availability.* The sequencing data that support the findings of this study are made available in the database of Genotypes and Phenotypes (dbGaP), accession number phs001564.v1.p1.

Supplementary material pertinent to this data can be found at <https://www.nature.com/articles/s41467-018-04745-0#Sec30>.

### 1.3.10 Urinary cfDNA accurately predicts infections and host cell damage

We have presented a strategy to identify and assess infections of the urinary tract based on profiling of urinary cfDNA and ‘omics analysis principles. We show that different layers of clinical information are accessible from a single assay that are either inaccessible using current diagnostic protocols, or require parallel implementation of a multitude of different tests. In nearly all samples with clinically reported viral or bacterial infection of the urinary tract, cfDNA sequencing identified the suspected causative agent of infection. In addition, cfDNA sequencing revealed the frequent occurrence of cfDNA from bacteria that remain undetected in current clinical practice. In many samples, including those from subjects regarded as clinically stable, we detected cfDNA from viruses that may be clinically relevant but not routinely assayed in the screening protocol at our institution. The assay we present has the potential to become a valuable tool to monitor bacteriuria and viruria in kidney transplant cohorts, and to ascertain their potential impact on allograft health.

Beyond measurement of the abundance of different components of the microbiome, urinary cfDNA provides a wealth of information about bacterial phenotypes. We show, for the first time, that analyses of the structure of microbial genomes from cfDNA sequencing allows for the estimation of bacterial population growth rates, thereby providing information about dynamics from a single snapshot. We compared the bacterial growth rates in samples with clinically-diagnosed UTI to those without diagnosed UTI and we observed higher growth rates for clinically-reported bacteria in subjects diagnosed with UTI. We further show that metagenomic analysis of urinary cfDNA can be used to infer antimicrobial susceptibility. We mined cfDNA sequencing data for AR genes, and found a good agreement between the presence of AR genes and in vitro phenotypic antimicrobial susceptibility testing results. cfDNA resistome profiling may have added potential over conventional AR testing methods as these methods typically use one or a few cultured colonies. cfDNA profiling can potentially capture AR gene fragments from the entire bacterial population which may be particularly important since cfDNA profiling revealed frequent putative co-infections within the UTI group.

Several new methodologies have been introduced in recent years to characterize the urinary microbiome and to diagnose UTI, including 16S ribosomal RNA sequencing [79, 113, 114] and expanded culture techniques [67, 115]. These approaches have challenged the clinical dogma that urine from healthy individuals is sterile [68], and have revealed potential deficiencies in the culture protocols that are used in clinical practice today [67]. The cfDNA shotgun sequencing assay described here provides a versatile alternative that will be particularly useful for monitoring kidney transplant recipients, given the potential to enable viral and bacterial pathogen detection, AR profiling, and graft injury assessment from a single assay.

More than 15,000 patients receive lifesaving kidney transplants in the United States each year [116]. Viral and bacterial infections of the urinary tract occur frequently in this patient group and often lead to serious complications, including graft loss and death. In the general population, UTI is one of the most frequent medical problems that patients present with [59]. Shotgun DNA sequencing of urinary cfDNA offers a comprehensive window into infections of the urinary tract and could be a valuable diagnostic tool to monitor and diagnose bacterial and viral infections in kidney transplantation as well as in the general population. The assay we have presented is compatible with a short assay turnaround time (one to two days), and will benefit from continued technical advances in DNA sequencing that will reduce cost and increase throughput in years to come.

## Chapter 4: Pathogen screening and microbiome profiling from low-biomass isolates of cell-free DNA

“Cell-free DNA (cfDNA) in biological fluids provides a rich window into human health. A small fraction of cfDNA is derived from the genomes of bacteria, viruses and fungi, creating opportunities for pathogen screening and microbiome profiling from blood via ‘omics approaches. However, the total mass of microbial-derived cfDNA in blood is low, making such metagenomic analyses prone to environmental contamination. We present a scheme to identify and remove contaminating microbial DNA, as well as falsely-assigned taxa, thereby greatly improving the resolution of metagenomic assays of cfDNA. The scheme takes advantage of the coefficient of variation in genomic coverage, anti-correlation between the proportion of environmental DNA and the total biomass of the sample, and the natural and large variation in biomass across samples assayed in a single batch. We analyzed profiles of microbial cfDNA in the urine of kidney transplant patients. We compare results against clinical gold standards and demonstrate that background subtraction leads to dramatic reductions in false positive rates while minimally affecting the true positive rates. Following parameter optimization, we applied the background correction algorithm to deduce the microbiome in a novel cohort of forty-four samples derived from pregnant women, with and without chorioamnionitis. Our results support the identification of a subgroup of patients with culture-negative chorioamnionitis. We make a bioinformatics toolkit that implements the approach available as an open access R library.”

Chapter is being prepared for submission as a journal article.

#### 1.4.0 Factors contributing to false-positive microbial identification

The use of metagenomic sequencing to uncover the microbial composition of environments has become an invaluable tool to medical microbiologists and ecologists in recent years. Researchers have revealed cause, response, and association relating the microbiome, the collection of microbial organisms, to several diseases and disorders [117, 118]. Other work has determined the persistence of these microbiomes over time and when introduced to extrinsic factors, including the use of pharmacological agents [119]. However, only recently has contamination been thoroughly addressed in these samples [120]. In order to produce sequencing data representative of the community, samples must be collected and stored and their DNA needs to be extracted, prepared, and sequenced. The ubiquitous nature of bacteria and other microorganisms can introduce external agents to the samples at any of these steps. This may lead to the false-positive identification in communities.

Microbial contamination could be introduced at one of several steps in the process of extracting biological fluids to eventually producing sequencing libraries. For example, microbes existing on the skin from the patient could be coincidentally extracted into a vial at the collection point. We have observed the abundance of gut-associated bacteria more strongly in urine samples received from women, compared to similar samples taken from men or blood samples acquired from either individual - likely from the transfer of bacteria from the rectum to the genitourinary tract [121]. Bacteria, or fragmented bacterial fragments, could also be present in contaminated tubes at collection or introduced accidentally by medical or laboratory staff after collection.

Processing steps may also introduce microbial contaminants into samples. The production process of some enzymes and kits necessitates the use of microbial components (e.g. replication enzymes). A recent study highlights the perils of associating newly observed microbes with diseased patients [122, 123]. In this work, a nonhuman parvovirus was identified at low abundance in many patient samples, leading the authors to suggest the discovery of a virus in the human population [123]. However, a careful investigation of the materials in the extraction process determined that the viral genetic sequences were introduced into their

samples via silica columns [122]. These columns capture DNA extracted from cells and fluid, and they are sourced from ocean diatoms, which are the natural host of these parvoviruses [124].

It is difficult to determine the exact point in sample processing pipelines wherein environmental contamination occurs. For example, in the library preparation of cfDNA sequencing libraries there are dozens of steps to go from cfDNA suspended in biological fluid to a representative library prepared for next generation sequencing. These steps involve the introduction of reagents, containment vessels, pipette tips, etc..., all of which may introduce contaminating microbial genome fragments into samples. In addition to contamination, algorithms used to assign sequences to microbial taxa, via alignment or k-mer matching, may incorrectly assign identity due to genetic similarity. Furthermore, the nature of horizontal gene transfer among various species of bacteria can make it difficult to determine which organism is present. Finally, recent evidence has indicated the phenomenon of “barcode hopping” on next generation sequencing platforms [108]. In such cases, samples multiplexed and sequenced on the same flow cell may be incorrectly assigned to the respective samples. These modes of digital cross-contamination have not been thoroughly addressed.

While many of these phenomena have been identified and addressed separately, few approaches have integrated corrective measures for all contamination and false assignment issues, particularly in cases of low microbiome biomass metagenomic sequencing. In this section, we comprise a series of filtering steps taken to address background contamination and taxonomic assignment failures in cell-free DNA sequencing data sets. Cell-free DNA from humans has been shown to be a valuable biomarker for disease in pregnancy and following organ transplantation [2, 3, 12]. We have recently shown that bacterial and viral cell-free DNA may be obtained from the urine to diagnose urinary tract infections [5], and recent work has shown the diversity of viral cell-free DNA detectable in blood plasma [1]. Non-host cell-free DNA can often make up over 99% of identifiable sequencing reads, even in the case of infections in immunocompromised patients. Thus, microbial cell-free DNA represents a case of low-biomass

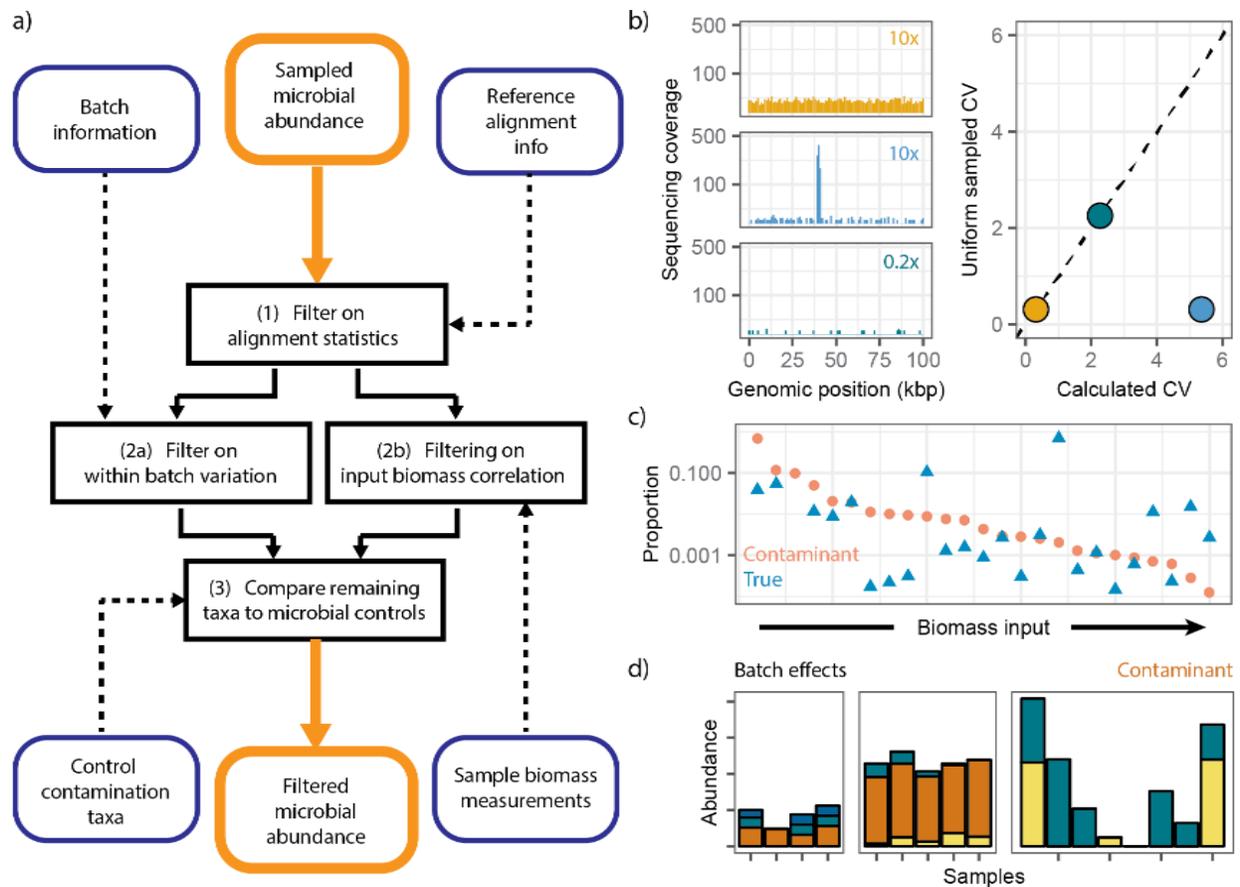
metagenomics, in which positive biomarkers of infection may be overwhelmed with the presence of the host and contaminating sequences.

Here we present our approach to identification and removal of contaminants and improperly assigned microbes in metagenomic cell-free DNA sequencing datasets, termed low-biomass background correction (LBBC). Our pipeline utilizes extraction and library preparation information, cfDNA biomass input, and microbial sequencing coverage statistics to filter false-positive identifications. We implemented a scoring system to optimize filtering parameters and applied the filtering scheme to a subset of samples from a previously published dataset for kidney transplant recipients with and without urinary tract infection.

As a follow-up study, we then applied the filtering scheme to a novel cfDNA dataset collected via amniocentesis from a cohort of pregnant women with and without chorioamnionitis. Chorioamnionitis is an inflammatory response during pregnancy that is associated with preterm birth and long term complications for the fetus. Often the underlying cause of chorioamnionitis has been determined as a bacterium, through culture and/or 16S sequencing. However, there exist a large number of cases without etiological agent. Our results confirm the sterile conditions by a new method of diagnosis, namely microbial cfDNA sequencing that correspond with clinical determination of bacterial-negative chorioamnionitis. Furthermore, cfDNA sequencing with background correction is able to identify most cases of bacterial-positive chorioamnionitis, and supports a broad range of work emphasizing a microbe-free amniotic fluid during healthy, full-term pregnancies.

#### 1.4.1 The low-biomass background correction pipeline

We built a background correction pipeline, implemented as a package in R that identifies contaminant or falsely-assigned taxa for different information inputs (illustrated in Fig. 1.4.1.1a). In all cases detailed in this chapter, our input microbial abundance was determined from the relative genomic abundance (proportion of microbial genomes per human genome in sample) from cfDNA originating in



**Figure 1.4.1.1** A pipeline for sparse metagenomic background correction with simulated cases of false identification. (a) Overview of pipeline. An initial file containing the abundance of microbes across samples is compared to external information (blue boxes) regarding alignment statistics, batch information, measured biomass, and taxa in negative control samples. (b-d) Simulated examples to identify contaminant through different means. (b) Sequencing coverage across three artificial microorganisms was simulated in the case of high and homogenous coverage (yellow), high and inhomogeneous coverage (blue), and low coverage (green). At right, the coefficient of variation (CV) was calculated for a uniformly sampled, uniformly sequenced at the same depth of sequencing for each organism and compared to the calculated CV was compared to the theoretical CV. Dotted line shows 1:1 correspondence. (c) Simulated cases for a microbe as a contaminant (orange) and true identification (blue) show the theoretical relationship between measured biomass and proportion of all sampled microbes. (d) Simulated cases of bacterial abundance for four microbes. Orange shows low variability within batches, so is likely a contaminant. [21]

human biological fluid and prepared using our extraction and library preparation pipeline [4]. However,

one could apply the pipeline to any type of microbial abundance measurements. We have employed this metric in various studies and found it accurately depicts cases of infection [5]. The measurement also normalizes microbiota both within and between samples to enable identification of enrichment. Relative abundance measurements are calculated by aligning nonhuman reads to a curated reference genome database via NCBI BLAST. The number of alignments is adjusted using GRAMMy, a maximum likelihood

estimator that adjusts reads based on sequence similarity [54]. Finally, the total number and length of assigned reads is compared to the genome length of the particular microbe, and a depth of sequencing for that microbe is determined. By dividing this value by the sequencing coverage of human chromosome 21 (and adjusting for ploidy), we determine the relative genome abundance.

While the collection of known microbiota is extremely diverse, there exists high amounts of genetic similarity between individual taxa. Even in cases of distinct genera, genomic similarity can exceed 80% [125]. Bacteria are able to transmit whole gene segments to genetically distinct organisms, to which metagenomic sequencing is sensitive [126]. For these reasons, it can be difficult to assign sequencing reads to the correct bacterial taxon through either k-mer or direct alignment strategies. Following alignment of nonhuman cfDNA reads to a microbial reference genome database, we observed that the distributions of reads across a microbial reference genome could be highly heterogeneous. While cfDNA sequencing is sensitive to protein-DNA interactions, the degree of coverage heterogeneity is not generally observed across the human genome, except in cases of low mappability. We reasoned that the pattern of genome coverage in these samples was due to either a region of low complexity or high similarity to that in the reference genome of another microbe present in the dataset.

To distinguish cases of false assignment in a generalizable manner we calculated the coefficient of variation ( $CV$ ) for the number of reads originating from 1 kbp bins across the genome. We then simulated random sampling across a uniformly sequenced genome for the same number of reads as the direct calculation case. In general, we calculated the following from the sequencing coverage across individual species:

$$CV = \sigma/\mu,$$

$$\Delta CV = CV - CV_{\text{uniform}},$$

where  $\sigma$  and  $\mu$  are the standard deviation and mean in the sequencing depth within 1 kbp bins, respectively.

Utilizing these values discriminated cases of high homogenous genomic coverage, low homogenous genomic coverage, and heterogeneous coverage (Fig. 1.4.1.1b). Coefficient of variation was chosen over other distribution metrics, such as the Gini coefficient, due to its sensitivity to heavy tails and the scales invariance [127]. We simulated a case of three equivalent bacterial genomes with low (0.2x) or high (10x) genomic coverage and differing coverage heterogeneity (Fig. 1.4.1.1b, left). In the case of a high coverage genome that was sequenced homogeneously,  $CV$  and  $\Delta CV$  were approximately zero. In the case of a low coverage genome that was sequenced homogeneously,  $\Delta CV$  was approximately zero. This indicates that the genome was sequenced as randomly, even with a low depth of sequencing. However, the genome covered with high heterogeneity shows a large discrepancy between the  $CV$  and  $CV_{\text{uniform}}$ , leading to a large  $\Delta CV$  (Fig. 1.4.1.1b, right).  $\Delta CV$  can be used therefore, to indicate and remove genomes with irregular coverage patterns.

Following removal of microbes with inhomogeneous genome coverage, we aimed to address bacteria that we likely present in the dataset, but were derived from sources of contamination. The physical mass of DNA added in library preparation can be used to distinguish contaminating taxa [120]. Following taxon assignment, aggregated to the desired taxonomic level, the relative abundance of each taxa present in the sample can be assessed. For each taxa, the proportional abundance may be compared to the total input mass for library preparation. A reduced DNA input creates more opportunity to incidentally prepare and sequence contaminant genomic fragments. In the extreme case, a pure water sample prepared for sequencing would only contain environmental microbes. Conversely, beginning with a large input would “crowd out” contaminating genomic fragments.

To utilize this method, we calculated the concentration of cfDNA following extraction from biological fluids, and marked the input volume of cfDNA into library preparation, allowing us to determine the total amount of cfDNA input. The relative proportion of each species was determined by dividing the relative genomic abundance of the given species by the aggregate relative genomic abundance of all species present in the sample. We determined if the correlation between the cfDNA input mass and the

proportionate abundance was negative and significant ( $p < 0.01$ ). The theoretical relationship between biomass input and proportionate abundance for contaminants and true microbes present in samples is depicted in Figure 1.4.1.1c. Permitted significant negative correlation for the particular taxa, the proportionate abundance was transformed using a Box-Cox transform [128], and a z-score was determined. Samples with a z-score exceeding 1.65 (representing greater than 95% above mean) are considered to have a species rising above background, while those with z-score below 1.65 are filtered out.

Simultaneously, we exploited the presence of batch effects to identify and remove bacteria with low variability within particular batches. Batch variation is a useful technique to remove species that could be introduced from reagents or broad cross contamination [129]. When calculating the variation of species within a particular batch, those with high variation can be interpreted as likely present within the sample, while those with low variation are likely an unintended contamination (simulated example in Fig. 1.4.1.1d). We annotated each sample with the batch number and calculated the within-batch variation of each taxon within the batch. Batches containing taxa below a variance threshold ( $\sigma^2$ ) were removed.

Finally, the use of negative controls in metagenomic sequencing, as previously mentioned, can directly yield evidence of contamination [130]. An easily adapted method for the identification of contaminant sequences is the inclusion of a water control. To more accurately simulate the case of low-biomass metagenomics we utilize a control which emulates the sample (e.g. in the length of molecules) but may be distinguished if there is cross contamination. We created a simple mixture of synthetic, short dsDNA oligos (25 bp, 40 bp, 55 bp, and 70 bp). Roughly 100 million oligos of this control were suspended in 1 mL of TE buffer and we processed the samples through cfDNA extraction and library preparation steps [5]. Each of these control libraries was sequenced at low depth (~1 million reads per sequencing run) and processed in the same bioinformatics pipeline as true cfDNA samples. The presence of microbiota assigned to these datasets should only be considered if they exceed a nominal abundance. Barcode-hopping, or misattributing indexing barcodes during Illumina sequencing, has been observed on multiple platforms [108]. Thus, microbial control samples, multiplexed with one or more samples containing an abundance of

infectious microbial cfDNA, could erroneously contain non-contaminants. To address this concern, we compared the relative number of BLAST hits in each sample and microbes were excluded if they exceeded an abundance threshold. We utilize this method as the last stage in the filtering pipeline.

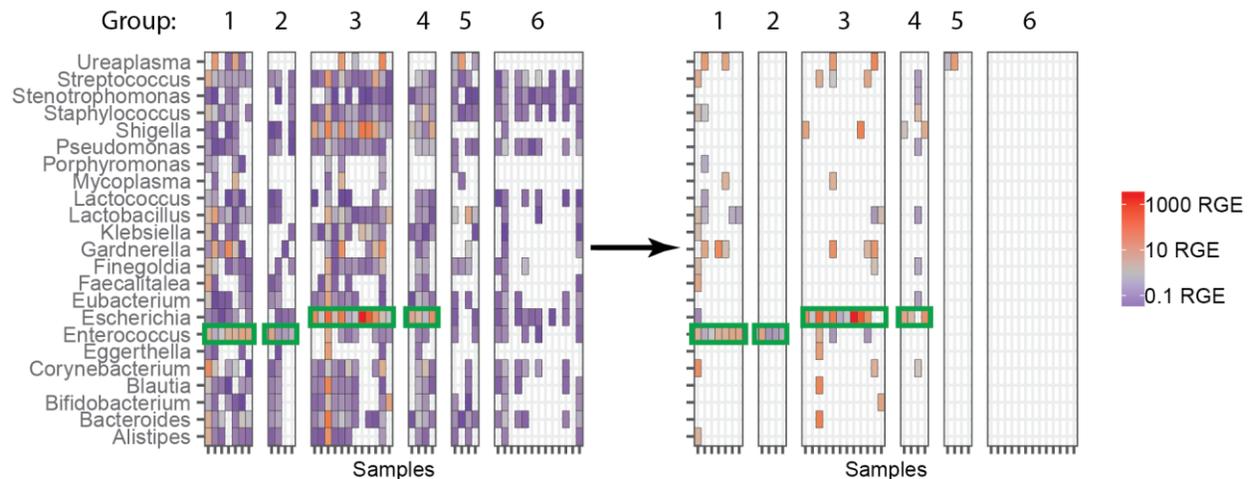
#### 1.4.2 Training correction parameters using clinically informed interpretation

In following the guidelines presented in Figure 1.4.1.1a, we have established the use of three critical thresholding parameters that affect the permissiveness of filtering, these are: (1) the maximum allowable coefficient of variation,  $CV_{\max}$ , (2) the maximum allowable difference in  $CV$  from that of a uniformly sequenced sample at the same depth,  $\Delta CV_{\max}$ , (3) and the minimum allowable within-batch variation,  $\sigma^2_{\min}$ . We determined a scoring metric which would optimize the parameters on a well-described cfDNA sequencing dataset. We selected a subset of the urinary cfDNA samples from a kidney transplant cohort. In this cohort, we included 16 samples with known *E. coli* UTIs, 11 samples with known *Enterococcus* UTIs, and 17 samples with no current UTI and no UTI during the time the patients were monitored.

We constructed a metric designed to evaluate the success of the filtering method. The score is calculated, as follows:

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

where  $\{TP, TN, FP, FN\}$  is the number of true positives, true negatives, false positives, and false negatives, respectively. We also include a term  $U$  to represent the total number of identified taxa for which a secondary methods of identification was not performed. The respective  $k$  coefficients for these values represent weights to optimize the filtering parameters based on assay interests. For example, if one wants to penalize the identification of untested bacteria,  $k_U$  should be a negative value. To optimize the filtering we chose  $\{k_{TP}, k_{TN}, k_{FP}, k_{FN}, k_U\} = \{4, 2, -1, -2, -0.25\}$ . Running a nonlinear minimization algorithm by gradient descent, we determined the optimal set of parameter thresholds as:  $\{CV_{\max}, \Delta CV_{\max}, \sigma^2_{\min}\} = \{4.267, 1.267, 0.398\}$ .



**Figure 1.4.2.1** Low-biomass background correction (LBBC) on microbial abundance reveals pathogens, commensals, and sterile biomes. We applied LBBC to the microbial abundance arrays calculated in 44 samples for *Enterococcus* UTI in females and males (Groups 1 and 2,  $n = 7$  and 4, respectively), *E. coli* UTI in females and males (Groups 3 and 4,  $n = 12$  and 4, respectively), and healthy females (Group 5,  $n = 4$ ) and males (Group 6,  $n = 13$ ). The genera detected were then subset from the microbial abundance matrix without background correction (left of arrow). Relative genome abundance is indicated by color gradient, on left. Green borders indicate positive clinical identification by standard urine culture. {22}

Applying these constraints on the kidney subset revealed true-positive rate (24/27) and true-negative rate (57/61). These filtering parameters also led to few false-positives and false-negatives (4 and 3, respectively). Moreover, there were relatively few bacteria identified outside of the *Escherichia* and *Enterococcus* genera that would be associated with contaminants or false-positives generated from sequence alignment (Fig. 1.4.2.1).

We also observed that the filtering method did not remove the presence of known commensal bacteria in the genitourinary tract as shown the abundance of *Gardnerella* and *Ureaplasma* in female samples (and not in male samples), reflecting an effect we discussed in the original publication [5]. In one sample, we identified an abundance of bacteria associated with the human gut microbiome (Group 3 in Fig 1.4.2.1). In many cases of UTI, particularly in transplant recipients, the causative uropathogen originated in the gut and traveled to genitourinary tract via external migration [121]. Metagenomic sequencing performed on the stool of this patient revealed that the *E. coli* strains between the stool and urine of this sample are more similar than comparative samples from other patients, and all genera present in urine were

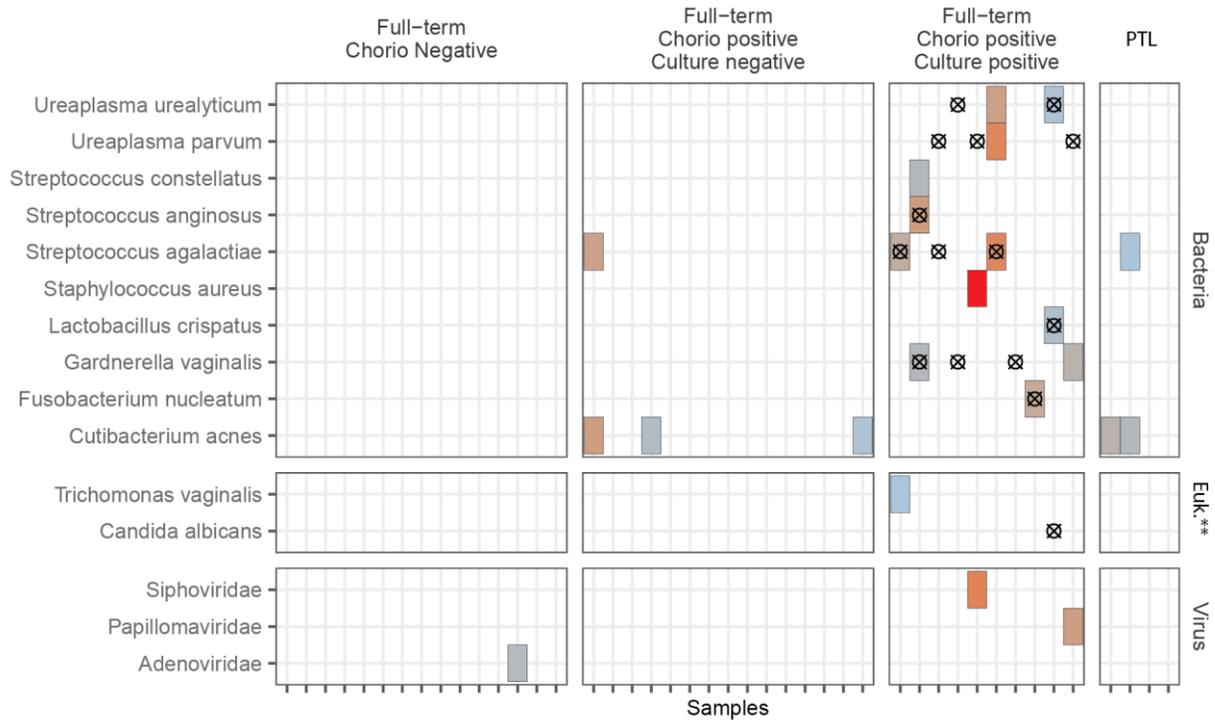
similarly present in stool. It is therefore likely that the patients' UTI was acquired from the patient's own gut microbiome (manuscript in review).

#### 1.4.3 Background corrected cfDNA sequencing confirms chorioamnionitis in sterile wombs

Acute chorioamnionitis, a host response to a chemotactic gradient in the amniotic fluid [131], is related to half of all preterm births and occurs in up to 4% of pregnancies [131, 132]. The inflammation of the fetal and maternal tissue is often instigated as a response to bacterial infection. In many cases, however, no microbial pathogen is observable through clinical techniques, such as bacterial culture or 16S sequencing.

Following the optimization of filtering parameters, we applied the low-biomass background correction pipeline to a novel dataset acquired from the amniotic fluid of a patient cohort with chorioamnionitis, as well as healthy controls. It has been shown in a multitude of cases that the amniotic fluid in healthy individuals is sterile, so it is important to reduce false positives in microbial identification [133]. We retrospectively selected 47 samples from patients with culture-positive and culture-negative chorioamnionitis. We extracted cfDNA from 175  $\mu$ L of amniotic fluid using the 1 mL urine protocol in the QiaAMP circulating nucleic acid kit, which was then prepared using a ssDNA library preparation and sequenced to about 1x depth (human genome coverage). We observed high sequencing coverage and low duplication rates in 44 of the 47 samples, allowing for further analysis. We identified microbial cfDNA through alignment to a comprehensive database using our previously described microbial alignment pipeline [1, 5] and we applied the filtering pipeline with parameters described above. Clinical identification of chorioamnionitis was determined through presentation of symptoms, and the causative pathogen was identified through culture and subsequent 16S sequencing on the IBIS system.

The results of the filtering pipeline, implemented with the optimized threshold parameters, are indicated in Figure 1.4.3.1. As expected, no bacteria were detected in the chorioamnionitis-negative group, reaffirming the utility of the filtering pipeline and supporting recent findings [134]. Furthermore, there was



**Figure 1.4.3.1** cfDNA sequencing with background correction reveals causative chorioamnionitis pathogen. Species-level abundance is shown across 44 samples in cohort. Abundance value occupy a range from  $10^{-2}$  RGE (blue) to  $10^3$  RGE (red). Samples are divided by clinical pathology and species are divided by superkingdom. Black crosshairs indicate clinical confirmation by culture and 16S sequencing. {23}

only one species detected outside of those determined through 16S sequencing, *Cutibacterium acnes*, a known skin contaminant [89]. The cfDNA sequencing assay with background correction failed to detect several cases of bacterial infection. We asked if these false negatives were due to the application of the filtering scheme. A relaxation of the thresholding parameters revealed that three of the false-negative *Ureaplasma parvum* calls were removed by the batch variation filters. *Ureaplasma* and *Mycoplasma* often inhabit the female genitourinary tract [80], and, in the absence of background correction, show a presence in many samples. The presence of these commensals likely forces their removal by background correction algorithms, an effect that is undesirable, as *Ureaplasma* are often the main cause of chorioamnionitis [135]. We did observe two patients in culture-negative chorioamnionitis group with an abundance of *Streptococcus agalactiae*, in these cases, this bacteria may have caused inflammation. However, in 17 of

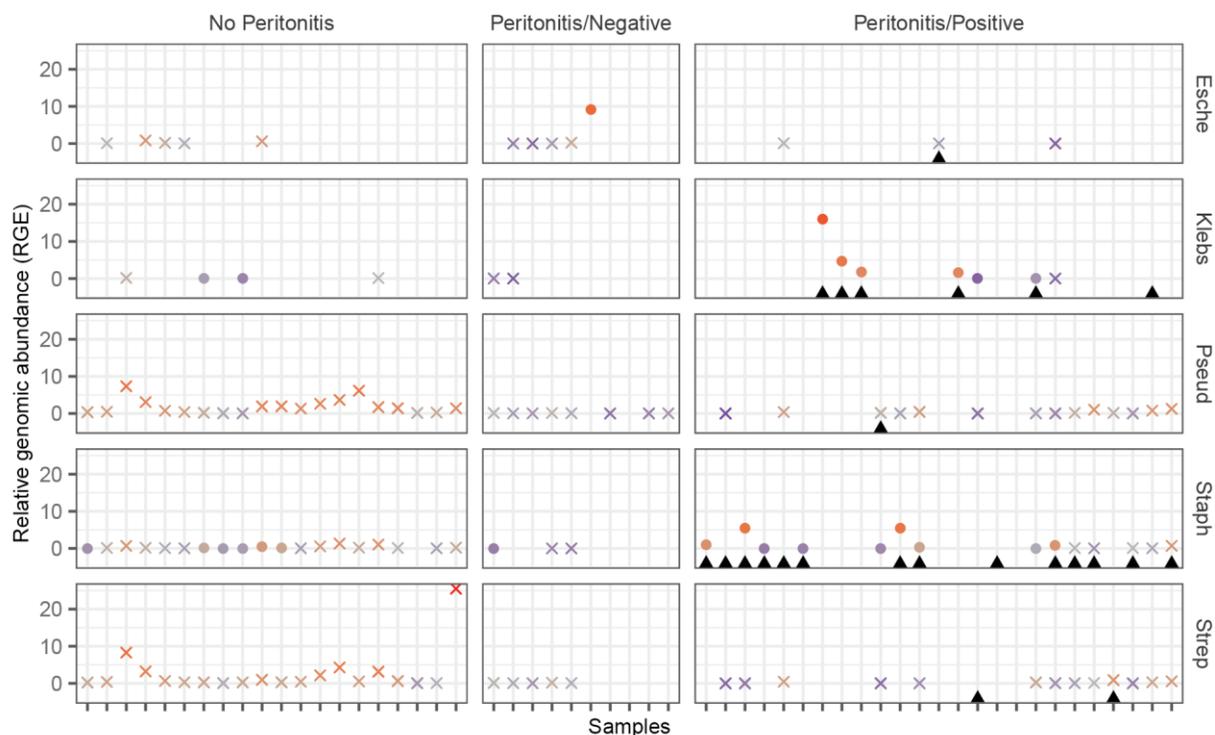
the 19 cases of culture-negative chorioamnionitis no bacterium (other than contaminant *C. acnes*) was identified.

When cfDNA is present from two individuals in a biological fluid and the individuals are of opposite sex, the fractional abundance from each individual can be determined by the ratio in coverage of the Y chromosome to the autosomes. We applied this scheme to patients pregnant with a male fetus and determined that the fractional abundance of maternal cfDNA was strongly correlated with inflammatory markers such as IL6. During chorioamnionitis, leukocytes from the mother, fetus, or both individuals may enter the amniotic sac to fight infection; this is the likely source of maternal cfDNA [136, 137]. We identified two samples outside of the culture-positive chorioamnionitis group (and four within the group) with high amounts of maternal cfDNA. We hypothesized these samples may include the presence of pathogenic bacteria or viruses that were not clinically detected; however, none were observed.

The results of background-corrected microbial cfDNA sequencing support previous reports of a sterile amniotic sac in the absence of inflammation and the occurrence of chorioamnionitis in the absence of microbial invasion of the amniotic cavity [135]. In this scenario, microbial cfDNA sequencing with background correction should be employed with other methods, such as standard culture or 16s sequencing, due to its tendency to report false negatives. cfDNA sequencing was additionally sensitive to viral species present in amniotic fluid, for which clinically testing was not performed. Our results indicate high amounts of papillomavirus and bacteriophage in two separate samples.

#### 1.4.4 Detection of microbial cfDNA in peritoneal dialysis effluent

In a follow-up study to the isolation and sequencing urinary cfDNA in kidney transplant recipients, we performed the microbial cfDNA sequencing pipeline on peritoneal dialysis (PD) effluent samples for patients awaiting renal transplant. In PD, a solution is introduced in the peritoneal cavity and filters toxins from the blood following kidney failure [138–140]. PD presents a more comfortable and accessible method of dialysis, as compared to hemodialysis, which requires patients to regularly visit centers to perform fluid



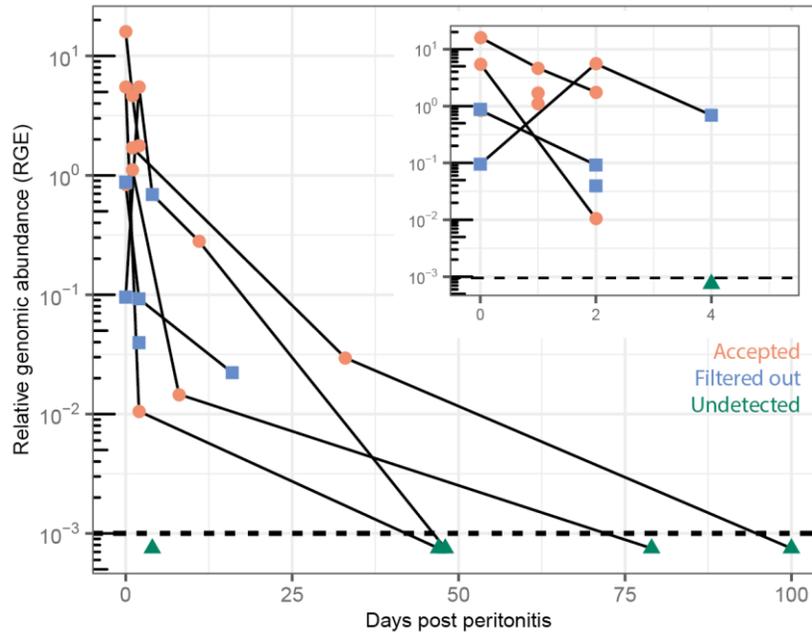
**Figure 1.4.4.1** Microbial cfDNA sequencing in PD effluent reveals pathogens and broad detection of skin flora. Relative genomic abundance is shown across samples (n = 55) for five genera tested using bacterial culture. Samples are grouped by clinical presentation of culture-positive peritonitis (Peritonitis/Positive), culture-negative peritonitis (Peritonitis/Negative), and no peritonitis. X = taxa identification removed by LBBC algorithm. ▲ = clinical confirmation of taxa in patient. Colored by log scale of relative genomic abundance. {24}

exchange. Cases of peritonitis, infection of the peritoneum, reduce the efficacy of PD and make users more susceptible to infection in the future; for these reasons, clinicians often recommend patients change to hemodialysis after a case of PD [141].

We examined the ability of cfDNA sequencing to detect and monitor cases of peritonitis by examining the microbiome in 1 mL PD effluent samples at or near cases of clinical peritonitis. We extracted cfDNA from PD effluent samples collected at New York Presbyterian Hospital where patients were monitored for peritonitis prior to kidney transplantation. If patients exhibited symptoms consistent with peritonitis, PD effluent was cultured for microorganisms. Our cohort (55 samples across 31 patients) consists of patients who had culture-positive and culture-negative peritonitis (n = 11 and 4, respectively) and patients without peritonitis (n = 16).

We observed a variable sequencing depth across the human genome in the samples ( $1.08x \pm 0.68x$  on chromosome 21) and a unique fragment length distribution (Fig. 1.2.6.1) of the human-aligned cfDNA. In particular, unlike plasma and urine cfDNA, the relative proportion of reads shorter than 100 bp was highly variable. When analyzing the microbial taxa across samples from peritonitis negative and positive patients, we observed a high amount of microbial promiscuity among reads. By subsampling the microbial genome abundance matrix to only those genera observed by PD effluent culture, we observe that there is a high amount of background. For example, while *Pseudomonas* bacteria was only isolated from one sample in culture, several samples without peritonitis and with culture-negative peritonitis exhibited high relative genomic abundance ( $RGE > 1$ , Fig. 1.4.4.1). To adjust for background, we applied the low-biomass background correction algorithm from Section 1.4.1, (though without the comparison to the microbial control). In applying the LBBC algorithm, we were able to reduce over 90% of assigned taxa, though this led to the inability to positively identify *Streptococcus* and *Pseudomonas* reads (Strep and Pseud, Fig. 1.4.4.1). Other genera, including *Escherichia*, *Klebsiella*, and *Staphylococcus*, confirmed through bacterial culture were present in high abundance after background correction.

For several patients in the cohort who had culture-positive peritonitis, we acquired multiple samples at multiple time points following clinical confirmation of peritonitis. Clinicians treated these patients immediately once peritonitis is diagnosed and the causative pathogen is determined. To demonstrate the effect of treatment on the causative pathogen, we isolated the reads originating from the pathogen determined by culture and observed the change in relative genomic abundance over time (Fig. 1.4.4.2). Notably, the relative genomic abundance did not decrease for all samples after peritonitis onset (Fig. 1.4.4.2, inset). In most cases, we did observe a decrease in the relative genomic abundance of the causative pathogen by an order of magnitude within two days of peritonitis diagnosis. Microbial cfDNA sequencing was able to detect the causative pathogen up to five weeks after peritonitis diagnosis and down to a relative genomic abundance of  $\sim 0.01$  RGE, or one bacterium for every one hundred human cells (Fig. 1.4.4.2).



**Figure 1.4.4.2** Microbial cfDNA sequencing in PD effluent is able to detect causative pathogens at low abundance and weeks after treatment. Relative genomic abundance of the clinically determined pathogen is shown for culture-positive peritonitis patients. Inset: Days 0 to 4 post-peritonitis are shown. Dotted line is the absolute limit of detection,  $10^{-3}$  RGE. Red points are both detected and remain in abundance profiles after filtering. Blue points are detected but are filtered out through LBBC algorithm. Green points were not detected through microbial cfDNA sequencing. Lines connect samples from same patient. {25}

To our knowledge, our pilot study represents the first time that cfDNA sequencing was performed on the fluid from the peritoneum. For this reason we also explored the presence of potential pathogens outside of those detected using culture. When relaxing our filtering parameters we detected the presence of *Enterococcus* (0.38 RGE), of which only one other sample had presentation at low abundance (0.01 RGE). The sample was obtained four days prior to the onset of a gallbladder infection by *Enterococcus*. Additionally, we detected the presence of several clinically-relevant viruses present in the peritoneum of several patients including cytomegalovirus (CMV, 22.37 RGE), betapapillomavirus 1 (HPV, 4.76 RGE), and herpesvirus 6A (1.28 RGE).

We hypothesize from the vignettes illustrated above (*Enterococcus* in gall bladder, CMV, and HPV), that cfDNA sequencing of PD effluent contains microbial cfDNA from both infections within the peritoneal cavity and infections in organs neighboring the peritoneum. In cases of neighboring infections, cfDNA from the microbial organisms could diffuse across barriers in the plasma or directly move the organ

into the peritoneal cavity if there are lesions. Furthermore, cfDNA originating from enteric microbiota may have originated in the intestines, colon, or stomach, all of which border the peritoneal cavity. Due to the small scale nature of this study, we cannot confirm this hypothesis beyond these vignettes. Larger studies will need to be performed to establish the utility of PD effluent cfDNA sequencing to detect infections in a wide variety of organs. If this is proven to be true, monitoring the effluent of patients on PD via cfDNA sequencing could provide a comprehensive method to detect and treat infection.

## Chapter 5: Applications of cfDNA sequencing for global health

“Metagenomic sequencing has proven to be a powerful tool in clinical settings to identify oncogenesis, fetal aneuploidy, and graft failure and infection in transplant recipients. However, cfDNA sequencing has not yet been applied to diseases of particular interest to global health, including tuberculosis, malaria, and neglected tropical diseases. To describe the need and scope of cfDNA sequencing in issues related global health, we applied our analysis of genome replication dynamics and microbiome abundance measurements to patients in rural settings in South America. We performed whole genome sequencing of cultured *Mycobacterium tuberculosis* (MTB). We calculated the proportion of MTB under active replication using the sequencing coverage across the MTB genome and observed disruption of this signal when various antibiotics were introduced. By comparing genome replication dynamics to total DNA measurements, we propose a novel observation of high ploidy genomes in static MTB after introduction with ethambutol. We present the results of sequencing plasma cfDNA from TB individuals from a cohort in Ecuador. In a separate study, we sequenced plasma cfDNA from sixty pediatric patients in Peru with suspected environmental enteropathy. Analysis of the plasma microbiome following low-biomass background correction revealed the presence of enteric microbiota in the plasma in patients with high L:M sugar ratios, a measure of disease severity. We believe this supports a model of gut perfusion during EE, allowing the diffusion of enteric microbiota into the circulatory system, which can lead to systemic disease.”

Experiments and sample collection in this chapter thanks to: Evgeniya Nazarova, David Russell, Margaret Kosek, Joan Sasing Lenz, Fanny, Chen, Jansy Sarathy, and Veronique Dartois.

### 1.5.0 cfDNA sequencing to study infectious disease in global health

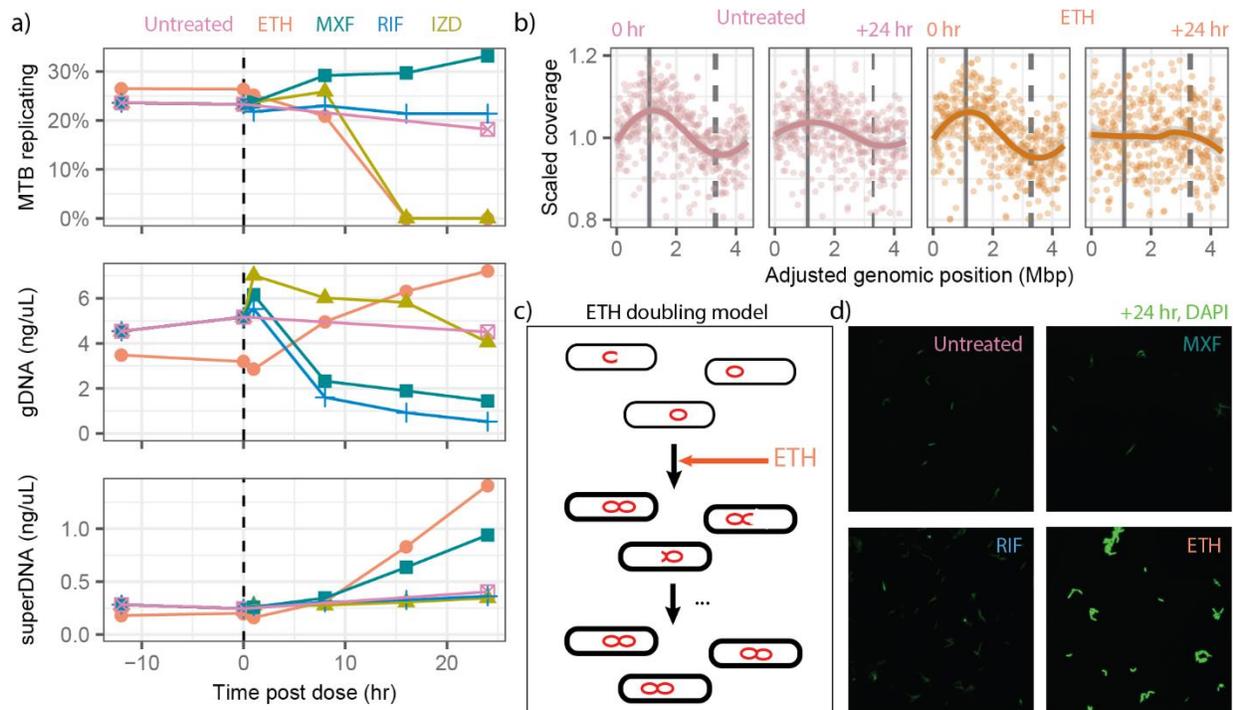
The patient groups mentioned up to this point, pregnant women and transplants recipients, are two groups for which the rapid and accurate diagnosis of infectious disease is paramount. The research on these groups, however, has taken place in the United States, where advanced resources are accessible in most healthcare settings. The overwhelming burden of infectious disease affects those from lower income nations [142]. For example, a 2017 survey by the WHO found 67% increase in the likelihood of dying from a lower respiratory tract infection as a citizen in sub-Saharan African nations compared with those in high income countries [142]. Furthermore, the incidence of outbreaks from emergent infectious diseases is increased in low-income nations and emerging economies [143]. With increased globalization, deforestation, and climate change, these countries have become the epicenter of recent pandemics including Ebola virus in West Africa (2014), Zika virus in Latin America (2016), and bubonic plague (caused by *Yersinia pestis*) in Madagascar (2017). It is therefore important for the development and implementation of broad diagnostic tests to track pathogens.

Collaborating with research and global health companies through the Gates Foundation, we applied metagenomic sequencing with background correction to determine its sensitivity in detecting neglected tropical diseases. We explored the use of plasma cfDNA sequencing for the discrimination of latent and active *Mycobacterium tuberculosis* (MTB). Our experiments in culture and in an animal model revealed the ability to detect changes in the growth dynamics following treatment with a panel of antibiotics. As a follow-up we implemented microbial cfDNA sequencing in patients who were sputum-positive for tuberculosis, though our results indicate shortcomings of the technique. We also surveyed a pediatric cohort of several dozen patients with suspected environmental enteropathy, a syndrome caused by malnutrition which leads to a “leaky gut” or the migration of gut-specific microbiota into the circulatory system [144]. Our work in this area has shown both the limitations of implementing microbial cfDNA sequencing in these settings and the reward of more comprehensive understanding of the disease features.

### 1.5.1 Analysis of MTB genome replication dynamics

Tuberculosis has been a prominent cause of death for centuries, and in the developing world is still responsible for over a million deaths a year worldwide [145]. Despite advances in antibiotics and vaccination, hundreds of millions of people around the world are seropositive for tuberculosis and harbor a latent form of the bacteria, which in 5% to 10% of individuals will reactivate into the deadly, infectious form [146]. The need to monitor these latently-infected individuals, and to distinguish them from those with actively-replicating MTB is great. We explored the use of cfDNA sequencing to indicate those infected by MTB and to distinguish latent and active MTB using genomic replication dynamics [87]. Before analyzing cfDNA samples, we performed sequencing directly on MTB cultures treated with a variety of pharmacological agents to see if MTB growth could be detected by sequencing coverage.

A stock of cultured MTB was placed into one of five flasks and an initial aliquot of MTB was extracted from each of the stocks. We incubated stocks at 37 °C for 12 hours to allow for MTB replication, at which point another aliquot of suspended MTB was removed and replaced with growth medium. Immediately following isolation of the aliquot a clinically relevant dose of the following antibiotics was applied: ethambutol (ETH), rifampicin (RIF), moxifloxacin (MXF), and isoniazid (IZD). Following the introduction of the antibiotics, aliquots from each flask were removed at one, eight, sixteen and 24 hours. We removed an additional aliquot at 24 hours post antibiotic introduction from a control flask, which had no antibiotics introduced. The optical density of each sample was calculated to estimate the number of bacteria. Samples were pelleted and the medium supernatant was isolated for each sample. Pelleted MTB samples and media supernatant samples were digested and washed with ethanol to precipitate DNA. We determined the concentration of genomic DNA (gDNA) and supernatant DNA (superDNA) using a Qubit 3.0 fluorometer.



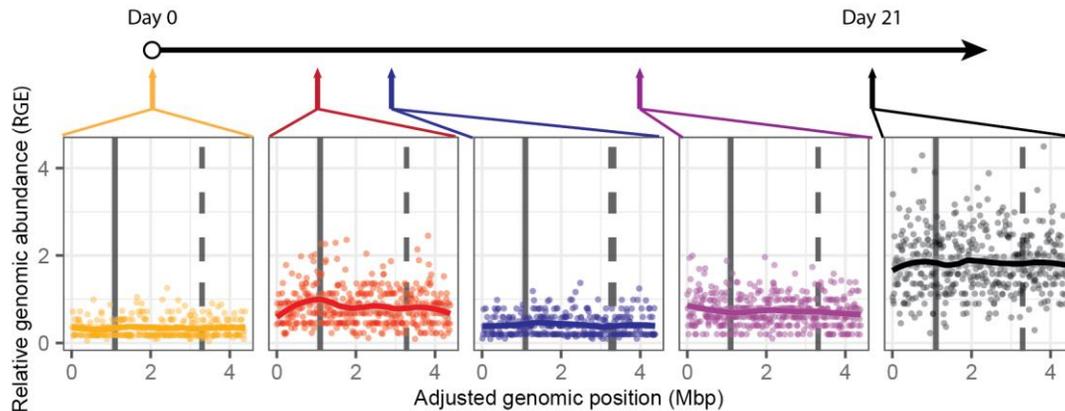
**Figure 1.5.1.1** Whole genome sequencing of MTB reveals effects of antibiotics. (a) MTB cultures were treated with one of several drugs (Untreated, ethambutol (ETH), moxifloxacin (MXF), rifampicin (RIF), isoniazid (IZD)). The percentage of replicating bacteria was determined using the replication score, and the concentration of genomic DNA (gDNA, normalized for volume and optical density) and supernatant DNA (superDNA) were calculated for time 12 hours before and up to 24 hours after dosing. Dotted line indicates time of drug inoculation. (b) The sequencing coverage (normalized to mean coverage) across the MTB genome (binned to 10 kbp) is shown for untreated and ETH-treated MTB prior to dosing and 24 hours after dosing. Solid and dashed vertical lines represent positions of replication origin and terminus, respectively. (c) A model of ETH effects on MTB in culture is shown. When ETH is inoculated into culture, MTB cell walls do not grow. Concurrently, genomes duplicate until there are two copies. (d) Confocal fluorescent imaging of a MTB after 24 hours of dosing. Bacteria were stained with DAPI. {26}

We prepared sequencing libraries of genomic DNA for each MTB sample using the Illumina Nextera protocol. This library preparation technique uses tagmentation enzymes to simultaneously cut and adapt genomic DNA with sequencing primers. MTB libraries were multiplexed and sequenced to a depth of roughly 100-800x across the MTB genome. Following sequencing alignment, the MTB genome was binned to 10 kbp and the mean coverage within each bin was determined. We observed the enrichment in coverage between the replication origin and terminus via the same method as our earlier urinary cfDNA work (Section 1.3.5) and described in other publications [5, 87, 88]. At each time point of sample collection

we determined: (1) the fraction of replication MTB bacteria, (2) the concentration of genomic DNA (normalized by optical density), and (3) the concentration of DNA in the media supernatant (Fig. 1.5.1.1a).

We compared these three measurements over time and between drug treatment conditions to observe how differences in how the drugs affected bacterial growth (Fig. 1.5.1.1a). Bacteriolytic antibacterial drugs, MXF and RIF, did not reduce the fraction of replicating MTB bacteria but did cause a reduction in genomic DNA over time. These results are consistent with bacteria lysis, as incomplete genomes would not finish replication before lysis. In comparison, bacteriostatic drugs affecting cell wall synthesis, ETH and IZD, showed a drastic reduction in the replicating fraction 24 hours after treatment.

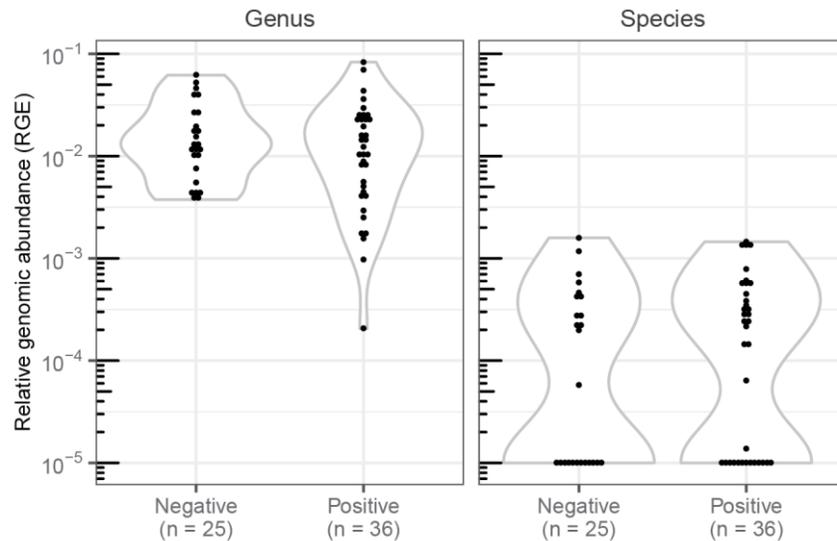
Interestingly, ethambutol-treated MTB also exhibited an increase in genomic DNA that doubled in an asymptotic manner after 24 hours (Fig. 1.5.1.1a). To assure the effect of replication was not a sequencing artifact at the terminus of replication, we directly analyzed the coverage pattern and compared to untreated samples (Fig. 1.5.1.1b); no artifact was observed. Since ethambutol inhibits arabinogalactan synthesis [147], we proposed a model for ethambutol action on MTB summarized in Figure 1.5.1.1c. In our model, ethambutol immediately acts on MTB bacteria, inhibiting new cell wall synthesis and preventing the formation of daughter cells. However, this effect does not restrict genome replication. Rather, all bacteria with a single or both a single and partial genome are permitted to continue genome replication until an entire copied genome is present within the cell (leaving two copies total). Because the cell cannot divide, the ethambutol treatment effectively creates a population of diploid MTB bacteria that are no longer replication active. We confirmed the excess of genomic DNA within single cells using fluorescent imaging of DAPI staining 24 hours after treatment with ETH, RIF, and MXF (Fig. 1.5.1.1d). While RIF- and MXF-treated showed similar amounts of fluorescence compared to an untreated control, we observed ETH-treated bacteria had higher DNA concentrations within single cells (Fig 1.5.1.1d).



**Figure 1.5.1.2** WGS of rabbit caseum shows presence of MTB, but lack of active replication. Samples collected from days 0, 5, 7, 14, and 21 post-inoculation are shown. Relative genomic abundance was calculated by dividing MTB genome coverage by the sequencing coverage of rabbit genome. Mean coverage shown for 10 kbp genome bins. Solid and dashed vertical lines represent positions of replication origin and terminus, respectively. {27}

We applied the genome replication dynamics analysis to an *in vivo* MTB infection model. A rabbit was infected with *Mycobacterium tuberculosis* HN878 using a nose-only aerosol exposure system. The caseum in the rabbit cavity was sampled over a three week period (days 0, 5, 7, 14, 21). DNA was extracted from the caseum samples and sequenced on the Illumina MiSeq platform (2x75 bp) after Illumina Nextera library preparation. Sequencing reads were aligned to both the rabbit and MTB reference genomes, iteratively. We detected an MTB relative abundance of 1.7 RGE, corresponding to 0.17% of all assigned reads. The data presented an increase in bacterial abundance after three weeks but did not detect active replication at any time point (Fig. 1.5.1.2). These results were consistent with matching CFU/CEQ measurements [148] of the caseum samples in the same rabbit..

We extended our analysis to the analysis of cfDNA fragments in the sputum and blood plasma of patients in limited resource areas. In the case of sputum, six samples were collected from patients from a clinic in Malawi and cfDNA was extracted in a BSL3 facility using ethanol precipitation. For plasma cell-free DNA samples, plasma was extracted from samples collected from patients visiting clinics in Peru who were determined positive or negative for MTB by sputum testing. cfDNA was extracted from plasma in a



**Figure 1.5.1.3** Comparison of *Mycobacterium* (genus) and MTB cfDNA in plasma samples from patients with sputum diagnosis reveals no difference in relative genomic abundance. For 61 patients, relative genomic abundance for *Mycobacterium* genus and *Mycobacterium tuberculosis* species was calculated and compared by clinical diagnosis. Samples registering  $10^{-5}$  RGE showed no cfDNA. {28}

BSL2 facility. We performed single-stranded library preparation on cfDNA from sputum and plasma (Illumina NextSeq, 2x75 bp) and analyzed data using the pipelines described above.

Few *M. tuberculosis* cfDNA molecules in either sample source were detected in the sputum or plasma. Furthermore, when comparing MTB cfDNA loads in plasma from patients with positive and negative sputum, no distinction could be made (Fig. 1.5.1.3). Past work has shown the presence of MTB DNA in human bodily fluids of those who were clinically diagnosed; though qPCR was used for diagnosis [149]. We believe that several factors could explain the dearth of MTB cfDNA fragments. In addition to host factors, the doubling time of MTB is very slow compared to other pathogenic bacteria (14-24 hours compared to roughly 20 minutes for *E. coli*) [150]. Furthermore, MTB has a high seroprevalence throughout the world; likely most individuals have been infected with MTB or are latently infected [146]. Low levels of MTB bacteria, and thus MTB cfDNA may persist as a result of these past or latent infections [148], making it difficult to discern positive and negative MTB cases, particularly when using background removal techniques. Finally, within the granulomas, a large amount of necrosis among WBCs takes place.

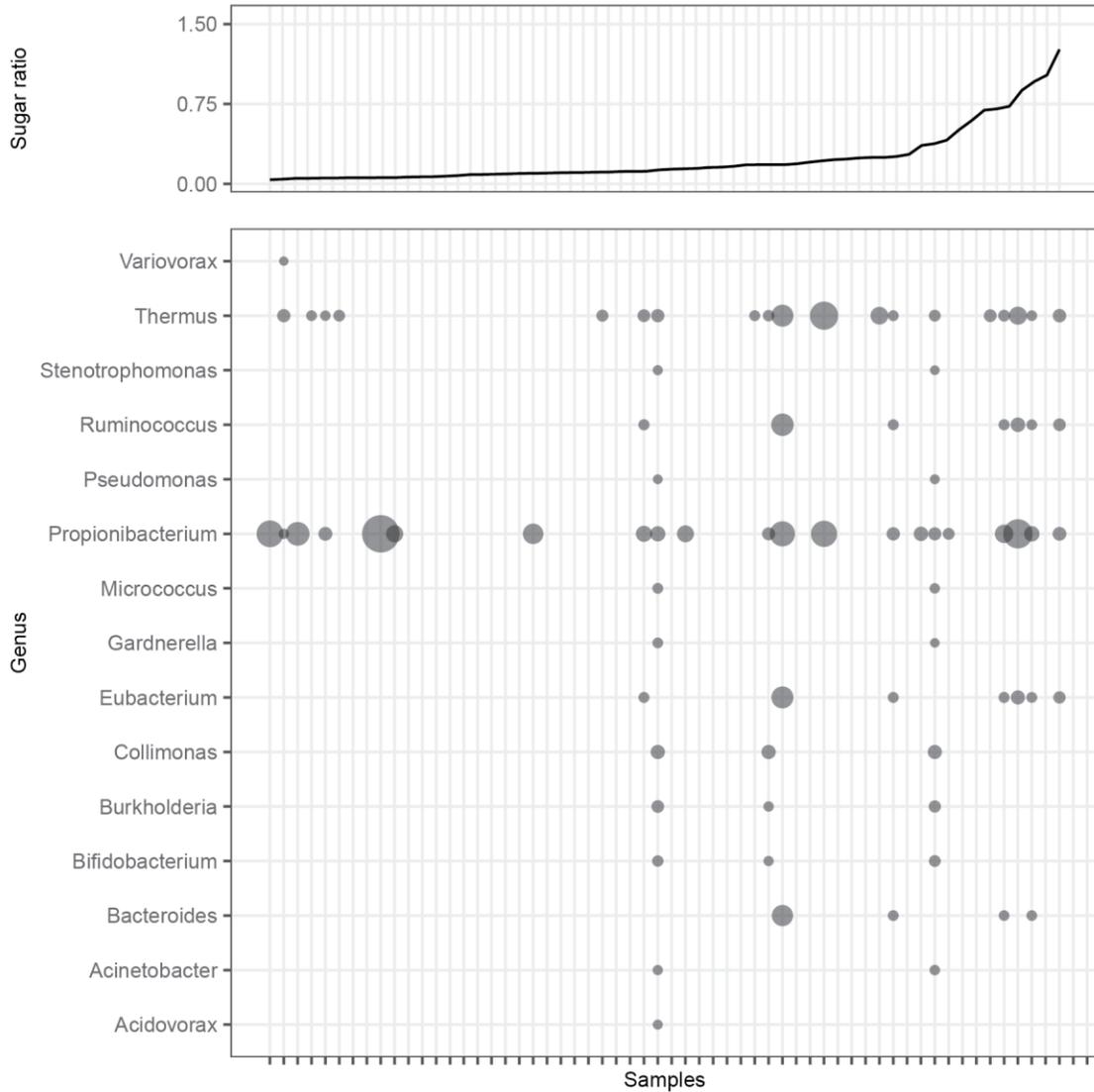
Myeloblast-derived cells in healthy individuals have a relatively high turnover rate (1-5 days for neutrophils to completely renew) [151]. The necrosis of these cells by MTB would add to this already high cellular turnover and, in turn, create a large amount of host cfDNA, which would obscure the MTB cfDNA signal.

The challenges presented in our pilot study in cfDNA sequencing of MTB present issues that need to be addressed in future studies and for those considering unsupervised use of microbial cfDNA sequencing.

#### 1.5.2 Enteric bacterial cfDNA detected in plasma during environmental enteropathy

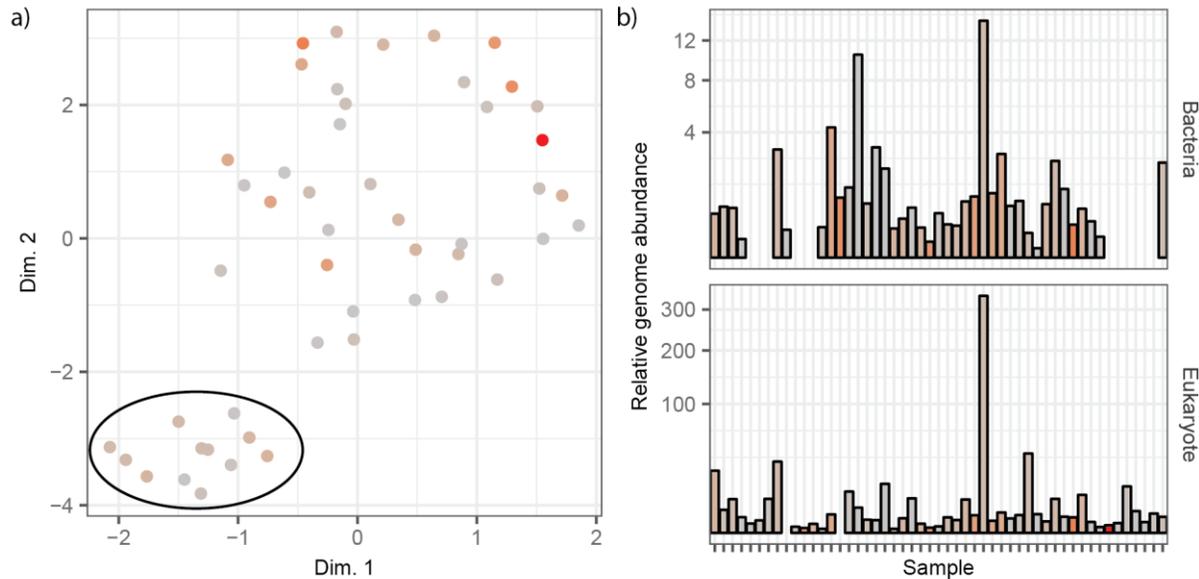
Environmental enteropathy (EE) is a debilitating disorder that still lacks full explanation. A consequence of malnutrition, villi in the gut become blunted, creating chronic challenges for nutrient absorption, as well as the migration of gut bacteria into the circulatory system [144]. This migration can result in systemic inflammation from the presence of pathogen-associated molecular patterns [152]. There are very few diagnostic methods available to access environmental enteropathy, in part due to ignorance that still exists surrounding the disease, but also due to expense. While endoscopies are considered the gold standard in accessing villus blunting, the test is not appropriate for testing in the rural settings where EE is likely to occur [152]. One rudimentary method to access malabsorption is the dual sugar absorption test [153]. In this test, patients are given a mixture of mannitol (a monosaccharide) and lactulose (a disaccharide), at a known ratio. The ratio of these sugars is monitored in the voided urine up to 24 hours after intake [153]. High lactulose-to-mannitol ratios can reflect leakiness of the intestinal tract. However, the interpretation of this test is subjective to the time of sampling and fluctuations in the measurements can describe result from different pathologies [153].

Unlike our previous studies to evaluate the utility of cfDNA sequencing for infectious disease, in environmental enteropathy, there is not one particular pathogen responsible for disease. Rather, the



**Figure 1.5.2.1** cfDNA sequencing of plasma microbiome in pediatric patients with EE reveals presence of gut flora. Genus-level identifications are presented for the bacteria identified across all samples in the EE cohort after sparse metagenomic background correction. Size of each point is relative to log10 measurement of relative genomic abundance. Samples organized by sugar ratio measurement (top). {29}

identification of taxa related to the gut microbiome are of interest. We received sixty plasma samples from pediatric patients (two years of age) in resource-limited settings in Peru. We processed samples through cfDNA extraction and single-stranded library preparation and sequencing in the manner discussed above in Section 1.2 [4]. We aligned sequencing data and compared the samples to the sugar ratio, which was collected within one day of plasma extraction. Unlike our previous observations of cfDNA concentrations in plasma for patients suffering infection, cfDNA concentration plasma in the EE cohort were near baseline



**Figure 1.5.2.2** Plasma virome in patient-samples with EE indicate diversity and breadth of viral infections. (a) UMAP clustering was performed on the family-level relative genomic abundance of viruses identified across samples in EE cohort. Ellipse indicates cluster with a high amount of human-tropic viruses. (b) Relative genomic abundance of viruses aggregated by host tropism (Bacteria, top; Eukaryote bottom). Color gradient (grey to red) corresponds to sugar ratio (low to high). {30}

(median concentration 30.4 ng/mL plasma). We applied the LBBC algorithm (Section 1.4) to the microbial abundance estimations of the cohort and identified the presence of enteric bacteria (*Eubacterium*, *Bacteroides*, and *Ruminococcus*) as well as a high presence of skin contaminants (Fig. 1.5.2.1). We did not observe the aforementioned bacteria in patients with sugar ratio less than 0.2.

In our analysis, an abundance of viruses were present across nearly all samples in the cohort. We investigated the abundance of bacteriophages as an indicator of enteric dysfunction. We aggregated viral reads according to family, and compared the viral family abundance between samples and to the sugar ratios. The two most prominent viral families across our cohort were *Siphoviridae* (n = 55), a bacteriophage, and *Anelloviridae* (n = 59), a family of small human viruses shown to increase in abundance when patients are under immunosuppressants [1]. We also observed the presence of parvoviruses (n = 18), herpesviruses (n = 6), and polyomaviruses (n = 18). There was no significant correlation between a specific viral family and the sugar ratio. In comparing viral family abundance between samples, we were able to cluster data based on the plasma virome. We did, however, observe a strong effect of clustering according to viral

tropism when using UMAP dimension reduction algorithm [154]. Additionally, while samples originating from patients with high sugar ratios were not exclusive to one cluster, samples with the four highest sugar ratio values clustered together in a group marked by low overall viral load (Fig. 1.5.2.2). We aggregated viral family abundance according to host tropism (Eukaryote or Bacteria) for each sample, and compared these aggregate measurements on a sample-by-sample basis with the sugar ratio. Again, we did not observe any significant correlation or anti-correlation between the clinical measurement and cfDNA sequencing (Fig. 1.5.2.2).

Further analysis needs to be performed to elucidate the relationship between the various viruses and bacteria present in this sampling cohort in order to identify relationships with disease. It has been suggested that benchmarking cfDNA sequencing against the sugar ratio is not proper for the proof-of-principle pilot study. Secondary clinical measurements, including qPCR panels of known enteric microbiota and confirmation by endoscopy may increase confidence in measurements.

## Chapter 6: Perspective in cell-free DNA diagnostics

cfDNA sequencing, combined with bioinformatics approaches to remove background contamination, detects a wide array of microbial life in a variety of biological fluids. Since 2015, we have created over five hundred cfDNA sequencing datasets from ssDNA library preparation to comprehensively study infectious disease (Table 1.6.0.1). Our work has shown that our novel library preparation technique (Section 1.2) is capable of creating cfDNA sequencing libraries from low sample volumes ( $< 250 \mu\text{L}$ ) and low sample biomasses ( $< 5 \text{ ng cfDNA}$ ), while accessing previously uncaptured sizes of cfDNA. The robustness of the technique has allowed us to uncover microbial cfDNA related to bacterial and viral infections for immunocompromised patients in the United States and children suffering from malnutrition in Peru. As we have shown, though, the presence of microbial cfDNA in a milliliter of biological fluid can span orders of magnitude and is highly dependent on the lifecycle and physiology of the microbe as well as the proximity of the infected tissue to the fluid collected for testing.

More studies need to be performed in order to determine the efficacy of cfDNA sequencing to monitor other types of infectious disease, including tuberculosis (Section 1.5.1) and parasitic infections, such as those from soil transmitted helminths. The case of cfDNA to monitor eukaryotic pathogens is of particular interest and not well characterized. In our work in the De Vlaminck lab, we have prepared sequencing libraries from the plasma and urine of pigs inoculated with a roundworm, *Ascaris suum*, in both the larval and adult stages (Table 1.6.0.1). Our cfDNA sequencing pipeline did not uncover any reads aligning to this organism. We have also had difficulty in identifying clinical cases of funguria, the presence of fungus in urine. These cases of type II error could be explained by a high host cfDNA background coupled with low amounts of pathogen cfDNA in the biofluid. The lack of well-annotated and assembled reference genomes for complex eukaryotic pathogens may also drive the challenge in identifying eukaryotic parasites via cfDNA sequencing [155].

Disease	Interest	Host	Biological fluid	Country of origin	Infectious organism	cfDNA detection?	No. samples
<b>UTI / BKVN</b>	Kidney Tx	Human	Urine	USA	Bacterial / Viral	Yes	149
<b>Post-Tx complications</b>	Lung Tx	Human	Plasma	USA	Bacterial / Viral	Yes	40
<b>Peritonitis</b>	Kidney Tx	Human	PD effluent	USA	Bacterial	Yes	56
<b>Chorio-amnionitis</b>	Pregnancy	Human	Amniotic fluid	USA	Bacterial	Yes	44
<b>Tuberculosis</b>	Global health	Human	Plasma	Ecuador	Bacterial	No	100
<b>EE</b>	Global health	Human	Plasma	Peru	Bacterial	Yes	60
<b>EE</b>	Global health	Human	Urine	Peru	Bacterial	Yes	12
<b>Viral infection</b>	Global health	Human	Plasma	Switz.	Viral	Yes	15
<b>Febrile illness</b>	Global health	Human	Plasma	Tanzania	Viral	No	15
<b><i>Ascaris suum</i> infection</b>	Global health	Pig	Plasma	Belgium	Eukaryotic	No	42
<b><i>Ascaris suum</i> infection</b>	Global health	Pig	Urine	Belgium	Eukaryotic	No	6

**Table 1.6.0.1** Summary of datasets acquired using cfDNA sequencing with ssDNA library preparation. cfDNA sequencing datasets acquired in the De Vlaminck lab, according to: disease, interest, host organism, fluid from which biological fluid was extracted, country of sample origin, infectious organism superkingdom, whether cfDNA from the organism was general detectable, and the number of samples processed. Tx = Transplantation. Every sample listed in the last column was prepared and sequenced, but not all samples were used in publications or analysis.

Without curated reference genomes for these organisms, metagenomic sequencing alignment algorithms are unable to properly assign reads; our efforts to normalize reads to sequencing depth across the respective genome is also difficult. Some approaches to solve this problem include the assembly of large contigs by comparing cfDNA fragments for overlap patterns, and then comparing these large synthetic structures to known gene sequences unique to the pathogen[157]. Application of contig assembly to cfDNA sequencing datasets has recently revealed the presence of large amounts of microbial “dark matter” and dozens of novel anelloviruses in blood [158]. These methods, along with the exponentially increasing number of eukaryotic reference genomes in publically available databases, will likely remedy these technical shortcomings in the near future.

A second consideration regarding cfDNA sequencing to monitor infectious disease involves the cost of testing and the time to diagnosis. New work in the private sector is showing the range and sensitivity of the technique to determine the presence of pathogen cfDNA, with the ability to issue a diagnosis in 40% of cases within 53 hours of sample collection [6]. By comparison, less than 10% of cases are diagnosed by clinical microbiology techniques in the same amount of time [6]. As I will discuss in the Conclusions section, the cost of sequencing for such tests will also likely decrease significantly in the coming years with the advent of new sequencing technologies.

Another consideration when designing studies using cfDNA sequencing should be the information required to make a positive diagnose or to intervene with therapy. Many clinicians rightfully argue that there is a difference between infection and infectious disease [156]. These arguments support the use of measurements to detect host tissue damage, as we have presented using donor-fraction and tissue-of-origin measurements (Section 1.3), and microbial cfDNA. These techniques often require deeper sequencing ( $> 2x$  coverage of human genome for methylation markers and  $> 8x$  coverage for nucleosome depletion) than what is required to accurately determine pathogenic species alone ( $< 0.5x$ ). New techniques are being developed, both in the De Vlaminck lab and in the general community, that make use of high-depth measurements at a limited number of predetermined sites to determine a tissue-of-origin profile for cfDNA. These techniques significantly reduce the cost and time to detect tissue damage, and make use of non-sequencing based techniques including dPCR and Crispr-Cas-based identification methods [157]. Using tissue-of-origin measurements in clinical practice could be done in tandem with low-depth microbial cfDNA sequencing to determine the pathogen and site of infection without incurring heavy costs.

We have shown in this dissertation that cfDNA sequencing can also be used as a tool for molecular epidemiology. Our analysis of renal transplant recipients suffering BK polyomavirus nephropathy highlights the ability to determine phylogenetic information from viral cfDNA. Since microbial cfDNA sequencing is capable of single nucleotide resolution, consensus sequences for viruses within patient

samples can be determined. The sequences can then be compared to one another to determine genetic relatedness and to understand the adaptation of the pathogen within its host.

Microbial cfDNA sequencing can go further than describing genetic properties, by providing protein binding information, similar to CHIP-seq. Ultrashort cfDNA molecules were previously determined to reflect sites of eukaryotic transcription factor binding [23]. It is likely that microbial cfDNA also represents molecules stabilized by proteins bound to the genomic DNA at time of cellular lysis. We have observed these protein binding signals in BK polyomavirus cfDNA, where fragment length distributions support previous observations that polyomavirus DNA is bound to nucleosomes in minichromosome complexes (Fig. 1.3.1.1). Such a signal suggests that, like eukaryotes, microbial cfDNA is not the product of random fragmentation. Because there is no sequence bias (based on degradation) in ssDNA library preparation, cfDNA sequencing could be used to validate protein binding sites on bacterial and viral DNA and associate patterns of microbial genome coverage to disease phenotype. More directly, the coverage patterns will be helpful to determine the optimum sequences for hybridization assays and PCR, as fragments associated with these coverage maxima would be most prevalent in the biological fluid under observation.

In summary, our work indicates that performing next generation sequencing on cfDNA isolated from biological fluids can be used to identify viral and bacterial infections with high sensitivity. Furthermore, as the cost of sequencing decreases, cfDNA sequencing will prove to be a valuable diagnostic analyte in rural and resource-limited settings, particularly to monitor the spread of infections in communities and reconstruct novel genomes.

PART II: Single-cell sequencing in infected cell populations – understanding the innate and adaptive  
immune response

“... Except in the case of viruses. They can turn off and go dead. Then, if they come in contact with a living system, they switch on and multiply.”

— Richard Preston, *The Hot Zone: The Terrifying True Story of the Origins of the Ebola Virus*

## Chapter 1: Introduction to virus-inclusive single-cell RNA sequencing

Paintings among Parisian artists in the latter half of the 19th century emphasized a focus on ordinary subjects and, while masterful, were painted in the method common of the previous centuries - the application of oil-based paint blended across their subjects. Works at the end of the century, however, did not copy this style. A new group of young artists realized that the beauty within each scene was often masked by the application of colors in splotches and the habit of painting with dark colors. Rather, these artists understood the importance of emphasizing the distinction of each brushstroke, making them individually noticeable, yet also part of the ensemble. When one views a work from the *Impressionist* artists, it is immediately clear that each moment the artist applied paint is distinguishable from the last and that appreciation of heterogeneity reflects an illuminated scene. Since the beginning of the 21st century, genome scientists have been part of a movement that is similar to what was achieved by artists during the *Impressionist* period. Scientists have started to appreciate the importance of understanding an organism's larger life processes by focusing on changes occurring at the single-cell level [158].

The single cell is the basic unit of biology. Yet, until recently, sequencing studies have focused on sequencing the genomes or transcriptomes of thousands of cells in a single assay. This technology, bulk RNA and DNA sequencing, has successfully identified the heterogeneity between individuals from different ethnic groups and the changes in transcriptomes for people with various genetic diseases [159]. However, bulk sequencing often neglects the biological contribution of rare cell types, since bulk sequencing effectively averages sequencing information from a group of cells.

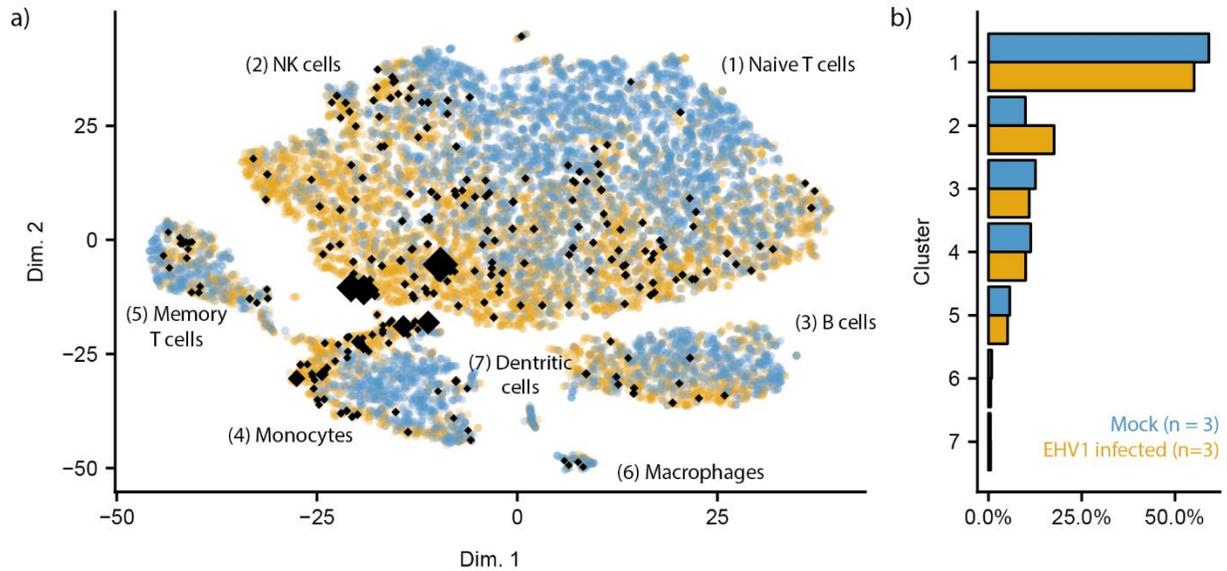
The field of single-cell studies has grown at a remarkable pace since the first individual single-cell transcriptome was isolated, amplified, and sequenced in 2009 [160]. Studies immediately following the pivotal work in single-cell RNA-seq (scRNA-seq) by Tang et al. [160] used microfluidics or fluorescence activated cell sorting (FACS) to isolate tens to hundreds of cells [161]. Often researchers interested in performing scRNA-seq on tissues or organisms are interested in observing the subpopulations of cells

making up the tissue; this requires many more cells in order to observe members present at low abundance in populations. The cellular throughput needed to be increased in order to make single-cell sequencing an established tool for groups studying cellular heterogeneity.

In 2015, the development of droplet microfluidic platforms supporting scRNA-seq allowed researchers to prepare tens of thousands of single cells for sequencing in the matter of hours [7]. In this method, Drop-seq, the authors created functionalized beads to capture polyadenylated messenger RNAs (mRNAs) once cells were lysed within droplets. The oligos on the bead surfaces included a bead-specific cell barcode (12 bp) and an oligo specific unique molecular identifier (UMI, 8 bp), which allowed for the deconvolution of sequencing reads following analysis. A digital expression matrix can then be generated for the entire sample from the demultiplexed reads, according to the cell barcode, which depicts the number of host transcripts across all genes and all cells.

Drop-seq, and other techniques, allowed for deep sequencing across the transcriptomes of cells at a low cost (< \$0.10 per cell) [7, 162]. The access to this level of resolution across tissues has permitted the discovery of novel cell types, such as cells within mammalian lungs that are related to those in the respiratory system of fish [163]. These cells exist at low abundance, and would not have been detected by bulk sequencing assays or low throughput single-cell sequencing techniques. Moreover, new bioinformatics approaches have been developed to describe the unique properties of individual cells and their relationship to others in the sample. For example, recent works have detailed lineage tracing, understanding the developmental process that cells undergo, and prediction of future cell states in single-cell transcriptomic datasets [164, 165]. Since 2015, optimization of droplet microfluidics and the establishment of new technologies has allowed for a continuing trend of higher throughput in single-cell studies, now reaching into millions of cells [162, 166].

These assays, while fast and cost-effective, have targeted eukaryotic mRNAs, and have largely ignored the abundance of other classes of RNA molecules. Eukaryotic mRNAs are polyadenylated, so that the 3'-end of the molecule is occupied with tens of adenines [167]. The polyadenylation is directly targeted



**Figure 2.1.0.1** Clustering of infected cells reflects dynamics of cell populations during EHV1. (a) t-SNE representation of 13,156 PBMCs from mock (blue) and EHV1-infected (yellow) horses. Cell subtypes, listed beside clusters, were determined by marker genes after k-nearest neighbor clustering. Black diamonds represent cells with detectable amounts of EHV1 transcripts; size of diamond corresponds to relative viral transcript abundance. (b) Percentage of cells from mock (blue) and EHV1-infected (yellow) horses making up each cluster (numbers corresponding to groups in (a)). {31}

for hybridization with poly(dT) capture probes, making transcript capture efficient for these molecules, but restricting the capture of molecules that do not have a similar 3'-end. Eukaryotic noncoding RNAs and transfer RNAs, for example, do not have this feature and are not captured. Additionally, many prokaryotes and viruses do not have polyadenylated tails on their messenger RNA [168, 169], making generalizable scRNA-seq approaches to study host-pathogen interactions difficult.

The study of viruses using scRNA-seq has so far been limited, in part because Drop-seq, 10x Genomics, and other library preparation platforms are unable to generally capture and sequence viral mRNAs. In this part of the dissertation, I will describe a modification that can be made to existing scRNA-seq technologies allowing for simultaneous capture of polyadenylated and targeted, non-polyadenylated mRNAs, particularly as it relates to virology. We identified four aims of virus-inclusive scRNA-seq:

1. Determine host cell transcriptional profiles and identify of subpopulations through data structure.
2. Identify which cells are infected with viral particles.
3. Understand how viral gene expression is correlated to subsets of host gene expression.

4. Analyze viral intracellular heterogeneity and determine how it relates to host gene expression.

Differential gene expression is used broadly to understand complex cell populations, as is described in Aim 1, and does not require innovation beyond current technologies. To properly address the other three goals, though, researchers needed to focus on viruses with polyadenylated mRNAs. A typical result of an analysis of a complex collection of cells infected with a virus transcribing polyadenylated mRNA is shown in Figure 2.1.0.1a. In this pilot study, we isolated peripheral blood mononuclear cells (PBMCs) from three horses infected with Equid herpesvirus 1 (EHV1) and three control horses. We then processed isolated cells from the six samples using the Drop-seq [7]. The expression of *E. caballus* and EHV1 genes was quantified and cells of low quality (mitochondrial gene expression over 5%, number of genes detected under 200) were removed. Through dimensional reduction and k-nearest neighbor clustering (Seurat pipeline [170]), we identified several distinct clusters across 13,156 cells from the six samples. We then related these clusters to cell type in the blood through known marker genes (e.g. *MS4A1* upregulation in B cell lymphocytes). Additionally, we observed that nearly twice as many cells in EHV1-infected samples were classified as NK cells as compared to mock samples (17.6% versus 9.9%, Fig 2.1.0.1b). NK cells are known to increase in abundance during viral infection and have an antagonistic effect on herpesvirus infected cells [171, 172].

As expected, we did not capture transcripts from EHV1 in the three uninfected samples. However, very few cells in infected horses had detectable viral transcripts (0.6% to 7.6% of cells in samples), and of these infected cells, only a small proportion of the transcripts originated from EHV1 (0.02% to 0.04% of transcripts). Across all filtered cells in all samples, seven of the sixty-eight EHV1 transcripts made up more than half of all captured viral transcripts. The lack of EHV1 gene coverage indicates the shortcomings of using standard scRNA-seq of the viral life cycle within infected cells.

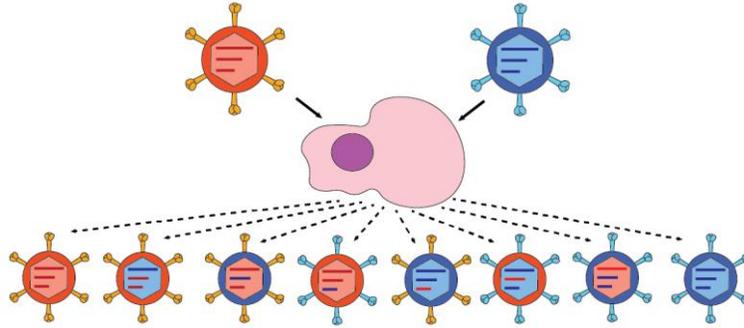
It cannot be assumed, due to the low viral transcript abundance, that cells without detected viral transcripts are not infected; however, we assume that cells expressing viral transcripts are infected with EHV1. These infected cells were not restricted to one PBMC subtype. EHV1-infected cells were observed

in all clusters with the exception of a small group of dendritic cells. The most highly expressed viral gene across the six samples, EHV1 ORF34, was only expressed in cells within the monocyte and T cell cluster. These clusters also showed the highest fraction of viral transcripts with infected cells (represented by size of black diamonds in Fig. 2.1.0.1a). As EHV1 ORF34 is associated with viral egress [172], monocytes and T cells may be the most fruitful environments for EHV1 replication.

This brief study illustrates both the promise and challenges of scRNA-seq. Our pilot dataset identifies changing cell populations in response to infection. Furthermore, we resolved viral transcripts at single-cell resolution and associated partial gene expression in distinct clusters. Of the four aims listed above, however, only the first can be achieved using the Drop-seq platform [7]. Though in this case, EHV1 mRNA is polyadenylated, we have shown that it is sparse within these cells. Thus, sampling of the transcriptome does not produce adequate read coverage at single-cell resolution to assess the number of cells in each subtype that are infected or the viral intracellular heterogeneity [173].

cDNA amplification of viral sequences can also be used to enrich the libraries, and has been shown in recent studies for viruses lacking polyadenylated mRNAs [174, 175]. However, this method still relies on initial capture by the poly(dT) probe. Furthermore, in complex tissues and communities, viral tropism for cells making up only a small portion of the sample may lead to relatively few viral reads in sequencing datasets, regardless of capture efficiency. To adjust for this scenario, some groups have used fluorescence activated cell sorting (FACS) so that low abundance cell types are sequenced at similar depth to high abundance subpopulations [175]. Such a system requires the user to know cell surface markers *a priori*, but has been shown to work well for peripheral blood mononuclear cells (PBMCs) infected with dengue virus [175].

Beyond accessing non-polyadenylated transcripts, we envisioned that a targeting approach could enrich for low abundance reads, as in the case of the EHV1 infected PBMCs or viral models exploring questions of persistence or latency (both states of relative little viral replication). To meet our goals, we



**Figure 2.1.0.2** Representation of viral genetic reassortment occurring when two genetically distinct viruses infect and replicate within a cell. Progeny viruses have a likelihood of encapsidating gene segments from both parent viruses, allowing for the formation of novel viral genotypes. {32}

developed an adaptable ligation chemistry that could be applied to existing high throughput scRNA-seq technologies. We chose to optimize our method on segmented RNA viruses.

Segmented viruses, those with genomes partitioned into multiple, separable molecules, are of interest in viral evolution, particularly on short time scales. In addition to highly mutable genomes, segmented viruses may evolve through genetic reassortment. In these cases, two or more separate virions infect the same cell, and progeny virions may assemble a heterogeneous mixture of genomic segments corresponding to the original infecting virions [176] (Fig. 2.1.0.2). Reassortment can therefore allow viruses to gain new phenotypic traits over one infection cycle and permit the escape from the adaptive immune system [177]. Seasonal, potentially epidemic viruses such as influenza A virus benefit from this type of evolution [178].

Due to the potential consequences of evolution by reassortment, it is important to understand how cells react to progeny virions. Single-cell sequencing provides a tool to observe the cellular response to observe these viral properties within individual cells. While some species of segmented viruses produce messenger RNAs with polyadenylated tails (such as influenza), others do not, including reovirus and rotavirus. Reovirus and rotavirus, both members of the *Reoviridae* family, in addition to being segmented, have double-stranded RNA (dsRNA) genomes [179]. Upon infecting the cells, proteases partially uncoat the viral capsid and move the core particles into the cytoplasm. The genome segments are denatured to allow transcription of the positive-sense mRNA. Virions in the *Reoviridae* family have an innate RNA-

dependent RNA polymerases that are used to transcribe their mRNA. The viruses create negative-sense RNA from positive-sense mRNA to allow for the creation of dsRNA genomes for progeny viruses. Host-derived protein complexes are used to translate mRNA into viral proteins [179].

We performed a series of experiments to characterize a novel technology capable of simultaneously sequencing polyadenylated mRNA and targeted amplicons in a high-throughput, single-cell manner. We termed the technology DART-seq, droplet-assisted RNA targeting by single-cell sequencing. DART-seq uses a ligation chemistry to append targeted oligonucleotide probes onto existing poly(dT) capture technologies, such as Drop-seq beads. We applied DART-seq to enrich both non-polyadenylated transcripts in the case of reovirus-infected cells and polyadenylated transcripts in B cell lymphocytes.

In applying DART-seq to murine fibroblasts infected with T3D orthoreovirus, we showed that DART-seq was able to capture non-polyadenylated viral transcripts with high targeting precision. Using this technique we measured the abundance of infected fibroblasts and compare the relative abundance of viral transcripts with respect to cell state. Using a second capture array, we showed that an entire viral transcript sequence can be assembled using DART-seq. We utilized these sequencing features to determine the mutational profile of viral mRNAs after infection. In a separate study, we applied DART-seq and non-targeted scRNA-seq to characterize the host-pathogen interactions in rotavirus-infected primate fibroblasts. From these experiments, we were able to recapitulate rotavirus infection biology. We are actively developing infection models using *Reoviridae* in organoid systems to understand the innate immune system at higher complexity.

We also used DART-seq to enable the capture of heavy chain and light chain isotype transcripts in populations of B cell lymphocytes, proving that DART-seq can improve the capture of polyadenylated mRNAs present at low abundance within single cells. Our results indicated that even at a modest depth-of-sequencing, DART-seq greatly enriches the abundance of heavy and light chains in B cells. We used the transcript enrichment to validate known phenomenon in these isotypes, and show the ability to pair heavy and light chain variable regions to characterize the immune repertoire in a population of PBMCs.

Taken together, our exploration of DART-seq applied to infected cell populations and PBMCs indicate targeted scRNA-seq is a potent tool understand the innate and adaptive immune system.

## Chapter 2: Simultaneous multiplexed amplicon sequencing and transcriptome profiling in single cells

“We describe droplet-assisted RNA targeting by single-cell sequencing (DART-seq), a versatile technology that enables multiplexed amplicon sequencing and transcriptome profiling in single cells. We applied DART-seq to simultaneously characterize the non-A-tailed transcripts of a segmented dsRNA virus and the transcriptome of the infected cell. In addition, we used DART-seq to simultaneously determine the natively paired, variable region heavy and light chain amplicons and the transcriptome of B lymphocytes.”

Chapter adapted from [8]:

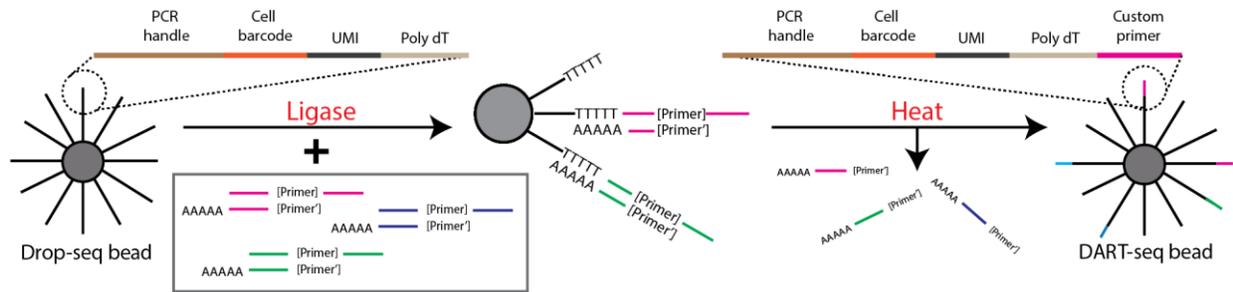
Saikia M\*, Burnham P\*, Keshavjee SH, Wang MFZ, Heyang M, Moral-Lopez P, Hinchman MM, Danko CG, Parker JSL, De Vlaminc I. Simultaneous multiplexed amplicon sequencing and transcriptome profiling in single cells. *Nature Methods*. 2019.

## 2.2.0 Targeted amplicon sequencing in single cells to fully describe cell heterogeneity

High-throughput single-cell RNA-seq (scRNA-seq) is being widely adopted for phenotyping of cells in heterogeneous populations [7, 162, 180, 181]. The most common implementations of this technology utilize droplet microfluidics to co-encapsulate single cells with beads that are modified with barcoded oligos to enable capturing the ends of RNA transcripts [7, 162, 181]. Although these approaches provide a means to perform inexpensive single-cell gene expression measurements at scale, they are limited to assaying the ends of mRNA transcripts. Therefore, they are ill-suited for the characterization of non-A-tailed RNA, including the transcripts of many viruses, viral RNA genomes, and non-coding RNAs. They are also uninformative of RNA segments that are located at a distance greater than a few hundred bases from transcript ends that often comprise essential functional information, for example the complementarity determining regions (CDRs) of immunoglobulins (B cell antibody) [182]. Additionally, these techniques are often unable to provide information on low copy number transcripts and splice variants [173].

Here we report DART-seq, a method that combines enriched measurement of targeted RNA sequences with unbiased profiling of the poly(A)-tailed transcriptome across thousands of single cells in the same biological sample. DART-seq achieves this by implementing a simple and inexpensive alteration of the Drop-seq strategy [7]. Barcoded primer beads that capture the poly(A)-tailed mRNA molecules in Drop-seq are enzymatically modified using a tunable ligation chemistry [31]. The resulting DART-seq primer beads are capable of priming reverse transcription of poly(A)-tailed transcripts as well as other RNA species of interest.

DART-seq is easy to implement and enables a range of new biological measurements. Here, we explored two applications. We first applied DART-seq to profile viral-host interactions and viral genome dynamics in single cells. We implemented two distinct DART-seq designs to investigate murine L929 cells (L cells) infected by the reovirus strain Type 3 Dearing (T3D). We demonstrate the ability of DART-seq to profile all 10 non-A-tailed viral gene transcripts of T3D reovirus individually, as well as to recover a complete genome segment, while simultaneously providing access to the transcriptome of the infected L

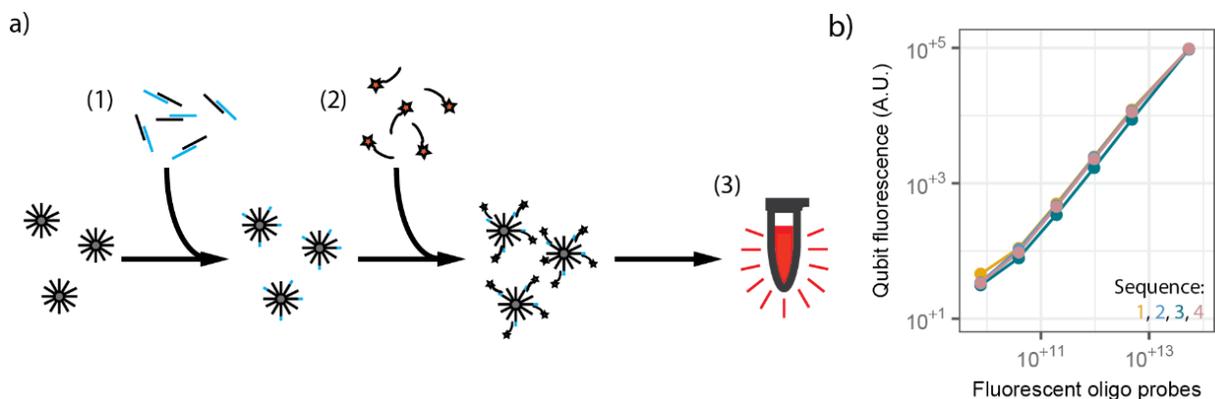


**Figure 2.2.1.1** DART-seq is an easily implemented and multiplex technology for single cell studies. Drop-seq beads (with oligos containing a poly(dT) tail) are combined with a diverse mixture of toehold molecules including the targeting primer and a splint oligo with a polyA overhang. A DNA ligase is added to the suspension, which binds the toehold oligos to the Drop-seq beads. A light heat treatment (65 °C) is used to denature splint oligos which are subsequently washed away. The procedure retains all oligo information present on the original bead. {33}

cells. In the second application, we applied DART-seq to determine natively paired antibody sequences of human B cells. DART-seq was able to determine B cell clonotypes, as well as variable heavy and light (VH:VL) pairings, even in mixed human peripheral blood mononuclear cells (PBMCs), highlighting the versatility of the approach.

### 2.2.1 DART-seq primer bead synthesis

Droplet microfluidics based scRNA-seq approaches rely on co-encapsulation of single cells with barcoded primer beads that capture and prime reverse transcription of mRNA molecules expressed by the cell [7, 162, 181]. In Drop-seq, the primers on all beads comprise a common sequence used for PCR amplification, a bead-specific cell barcode, a unique molecular identifier (UMI), and a poly-dT sequence for capturing polyadenylated mRNAs and priming reverse transcription. To enable simultaneous measurement of the transcriptome and multiplexed RNA amplicons in DART-seq, we devised a scheme to enzymatically attach custom primers to a subset of poly-dTs on the Drop-seq bead (Fig. 2.2.1.1). Conversion is achieved by annealing a double-stranded toehold probe with a 3' ssDNA overhang that is complementary to the poly-dT sequence of the Drop-seq primers. The toehold is then ligated to the bead

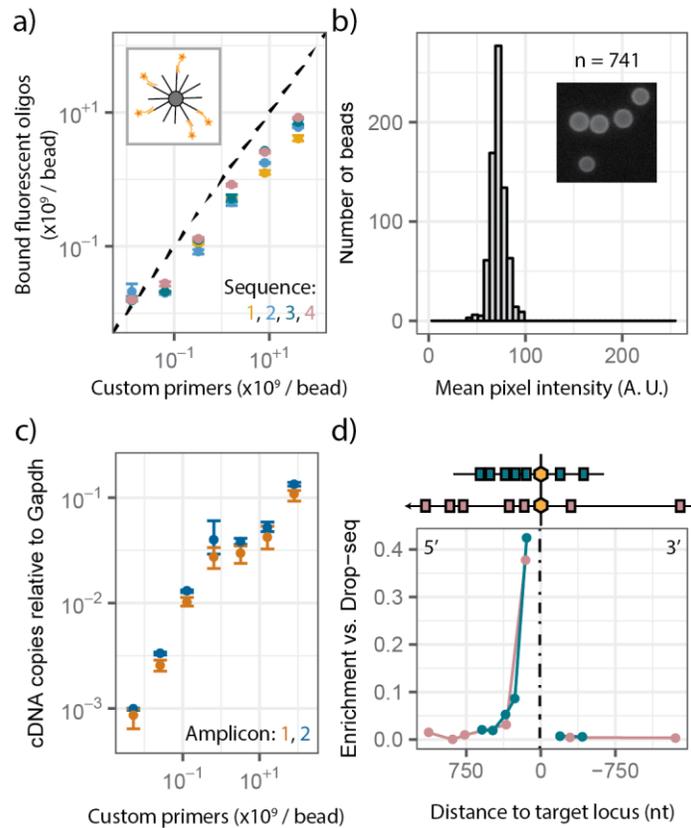


**Figure 2.2.1.2** Qubit fluorometer measurements can resolve the number of fluorescent oligos bound to DART-seq beads. (a) To evaluate DART-seq ligation efficiency we designed an assay able to indirectly measure probe binding. (1) Toehold molecules are added to Drop-seq beads via the DART-seq conversion protocol. (2) Fluorescent oligos with sequence complementary to DART-seq probes are added to DART-seq beads. (3) The fluorescence of suspensions of 2000 beads are measured via Qubit 3.0 fluorometer in the 647 nm channel. (b) For four different toehold probes we compared the fluorescence (in A.U., arbitrary units) as a response to the number of fluorescent probes added to the Qubit measurement. {34}

using T4 DNA ligase. Toeholds with a variety of different sequences can be attached to the same primer beads in a single reaction in this manner. The complementary toehold strand is removed after ligation.

After synthesis of DART-seq primer beads, DART-seq follows the Drop-seq workflow without modification (see Section 2.2.4). Briefly, cells and barcoded primer beads are co-encapsulated in droplets using a microfluidic device. Cellular RNA is captured by the primer beads, and is reverse transcribed after breaking the droplets. The DART-seq beads prime reverse transcription of both A-tailed mRNA transcripts and RNA segments complementary to the custom primers ligated to the beads. The resulting complementary DNA (cDNA) is PCR-amplified, randomly fragmented via tagmentation, and again PCR amplified to create libraries for sequencing. Sequences of mRNAs and RNA amplicons derived from the same cells are identified by decoding cell-specific barcodes, allowing for gene expression and amplicon measurements across individual cells.

We characterized the efficiency, tunability and variability of the ligation reaction using fluorescence hybridization assays based on emission intensity measured in a Qubit fluorometer (Fig. 2.2.1.2a). We found the fluorescence measurements were capable of measuring the number of oligos



**Figure 2.2.1.3** Conversion of Drop-seq beads to DART-seq beads is efficient and uniform, and DART-seq beads enrich targeted RNAs. (a) DART-seq beads were created from a mixture of four probes and complementary fluorescent oligos were bound to beads and measured by Qubit fluorometer. Dotted line represents 100% conversion efficiency. Inset: diagram of fluorescent oligos bound to beads. (b) DART-seq beads with bound fluorescent oligos were imaged using a fluorescence microscope and the average pixel intensity across 741 beads was determined. (c) Enrichment of targeted RNAs with respect to *Gapdh*, as measured by qPCR on cDNA from bulk RNA samples, is shown for various concentrations of probes added ( $10^6$  to  $10^{12}$  probes per bead). (d) DART-seq was used to capture two viral mRNAs at seven loci and qPCR was performed at positions above the plot and compared to *Gapdh*. {35}

over many orders of magnitude (Fig. 2.2.1.2b). By comparing the number of input fluorescent oligos with those determined via fluorescent measurements, we found that the primer ligation reaction is highly efficient (25–40%), and the number of custom primers ligated to the beads is directly proportional to the number of primers included in the ligation reaction (Fig. 2.2.1.3a). This was true for four primer sequences tested over a wide range of primer concentrations and in mixed proportions. The efficiency of probe ligation decreased for ligation reactions with more than  $10^{10}$  molecules per bead, indicating saturation of available oligonucleotide(dT)s. We compared the fluorescence hybridization signal across individual beads under a fluorescent microscope and found that the bead-to-bead variability in

fluorescence signal was small (s.d. 3.0%; Figure 2.2.1.3b). The homogeneity of coverage was replicated for various input concentrations of toehold probes. We observed that failures to keep beads well mixed during the ligation step revealed differential toehold coverage both between beads and across the surface of individual beads.

We assessed reverse transcription priming efficiency as a function of the number of custom primers ligated to DART-seq beads. We used quantitative PCR (qPCR) to measure the yield of cDNA copies of a non-polyadenylated viral mRNA in reovirus-infected murine fibroblasts (L cells, Fig. 2.2.1.3c). Two distinct primers were ligated, targeting the same viral genome segment (S2). The yield of cDNA copies of viral mRNA, relative to cDNA copies of a host transcript (*Gapdh*), increased with the number of primers included in the ligation reaction, and saturated for reactions with over  $10^9$  primers per bead (Fig. 2.2.1.3c). Reverse transcription of *Gapdh* was not affected for DART-seq beads prepared with up to  $10^{10}$  primers per bead. These measurements allowed us to determine the optimal range of custom primers to impart onto beads for scRNA-seq experiments. Between  $10^9$  and  $10^{10}$  custom primers per bead, we observe no change in the relative amount of capture of viral transcripts to *Gapdh*, implying that we are not affecting host capture and have likely captured all targeted viral transcripts.

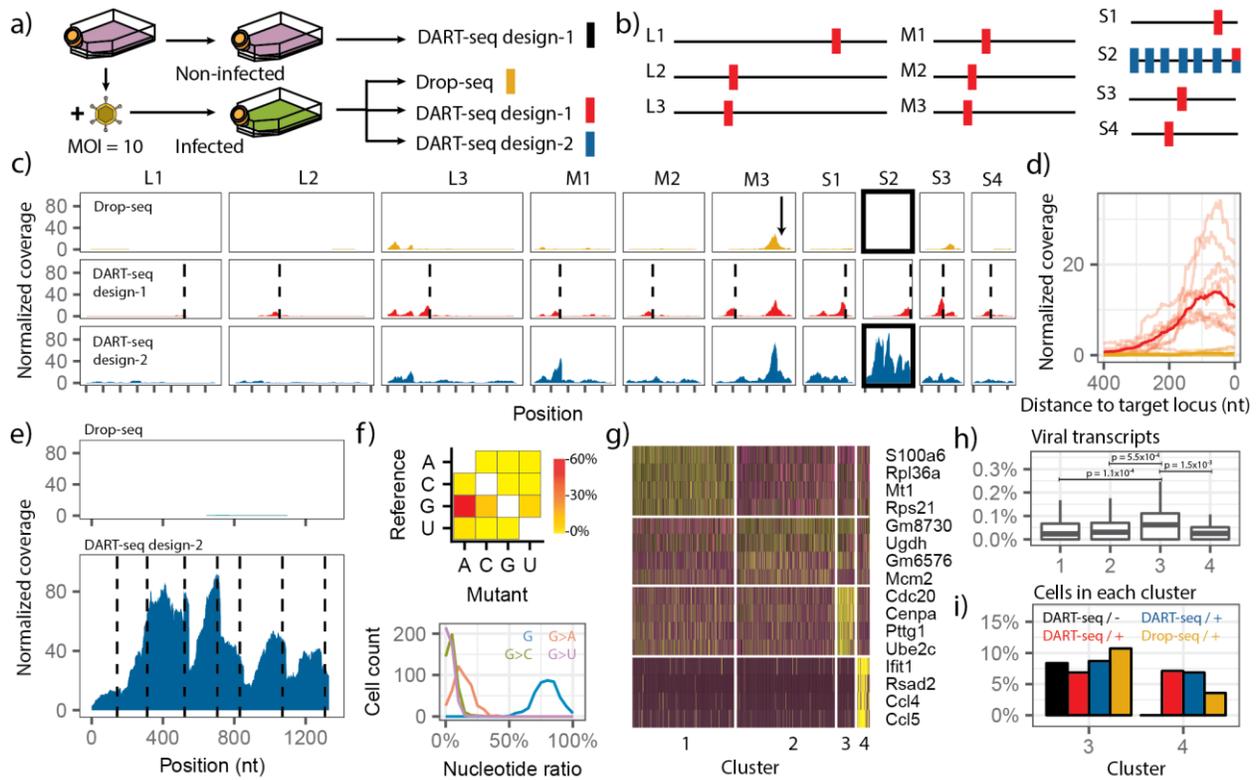
Next, we evaluated the abundance of amplicons in sequencing libraries of reovirus-infected cells generated by Drop-seq and a DART-seq assay targeting all ten viral genome segments. We designed seven qPCR assays with amplicons distributed across two viral genome segments (S3 and L3). To account for assay-to-assay and sample-to-sample variability, we normalized the number of molecules detected in DART-seq and Drop-seq libraries to the number of *Gapdh* transcripts. We observed substantial enrichment upstream (5'-end), but not downstream (3'-end), of the custom primer sites (Fig. 2.2.1.3d). Consistent with sequencing library preparation via tagmentation, we found that the degree of enrichment decreased with distance from the primer site.

Utilizing these proof-of-principle experiments allowed us to understand the limits and performance of the DART-seq assay. In general, the use of DART-seq should include the addition of between one and ten billion custom primer toeholds per bead. Toehold beads should be designed for genetically unique targeting regions, include few thymine and adenine repeats, and target at least 250 nt upstream of the 5'-end of the RNA molecules of interest. Last, during DART-seq bead synthesis, beads should remain well mixed during enzymatic reactions. With these considerations, we performed DART-seq on two unique systems related to innate and adaptive immunity.

## 2.2.2 DART-seq reveals heterogeneity of cellular phenotypes and viral genotypes

There is a great need for novel single-cell genomics tools that can dissect the heterogeneity in viral genotypes and cellular phenotypes during viral infection [183]. We used DART-seq to examine infection of murine L cells with T3D reovirus. The reovirus polymerase transcribes non-polyadenylated mRNAs from each of its ten dsRNA genome segments [184, 185]. We infected L cells at a multiplicity of infection of 10 (MOI 10), and allowed the virus to replicate for 15 hours after inoculation, creating a condition for which nearly all cells are infected (Fig. 2.2.2.1a). We performed Drop-seq and DART-seq experiments on infected L cells and non-infected L cells as control. We implemented two distinct DART-seq designs. The first DART-seq design targeted each viral genome segment with a single amplicon. The second DART-seq design was comprised of seven amplicons targeting loci distributed relatively evenly across the S2 genome segment (Fig. 2.2.2.1b).

To determine the efficiency by which DART-seq retrieves viral transcripts near the target sequence, we analyzed the per-base coverage of positions upstream of the DART-seq target sites. For DART-seq design-1, we observed a mean enrichment of 34.7x in the gene regions 200 nt upstream of the ten toeholds. In both DART-seq design-1 and 2, all targeted sites were enriched compared to standard Drop-seq beads (Fig. 2.2.2.1c,d). Viral transcripts were detected in Drop-seq libraries upstream of A-rich sequences in the



**Figure 2.2.2.1** DART-seq reveals heterogeneity in viral genotypes and host response to infection. (a) Experimental design. MOI, multiplicity of infection. b, Schematic of DART-seq designs (design-1, red bars; design-2, blue bars). c, Comparison of sequence coverage (normalized to host UMI detected  $\times 106$ ) of the ten reovirus gene segments (columns) for three library preparations (rows). An A5 pentanucleotide sequence part of segment M3 is shown (arrow). Dotted lines, DART-seq target positions. d, Per-base coverage upstream (5' end) of ten custom primers of DART-seq design-1 (light red; average shown in dark red), and mean coverage achieved with Drop-seq (yellow). e, Per-base coverage of S2 gene segment achieved with DART-seq design-2 (bottom; dashed lines indicate custom primer positions) and Drop-seq (top). f, Frequency and pattern of base mutations (top); histogram of nucleotide ratios for positions with reference nucleotide G detected in single cells (bottom). g, Clustering analysis for variable gene expression of reovirus-infected L cells (DART-seq design-1; yellow/purple indicates higher/lower expression). Similar clustering was observed in all three experiments with infected cells. h, Relative abundance of viral transcripts in L cell clusters (P values determined by two-tailed Wilcox rank-sum test). Lower and upper hinges correspond to 25th and 75th percentiles, respectively. Lower/upper whisker corresponds to smallest/largest value within 150% of the interquartile range from the nearest hinge (cluster 1,  $n = 411$ ; cluster 2,  $n = 397$ ; cluster 3,  $n = 50$ ; cluster 4,  $n = 69$ ). i, Fraction of cells in metaclusters for four experiments depicted in panel a with assay type and infection status (+ or -) indicated. {36}

viral genome, consistent with spurious priming of reverse transcription by poly-dT sequences on the oligo, as expected for Drop-seq. For example, a 200 nt gene segment upstream of an A<sub>5</sub> sequence on segment M3 (position 1952) was significantly enriched in the Drop-seq dataset (Fig. 2.2.2.1c; marked by arrow).

To test the utility of DART-seq to measure the heterogeneity of viral genotypes in single infected cells, we used DART-seq design-2 (Fig. 2.2.2.1b), which was tailored to retrieve the complete S2 viral gene segment. The S2 segment encodes inner capsid protein  $\sigma 2$ . Across cells with at least 1500 UMIs, DART-seq design-2 increased the mean coverage across the S2 segment 430-fold compared to Drop-seq, thereby enabling the investigation of the rate and pattern of mutations (Fig. 2.2.2.1e). 176 single-nucleotide variants (SNVs) were identified across the S2 segment (minor allele frequency greater than 10%, and per-base-coverage greater than 50x). Mutations from guanine-to-adenine (G-to-A) were most common (58%; Figure 2.2.2.1f, top). We did not observe such a mutation pattern in a highly-expressed host transcript (*Actb*). We examined the mutation load of viral transcripts at the single-cell level, and observed a wide distribution in mutation load, with a mean G-to-A conversion rate of 13%, and up to 41% (Fig. 2.2.2.1f, bottom).

The reason for such a high rate of mutation is unclear. G-to-A transamidation is an uncommon post-transcriptional modification that is not been previously seen as a host response to reovirus infection [186, 187]. The high rate of G-to-A transition in the viral transcript could also be secondary to a defect in the fidelity of viral transcription. The T3D orthoreovirus strain used in this study has strain-specific allelic variation in the viral polymerase co-factor,  $\mu 2$ , that has been shown to affect the capacity of  $\mu 2$  to associate with microtubules and the encapsidation of viral mRNAs within capsids [188, 189]. Following the publication of DART-seq, experiments to determine the origin the G-to-A hypermutation were performed<sup>1</sup>. Negative-stranded genomic RNA and viral mRNA were isolated individually from the S2 and S4 segments for T3D and T1L reovirus infected murine fibroblasts under the same conditions as described above. These eight unique RNA libraries were reverse-transcribed and sequenced to a depth of over 5000x across 95% of the segment (Illumina Miseq 2x75bp). We found strong concordance between the targeted sequence and the most abundant segment determined through sequence alignment (> 95% of reads align to targeted segment). In one of the eight libraries, representing genomic RNA from the S2 segment of T3D reovirus,

---

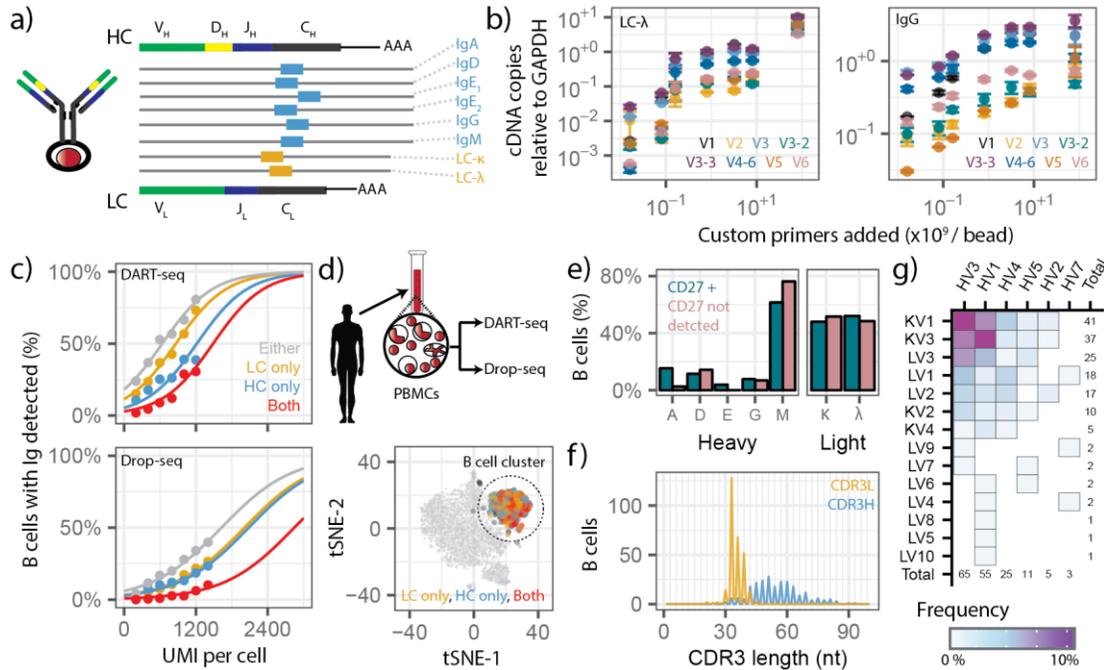
<sup>1</sup> Experiments performed by Mercedes Lewandrowski of John Parker's laboratory.

we observed an increased mutation burden from guanine to adenine of 2% (across 73 sites). No other libraries yielded significant mutational burden deviant from the reference sequences. Additional experiments are currently being performed to confirm our observations in scRNA-seq data.

To identify distinct host cell populations based on patterns of gene expression, we performed dimensional reduction and unsupervised clustering using approaches implemented in Seurat [170, 190]. We identified four distinct cell clusters for the monoculture infection model (DART-seq design-1, Figure 2.2.2.1g). Two major clusters comprised of cells with elevated expression of genes related to viral RNA transcription and replication (*Rpl36a*, cluster 1) and metabolic pathways (*Ugdh*, cluster 2). Two additional clusters were defined by the upregulation of genes related to mitotic function (*Cdc20*, *Cenpa*; cluster 3) and innate immunity (*Ifit1*, *Rsad2*; cluster 4), respectively (Fig. 2.2.2.1g). The abundance of viral gene transcripts relative to host transcripts was significantly elevated for cells in cluster 3 ( $n = 69$  of 927 total cells) compared to cells in all other clusters (Fig. 2.2.2.1h; two-tailed Mann Whitney U test,  $p = 1.0 \times 10^{-4}$ ). We merged all datasets for the Drop-seq and three DART-seq assays and quantified the cell type composition for each experiment. We did not observe cells related to cluster 4 (immune response), for the non-infected control, as expected (Fig. 2.2.2.1i). These results support the utility of DART-seq to study single-cell heterogeneity in viral genotypes and cellular phenotypes during viral infection.

### 2.2.3 DART-seq allows high-throughput paired repertoire sequencing of B lymphocytes

As a second application of DART-seq, we explored the biological corollary of viral infection, the cellular immune response. The adaptive immune response is reliant upon the generation of a highly diverse repertoire of B lymphocyte antigen receptors, the membrane-bound form of antibodies expressed on the surface of B cells, as well as antibodies secreted by plasmablasts [191, 192]. We applied DART-seq to investigate the B cell antibody repertoire in human PBMCs. We compared the performance of DART-seq and Drop-seq to describe the antibody repertoire (Fig. 2.2.3.1a). Antibodies are comprised of heavy ( $\mu$ ,  $\alpha$ ,



**Figure 2.2.3.1** DART-seq measures paired heavy and light chain B cell transcripts at single-cell resolution. (a) DART-seq custom primer design targeting the constant region of human heavy and light isotypes. (b) cDNA copies of immunoglobulin (Ig) transcripts relative to *GAPDH* as a function of the number of custom primers included in the ligation reaction (left panel, LC- $\lambda$  + V primers; right panel, IgG + V primers; 62,500 cells, 12,000 beads, bulk assay). Points are mean of two replicate measurements; bars indicate minimum and maximum. (c) Percentage of B cells for which heavy and/or light chain transcripts were detected as a function of the UMI count per cell. Cells were binned by the number of UMI detected (bin width 200 UMI, 0–2,400 UMI per cell, bins with fewer than 20 cells omitted, 26–2,396 cells per bin). Distributions were fit with a sigmoid curve (Methods). (d) Drop-seq and DART-seq assays of human PBMCs. Experiments were performed on two distinct PBMC samples ( $n = 2$ ). Representative t-SNE for one DART-seq assay shown here (4,997 single cells). Cells are colored on the basis of heavy and/or light chain transcript detection. € Bar graph of isotype distribution for CD27+ B cells and B cells for which CD27 was not detected. (f) CDR3L and CDR3H length distribution.  $n = 818$  B cells. (g) Paired heavy (IGHV) and light (IGKV and IGLV) variable chain usage in B cells;  $n = 164$  single cells. {37}

$\gamma$ ,  $\delta$ ,  $\epsilon$ ) and light chains ( $\kappa$ ,  $\lambda$ ), linked by disulfide bonds (Fig. 2.2.3.11). Each chain contains variable and constant domains. The variable region of the heavy chain is comprised of variable (V), diversity (D) and joining (J) segments, whereas the variable region of the light chain consists of a V and J segment (Fig. 2.2.3.1a). We designed DART-seq to target the site where the constant domain is joined to the VDJ gene segment in both heavy and light chain loci [193] (Fig. 2.2.3.1a). This design allows us to investigate the complementarity-determining region 3 (CDR3), which plays a key role in antigen binding. This region often goes undetected in regular scRNA-seq methods due to its distance from the 3'-end of the transcript.

We examined the efficiency of heavy and light chain reverse transcription by qPCR (CD19+ B cells) and observed an enrichment of transcripts for all isotypes tested, as the number of custom primers on DART-seq beads was increased (Fig. 2.2.3.1b). The response in enrichment to the addition of custom primers reflected the same pattern seen for targeted viral mRNAs in reovirus-infected cells (Fig. 2.2.1.3b). Next, we compared the performance of DART-seq and Drop-seq to describe antibody repertoires (Fig. 2.2.3.1c). Approximately 120,000 B cells were loaded in each reaction, yielding 4,909 and 4,965 transcriptomes for DART-seq and Drop-seq, respectively. The number of UMIs and genes detected per cell was similar for DART-seq and Drop-seq. We mapped transcript sequences to an immunoglobulin sequence database (see Section 2.2.4). For both DART-seq and Drop-seq, the percentage of cells for which immunoglobulin transcripts were detected scaled with the number of UMIs detected in the cells (Fig. (Fig. 2.2.3.1c). The immunoglobulin transcript recovery rate was substantially greater for DART-seq. For cells with 1,000–1,200 UMIs, we identified both heavy and light chain transcripts in 29% of cells using DART-seq, but in only 3% of cells using Drop-seq.

Next we applied DART-seq to determine the B cell antibody repertoire within human peripheral blood mononuclear cells (PBMCs) (120,000 PBMCs, 4,997 single-cell transcriptomes). [170] (Fig. 2.2.3.1d, top). We identified B cells based on expression of the B cell specific marker *MS4AI* [194]. We mapped transcript sequences obtained from B cells to the immunoglobulin (IG) sequence database, to find matches for the heavy and light chain transcripts in these cells, using MiXCR 2.1.5 [195]. We visualized B cells for which heavy and/or light chain transcripts were detected using t-SNE [190] (Fig. 2.2.3.1d, bottom). We detected immunoglobulin transcripts in 564 of the 818 cells in the B cell cluster, and immunoglobulin expression mapped accurately onto the B cell population.

We performed isotype distribution analysis on CD27+ B cells (Fig. 2.2.3.1e). As expected, CD27+ B cells were a mixed population of heavy chain isotypes, with IgM most frequently observed, followed by

IgD and IgA [196] (Fig. 2.2.3.1e). Kappa and lambda light chain isotypes were equally represented, as expected [197–199] (Fig. 2.2.3.1e). B cells for which we did not detect CD27 were predominantly of the IgM isotype [200] (Fig. 2.2.3.1e).

B cells derive their repertoire diversity from the variable regions of their heavy (IGHV) and light chains [201] (IGKV, IGLV). DART-seq captured a more diverse population of variable isoforms than Drop-seq. DART-seq can pair variable heavy and light chain transcripts in single cells. Out of 564 immunoglobulin-transcript-positive cells, we mapped the complete CDR3L in 339 cells and the complete CDR3H in 236 cells. The complete CDR3L+ CDR3H region was detected in 120 B cells. The number of variable heavy chain ( $V_H$ ) and variable light chain ( $V_L$ ) transcripts in single cells was correlated, as expected (corr. = 0.683, Pearson,  $p \ll 10^{-10}$ ,  $n = 120$ ). The CDR3L and CDR3H length distributions had maxima around 30 and 50 nucleotides, respectively, as described previously [193, 202] (Fig. 2.2.3.1f). In line with previous reports, promiscuous light chain pairing was observed in 73.5% of the repertoires in CD27– B cells [202]. Finally, we measured clone-specific pairing for the heavy (IGHV) and light chain variable regions (IGKV, IGLV) in 164 single B cells (Fig. 2.2.3.1g). The highest pairing frequency was observed between the most highly expressed heavy and light chain transcripts, in agreement with previous reports [192, 203]. The observed trend for preferred pairings in single cells was similar to published data [203].

#### 2.2.4 Detailed methods of DART-seq assay

*Primer bead synthesis.* Single-stranded DNA (ssDNA) probe sequences were designed to complement regions of interest. The probes were annealed to the complementary splint sequences that also carry a 10-12 bp overhang of A-repeats. All oligos were resuspended in Tris-EDTA (TE) buffer at a concentration of 500  $\mu$ M. Double-stranded toehold adapters were created by heating equal volumes (20  $\mu$ L) of the probe and splint oligos in the presence of 50 mM NaCl. The reaction mixture was heated to 95 °C and cooled to 14 °C at a slow rate (-0.1 °C/s). The annealed mixture of toehold probes was diluted with TE buffer to obtain

a final concentration of 100  $\mu$ M. Equal amounts of toehold probes were mixed and the final mixture diluted to obtain the desired probe concentration (2 pmoles for reovirus DART-seq design-1 and B-cell DART-seq, and 10 pmoles for reovirus DART-seq design-2). 16  $\mu$ L of this pooled probe mixture was combined with 40  $\mu$ L of PEG-4000 (50% w/v), 40  $\mu$ L of T4 DNA ligase buffer, 72  $\mu$ L of water, and 2  $\mu$ L of T4 DNA Ligase (30 U/ $\mu$ L, Thermo Fisher). Roughly 12,000 beads were combined with the above ligation mix and incubated for 1 hr at 37 °C (15 second alternative mixing at 1800 rpm). After ligation, enzyme activity was inhibited (65 °C for 3 minutes) and beads were quenched in ice water. To obtain the desired quantity of DART-seq primer beads, 6-10 bead ligation reactions were performed in parallel. All reactions were pooled, and beads were washed once with 250  $\mu$ L Tris-EDTA Sodium dodecyl sulfate (TE-SDS) buffer, and twice with Tris-EDTA-Tween 20 (TE-TW) buffer. DART-seq primer beads were stored in TE-TW at 4 °C.

*Cell preparation.* Murine L929 cells (L cells) in suspension culture were infected with recombinant Type 3 Dearing reovirus at MOI 10. After 15 hours of infection, the cells were centrifuged at 2300 rpm for 10 minutes and resuspended in PBS containing 0.01% BSA. Two additional washes were followed by centrifugation at 1200 rpm for 8 min, and then resuspended in the same buffer to a final concentration of 300,000 cells/mL. Human PBMCs were obtained from Zen-Bio. Cells were washed three times with PBS containing 0.01% BSA, each wash followed by centrifugation at 1500 rpm for 5 min, and then resuspended in the same buffer. The cell suspension was filtered through a 40 micron filter and resuspended to a final concentration of 120,000 cells/mL.

*Single cell library preparation.* Single cell library preparation was carried out as described previously<sup>2</sup>. Briefly, single cells were encapsulated with beads in a droplet using a microfluidics device (FlowJEM, Toronto, Ontario). After cell lysis, cDNA synthesis was carried out (Maxima Reverse Transcriptase, Thermo Fisher), followed by PCR (2X Kapa Hotstart Ready mix, VWR, 15 cycles). cDNA libraries were tagged and PCR amplified (Nextera tagmentation kit, Illumina). Finally, libraries were pooled and sequenced (Illumina Nextseq 500, 20x130 bp).

*qPCR measurement of viral gene segments.* 0.1 ng DNA from sequencing libraries was used per qPCR reaction. Each reaction was comprised of 1  $\mu$ L cDNA (0.1 ng/ $\mu$ L), 10  $\mu$ L of iTaq™ Universal SYBR® Green Supermix (Bio-Rad), 0.5  $\mu$ L of forward primer (10  $\mu$ M), 0.5  $\mu$ L of reverse primer (10  $\mu$ M) and 13  $\mu$ L of DNase, RNase free water. Reactions were performed in a sealed 96-well plate using the following program in the Bio-Rad C1000 Touch Thermal Cycler: (1) 95 °C for 10 minutes, (2) 95 °C for 30 seconds, (3) 65 °C for 1 minute, (4) plate read in SYBR channel, (5) repeat steps (2)-(4) 49 times, (6) 12 °C infinite hold. The resulting data file was viewed using Bio-Rad CFX manager and the Cq values were exported for further analysis. Each reaction was performed with two technical replicates.

*Toehold ligation measurement via fluorescent hybridization.* Roughly 6000 DART-seq beads were added to a mixture containing 18  $\mu$ L of 5M NaCl, 2  $\mu$ L of 1M Tris HCl pH 8.0, 1  $\mu$ L of SDS, 78  $\mu$ L of water, and 1  $\mu$ L of 100  $\mu$ M Cy5 fluorescently labeled oligo. The beads were incubated for 45 minutes at 46 °C in an Eppendorf ThermoMixer C (15", at 1800 RPM). Following incubation, the beads were pooled and washed with 250  $\mu$ L TE-SDS, followed by 250  $\mu$ L TE-TW. The beads were suspended in water and imaged in the Zeiss Axio Observer Z1 in the Cy5 channel and bright field. A custom Python script was used to determine the fluorescence intensity of each bead.

*Single cell host transcriptome profiling in viral infected cells.* We used previously described bioinformatics tools to process raw sequencing reads [7], and the Seurat package for downstream analysis [170]. Cells with low overall expression or a high proportion of mitochondrial transcripts were removed. For clustering, we used principal component analysis (PCA), followed by k-means clustering to identify distinct cell states. For meta-clustering, host expression matrices from all four experiments were merged using Seurat [170]. Cells with fewer than 2000 host transcripts were excluded. k-means clustering on principal components was used to identify cell clusters.

*Viral genotype analysis.* Sequencing reads that did not align to the host genome were collected and aligned to the T3D reovirus genome [204] (GenBank Accession EF494435-EF494445). Aligned reads were tagged

with their cell barcode and sorted. The per-base coverage across viral gene segments was computed (Samtools [50] depth). Positions where the per-base coverage exceeded 50, and where a minor allele with frequency greater than 10% was observed, were labeled as SNV positions. The frequency of SNVs was calculated across all cells. For the combined host virus analysis, the host expression matrix and virus alignment information were merged. The per-base coverage of the viral genome was normalized by the number of host transcripts. Cells with fewer than 1500 host transcripts were excluded from the analysis.

*IG heavy and light chain identification.* Sequences derived from B cells (cells that are part of the cluster of B cells identified in Seurat, and that have nonzero expression of the *MS4A1* marker gene) were collected and aligned to a catalog of human germline V, D, J and C gene sequences using MiXCR version 2.1.5 [195]. For each cell, the top scoring heavy and light chain variable regions were selected for subtyping and pairing analyses (Fig. 2.2.3.1g).

*Sigmoidal fitting heavy/light chain capture.* The mapping for the fractions of B cells containing heavy chains or light chains was fit with the following sigmoidal function:

$$y = \frac{1}{1 + e^{-b/(x-c)'}}$$

where the parameter  $b$  was a free parameter for the fit of the light chain or heavy chain data, and then fixed for the light chain only, heavy chain only, and combined light chain and heavy chain data.

*Statistical analysis.* Statistical tests were performed in R version 3.3.2. Groups were compared using the nonparametric Mann-Whitney U test.

Supplementary materials available at <https://www.nature.com/articles/s41592-018-0259-9#Sec20>. Code for sequencing analysis pipelines available at <https://github.com/pburnham50/DART-seq>.

## 2.2.5 Conclusions from DART-seq proof-of-principle experiments

We have presented an easy-to-implement, high-throughput scRNA-seq technology that overcomes the limitation of 3'-end focused transcriptome measurements. DART-seq allows sequencing of all RNA types and all regions of the polyadenylated transcriptome in a single cell while maintaining the ability to perform single-cell transcriptome profiling. A straightforward and inexpensive ligation assay is used to synthesize DART-seq primer beads (Fig. 2.2.1.1). The additional experiment time required for DART-seq compared to Drop-seq is minimal (2 hours) as is the cost per experimental design (~ \$100 per experiment). DART-seq is compatible with simultaneous querying of many amplicons. Here, we present example designs with 7-10 amplicons. The design and ratio of probes can be tailored to individual applications allowing researchers the flexibility to use their existing scRNA-seq set-up for a wide variety of biological measurements.

We have highlighted two potential applications of DART-seq technology. First, we demonstrated that DART-seq provides a means to study the heterogeneity in viral genotypes and cellular phenotypes during viral infection. We were able to recapitulate a full segment of a dsRNA viral genome, while simultaneously profiling the transcriptome of the infected host cells (Fig. 2.2.2.1). DART-seq opens new avenues for studies of host-virus interactions. We further applied DART-seq to measure endogenously paired, heavy and light chain amplicons within the transcriptome of human B lymphocyte cells in a mixed human PBMC population, while having access to full transcriptome data of all other cell types (Fig. 2.2.3.1). Determination of the paired antibody repertoire at depth can provide insights into several medically and immunologically relevant issues, including vaccine design and deployment.

### Chapter 3: Virus-inclusive scRNA-seq to understand enteric viral infections

“Enteric viruses cause broad morbidity and mortality across the world and particularly in resource-limited countries without access to adequate treatment. As vaccines for many of these viruses become more available, questions about their biology, particularly as it pertains to evolution, infectivity, and proliferation remain unanswered. Here we apply scRNA-seq technologies to describe host-virus interactions for the *Reoviridae* family viruses in two different infection models. We infected a monoculture of green monkey kidney fibroblasts with rotavirus and used DART-seq to profile single cells during the infection lifecycle. By using the relative viral transcript abundance as a pseudo-time measurement, we were able to construct a timeline for the transcription of various rotavirus genes. In a second experiment, we performed scRNA-seq on murine intestinal organoids infected with T1L orthoreovirus. We describe an initial dataset originating from this system and technical challenges that need to be addressed in performing scRNA-seq on organoids. We propose that marrying these two studies, performing scRNA-seq on a rotavirus infection in an intestinal organoid system, will yield novel insights into the host-pathogen biology.”

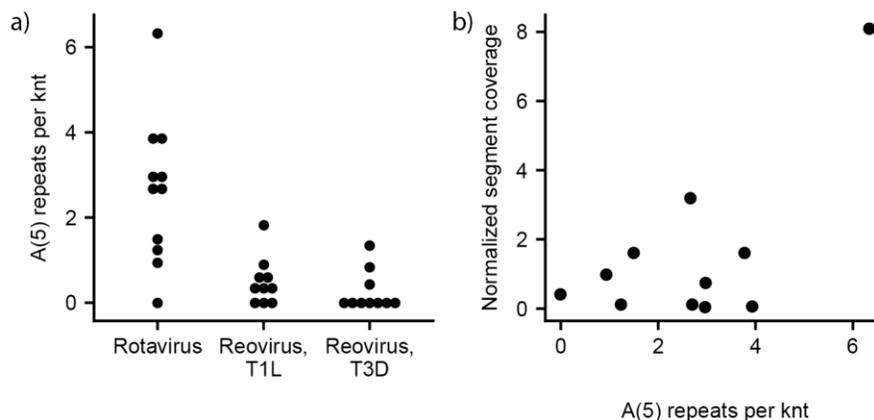
Experiments and sample collection in this chapter thanks to: Meleana Hinchmann, John S. L. Parker, Oyebola Oyesola, Elia Tait Wojno, and Mridusmita Saikia.

### 2.3.0 Motivations to expand studies to greater pathogenicity and complexity

Diarrheal diseases, many of which are caused by enteric viruses, are among the highest causes of death worldwide and particularly burdensome for young children [205]. Here, we described the enteric pathogen, rotavirus, in order to understand one of the drivers of enteric disease and due to its similarity to reovirus. Similar to orthoreovirus, rotavirus virions contain a double-stranded RNA genome made up of many segments (eleven segments in rotavirus compared to ten segments in the reovirus) and does not modify mRNA with polyadenylated tails [179]. Because the genome is segmented, it is possible for progeny virions to contain a mixture of genomes from parent viruses that co-infected the same cell (Fig. 2.1.0.2). While the genetics of rotavirus and reovirus are similar, rotavirus is more virulent. Rotavirus infects hundreds of thousands of children a year and is capable of causing severe gastrointestinal distress [206]. From the discovery of the virus in 1973 up to the development of a vaccine in the late 2000s, rotavirus was the primary cause for childhood death in sub-Saharan Africa, and persists as a burden on within communities lacking strong medical infrastructure [206].

Significant questions surrounding rotavirus biology and evolution remain. For example, though rotavirus is associated with gastrointestinal distress following infection of enterocytes, there are examples of systemic rotavirus infections with viral particles found in multiple organs [206]. This behavior suggests a form of cooperation among various viral subtypes, as has previously been noted in other enteric viruses [207]. We applied viral DART-seq to a monoculture of cells exposed to simian rotavirus (SA11).

We sought questions related to viral response within a community of diverse cells. As shown in the case of the EHV1-infected cells (Section 2.1), viral infection in an organism affects multiple cell types even if they are not directly infected. Furthermore, recent studies exploring cellular heterogeneity during viral infection in PBMCs and lung tissue have uncovered the actions of bystander cells due to infections [175, 208]. These cells, while not infected, are affected by the infection of neighboring cells, and likely play a role in the immune response [209]. Infections in monoculture are important to characterize viruses and for



**Figure 2.3.1.1** Targeted sequencing is not necessary for all viral transcripts. (a) The number of A(5) repeats per kilonucleotide (knt) is shown for three dsRNA, segmented viruses. (b) Mean normalized sequencing coverage across eleven rotavirus segments was calculated based on experimental results. {38}

---

technical development. However, it is only by understanding systems closer to natural infection that we can uncover the true behavior of cells, tissues, and the entire organism during infection.

Intestinal organoids (enteroids) represent a tractable system to study gastrointestinal diseases and are typically made up of several cell types [210]. The enteroids have a unique morphology bearing structural similarity to crypt-villi structure *in vivo* [210]. We developed a viral infection system using T1L orthoreovirus at a high MOI to infect murine enteroids. We will present the challenges with single-cell sequencing in organoid systems and describe the results of our preliminary experiments. We conclude this section by describing future directions in this field of research.

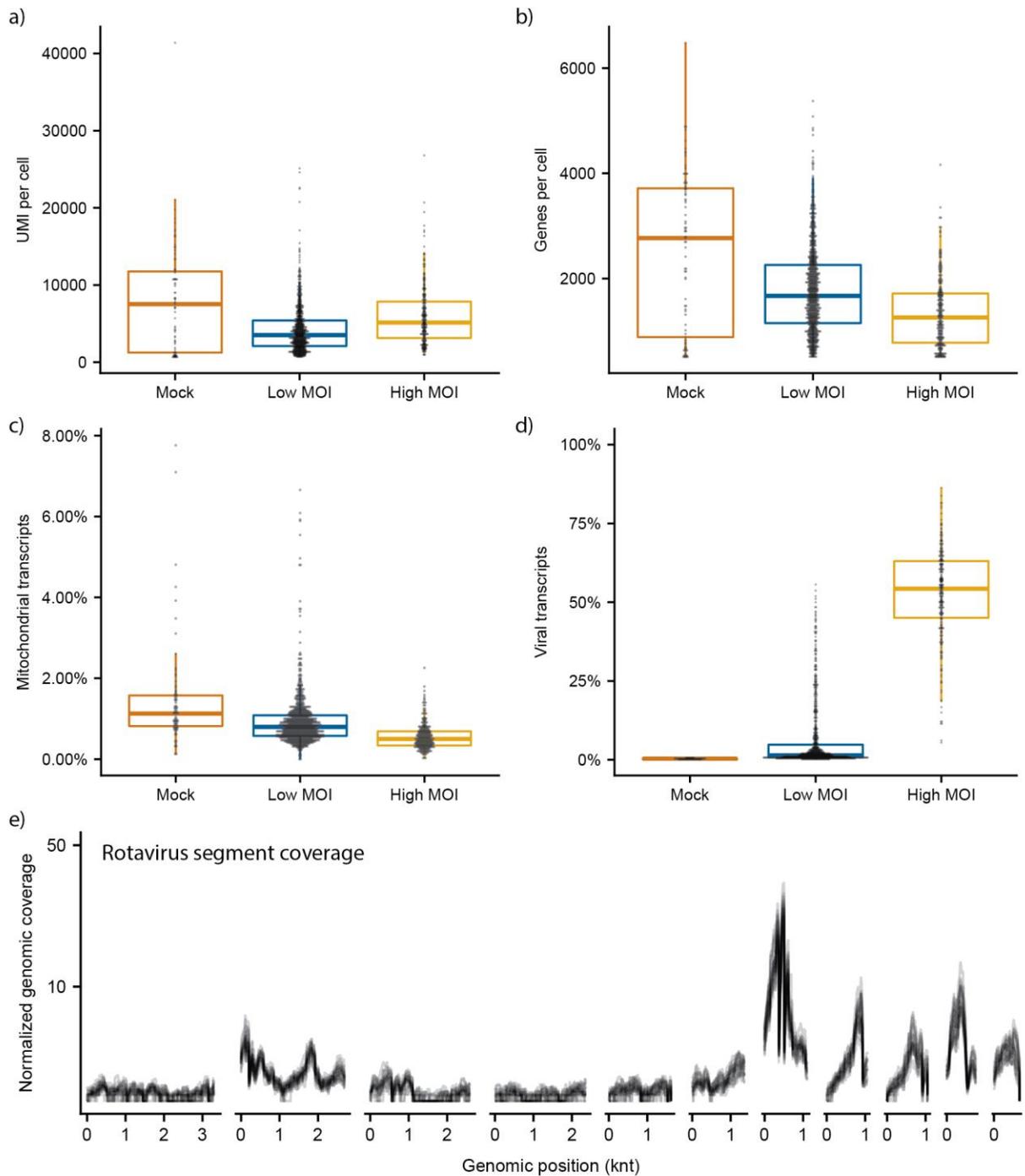
### 2.3.1 Rotavirus infections in monkey fibroblasts

Reovirus and rotavirus are part of the *Reoviridae* family of viruses and have similar genomic architecture and transcriptional mechanisms. However, the GC content of rotavirus is much lower than that of T3D reovirus across all segments (mean GC: rotavirus =  $34.7\% \pm 3.4\%$ , T3D reovirus =  $47.3\% \pm 1.4\%$ ). As a result, rotavirus mRNAs have accumulated multiple regions with adenine repeats. We previously identified a correspondence between A(5) repeats in reovirus positive-strand RNA (mRNA) molecules and

upstream sequencing coverage, due to the spurious binding of these regions to the poly(dT) regions of standard Drop-seq beads. The distribution of A(5) repeats per kilobase across the segments of T3D reovirus, T1L reovirus, and rotavirus is depicted in Figure 2.3.1.1a. We believe that due to the high density of A(5) repeats, targeted and untargeted scRNA-seq methods would both effectively capture rotavirus mRNAs. To test the efficiency of viral capture across various systems we prepared DART-seq (targeting each transcript's 3' region) for infected and control samples.

Green monkey kidney fibroblast (MA104) cells were suspended in a deep well plate and incubated with rotavirus A for one hour. Unlike reovirus, rotavirus was highly cytopathic in the MA104 cells at similar times and multiplicity. To avoid cytopathic effects in the fibroblasts, we used low multiplicity of infection and shorter incubation times. We trypsinized the cells prior to washing with PBS + 0.01% BSA; infecting at lower MOI (0.1, 1.0, 5.0); and incubating infected cells shorter time (3 to 5 hours). Cells were incubated at 37 °C for five hours. Following incubation, cells were trypsinized from the plate (TrypLE) and washed three times with PBS + 0.01% BSA to prepare for scRNA-seq library prep.

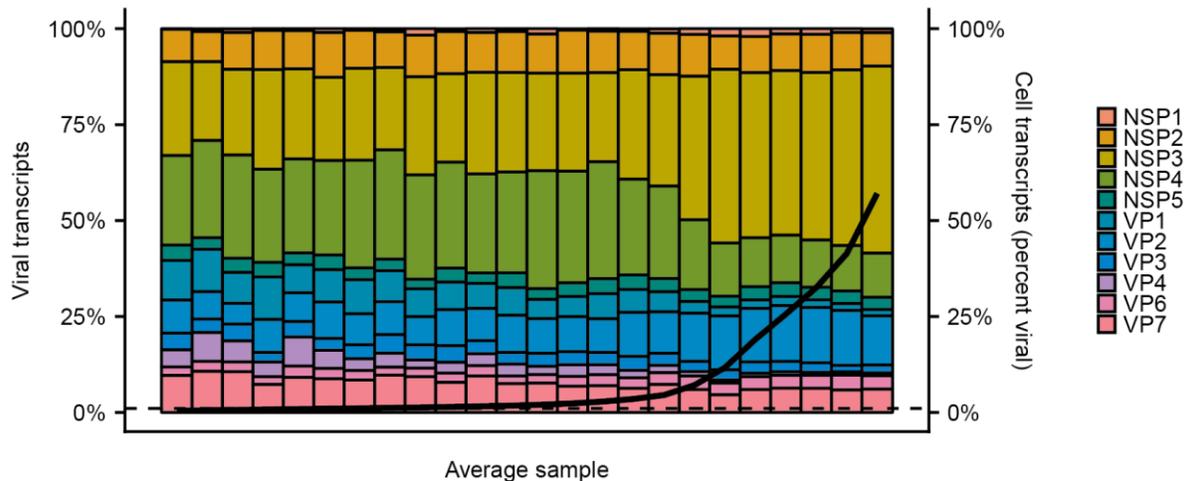
Following alignment of single-cell sequencing data to a hybrid reference transcriptome (Green Vervet Monkey [211] and Rotavirus A), we determined sequencing quality based on various single-cell metrics (Fig. 2.3.1.2a-d). We collected few cells in the Mock and High MOI samples (Fig. 2.3.1.2). However, the metrics reveal single-cell sequencing of the rotavirus-infected cells yielded a high amount of viral transcripts (Fig. 2.3.1.2d), which likely caused a reduction in the number of total genes and mitochondrial transcripts detected in cells, since the average number of UMIs per cell did not change between infection conditions. Following filtering we examine the read coverage across the viral genome segments in the cells with high viral fitness in the Low MOI sample. We examined the viral genome coverage across twenty-five high quality cells with a large amount of virus present in the cell (> 10% of transcripts originating from the virus). The sequencing coverage of the viral genome was significantly enriched upstream of A(5) repeats, while enrichment at the 3'-end of the DART-seq probe had varying efficiency (Fig. 2.3.1.2e).



**Figure 2.3.1.2** Rotavirus infected fibroblast quality analysis in single cells sequencing. (a-d) For Mock, Low MOI (MOI 0.1), and High MOI (MOI 5.0) single-cell sequencing analysis revealed the number of (a) UMI per cell, (b) genes per cell, (c) percentage of transcripts from the mitochondria, and (d) the percentage of transcripts from rotavirus. (e) Sequencing coverage across the rotavirus genome segments (ordered largest to smallest) 25 cells with the highest percentage of viral reads. Traces of each cell are in grey and mean of the traces is shown in black. Scale is square-root transformed. {39}

To our knowledge, the distribution of rotavirus gene transcripts during the virus lifecycle has not been reported. The rate of transcription by the viral RNA-dependent RNA polymerase (RdRp) is constant across segments for *Reoviridae* [179]. However, confounding factors such as mRNA protection via RNA-binding proteins and mRNA decay make it difficult to assess the veracity of the mean relative abundance for the eleven viral gene segments. We observed a pronounced variability in the expression of viral genes in the Low MOI samples. Most notably, as the relative proportion of viral to host transcripts decreased, an increasing proportion of rotavirus NSP3 gene expression was observed.

During the rotavirus infection lifecycle, NSP3 proteins recognize and bind to the 3'-end viral mRNA motif (UGACC), synonymous to the binding of polyA binding protein (PABP) to eukaryotic polyadenylated mRNAs [212]. NSP3 has a higher affinity for eIF4G than PABP, another protein in the host translational machinery, allowing viral mRNAs to be preferentially translated over host mRNAs [212, 213]. This process coincides with a forced relocalization of PABP into the nucleus, and occurs with fairly few NSP3 proteins, not long after the initial infection (within three hours) [214]. We demarcated cells that likely had no viral infection by comparing to the fraction of viral transcripts in the Mock dataset. Viral mRNAs detected in this dataset are likely from barcode hopping (cut-off of 1.00%) [108]. Thirty-seven genes were determined to be significantly correlated (Spearman,  $p < 10^{-10}$ ) with the relative abundance of viral transcripts. Most of these genes corresponded to binding and structural activity. To find transcripts significantly affected between the low and high viral abundance groups, we compared cells for which fewer than 5% of mRNAs originated from rotavirus to those with higher than 5% of mRNAs from rotavirus. A single gene, *PABPCI*, was significantly altered, having a 2-fold decrease in the high versus the low group; *PABPCI* encodes a protein that is part of the PABP complex.



**Figure 2.3.1.3** Viral gene transcription is altered through infection progression. Viral mRNAs are averaged in groups of thirty cells after ordering by the percent of transcripts of viral origin in each cell. The fractional abundance of each transcript in the cell bin (30 cells) is shown as a stacked barplot. Line indicates the fraction of reads originating from the virus in each cell bin. Dotted line represents the cutoff of uninfected cells. Bar color represents viral gene segment (annotated on right). {40}

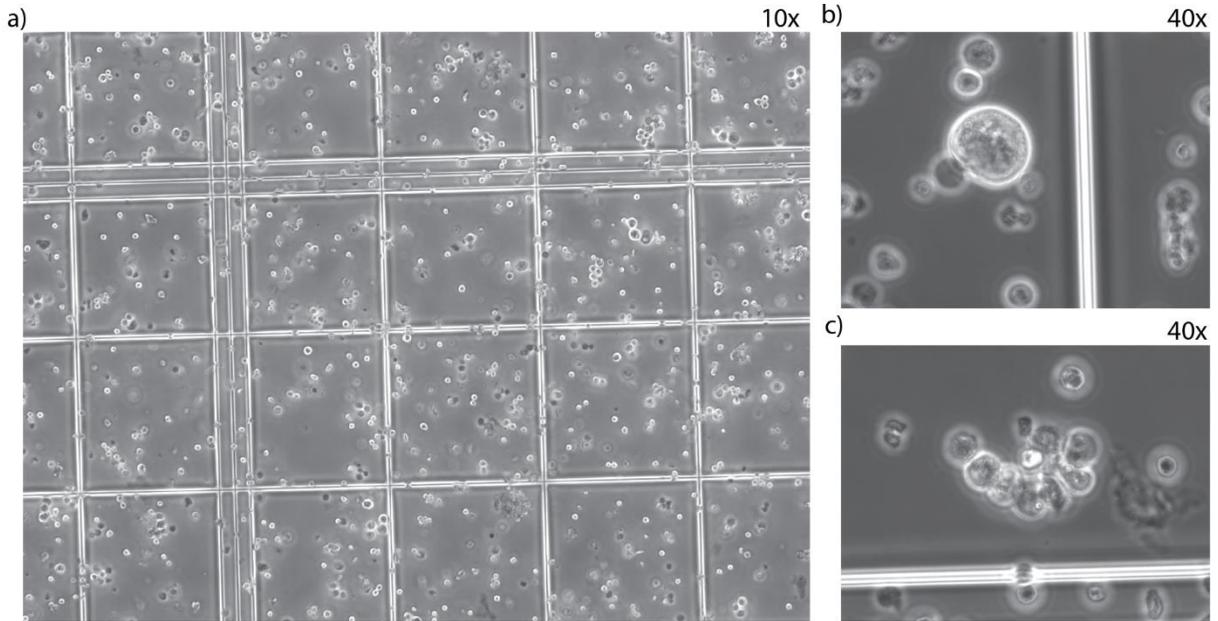
In addition to the higher relative abundance of NSP3 with increased viral abundance (29.1% of viral mRNAs to 48.8% of viral mRNAs), we also observed a corresponding decrease in the abundance of VP1 (5.2% to 1.6% of viral reads) and NSP4 (24.1% to 11.5% of viral reads). The relative abundance of other viral genes was not significantly altered. VP1 and NSP4 genes encode the viral RdRp and a nonstructural enterotoxin, respectively [215]. We hypothesize that the decrease in mRNAs corresponding to the RdRp could be an effect of the viral lifecycle. As the cell is preparing to reduce progeny virions, there is no longer a need for active transcription, leading to a decrease in VP1 protein. Similarly, infections at this stage of the viral lifecycle could be downregulating enterotoxin production, leading to a decrease in the relative abundance of NSP4. Indeed, NSP4 has been shown to be a transcription regulator [216].

This pilot study revealed a heterogeneity in infection likely caused by the time of infection. In our experiments, the only condition altered was the multiplicity of infection. However, it is likely that there is some heterogeneity in specific time of infection and number of virions infecting individual cells. At an MOI of 0.1, for instance, it is likely that 1 in 200 cells is infected by multiple virions. Because we use a stock of pure viruses on a monoculture, the transcriptome of cells likely represents a look into cellular progression

through infection. Further experiments need to be conducted to confirm these observations and access deeper information about rotavirus infection. Due to the high A(5) repeat density across the rotavirus segments, non-targeted sequencing strategies such as 10x genomics sequencing can be used. This will greatly improve the depth of sequencing per cell. These experiments are currently underway.

### 2.3.2 Single-cell sequencing of complex cellular communities

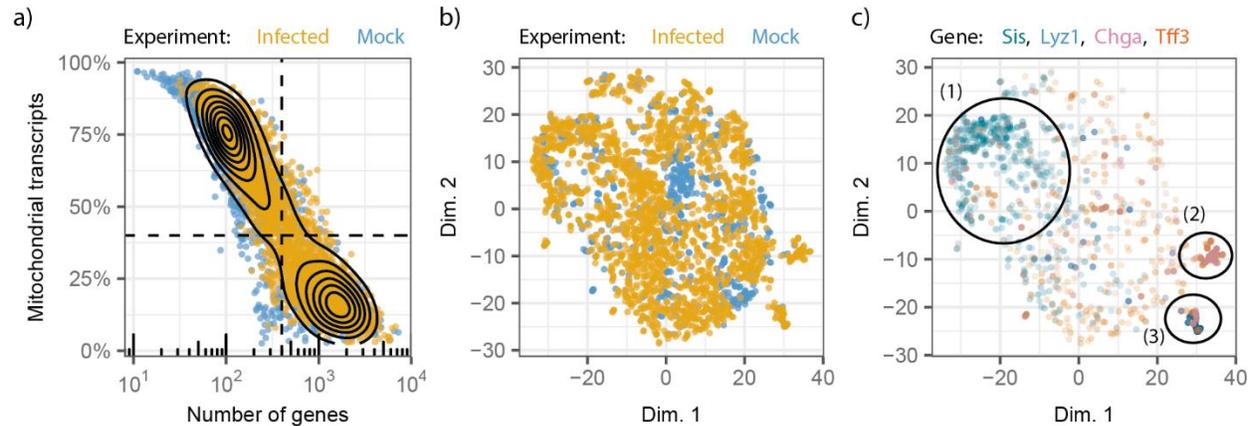
It is appreciated that virus-inclusive scRNA-seq can reveal the extent of heterogeneity in infected cells and their associated viruses [174, 217]. Uninfected cells within infected hosts and those neighboring infected cells can may also exhibit extreme heterogeneity driven by paracrine signaling and other intercellular communication [208]. Having shown the ability to perform Drop-seq and DART-seq in single cells to describe the changes induced by viral infections, we sought to expand our host system to one with greater cellular heterogeneity. The formation of intestinal organoids from *Lgr5+* stem cells has become an accessible system to study gastrointestinal disorders outside of live animal models [210, 218]. Enteroids contain several cell types including enterocytes, goblet cells, enteroendocrine cells, stem cells, and Paneth cells [219]. Viruses in the *Reoviridae* family have been shown to infect the enterocyte and enteroendocrine cells, the first and second most prominent cell type in these samples, respectively [219].



**Figure 2.3.2.1** Dissociation of organoids leaves cellular aggregates but represents all cell types. (a) 10x phase contrast image of dissociated enteroid on Fuchs-Rosenthal hemocytometer. (b-c) The diversity of cell types is observed by morphological features, as shown at 40x by the presence of (b) goblet cells and (c) remnants of crypts. Images taken on Zeiss Axio Observer Z1 under phase contrast. {41}

We explored the use of enteroids to describe infection with T1L reovirus. Enteroids were created by isolating the intestinal crypts from mice and allowed to develop in a nutrient-rich Matrigel for three weeks. We infected murine enteroids with reovirus (MOI 50), allowing binding to occur for one hour and followed by incubation for 24 hours. Following incubation, we dissociated enteroids by adding TrypLE enzyme pipetting vigorously. Following dissociation, the cell suspension was washed with PBS + 0.01% BSA and centrifuged at 200xg for five minutes between wash steps. After the final wash, cells were passed through a 100 micron filter to remove cell clusters (Fig. 2.3.2.1a-c). We processed filtered cells using droplet microfluidics for scRNA-seq as described previously [7, 8].

We performed Drop-seq as a proof-of-principle sequencing assay, with the intent to perform DART-seq. Drop-seq sequencing data from infected and mock organoids revealed a failure to capture viral transcripts (no T1L reovirus detected in both cases). The T1L reovirus genome has several A(5) repeat regions across the genome, so it is likely if the virus was replicating in the cells we would detect viral transcripts. We believe the lack of reads from viral mRNAs could be a result of a failed infection of the



**Figure 2.3.2.2** Differential gene expression and clustering analysis of low quality cells from enteroids. (a) The percentage of transcripts originating from the mitochondria is compared to the total number of genes expressed. Dotted lines represent cutoffs for filtering. Density plot overlaid on cells (colored by experiment). (b,c) t-SNE dimensional reduction of cell-cycle regressed gene expression. (b) Colors indicate Mock and Infected datasets. (c) Colors indicate increased expression of *Sis*, *Lyz1*, *Chga*, and *Tff3*. Intensity of color corresponds to intensity of gene expression. (1) Enterocyte cluster. (2) enteroendocrine + goblet cell cluster. (3) Paneth cell + goblet cell cluster. {42}

enteroids or a high amount of host transcripts leading to under-sampling of the viral mRNAs. However, we were able to perform single-cell sequencing analysis based on differential gene expression on the data aggregated from the two experiments [170].

Our results indicated that the scRNA-seq pipeline worked efficiently, though the enteroid system was likely under stress at the time of sample processing. It is possible that there was a high amount of cell lysate present in the cell suspension due to the dissociation technique. We observed a high relative abundance of mitochondrial transcripts compared to other single-cell sequencing datasets; the mitochondrial gene expression was bimodal distributed (Fig. 2.3.2.2a). Cells with high mitochondrial transcription had a corresponding low number of expressed genes (Fig. 2.3.2.2a) and were removed from the dataset. Following this quality control measure (reducing total number of cells to 2,696), we regressed the effect of 97 genes related to the murine cell cycle from our data and clustered based on the principal components [220]. Cells clustered on the cell cycle regressed gene expression did not segregate by experiment (Fig. 2.3.2.2b). However, we did observe small clusters of cells corresponding to the four major cell types based on enrichment of known gene markers (enterocytes, *Sis*; enteroendocrine cells, *Chga*;

goblet cells, *Tff3*; and Paneth cells, *Lyz1*) (Fig. 2.3.2.2c) [219, 221]. We defined enrichment as any gene expression value greater than one after scaling the UMI counts. Enterocytes made up the plurality of cell types (20.8%), followed by goblet cells (13.3%), Paneth cells (5.0 %), and enteroendocrine cells (1.7%). We found no gene enrichment of the four marker genes in 64.7% of cells.

There are several factors that could account for the unassigned majority of cells. High expression of mitochondrial, ribosomal, and cell cycle mRNA contributed to the majority of reads in all cell clusters. Lower abundance reads that likely carried cell markers were under-sampled and did not appear in datasets. The unassigned cells could also be an effect of proliferation and stem-ness among the cell population. We detected nonzero expression of *Lgr5*, the gene involved in stem-ness in the organoid system, and *Mki67*, a cell marker of proliferation, in many cells that were not assigned to the four cell types mentioned above [210, 222]. Unassigned cells could also be an effect of cell clusters trapped inside single droplets within the microfluidic device. As shown in Figure 2.3.2.1, we have observed cell aggregates resemble the crypts and villi, likely leading to larger numbers of cell doublets (triplets, quadruplets, etc...) in the datasets. Finally, the low number of enteroendocrine cells, coupled with the high amount of mitochondrial gene expression might explain the low abundance of viral reads in our dataset.

In summary, while our results do indicate the ability to sequence murine intestinal organoids at single-cell resolution, technical hurdles must be overcome to develop an accurate understanding of organoid systems during infection. In particular, the aspects that need to be understood and addressed concern: (1) the high number of mitochondrial gene transcripts, (2) the lack of transcripts originating from the reovirus, (3) the high number of nondescript cells (provided they are not accounted for after reduction of mitochondrial gene expression), and (4) enrichment of cells infected with reovirus.

### 2.3.3 Future experiments: rotavirus infections in organoid systems

In this section we have described virus-inclusive scRNA-seq from the highly pathogenic rotavirus and for reovirus in a complex organoid system. We have shown that rotavirus infections are well characterized in a DART-seq experiment, but, importantly have found that standard scRNA-seq methods will likely capture the heterogeneity of infection due to A(5) repeats in the rotavirus messenger RNA. In contrast, our results in the organoid system and in previous experiments indicate the need for targeted amplicon sequencing to recover reovirus transcripts. These observations indicate that a more sophisticated probe design algorithm can be used to create future DART-seq libraries. For example, non-polyadenylated mRNA with A(5) repeats near the 3'-end of the molecule do not need to be probed. The reduction in the probe pool complexity would, in turn, increase capture of other targets as well.

Intestinal organoids have been shown to be a powerful tool in which rotavirus infection can be studied [219]. Human rotavirus infection of intestinal organoids revealed tropism for multiple cell types, including enterocytes and enteroendocrine cells, and showed structural changes in the organoids after infection, such as luminal swelling [219]. Infections in the organoid system, compared to monoculture, also revealed an increase in the viral titer [219]. Given the ability to robustly capture rotavirus using any scRNA-seq technology, merging the studies discussed in this section would yield a robust system to study *Reoviridae* in a complex community. Though previous studies characterized gross changes to the intestinal organoids during human rotavirus infection, analysis at the single-cell level has not been documented. Analyses on viral mRNAs at single-cell resolution could therefore identify novel host defense mechanisms depending on cell type.

Infecting intestinal organoids with rotavirus could also reveal a new perspective on how bystander cells react during infection. In their analysis of bystander cells in influenza infections, Steuerman et al. observed that the bystander cells produce a strong interferon response [208]. Such an effect may be observed in Paneth and goblet cells in the organoid. Lastly, the addition of non-enteroid cell types into the Matrigel during development, the collective group called assembloids, have yielded systems that more

strongly resemble organs *in vivo* [223]. Adapting this method to include, for example, white blood cells into the enteroids might yield insights into viral adaptation that can describe the mechanism that leads to viremia and extraintestinal infections during some cases of infection, which is not well understood [224, 225].

## Chapter 4: Afterwards on virus-inclusive single-cell sequencing

Virus-inclusive scRNA-seq, while nascent, has provided a unique perspective to understanding the host-pathogen biology. However, there are still many aspects of virology which have not yet been described at single-cell resolution, including viral evolution. In this part we have described the implementation of a novel and facile scRNA-seq modification to interrogate viral infections in any system. We have applied the technique, DART-seq, to describe intercellular heterogeneity in both host and viral mRNA during infection with mammalian orthoreovirus and rotavirus. There are many questions that can be interrogated in these systems, but which will likely require technical developments. In particular, we believe that targeted scRNA-seq could be used to describe: (1) differences in genomic and messenger RNA in viruses, (2) polymicrobial infections, and (3) full viral transcripts, using long-read sequencing technologies.

Many viruses, including influenza, utilize a negative-stranded RNA to synthesize positive-stranded messenger RNAs. By capturing and analyzing SNPs between the genomic and messenger RNA for single virions, one could determine host defense mechanisms employed to, for instance, hypermutate viral mRNA [226]. We have previously attempted to use DART-seq to describe differences in the gRNA and mRNA for reovirus. Like influenza, dsRNA viruses use the negative-strand as a template to create positive-stranded mRNAs. We targeted one locus on the positive-stranded RNA and one locus on the negative-stranded gRNA for three segments in T3D orthoreovirus following murine fibroblast infection. We performed the DART-seq pipeline on these samples as described in Section 2.2.4. Our results indicated capture of positive-stranded mRNA at all three targeted loci, similar to what was depicted in Figure 2.2.3.1. We did not observe capture and enrichment at the three loci targeted on the negative-stranded gRNA. Moreover, the capture rate of sequencing reads aligning to the negative strand across all ten segments was similar to other experiments where the negative strand was not targeted.

The inability to reliably capture negative-stranded RNA is likely caused by the relative abundance of negative-stranded RNA to positive-stranded RNA. Since there is an abundance of positive-stranded mRNA relative to gRNA, negative-stranded RNA is almost always hybridized to the complementary positive-strand. dsRNA genomes are difficult to denature, and often require high temperatures and some form of chemical treatment [227, 228]. Therefore, it is unlikely that the DART-seq probe is able to displace the positive-stranded RNA to capture the negative-stranded RNA. Adaptations to scRNA-seq protocols are likely necessary to perform simultaneous strand capture. We are currently working to evaluate gRNA capture in single-stranded RNA viruses.

Another area of interest, as mentioned in the introduction of this part, involves the sequencing of single cells following infection by multiple virions. For example, Russell et al. previously performed synonymous coinfections with influenza in cell culture [217]. Furthermore, a recent discovery that enteric viruses, including rotavirus and norovirus, can achieve high titers through assembly within vesicles provides further impetus to study coinfections at single-cell resolution [229]. DART-seq can be used to target transcripts from various viral genotypes and explore the effect of multiple virions within single cells.

A third direction for future viral scRNA-seq studies involves the sequencing of full length cDNA to identify genetic variants on a cell-by-cell basis. DART-seq allows users to sequence transcripts at specific positions, and is not relegated to the 3'-end of the molecule like Drop-seq and 10x genomics based library preparation. However the sequence coverage is limited to the chosen read length on Illumina sequencing platforms, with a maximum length of roughly 300 bp. In contrast, long-read sequencing technologies have recently gained popularity and broad user accessibility, with no theoretical limit on the length of the fragment to be sequenced. Long-read scRNA-seq has recently been achieved using nanopore sequencing [230] and was applied to study viral genomic diversity in influenza virus [231]. Long read sequencing revealed that the host response in infected cells showed a varied expression based on SNPs and structural variants during infection [231]. These insights could not be easily resolved with standard scRNA-seq pipelines. It is likely that combining short-read scRNA-seq technologies, to broadly determine

differential gene expression in the host, with long-read scRNA-seq technologies on viral reads will further describe viral infections in the future.

Finally, we believe that as the field of scRNA-seq moves towards an understanding of spatial heterogeneity of gene expression in tissues, so too will research in viral scRNA-seq. It is well understood that a viral infection can begin with countably few virions [232] and that can trigger a wide-ranging immune response through paracrine signaling [209, 233]. Application of fluorescence in situ hybridization to localize transcripts of interest has become a reliable tool to measure single-cell heterogeneity and spatial distribution of transcripts [173, 234]. We envision that the next stage of understanding intracellular infections will employ these techniques to understand viral dissemination and intercellular signaling effects on cell state.

## Conclusions

In this dissertation, I have described a new series of biomolecular and bioinformatics techniques that can be used to better surveil and understand infectious disease.

We have shown the ability of cfDNA sequencing to accurately detect and describe viral and bacterial pathogens over time and space. In Section 1.2, I described how the implementation of a novel single-stranded library preparation strategy increased the abundance of ultrashort cfDNA molecules (< 100 bp in length). These molecules represent highly degraded forms of cell-free DNA in biological fluids, but have been neglected in past sequencing assays. We compared matched plasma cfDNA samples prepared using both single-stranded and double-stranded library preparation approaches. Our results indicated a higher relative abundance of microbial cfDNA in these samples, leading us to hypothesize that cfDNA sequencing could be employed to monitor infections.

We applied our single-stranded library preparation to cell-free DNA extracted from the urine supernatant of 141 samples from renal allograft recipients (Section 1.3). These patients are particularly prone to both bacterial and viral urinary tract infections. We compared the microbial abundance detected in cell-free DNA with clinical diagnoses determined through quantitative PCR and urine culture; we found high concordance between cfDNA sequencing and these gold standard techniques. Moreover, urinary cfDNA sequencing was sensitive to an array of viral and bacterial pathogens that were not detected in standard screens, including herpesvirus and *Haemophilus influenzae*. cfDNA sequencing also provided functional information regarding infections, including growth kinetics and antibiotic resistance of bacteria, as well as a measure of the extent of host damage.

As we observed in our analysis of urinary cfDNA from renal allograft recipients, cfDNA sequencing was sensitive to the matched, clinically diagnosed uropathogen in nearly every case. However, our analysis showed an abundance of nonpathogenic bacteria, including commensals and contaminants. To address the issue of contamination, we developed a pipeline to determine falsely-assigned microbial sequencing reads in datasets from samples with low microbial biomass (Section 1.4). We optimized our

background correction algorithm on a refined dataset from patients with monomicrobial UTIs; this allowed us to apply cfDNA sequencing with background correction to several novel datasets, including cfDNA extracted from amniotic fluid and peritoneal dialysis effluent.

The development of our cfDNA sequencing assay allowed us to pursue analyte validation in neglected tropical diseases for which current diagnostics may be insufficient (Section 1.5). We implemented cfDNA sequencing with background correction to analyze the presence of enteric microbiota in the plasma of pediatric patients in rural settings. Comparison to the standard dual-sugar assay revealed a group of bacteria associated with gut flora that was abundant in more severe cases of environmental enteropathy. In a separate study, we hypothesized that genome replication dynamics, ascertained by sequencing coverage across the *Mycobacterium tuberculosis* genome, could discrepant cases of latent and active MTB replication. This hypothesis was validated using whole genome sequencing of antibiotic-treated MTB in culture and MTB from the caseum in a rabbit model. We sequenced plasma cfDNA from adult patients with active MTB, but detected very few MTB molecules.

While cell-free DNA sequencing is able to capture the causative agent of infection in many cases of disease, several shortcomings in the processing of samples need to be resolved before it becomes a broad tool for clinical labs. In our own experiences, the ability to detect microbial cfDNA changes dramatically between biological fluids. Blood plasma, while having the ability to provide infection information from any vascularized tissue, often has low levels of microbial cfDNA, even in cases of infection. This effect is likely due to the presence of immune cells and nucleases clearing out cfDNA molecules. In contrast, urine and peritoneal effluent, have proven to be excellent reservoirs of microbial cfDNA during UTI and peritonitis, respectively. Other fluids, such as cerebrospinal and synovial fluid, have proven to be of interest to analyze host cell-free DNA, but have not been thoroughly analyzed for the presence of healthy and infected individuals [15, 235, 236].

There exist time and cost barriers for using cell-free DNA sequencing to monitor and diagnose infections. At the time of writing this dissertation, cfDNA sequencing can cost hundreds to thousands of

dollars to properly implement and requires an investment of large amount of capital in next generation sequencing platforms. Improvements are being made in the turn-around time to produce a diagnosis. A recent report illustrated that patients submitting plasma samples for cfDNA sequencing would have diagnostic reports in 24-36 hours [6]. While the costs of next generation sequencing are not prohibitive in many Western countries, the ability to implement the technique in rural areas and emerging economies, which are those most affected by infectious disease, is currently out of reach.

However, the decreasing cost of genome sequencing and the emergence of new sequencing technologies are moving cfDNA measurements to the realm of effective, field-deployable diagnostics. For example, in the last five years, the use of nanopore sequencing has been employed to rapidly sequence the genomes of hundreds of organisms in real-time. Nanopore sequencing works by detecting changes in electrical impedance as DNA and RNA molecules are pulled through enzymatic pores in a membrane [237]. The technique does not require prior sample amplification, is not restricted to a maximum molecule size, and can produce results in real time [238]. Microbiologists have recently used nanopore sequencing to detect the presence of bacteria in a variety of surveillance setting, including foodborne illnesses [239] and pandemic virus outbreaks [240]. Furthermore, the size and cost of the nanopore sequencers allow for easy transport to resource-limited settings including the International Space Station [241]. While nanopore sequencing of cell-free DNA has not yet been realized, short DNA sequencing via the ON Minion system has recently been demonstrated [242].

As described in Section 1.3.8, our work has shown that non-standard methods of cell-free DNA sequencing can be used to more deeply understand the origin of cfDNA molecules. Treating single-stranded cfDNA molecules with bisulfite can be used to identify the tissue-of-origin of molecules via their patterns of CpG methylation. Work in our lab has shown that non-human cfDNA can still be accurately mapped to microbes after bisulfite treatment, allowing for the simultaneous measure of host tissue and microbial cfDNA in biological fluids. While the ability to, for example, observe SNP changes in the microbial genomes is obscured by the bisulfite treatment, this technique gives unprecedented molecular detail about

the host-pathogen dynamics from cfDNA. We have used this technique to indicate kidney damage and white blood cell recruitment in viral and bacterial UTIs, respectively, which has expanded on standard metagenomic sequencing.

Cell-free RNA (cfRNA) sequencing provides another exciting potential biomarker to diagnose and monitor infectious disease. Somewhat analogous to bisulfite-treated cfDNA sequencing, cfRNA sequencing allows for the identification of tissue-level and cell-level histories of cfRNA molecules, which can be used to describe cellular migration, damage, and signaling [243, 244]. cfRNA is, theoretically, more comprehensive than cfDNA in identifying causative pathogens during infection (e.g. RNA viruses). In cases of bacterial infection, cfRNA could be used to identify active transcription of bacteria at the time of their lysis, which could determine if the bacteria are expressing genes related to growth, virulence, or antibiotic resistance. To date, few studies evaluating the clinical efficacy of cfRNA sequencing have been published, likely due to the technical challenges in isolating cfRNA and preventing sample degradation.

In the second part of the dissertation, I outlined the state of virus-inclusive scRNA-seq and described our efforts to further scRNA-seq technology in the context of infectious disease. High throughput scRNA-seq has produced novel insights into development, immunology, and neuroscience. However, many popular approaches to scRNA-seq use poly(dT) probes to capture polyadenylated RNA molecules. Polyadenylation, however, is not present on many noncoding RNAs or the messenger RNAs of many prokaryotes or viruses. To address this issue, we developed a straightforward modification protocol that allows for simultaneous capture of the polyadenylated transcriptome and specific RNA molecules using a multiplexible, custom probe array (Section 2.2). We validated our new technology, DART-seq, on cells infected with mammalian orthoreovirus, a segmented virus that does not modify its mRNA with polyadenylated tails. Our results indicated that DART-seq can more fully describe host-pathogen interactions, allowing for the depth of sequencing to categorize viral mutations and determine host transcriptomic programming that alters viral fitness.

We also illustrated that DART-seq may be employed to enrich the abundance of low abundance polyadenylated mRNAs in sequencing datasets. We targeted the constant regions of mRNAs corresponding to the heavy and light chain in antibodies in commercially available B-cells and a population of PBMCs. The capture of heavy and light chain targets was validated, and we showed that these molecules reciprocate known features of the immune repertoire, such as CDR3 length and relative abundance of various variable regions. The enrichment in heavy and light chain variable regions allowed us to produce a paired immune repertoire consistent with previous reports.

We extended our research into host-virus interactions at single-cell resolution using other models of infection by viruses in the *Reoviridae* family (Section 2.3). Rotavirus provides a more clinically relevant infectious organism compared to reovirus, as it is responsible for tens of thousands of deaths worldwide, annually. We performed rotavirus infection in cell culture at various multiplicities of infection to observe response in the host cell population. Differential expression among viral and host genes revealed the progressive life cycle of the virus through infection. To increase cellular complexity and analyze the infection response among bystander cells, we performed infection of murine enteroids (composed of four cell types), with T1L orthoreovirus. In our first experiment we were unable to detect viral transcripts; however, this pilot study has allowed us to establish guidelines for proper experimentation and analysis. We envision the expansion of virus-inclusive scRNA-seq studies using rotavirus will yield novel insights into viral cooperation and the innate immune response. Taken together, our experiments reveal the utility of the DART-seq platform to more completely describe the innate and adaptive immune system during infection.

Though preliminary, several studies have described the application of single-cell sequencing as a platform for novel, high-resolution biomarker discovery. Virus-inclusive scRNA-seq has recently been implemented to discern RNA analytes of severe dengue fever, a disease which can have as high as 20% mortality rate if untreated [175]. PBMCs were isolated by fluorescence-activated cell sorting into a 384-well plate and sequencing libraries were prepared. By sorting patient PBMC cells prior to scRNA-seq

library preparation, the authors enriched underrepresented populations of cells. scRNA-seq analysis determined several viable markers to predict severe dengue analysis including *MX2* in naive B cells and *CD163* in CD14<sup>+</sup> CD16<sup>+</sup> monocytes. In a separate study, scRNA-seq was applied to patient-derived placental tissue to explore biomarkers for preeclampsia [245]. Differential expression among single cells allowed the researchers determine cell population trajectories around the time of preeclampsia onset. cfRNA datasets reanalyzed by the group indicated a previously unseen enrichment in transcripts related to extravillous trophoblasts [245]. While both the works described above are preliminary, the studies illustrate potential for single-cell sequencing to become a platform for analyte discovery.

Future work in the De Vlaminck lab will expand the community's knowledge of infectious disease using single cells and circulating nucleic acids, and it is likely these two perspectives will overlap. In summary, I hope that my efforts in this field has contributed novel tools and approaches to understanding host-microbe interactions. Microbes are the most highly represented organisms on Earth [246]. This dissertation has focused mainly on their role in disease and disruption; however, our story and wellbeing is intricately woven into theirs. It will be exciting to see how our understanding of microbiology and immunology changes in the coming years.

## REFERENCES

1. De Vlaminc I, Khush KK, Strehl C, Kohli B, Luikart H, Neff NF, Okamoto J, Snyder TM, Cornfield DN, Nicolls MR, Weill D, Bernstein D, Valantine HA, Quake SR: **Temporal Response of the Human Virome to Immunosuppression and Antiviral Therapy**. *Cell* 2013, **155**:1178–1187.
2. De Vlaminc I, Valantine H a, Snyder TM, Strehl C, Cohen G, Luikart H, Neff NF, Okamoto J, Bernstein D, Weisshaar D, Quake SR, Khush KK: **Circulating cell-free DNA enables noninvasive diagnosis of heart transplant rejection**. *Sci Transl Med* 2014, **6**:241ra77.
3. De Vlaminc I, Martin L, Kertesz M, Patel K, Kowarsky M, Strehl C, Cohen G, Luikart H, Neff NF, Okamoto J, Nicolls MR, Cornfield D, Weill D, Valantine H, Khush KK, Quake SR: **Noninvasive monitoring of infection and rejection after lung transplantation**. *Proc Natl Acad Sci* 2015, **112**:13336–13341.
4. Burnham P, Kim MS, Agbor-Enoh S, Luikart H, Valantine HA, Khush KK, De Vlaminc I: **Single-stranded DNA library preparation uncovers the origin and diversity of ultrashort cell-free DNA in plasma**. *Sci Rep* 2016, **6**:27859.
5. Burnham P, Dadhania D, Heyang M, Chen F, Westblade LF, Suthanthiran M, Lee JR, De Vlaminc I: **Urinary cell-free DNA is a versatile analyte for monitoring infections of the urinary tract**. *Nat Commun* 2018, **9**:2412.
6. Blauwkamp TA, Thair S, Rosen MJ, Blair L, Lindner MS, Vilfan ID, Kawli T, Christians FC, Venkatasubrahmanyam S, Wall GD, Cheung A, Rogers ZN, Meshulam-Simon G, Huijse L, Balakrishnan S, Quinn J V, Hollemon D, Hong DK, Vaughn ML, Kertesz M, Bercovici S, Wilber JC, Yang S: **Analytical and clinical validation of a microbial cell-free DNA sequencing test for infectious disease**. *Nat Microbiol* 2019.
7. Macosko EZ, Basu A, Satija R, Nemes J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, Trombetta JJ, Weitz DA, Sanes JR, Shalek AK, Regev A, McCarroll SA: **Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets**. *Cell* 2015, **161**:1202–1214.
8. Saikia M, Burnham P, Keshavjee SH, Wang MFZ, Heyang M, Moral-Lopez P, Hinchman MM, Danko CG, Parker JSL, De Vlaminc I: **Simultaneous multiplexed amplicon sequencing and transcriptome profiling in single cells**. *Nat Methods* 2019, **16**:59–62.
9. Mandel P, Metais P: **Les acides nucléiques du plasma sanguin chez l'homme**. *C R Seances Soc Biol Fil* 1948, **142**:241–243.
10. Javillier M, Fabrykant M: **Recherches experimentales sur le phosphore sanguin et particulièrement sur variations de la phosphatemie**. *Bull Soc Chim Biol* 1931, **13**.
11. Lo YM, Corbetta N, Chamberlain PF, Rai V, Sargent IL, Redman CW, Wainscoat JS: **Presence of fetal DNA in maternal plasma and serum**. *Lancet (London, England)* 1997, **350**:485–487.
12. 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**:16266–16271.
13. Bettegowda C, Sausen M, Leary RJ, Kinde I, Wang Y, Agrawal N, Bartlett BR, Wang H, Luber B, Alani RM, Antonarakis ES, Azad NS, Bardelli A, Brem H, Cameron JL, Lee CC, Fecher LA, Gallia GL, Gibbs P, Le D, Giuntoli RL, Goggins M, Hogarty MD, Holdhoff M, Hong S-M, Jiao Y, Juhl HH, Kim JJ, Siravegna G, Laheru DA, et al.: **Detection of circulating tumor DNA in early- and late-stage human malignancies**. *Sci Transl Med* 2014, **6**:224ra24.
14. Snyder TM, Khush KK, Valantine HA, Quake SR: **Universal noninvasive detection of solid organ**

**transplant rejection.** *Proc Natl Acad Sci U S A* 2011, **108**:6229–6234.

15. Harker Rhodes C, Honsinger C, Sorenson GD: **PCR-Detection of Tumor-Derived p53 DNA in Cerebrospinal Fluid.** *Am J Clin Pathol* 1995, **103**:404–408.

16. Aucamp J, Bronkhorst AJ, Badenhorst CPS, Pretorius PJ: **The diverse origins of circulating cell-free DNA in the human body: a critical re-evaluation of the literature.** *Biol Rev* 2018, **93**:1649–1683.

17. Macher H, Egea-Guerrero JJ, Revuelto-Rey J, Gordillo-Escobar E, Enamorado-Enamorado J, Boza A, Rodriguez A, Molinero P, Guerrero JM, Dominguez-Roldán JM, Murillo-Cabezas F, Rubio A: **Role of early cell-free DNA levels decrease as a predictive marker of fatal outcome after severe traumatic brain injury.** *Clin Chim Acta* 2012, **414**:12–17.

18. Saukkonen K, Lakkisto P, Pettilä V, Varpula M, Karlsson S, Ruokonen E, Pulkki K: **Cell-Free Plasma DNA as a Predictor of Outcome in Severe Sepsis and Septic Shock.** *Clin Chem* 2008, **54**:1000 LP-1007.

19. Yao W, Mei C, Nan X, Hui L: **Evaluation and comparison of in vitro degradation kinetics of DNA in serum, urine and saliva: A qualitative study.** *Gene* 2016, **590**:142–148.

20. Gaffney DJ, McVicker G, Pai AA, Fondufe-Mittendorf YN, Lewellen N, Michelini K, Widom J, Gilad Y, Pritchard JK: **Controls of Nucleosome Positioning in the Human Genome.** *PLOS Genet* 2012, **8**:e1003036.

21. Jones S, van Heyningen P, Berman HM, Thornton JM: **Protein-DNA interactions: a structural analysis** Edited by K. Nagai. *J Mol Biol* 1999, **287**:877–896.

22. Cui F, Zhurkin VB: **Distinctive sequence patterns in metazoan and yeast nucleosomes: Implications for linker histone binding to AT-rich and methylated DNA.** *Nucleic Acids Res* 2009, **37**:2818–2829.

23. 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.

24. Ulz P, Thallinger GG, Auer M, Graf R, Kashofer K, Jahn SW, Abete L, Pristauz G, Petru E, Geigl JB, Heitzer E, Speicher MR: **Inferring expressed genes by whole-genome sequencing of plasma DNA.** *Nat Genet* 2016, **48**:1273–1278.

25. Jiang P, Chan CWM, Chan KCA, Cheng SH, Wong J, Wong VW-S, Wong GLH, Chan SL, Mok TSK, Chan HLY, Lai PBS, Chiu RWK, Lo YMD: **Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients.** *Proc Natl Acad Sci U S A* 2015, **112**:E1317-25.

26. Mouliere F, Robert B, Arnau Peyrotte E, Del Rio M, Ychou M, Molina F, Gongora C, Thierry AR: **High fragmentation characterizes tumour-derived circulating DNA.** *PLoS One* 2011, **6**:e23418.

27. Quake S: **Sizing up cell-free DNA.** *Clinical chemistry* 2012:489–490.

28. Lang BF, Gray MW, Burger G: **Mitochondrial Genome Evolution and the Origin of Eukaryotes.** *Annu Rev Genet* 1999, **33**:351–397.

29. Iborra FJ, Kimura H, Cook PR: **The functional organization of mitochondrial genomes in human cells.** *BMC Biol* 2004, **2**:9.

30. Mouliere F, Rosenfeld N: **Circulating tumor-derived DNA is shorter than somatic DNA in plasma.** *Proc Natl Acad Sci* 2015, **112**:3178–3179.

31. Gansauge M-T, Gerber T, Glocke I, Korlevic P, Lippik L, Nagel S, Riehl LM, Schmidt A, Meyer M: **Single-stranded DNA library preparation from highly degraded DNA using T4 DNA ligase.** *Nucleic Acids Res* 2017, **45**:e79.

32. Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prufer K, de Filippo C, Sudmant PH, Alkan C, Fu Q, Do R, Rohland N, Tandon A, Siebauer M, Green RE, Bryc K, Briggs AW, Stenzel U, Dabney J, Shendure J, Kitzman J, Hammer MF, Shunkov M V., Derevianko AP,

- Patterson N, Andres AM, Eichler EE, et al.: **A High-Coverage Genome Sequence from an Archaic Denisovan Individual.** *Science* 2012:222–226.
33. Lo YD, Tein MS, Pang CC, Yeung CK, Tong K-L, Hjelm NM: **Presence of donor-specific DNA in plasma of kidney and liver-transplant recipients.** *Lancet* 1998, **351**:1329–1330.
34. Miller FJ, Rosenfeldt FL, Zhang C, Linnane AW, Nagley P: **Precise determination of mitochondrial DNA copy number in human skeletal and cardiac muscle by a PCR-based assay: lack of change of copy number with age.** *Nucleic Acids Res* 2003, **31**:e61.
35. Hindson BJ, Ness KD, Masquelier DA, Belgrader P, Heredia NJ, Makarewicz AJ, Bright IJ, Lucero MY, Hiddessen AL, Legler TC, Kitano TK, Hodel MR, Petersen JF, Wyatt PW, Steenblock ER, Shah PH, Bousse LJ, Troup CB, Mellen JC, Wittmann DK, Erndt NG, Cauley TH, Koehler RT, So AP, Dube S, Rose KA, Montesclaros L, Wang S, Stumbo DP, Hodges SP, et al.: **High-Throughput Droplet Digital PCR System for Absolute Quantitation of DNA Copy Number.** *Anal Chem* 2011, **83**:8604–8610.
36. Gansauge M-T, Meyer M: **Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA.** *Nat Protoc* 2013, **8**:737–48.
37. Schwarzenbach H, Hoon DSB, Pantel K: **Cell-free nucleic acids as biomarkers in cancer patients.** *Nat Rev Cancer* 2011, **11**:426–437.
38. Tsui NBY, Jiang P, Chow KCK, Su X, Leung TY, Sun H, Chan KCA, Chiu RWK, Lo YMD: **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.
39. Lecuit M, Eloit M: **The potential of whole genome NGS for infectious disease diagnosis.** *Expert Rev Mol Diagn* 2015:1–3.
40. Wilson MR, Naccache SN, Samayoa E, Biagtan M, Bashir H, Yu G, Salamat SM, Somasekar S, Federman S, Miller S, Sokolic R, Garabedian E, Candotti F, Buckley RH, Reed KD, Meyer TL, Seroogy CM, Galloway R, Henderson SL, Gern JE, DeRisi JL, Chiu CY: **Actionable diagnosis of neuroleptospirosis by next-generation sequencing.** *N Engl J Med* 2014, **370**:2408–17.
41. Aquadro CF, Greenberg BD: **Human mitochondrial DNA variation and evolution: analysis of nucleotide sequences from seven individuals.** *Genetics* 1983, **103**:287–312.
42. Li WH, Sadler L a.: **Low nucleotide diversity in man.** *Genetics* 1991, **129**:513–523.
43. Linch CA, Whiting DA, Holland MM: **Human hair histogenesis for the mitochondrial DNA forensic scientist.** *J Forensic Sci* 2001, **46**:844–853.
44. Avise JC, Arnold J, Ball RM, Bermingham E, Lamb T, Neigel JE, Reeb C a, Saunders NC: **Intraspecific Phylogeography: The Mitochondrial DNA Bridge Between Population Genetics and Systematics.** *Annu Rev Ecol Syst* 1987, **18**:489–522.
45. Robin ED, Wong R: **Mitochondrial DNA molecules and virtual number of mitochondria per cell in mammalian cells.** *J Cell Physiol* 1988, **136**:507–13.
46. Karlsson K, Sahlin E, Iwarsson E, Westgren M, Nordenskjöld M, Linnarsson S: **Amplification-free sequencing of cell-free DNA for prenatal non-invasive diagnosis of chromosomal aberrations.** *Genomics* 2015, **105**:150–8.
47. Bolger AM, Lohse M, Usadel B: **Trimmomatic: A flexible trimmer for Illumina sequence data.** *Bioinformatics* 2014, **30**:2114–2120.
48. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25**:1754–60.
49. Gordon A, Hannon GJ: **Fastx-toolkit. FASTQ/A short-reads pre-processing tools.** *Unpubl*

[http://hannonlab.cshl.edu/fastx\\_toolkit](http://hannonlab.cshl.edu/fastx_toolkit) 2010.

50. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup 1000 Genome Project Data Processing: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25**:2078–2079.
51. Magoc T, Salzberg SL: **FLASH: fast length adjustment of short reads to improve genome assemblies.** *Bioinformatics* 2011, **27**:2957–2963.
52. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat Methods* 2012, **9**:357–359.
53. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic Local Alignment Search Tool.** *J Mol Biol* 1990:403–410.
54. Xia LC, Cram JA, Chen T, Fuhrman JA, Sun F: **Accurate genome relative abundance estimation based on shotgun metagenomic reads.** *PLoS One* 2011, **6**.
55. Zhang Q, Raoof M, Chen Y, Sumi Y, Sursal T, Junger W, Brohi K, Itagaki K, Hauser CJ: **Circulating mitochondrial DAMPs cause inflammatory responses to injury.** *Nature* 2010, **464**:104–107.
56. Oka T, Hikoso S, Yamaguchi O, Taneike M, Takeda T, Tamai T, Oyabu J, Murakawa T, Nakayama H, Nishida K, Akira S, Yamamoto A, Komuro I, Otsu K: **Mitochondrial DNA that escapes from autophagy causes inflammation and heart failure.** *Nature* 2012:292–292.
57. Dinakaran V, Rathinavel A, Pushpanathan M, Sivakumar R, Gunasekaran P, Rajendhran J: **Elevated levels of circulating DNA in cardiovascular disease patients: metagenomic profiling of microbiome in the circulation.** *PLoS One* 2014, **9**:e105221.
58. Van Der Vaart M, Pretorius PJ: **Circulating DNA: Its origin and fluctuation.** *Ann N Y Acad Sci* 2008, **1137**:18–26.
59. Foxman B: **Urinary tract infection syndromes. Occurrence, recurrence, bacteriology, risk factors, and disease burden.** *Infectious Disease Clinics of North America* 2014:1–13.
60. Abbott KC, Swanson SJ, Richter ER, Bohem EM, Agodoa LY, Peters TG, Barbour G, Lipnick R, Cruess DF: **Late urinary tract infection after renal transplantation in the United States.** *Am J Kidney Dis* 2004, **44**:353–362.
61. Ariza-Heredia EJ, Beam EN, Lesnick TG, Cosio FG, Kremers WK, Razonable RR: **Impact of urinary tract infection on allograft function after kidney transplantation.** *Clin Transplant* 2014, **28**:683–690.
62. Chuang P, Parikh CR, Langone A: **Urinary tract infections after renal transplantation: a retrospective review at two US transplant centers.** *Clin Transplant* 2005, **19**:230–235.
63. Hirsch HH, Brennan DC, Drachenberg CB, Ginevri F, Gordon J, Limaye AP, Mihatsch MJ, Nickleit V, Ramos E, Randhawa P, Shapiro R, Steiger J, Suthanthiran M, Trofe J: **Polyomavirus-associated nephropathy in renal transplantation: interdisciplinary analyses and recommendations.** *Transplantation* 2005, **79**:1277–1286.
64. Dadhania D, Snopkowski C, Ding R, Muthukumar T, Chang C, Aull M, Lee J, Sharma VK, Kapur S, Suthanthiran M: **Epidemiology of BK virus in renal allograft recipients: independent risk factors for BK virus replication.** *Transplantation* 2008, **86**:521–528.
65. J AG, Rama T, A. GS, Katherina M, James G, M. EJ, S. WM, H. SD, H. CP, Abdolreza H: **Infectious complications after kidney transplantation: current epidemiology and associated risk factors.** *Clin Transplant* 2006, **20**:401–409.
66. Schmiemann G, Kniehl E, Gebhardt K, Matejczyk MM, Hummers-Pradier E: **The Diagnosis of Urinary Tract Infection: A Systematic Review.** *Deutsches Ärzteblatt International* 2010:361–367.

67. Price TK, Dune T, Hilt EE, Thomas-White KJ, Kliethermes S, Brincat C, Brubaker L, Wolfe AJ, Mueller ER, Schreckenberger PC: **The Clinical Urine Culture: Enhanced Techniques Improve Detection of Clinically Relevant Microorganisms.** *J Clin Microbiol* 2016, **54**:1216–22.
68. Hilt EE, McKinley K, Pearce MM, Rosenfeld AB, Zilliox MJ, Mueller ER, Brubaker L, Gai X, Wolfe AJ, Schreckenberger PC: **Urine Is Not Sterile: Use of Enhanced Urine Culture Techniques To Detect Resident Bacterial Flora in the Adult Female Bladder.** *J Clin Microbiol* 2014, **52**:871–876.
69. Fan HC, Gu W, Wang J, Blumenfeld YJ, El-Sayed YY, Quake SR: **Non-invasive prenatal measurement of the fetal genome.** *Nature* 2012, **487**:320–324.
70. Botezatu I, Serdyuk O, Potapova G, Shelepov V, Alechina R, Molyaka Y, Ananev V, Bazin I, Garin A, Narimanov M, Knysh V, Melkonyan H, Umansky S, Lichtenstein A: **Genetic analysis of DNA excreted in urine: a new approach for detecting specific genomic DNA sequences from cells dying in an organism.** *Clin Chem* 2000, **46**(8 Pt 1):1078–1084.
71. Zhang J, Tong KL, Li PK, Chan AY, Yeung CK, Pang CC, Wong TY, Lee KC, Lo YM: **Presence of donor- and recipient-derived DNA in cell-free urine samples of renal transplantation recipients: urinary DNA chimerism.** *Clin Chem* 1999, **45**:1741–1746.
72. Su Y-H, Wang M, Brenner DE, Ng A, Melkonyan H, Umansky S, Syngal S, Block TM: **Human urine contains small, 150 to 250 nucleotide-sized, soluble DNA derived from the circulation and may be useful in the detection of colorectal cancer.** *J Mol Diagn* 2004, **6**:101–107.
73. Lo YMD, Chan KCA, Sun H, Chen EZ, Jiang P, Lun FMF, Zheng YW, Leung TY, Lau TK, Cantor CR, Chiu RWK: **Maternal Plasma DNA Sequencing Reveals the Genome-Wide Genetic and Mutational Profile of the Fetus.** *Sci Transl Med* 2010, **2**:61ra91 LP-61ra91.
74. Varshavsky AJ, Bakayev V V, Chumackov PM, Georgiev GP: **Minichromosome of simian virus 40: presence of histone H1.** *Nucleic Acids Research* 1976:2101–2113.
75. Ding R, Li B, Muthukumar T, Dadhania D, Medeiros M, Hartono C, Serur D, Seshan S V, Sharma VK, Kapur S, Suthanthiran M: **CD103 mRNA levels in urinary cells predict acute rejection of renal allografts.** *Transplantation* 2003, **75**:1307–1312.
76. Dadhania D, Snopkowski C, Ding R, Muthukumar T, Lee J, Bang H, Sharma VK, Seshan S, August P, Kapur S, Suthanthiran M: **Validation of Noninvasive Diagnosis of BK Virus Nephropathy and Identification of Prognostic Biomarkers.** *Transplantation* 2010, **90**:189–197.
77. JM R: **Haemophilus influenzae pyelonephritis in adults.** *Arch Intern Med* 1999, **159**:316.
78. Wolfe AJ, Brubaker L: **“Sterile Urine” and the Presence of Bacteria.** *Eur Urol* 2015, **68**:173–174.
79. Wolfe AJ, Toh E, Shibata N, Rong R, Kenton K, Fitzgerald M, Mueller ER, Schreckenberger P, Dong Q, Nelson DE, Brubaker L: **Evidence of uncultivated bacteria in the adult female bladder.** *J Clin Microbiol* 2012, **50**:1376–1383.
80. Chaban B, Links MG, Jayaprakash TP, Wagner EC, Bourque DK, Lohn Z, Albert AY, van Schalkwyk J, Reid G, Hemmingsen SM, Hill JE, Money DM: **Characterization of the vaginal microbiota of healthy Canadian women through the menstrual cycle.** *Microbiome* 2014, **2**:23.
81. Knowles WA: **The Epidemiology of BK Virus and the Occurrence of Antigenic and Genomic Subtypes.** *Human Polyomaviruses* 2001. [Wiley Online Books]
82. Morel V, Martin E, François C, Helle F, Faucher J, Mourez T, Choukroun G, Duverlie G, Castelain S, Brochet E: **A Simple and Reliable Strategy for BK Virus Subtyping and Subgrouping.** *J Clin Microbiol* 2017, **55**:1177 LP-1185.
83. Bohl DL, Storch GA, Ryschkewitsch C, Gaudreault-Keener M, Schnitzler MA, Major EO, Brennan DC: **Donor Origin of BK Virus in Renal Transplantation and Role of HLA C7 in Susceptibility to**

**Sustained BK Viremia.** *Am J Transplant* 2005, **5**:2213–2221.

84. Zhong S, Randhawa PS, Ikegaya H, Chen Q, Zheng H-Y, Suzuki M, Takeuchi T, Shibuya A, Kitamura T, Yogo Y: **Distribution patterns of BK polyomavirus (BKV) subtypes and subgroups in American, European and Asian populations suggest co-migration of BKV and the human race.** *J Gen Virol* 2009, **90**(Pt 1):144–152.

85. Drummond AJ, Rambaut A: **BEAST: Bayesian evolutionary analysis by sampling trees.** *BMC Evol Biol* 2007, **7**:214.

86. Domingo-Calap P, Schubert B, Joly M, Solis M, Untrau M, Carapito R, Georgel P, Caillard S, Fafi-Kremer S, Paul N, Kohlbacher O, González-Candelas F, Bahram S: **An unusually high substitution rate in transplant-associated BK polyomavirus in vivo is further concentrated in HLA-C-bound viral peptides.** *PLOS Pathog* 2018, **14**:e1007368.

87. Korem T, Zeevi D, Suez J, Weinberger A, Avnit-Sagi T, Pompan-Lotan M, Matot E, Jona G, Harmelin A, Cohen N, Sirota-Madi A, Thaïss CA, Pevsner-Fischer M, Sorek R, Xavier RJ, Elinav E, Segal E: **Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples.** *Science* 2015, **349**:1101–1106.

88. Brown CT, Olm MR, Thomas BC, Banfield JF: **Measurement of bacterial replication rates in microbial communities.** *Nat Biotech* 2016, **34**:1256–1263.

89. Møllerup S, Friis-Nielsen J, Vinner L, Hansen TA, Richter SR, Fridholm H, Herrera JAR, Lund O, Brunak S, Izarzugaza JMG, Mourier T, Nielsen LP, Hansen AJ: **Propionibacterium acnes: Disease-Causing Agent or Common Contaminant? Detection in Diverse Patient Samples by Next-Generation Sequencing.** *J Clin Microbiol* 2016, **54**:980–987.

90. McArthur AG, Waglechner N, Nizam F, Yan A, Azad MA, Baylay AJ, Bhullar K, Canova MJ, De Pascale G, Ejim L, Kalan L, King AM, Koteva K, Morar M, Mulvey MR, O'Brien JS, Pawlowski AC, Piddock LJ V, Spanogiannopoulos P, Sutherland AD, Tang I, Taylor PL, Thaker M, Wang W, Yan M, Yu T, Wright GD: **The Comprehensive Antibiotic Resistance Database.** *Antimicrobial Agents and Chemotherapy* 2013:3348–3357.

91. Grskovic M, Hiller DJ, Eubank LA, Sninsky JJ, Christopherson C, Collins JP, Thompson K, Song M, Wang YS, Ross D, Nelles MJ, Yee JP, Wilber JC, Crespo-Leiro MG, Scott SL, Woodward RN: **Validation of a Clinical-Grade Assay to Measure Donor-Derived Cell-Free DNA in Solid Organ Transplant Recipients.** *J Mol Diagn* 2016, **18**:890–902.

92. Smith RM: **Urinary Infection in Children.** *N Engl J Med* 1931, **205**:181–185.

93. Eirin A, Saad A, Tang H, Herrmann SM, Woollard JR, Lerman A, Textor SC, Lerman LO: **Urinary Mitochondrial DNA Copy Number Identifies Chronic Renal Injury in Hypertensive Patients.** *Hypertens (Dallas, Tex 1979)* 2016, **68**:401–410.

94. Lood C, Blanco LP, Purmalek MM, Carmona-Rivera C, De Ravin SS, Smith CK, Malech HL, Ledbetter JA, Elkon KB, Kaplan MJ: **Neutrophil extracellular traps enriched in oxidized mitochondrial DNA are interferogenic and contribute to lupus-like disease.** *Nat Med* 2016, **22**:146–153.

95. Sharon E, Shi H, Kharbanda S, Koh W, Martin LR, Khush KK, Valantine H, Pritchard JK, De Vlaminck I: **Quantification of transplant-derived circulating cell-free DNA in absence of a donor genotype.** *PLOS Comput Biol* 2017, **13**:e1005629.

96. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra MA, Beaudet AL, Ecker JR, Farnham PJ, Hirst M, Lander ES, Mikkelsen TS, Thomson JA: **The NIH Roadmap Epigenomics Mapping Consortium.** *Nat Biotechnol* 2010, **28**:1045.

97. Fernandez AF, Assenov Y, Martin-Subero JI, Balint B, Siebert R, Taniguchi H, Yamamoto H, Hidalgo

- M, Tan A-C, Galm O, Ferrer I, Sanchez-Cespedes M, Villanueva A, Carmona J, Sanchez-Mut J V, Berdasco M, Moreno V, Capella G, Monk D, Ballestar E, Ropero S, Martinez R, Sanchez-Carbayo M, Prosper F, Agirre X, Fraga MF, Graña O, Perez-Jurado L, Mora J, Puig S, et al.: **A DNA methylation fingerprint of 1628 human samples.** *Genome Res* 2012, **22**:407–419.
98. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR: **Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis.** *Cell* 2008, **133**:523–536.
99. Jühling F, Kretzmer H, Bernhart SH, Otto C, Stadler PF, Hoffmann S: **metilene: fast and sensitive calling of differentially methylated regions from bisulfite sequencing data.** *Genome Res* 2016, **26**:256–262.
100. Cheng THT, Jiang P, Tam JCW, Sun X, Lee W-S, Yu SCY, Teoh JYC, Chiu PKF, Ng C-F, Chow K-M, Szeto C-C, Chan KCA, Chiu RWK, Lo YMD: **Genomewide bisulfite sequencing reveals the origin and time-dependent fragmentation of urinary cfDNA.** *Clin Biochem* 2017.
101. Lehmann-Werman R, Neiman D, Zemmour H, Moss J, Magenheim J, Vaknin-Dembinsky A, Rubertsson S, Nellgård B, Blennow K, Zetterberg H, Spalding K, Haller MJ, Wasserfall CH, Schatz DA, Greenbaum CJ, Dorrell C, Grompe M, Zick A, Hubert A, Maoz M, Fendrich V, Bartsch DK, Golan T, Ben Sasson SA, Zamir G, Razin A, Cedar H, Shapiro AMJ, Glaser B, Shemer R, et al.: **Identification of tissue-specific cell death using methylation patterns of circulating DNA.** *Proc Natl Acad Sci* 2016, **113**:E1826–E1834.
102. Tanaka K, Okamoto A: **Degradation of DNA by bisulfite treatment.** *Bioorg Med Chem Lett* 2007, **17**:1912–1915.
103. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, Wiencke JK, Kelsey KT: **DNA methylation arrays as surrogate measures of cell mixture distribution.** *BMC Bioinformatics* 2012, **13**:86.
104. CLSI: *Performance Standards for Antimicrobial Susceptibility Testing. 27th Ed. CLSI Supplement M100-27.* Wayne, PA; 2017.
105. CLSI: *Performance Standards for Antimicrobial Disk Susceptibility Tests; Approved Standard. 12th Ed. CLSI Supplement M02-A12.* Wayne, PA; 2015.
106. CLSI: *Performance Standards for Antimicrobial Susceptibility Testing. 25th Ed. CLSI Supplement M100-25.* Wayne, PA; 2015.
107. CLSI: *Performance Standards for Antimicrobial Susceptibility Testing. 26th Ed. CLSI Supplement M100-26.* Wayne, PA; 2016.
108. Sinha R, Stanley G, Gulati GS, Ezran C, Travaglini KJ, Wei E, Chan CKF, Nabhan AN, Su T, Morganti RM, Conley SD, Chaib H, Red-Horse K, Longaker MT, Snyder MP, Krasnow MA, Weissman IL: **Index Switching Causes “Spreading-Of-Signal” Among Multiplexed Samples In Illumina HiSeq 4000 DNA Sequencing.** *bioRxiv* 2017.
109. Kent WJ: **BLAT--the BLAST-like alignment tool.** *Genome Res* 2002, **12**:656–664.
110. Ha G, Roth A, Lai D, Bashashati A, Ding J, Goya R, Giuliany R, Rosner J, Oloumi A, Shumansky K, Chin S-F, Turashvili G, Hirst M, Caldas C, Marra MA, Aparicio S, Shah SP: **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.
111. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL: **BLAST+: architecture and applications.** *BMC Bioinformatics* 2009, **10**:421.
112. Pedersen BS, Eyring K, De S, Yang I V., Schwartz DA: **Fast and accurate alignment of long bisulfite-seq reads.** 2014.

113. Fouts DE, Pieper R, Szpakowski S, Pohl H, Knoblach S, Suh M-J, Huang S-T, Ljungberg I, Sprague BM, Lucas SK, Torralba M, Nelson KE, Groah SL: **Integrated next-generation sequencing of 16S rDNA and metaproteomics differentiate the healthy urine microbiome from asymptomatic bacteriuria in neuropathic bladder associated with spinal cord injury.** *J Transl Med* 2012, **10**:174.
114. Nelson DE, Van Der Pol B, Dong Q, Revanna K V, Fan B, Easwaran S, Sodergren E, Weinstock GM, Diao L, Fortenberry JD: **Characteristic male urine microbiomes associate with asymptomatic sexually transmitted infection.** *PLoS One* 2010, **5**:e14116.
115. Thomas-White KJ, Hilt EE, Fok C, Pearce MM, Mueller ER, Kliethermes S, Jacobs K, Zilliox MJ, Brincat C, Price TK, Kuffel G, Schreckenberger P, Gai X, Brubaker L, Wolfe AJ: **Incontinence medication response relates to the female urinary microbiota.** *Int Urogynecol J* 2016, **27**:723–733.
116. Hart A, Smith JM, Skeans MA, Gustafson SK, Stewart DE, Cherikh WS, Wainright JL, Kucheryavaya A, Woodbury M, Snyder JJ, Kasiske BL, Israni AK: **OPTN/SRTR 2015 Annual Data Report: Kidney.** *Am J Transplant* 2017, **17**:21–116.
117. Turnbaugh PJ, Hamady M, Yatsunencko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP, Egholm M, Henrissat B, Heath AC, Knight R, Gordon JI: **A core gut microbiome in obese and lean twins.** *Nature* 2008, **457**:480.
118. Tilg H, Kaser A: **Gut microbiome, obesity, and metabolic dysfunction.** *J Clin Invest* 2011, **121**:2126–2132.
119. Blaser MJ: **Antibiotic use and its consequences for the normal microbiome.** *Science (80- )* 2016, **352**:544 LP-545.
120. 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.
121. Lee JR, Muthukumar T, Dadhania D, Toussaint NC, Ling L, Pamer E, Suthanthiran M: **Gut microbial community structure and complications after kidney transplantation: a pilot study.** *Transplantation* 2014, **98**:697–705.
122. Naccache SN, Greninger AL, Lee D, Coffey LL, Phan T, Rein-Weston A, Aronsohn A, Hackett J, Delwart EL, Chiu CY: **The Perils of Pathogen Discovery: Origin of a Novel Parvovirus-Like Hybrid Genome Traced to Nucleic Acid Extraction Spin Columns.** *J Virol* 2013, **87**:11966 LP-11977.
123. Xu B, Zhi N, Hu G, Wan Z, Zheng X, Liu X, Wong S, Kajigaya S, Zhao K, Mao Q, Young NS: **Hybrid DNA virus in Chinese patients with seronegative hepatitis discovered by deep sequencing.** *Proc Natl Acad Sci* 2013, **110**:10264 LP-10269.
124. Nagasaki K: **Dinoflagellates, diatoms, and their viruses.** *J Microbiol* 2008, **46**:235–243.
125. Brenner DJ, Fanning GR, Steigerwalt AG, Ørskov I, Ørskov F: **Polynucleotide Sequence Relatedness Among Three Groups of Pathogenic & Escherichia coli Strains.** *Infect Immun* 1972, **6**:308 LP-315.
126. Tamames J, Moya A: **Estimating the extent of horizontal gene transfer in metagenomic sequences.** *BMC Genomics* 2008, **9**:136.
127. Bendel RB, Higgins SS, Teberg JE, Pyke DA: **Comparison of Skewness Coefficient, Coefficient of Variation, and Gini Coefficient as Inequality Measures within Populations.** *Oecologia* 1989, **78**:394–400.
128. Box GEP, Cox DR: **An Analysis of Transformations.** *J R Stat Soc Ser B* 1964, **26**:211–252.
129. de Goffau MC, Lager S, Salter SJ, Wagner J, Kronbichler A, Charnock-Jones DS, Peacock SJ, Smith GCS, Parkhill J: **Recognizing the reagent microbiome.** *Nat Microbiol* 2018, **3**:851–853.

130. 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**:105–117.
131. Kim CJ, Chaemsaitong P, Chaiyasit N, Yoon BH, Kim YM: **Acute chorioamnionitis and funisitis: definition, pathologic features, and clinical significance.** *Am J Obstet Gynecol* 2015, **213**:S29–S52.
132. Gibbs RS, Duff P: **Progress in pathogenesis and management of clinical intraamniotic infection.** *Am J Obstet Gynecol* 1991, **164**:1317–1326.
133. Perez-Muñoz ME, Arrieta M-C, Ramer-Tait AE, Walter J: **A critical assessment of the “sterile womb” and “in utero colonization” hypotheses: implications for research on the pioneer infant microbiome.** *Microbiome* 2017, **5**:48.
134. Leiby JS, McCormick K, Sherrill-Mix S, Clarke EL, Kessler LR, Taylor LJ, Hofstaedter CE, Roche AM, Mattei LM, Bittinger K, Elovitz MA, Leite R, Parry S, Bushman FD: **Lack of detection of a human placenta microbiome in samples from preterm and term deliveries.** *Microbiome* 2018, **6**:196.
135. Park JW, Park KH, Jung EY: **Clinical significance of histologic chorioamnionitis with a negative amniotic fluid culture in patients with preterm labor and premature membrane rupture.** *PLoS One* 2017, **12**:e0173312.
136. Gomez-Lopez N, Romero R, Xu Y, Leng Y, Garcia-Flores V, Miller D, Jacques SM, Hassan SS, Faro J, Alsamsam A, Alhousseini A, Gomez-Roberts H, Panaitescu B, Yeo L, Maymon E: **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.
137. Gomez-Lopez N, Romero R, Xu Y, Miller D, Leng Y, Panaitescu B, Silva P, Faro J, Alhousseini A, Gill N, Hassan SS, Hsu C-D: **The immunophenotype of amniotic fluid leukocytes in normal and complicated pregnancies.** *Am J Reprod Immunol* 2018, **79**:e12827.
138. Popovich RP, Moncrief JW, Nolph KD, Ghods AJ, Twardowski ZJ, Pyle WK: **Continuous Ambulatory Peritoneal Dialysis.** *Ann Intern Med* 1978, **88**:449–456.
139. Fenton SSA, Schaubel DE, Desmeules M, Morrison HI, Mao Y, Copleston P, Jeffery JR, Kjellstrand CM: **Hemodialysis versus peritoneal dialysis: A comparison of adjusted mortality rates.** *Am J Kidney Dis* 1997, **30**:334–342.
140. Jain AK, Blake P, Cordy P, Garg AX: **Global Trends in Rates of Peritoneal Dialysis.** *J Am Soc Nephrol* 2012, **23**:533 LP-544.
141. Li PK-T, Szeto CC, Piraino B, Bernardini J, Figueiredo AE, Gupta A, Johnson DW, Kuijper EJ, Lye W-C, Salzer W, Schaefer F, Struijk DG: **PERITONEAL DIALYSIS-RELATED INFECTIONS RECOMMENDATIONS: 2010 UPDATE.** *Perit Dial Int* 2010, **30**:393–423.
142. Troeger C, Forouzanfar M, Rao PC, Khalil I, Brown A, Swartz S, Fullman N, Mosser J, Thompson RL, Reiner Jr RC, Abajobir A, Alam N, Alemayohu MA, Amare AT, Antonio CA, Asayesh H, Avokpaho E, Barac A, Beshir MA, Boneya DJ, Brauer M, Dandona L, Dandona R, Fitchett JRA, Gebrehiwot TT, Hailu GB, Hotez PJ, Kasaeian A, Khoja T, Kissoon N, et al.: **Estimates of the global, regional, and national morbidity, mortality, and aetiologies of lower respiratory tract infections in 195 countries: a systematic analysis for the Global Burden of Disease Study 2015.** *Lancet Infect Dis* 2017, **17**:1133–1161.
143. Grubaugh ND, Ladner JT, Lemey P, Pybus OG, Rambaut A, Holmes EC, Andersen KG: **Tracking virus outbreaks in the twenty-first century.** *Nat Microbiol* 2019, **4**:10–19.
144. Korpe PS, Petri WA: **Environmental enteropathy: Critical implications of a poorly understood condition.** *Trends in Molecular Medicine* 2012:328–336.

145. Glaziou P, Sismanidis C, Floyd K, Raviglione M: **Global Epidemiology of Tuberculosis**. *Cold Spring Harb Perspect Med* 2015, **5**.
146. Lin PL, Flynn JL: **Understanding Latent Tuberculosis: A Moving Target**. *J Immunol* 2010, **185**:15 LP-22.
147. Takayama K, Kilburn JO: **Inhibition of synthesis of arabinogalactan by ethambutol in Mycobacterium smegmatis**. *Antimicrob Agents Chemother* 1989, **33**:1493–1499.
148. Lin PL, Ford CB, Coleman MT, Myers AJ, Gawande R, Ioerger T, Sacchettini J, Fortune SM, Flynn JL: **Sterilization of granulomas is common in active and latent tuberculosis despite within-host variability in bacterial killing**. *Nat Med* 2013, **20**:75.
149. Green C, Huggett JF, Talbot E, Mwaba P, Reither K, Zumla AI: **Rapid diagnosis of tuberculosis through the detection of mycobacterial DNA in urine by nucleic acid amplification methods**. *Lancet Infect Dis* 2009, **9**:505–511.
150. Gill WP, Harik NS, Whiddon MR, Liao RP, Mittler JE, Sherman DR: **A replication clock for Mycobacterium tuberculosis**. *Nat Med* 2009, **15**:211.
151. Dancy JT, Deubelbeiss KA, Harker LA, Finch CA: **Neutrophil kinetics in man**. *J Clin Invest* 1976, **58**:705–715.
152. Harper KM, Mutasa M, Prendergast AJ, Humphrey J, Manges AR: **Environmental enteric dysfunction pathways and child stunting: A systematic review**. *PLoS Negl Trop Dis* 2018, **12**:e0006205.
153. van Elburg RM, Uil JJ, Kokke FT, Mulder AM, van de Broek WG, Mulder CJ, Heymans HS: **Repeatability of the sugar-absorption test, using lactulose and mannitol, for measuring intestinal permeability for sugars**. *J Pediatr Gastroenterol Nutr* 1995, **20**:184–188.
154. McInnes L, Healy J, Melville J: **UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction**. 2018.
155. Hupal DN, Bradic M, Carlton JM: **The impact of genomics on population genetics of parasitic diseases**. *Curr Opin Microbiol* 2015, **23**:49–54.
156. Casadevall A, Pirofski LA: **Host-pathogen interactions: basic concepts of microbial commensalism, colonization, infection, and disease**. *Infect Immun* 2000, **68**:6511–6518.
157. Gootenberg JS, Abudayyeh OO, Lee JW, Essletzbichler P, Dy AJ, Joung J, Verdine V, Donghia N, Daringer NM, Freije CA, Myhrvold C, Bhattacharyya RP, Livny J, Regev A, Koonin E V, Hung DT, Sabeti PC, Collins JJ, Zhang F: **Nucleic acid detection with CRISPR-Cas13a/C2c2**. *Science (80- )* 2017, **356**:438 LP-442.
158. Shapiro E, Biezuner T, Linnarsson S: **Single-cell sequencing-based technologies will revolutionize whole-organism science**. *Nat Rev Genet* 2013, **14**:618.
159. Consortium T 1000 GP, Durbin RM, Altshuler (Co-Chair) D, Durbin (Co-Chair) RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Collins FS, De La Vega FM, Donnelly P, Egholm M, Flicek P, Gabriel SB, Gibbs RA, Knoppers BM, Lander ES, Lehrach H, Mardis ER, McVean GA, Nickerson DA, Peltonen L, Schafer AJ, Sherry ST, Wang J, Wilson RK, Gibbs (Principal Investigator) RA, Deiros D, Metzker M, Muzny D, et al.: **A map of human genome variation from population-scale sequencing**. *Nature* 2010, **467**:1061.
160. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, Lao K, Surani MA: **mRNA-Seq whole-transcriptome analysis of a single cell**. *Nat Methods* 2009, **6**:377.
161. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, Cook K, Stepansky A, Levy D, Esposito D, Muthuswamy L, Krasnitz A, McCombie WR, Hicks J, Wigler M: **Tumour evolution inferred by single-cell sequencing**. *Nature* 2011, **472**:90.

162. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, Kirschner MW: **Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells.** *Cell* 2015, **161**:1187–1201.
163. Plasschaert LW, Žilionis R, Choo-Wing R, Savova V, Knehr J, Roma G, Klein AM, Jaffe AB: **A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte.** *Nature* 2018, **560**:377–381.
164. Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner HA, Trapnell C: **Reversed graph embedding resolves complex single-cell trajectories.** *Nat Methods* 2017, **14**:979.
165. La Manno G, Soldatov R, Zeisel A, Braun E, Hochgerner H, Petukhov V, Lidschreiber K, Kastrioti ME, Lönnerberg P, Furlan A, Fan J, Borm LE, Liu Z, van Bruggen D, Guo J, He X, Barker R, Sundström E, Castelo-Branco G, Cramer P, Adameyko I, Linnarsson S, Kharchenko P V: **RNA velocity of single cells.** *Nature* 2018, **560**:494–498.
166. Cao J, Packer JS, Ramani V, Cusanovich DA, Huynh C, Daza R, Qiu X, Lee C, Furlan SN, Steemers FJ, Adey A, Waterston RH, Trapnell C, Shendure J: **Comprehensive single-cell transcriptional profiling of a multicellular organism.** *Science (80- )* 2017, **357**:661 LP-667.
167. Edmonds M, Vaughan MH, Nakazato H: **Polyadenylic Acid Sequences in the Heterogeneous Nuclear RNA and Rapidly-Labeled Polyribosomal RNA of HeLa Cells: Possible Evidence for a Precursor Relationship.** *Proc Natl Acad Sci* 1971, **68**:1336 LP-1340.
168. Sarkar N: **Polyadenylation of mRNA in prokaryotes.** *Annu Rev Biochem* 1997, **66**:173–197.
169. Chizhikov V, Patton JT: **A four-nucleotide translation enhancer in the 3'-terminal consensus sequence of the nonpolyadenylated mRNAs of rotavirus.** *RNA* 2000, **6**:814–825.
170. Satija R, Farrell JA, Gennert D, Schier AF, Regev A: **Spatial reconstruction of single-cell gene expression data.** *Nat Biotechnol* 2015, **33**:495–502.
171. Biron CA, Byron KS, Sullivan JL: **Severe Herpesvirus Infections in an Adolescent without Natural Killer Cells.** *N Engl J Med* 1989, **320**:1731–1735.
172. Shakya AK, O'Callaghan DJ, Kim SK: **Comparative Genomic Sequencing and Pathogenic Properties of Equine Herpesvirus 1 KyA and RacL11 .** *Frontiers in Veterinary Science* 2017:211.
173. Torre E, Dueck H, Shaffer S, Gospocic J, Gupte R, Bonasio R, Kim J, Murray J, Raj A: **Rare Cell Detection by Single-Cell RNA Sequencing as Guided by Single-Molecule RNA FISH.** *Cell Syst* 2018, **6**:171–179.e5.
174. Zanini F, Pu S-Y, Bekerman E, Einav S, Quake SR: **Single-cell transcriptional dynamics of flavivirus infection.** *Elife* 2018, **7**:e32942.
175. Zanini F, Robinson ML, Croote D, Sahoo MK, Sanz AM, Ortiz-Lasso E, Albornoz LL, Rosso F, Montoya JG, Goo L, Pinsky BA, Quake SR, Einav S: **Virus-inclusive single-cell RNA sequencing reveals the molecular signature of progression to severe dengue.** *Proc Natl Acad Sci* 2018, **115**:E12363 LP-E12369.
176. Lowen AC: **It's in the mix: Reassortment of segmented viral genomes.** *PLOS Pathog* 2018, **14**:e1007200.
177. Rambaut A, Pybus OG, Nelson MI, Viboud C, Taubenberger JK, Holmes EC: **The genomic and epidemiological dynamics of human influenza A virus.** *Nature* 2008, **453**:615.
178. Zhou NN, Senne DA, Landgraf JS, Swenson SL, Erickson G, Rossow K, Liu L, Yoon K, Krauss S, Webster RG: **Genetic Reassortment of Avian, Swine, and Human Influenza A Viruses in American Pigs.** *J Virol* 1999, **73**:8851 LP-8856.

179. *Reoviruses: Entry, Assembly and Morphogenesis*. 2006.
180. Gawad C, Koh W, Quake SR: **Single-cell genome sequencing: current state of the science**. *Nat Rev Genet* 2016, **17**:175–88.
181. Gierahn TM, Wadsworth MH, Hughes TK, Bryson BD, Butler A, Satija R, Fortune S, Love JC, Shalek AK: **Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput**. *Nat Methods* 2017, **14**:395–398.
182. Xu JL, Davis MM: **Diversity in the CDR3 region of V(H) is sufficient for most antibody specificities**. *Immunity* 2000, **13**:37–45.
183. Dolan PT, Whitfield ZJ, Andino R: **Mapping the Evolutionary Potential of RNA Viruses**. *Cell Host Microbe* 2018, **23**:435–446.
184. Patton JT, Spencer E: **Genome replication and packaging of segmented double-stranded RNA viruses**. *Virology* 2000, **277**:217–25.
185. Joklik WK: **Structure and function of the reovirus genome**. *Microbiol Rev* 1981, **45**:483–501.
186. Niavarani A, Currie E, Rey Y, Anjos-Afonso F, Horswell S, Griessinger E, Luis Sardina J, Bonnet D: **APOBEC3A is implicated in a novel class of G-to-A mRNA editing in WT1 transcripts**. *PLoS One* 2015, **10**:e0120089.
187. Harris RS, Dudley JP: **APOBECs and virus restriction**. *Virology* 2015, **479–480**:131–145.
188. Parker JSL, Broering TJ, Kim J, Higgins DE, Nibert ML: **Reovirus core protein mu2 determines the filamentous morphology of viral inclusion bodies by interacting with and stabilizing microtubules**. *J Virol* 2002, **76**:4483–4496.
189. Ooms LS, Jerome WG, Dermody TS, Chappell JD: **Reovirus replication protein mu2 influences cell tropism by promoting particle assembly within viral inclusions**. *J Virol* 2012, **86**:10979–10987.
190. van der Maaten L, Hinton GE: **Visualizing data using t-SNE**. *J Mach Learn* 2008, **9**:2579–2605.
191. Georgiou G, Ippolito GC, Beausang J, Busse CE, Wardemann H, Quake SR: **The promise and challenge of high-throughput sequencing of the antibody repertoire**. *Nat Biotechnol* 2014, **32**:158–68.
192. DeKosky BJ, Kojima T, Rodin A, Charab W, Ippolito GC, Ellington AD, Georgiou G: **In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire**. *Nat Med* 2015, **21**:86–91.
193. Vollmers C, Sit R V, Weinstein JA, Dekker CL, Quake SR: **Genetic measurement of memory B-cell recall using antibody repertoire sequencing**. *Proc Natl Acad Sci U S A* 2013, **110**:13463–13468.
194. Tedder TF, Streuli M, Schlossman SF, Saito H: **Isolation and structure of a cDNA encoding the B1 (CD20) cell-surface antigen of human B lymphocytes**. *Proc Natl Acad Sci U S A* 1988, **85**:208–12.
195. Bolotin DA, Poslavsky S, Mitrophanov I, Shugay M, Mamedov IZ, Putintseva E V, Chudakov DM: **MiXCR: software for comprehensive adaptive immunity profiling**. *Nat Methods* 2015, **12**:380–1.
196. Kaminski D, Wei C, Qian Y, Rosenberg A, Sanz I: **Advances in Human B Cell Phenotypic Profiling**. *Frontiers in Immunology* 2012:302.
197. Smith K, Shah H, Muther JJ, Duke AL, Haley K, James JA: **Antigen nature and complexity influence human antibody light chain usage and specificity**. *Vaccine* 2016, **34**:2813–2820.
198. Abe M, Goto T, Kosaka M, Wolfenbarger D, Weiss DT, Solomon A: **Differences in kappa to lambda ( $\kappa:\lambda$ ) ratios of serum and urinary free light chains**. *Clin Exp Immunol* 1998, **111**:457–462.
199. Barandun S: **Immunsustitution BT - 84. Kongreß**. Edited by Schlegel B. Munich: J.F. Bergmann-Verlag; 1978:481–490.

200. Kugelberg E: **Making sense in humans.** *Nat Rev Immunol* 2015, **15**:133.
201. Mroczek ES, Ippolito GC, Rogosch T, Hoi KH, Hwangpo TA, Brand MG, Zhuang Y, Liu CR, Schneider DA, Zemlin M, Brown EE, Georgiou G, Schroeder HW: **Differences in the composition of the human antibody repertoire by B cell subsets in the blood.** *Front Immunol* 2014, **5**:96.
202. DeKosky BJ, Lungu OI, Park D, Johnson EL, Charab W, Chrysostomou C, Kuroda D, Ellington AD, Ippolito GC, Gray JJ, Georgiou G: **Large-scale sequence and structural comparisons of human naive and antigen-experienced antibody repertoires.** *Proc Natl Acad Sci* 2016, **113**:E2636 LP-E2645.
203. DeKosky BJ, Ippolito GC, Deschner RP, Lavinder JJ, Wine Y, Rawlings BM, Varadarajan N, Giesecke C, Dörner T, Andrews SF, Wilson PC, Hunicke-Smith SP, Willson CG, Ellington AD, Georgiou G: **High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire.** *Nat Biotechnol* 2013, **31**:166–9.
204. Kobayashi T, Antar AAR, Boehme KW, Danthi P, Eby EA, Guglielmi KM, Holm GH, Johnson EM, Maginnis MS, Naik S, Skelton WB, Wetzell JD, Wilson GJ, Chappell JD, Dermody TS: **A plasmid-based reverse genetics system for animal double-stranded RNA viruses.** *Cell Host Microbe* 2007, **1**:147–57.
205. Bryce J, Boschi-Pinto C, Shibuya K, Black RE: **WHO estimates of the causes of death in children.** *Lancet* 2005, **365**:1147–1152.
206. Greenberg HB, Estes MK: **Rotaviruses: From Pathogenesis to Vaccination.** *Gastroenterology* 2009, **136**:1939–1951.
207. Vignuzzi M, Stone JK, Arnold JJ, Cameron CE, Andino R: **Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population.** *Nature* 2005, **439**:344.
208. Steurman Y, Cohen M, Peshes-Yaloz N, Valadarsky L, Cohn O, David E, Frishberg A, Mayo L, Bacharach E, Amit I, Gat-Viks I: **Dissection of Influenza Infection In Vivo by Single-Cell RNA Sequencing.** *Cell Syst* 2018, **6**:679–691.e4.
209. Shalek AK, Satija R, Shuga J, Trombetta JJ, Gennert D, Lu D, Chen P, Gertner RS, Gaublomme JT, Yosef N, Schwartz S, Fowler B, Weaver S, Wang J, Wang X, Ding R, Raychowdhury R, Friedman N, Hacohen N, Park H, May AP, Regev A: **Single-cell RNA-seq reveals dynamic paracrine control of cellular variation.** *Nature* 2014, **510**:363.
210. Sato T, Vries RG, Snippert HJ, van de Wetering M, Barker N, Stange DE, van Es JH, Abo A, Kujala P, Peters PJ, Clevers H: **Single Lgr5 stem cells build crypt–villus structures in vitro without a mesenchymal niche.** *Nature* 2009, **459**:262.
211. Warren WC, Jasinska AJ, García-Pérez R, Svoldal H, Tomlinson C, Rocchi M, Archidiacono N, Capozzi O, Minx P, Montague MJ, Kyung K, Hillier LW, Kremitzki M, Graves T, Chiang C, Hughes J, Tran N, Huang Y, Ramensky V, Choi O-W, Jung YJ, Schmitt CA, Juretic N, Wasserscheid J, Turner TR, Wiseman RW, Tuscher JJ, Karl JA, Schmitz JE, Zahn R, et al.: **The genome of the vervet (*Chlorocebus aethiops sabaues*).** *Genome Res* 2015, **25**:1921–1933.
212. Vende P, Piron M, Castagné N, Poncet D: **Efficient Translation of Rotavirus mRNA Requires Simultaneous Interaction of NSP3 with the Eukaryotic Translation Initiation Factor eIF4G and the mRNA 3' End.** *J Virol* 2000, **74**:7064 LP-7071.
213. Piron M, Vende P, Cohen J, Poncet D: **Rotavirus RNA-binding protein NSP3 interacts with eIF4GI and evicts the poly(A) binding protein from eIF4F.** *EMBO J* 1998, **17**:5811 LP-5821.
214. Harb M, Becker MM, Vitour D, Baron CH, Vende P, Brown SC, Bolte S, Arold ST, Poncet D: **Nuclear Localization of Cytoplasmic Poly(A)-Binding Protein upon Rotavirus Infection Involves the Interaction of NSP3 with eIF4G and RoXaN.** *J Virol* 2008, **82**:11283 LP-11293.
215. Kanai Y, Komoto S, Kawagishi T, Nouda R, Nagasawa N, Onishi M, Matsuura Y, Taniguchi K,

- Kobayashi T: **Entirely plasmid-based reverse genetics system for rotaviruses.** *Proc Natl Acad Sci* 2017, **114**:2349 LP-2354.
216. Silvestri LS, Tortorici MA, Vasquez-Del Carpio R, Patton JT: **Rotavirus glycoprotein NSP4 is a modulator of viral transcription in the infected cell.** *J Virol* 2005, **79**:15165–15174.
217. Russell AB, Trapnell C, Bloom JD: **Extreme heterogeneity of influenza virus infection in single cells.** *Elife* 2018, **7**:e32303.
218. Adolph TE, Tomczak MF, Niederreiter L, Ko H-J, Böck J, Martinez-Naves E, Glickman JN, Tschurtschenthaler M, Hartwig J, Hosomi S, Flak MB, Cusick JL, Kohno K, Iwawaki T, Billmann-Born S, Raine T, Bharti R, Lucius R, Kweon M-N, Marciniak SJ, Choi A, Hagen SJ, Schreiber S, Rosenstiel P, Kaser A, Blumberg RS: **Paneth cells as a site of origin for intestinal inflammation.** *Nature* 2013, **503**:272.
219. Saxena K, Blutt SE, Ettayebi K, Zeng X-L, Broughman JR, Crawford SE, Karandikar UC, Sastri NP, Conner ME, Opekun AR, Graham DY, Qureshi W, Sherman V, Foulke-Abel J, In J, Kovbasnjuk O, Zachos NC, Donowitz M, Estes MK: **Human Intestinal Enteroids: a New Model To Study Human Rotavirus Infection, Host Restriction, and Pathophysiology.** *J Virol* 2016, **90**:43 LP-56.
220. Kowalczyk MS, Tirosh I, Heckl D, Rao TN, Dixit A, Haas BJ, Schneider RK, Wagers AJ, Ebert BL, Regev A: **Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells.** *Genome Res* 2015, **25**:1860–1872.
221. Yin Y, Bijvelds M, Dang W, Xu L, van der Eijk AA, Knipping K, Tuysuz N, Dekkers JF, Wang Y, de Jonge J, Sprengers D, van der Laan LJW, Beekman JM, ten Berge D, Metselaar HJ, de Jonge H, Koopmans MPG, Peppelenbosch MP, Pan Q: **Modeling rotavirus infection and antiviral therapy using primary intestinal organoids.** *Antiviral Res* 2015, **123**:120–131.
222. Wallach TE, Bayrer JR: **Intestinal Organoids: New Frontiers in the Study of Intestinal Disease and Physiology.** *J Pediatr Gastroenterol Nutr* 2017, **64**:180–185.
223. Paşca SP: **Assembling human brain organoids.** *Science (80- )* 2019, **363**:126 LP-127.
224. Gilger MA, Matson DO, Conner ME, Rosenblatt HM, Finegold MJ, Estes MK: **Extraintestinal rotavirus infections in children with immunodeficiency.** *J Pediatr* 1992, **120**:912–917.
225. Jalilvand S, Marashi SM, Tafakhori A, Shoja Z: **Extraintestinal Involvement of Rotavirus Infection in Children.** *Arch Iran Med* 2015, **18**:604–605.
226. Cattaneo R: **Biased (A→I) hypermutation of animal RNA virus genomes.** *Curr Opin Genet Dev* 1994, **4**:895–900.
227. Qian B, Kibenge FSB: **Observations on polymerase chain reaction amplification of infectious bursal disease virus dsRNA.** *J Virol Methods* 1994, **47**:237–242.
228. Maan S, Rao S, Maan NS, Anthony SJ, Attoui H, Samuel AR, Mertens PPC: **Rapid cDNA synthesis and sequencing techniques for the genetic study of bluetongue and other dsRNA viruses.** *J Virol Methods* 2007, **143**:132–139.
229. Santiana M, Ghosh S, Ho BA, Rajasekaran V, Du W-L, Mutsafi Y, De Jesús-Díaz DA, Sosnovtsev S V, Levenson EA, Parra GI, Takvorian PM, Cali A, Bleck C, Vlasova AN, Saif LJ, Patton JT, Lopalco P, Corcelli A, Green KY, Altan-Bonnet N: **Vesicle-Cloaked Virus Clusters Are Optimal Units for Inter-organismal Viral Transmission.** *Cell Host Microbe* 2018, **24**:208–220.e8.
230. Volden R, Palmer T, Byrne A, Cole C, Schmitz RJ, Green RE, Vollmers C: **Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA.** *Proc Natl Acad Sci* 2018, **115**:9726 LP-9731.
231. Russell AB, Kowalsky JR, Bloom JD: **Single-cell virus sequencing of influenza infections that**

**trigger innate immunity.** *bioRxiv* 2018:437277.

232. McCrone JT, Woods RJ, Martin ET, Malosh RE, Monto AS, Lauring AS: **Stochastic processes constrain the within and between host evolution of influenza virus.** *Elife* 2018, **7**:e35962.

233. Stetson DB, Medzhitov R: **Type I Interferons in Host Defense.** *Immunity* 2006, **25**:373–381.

234. Wang X, Allen WE, Wright MA, Sylwestrak EL, Samusik N, Vesuna S, Evans K, Liu C, Ramakrishnan C, Liu J, Nolan GP, Bava F-A, Deisseroth K: **Three-dimensional intact-tissue sequencing of single-cell transcriptional states.** *Science (80- )* 2018, **361**:eaat5691.

235. Hashimoto T, Yoshida K, Hashimoto N, Nakai A, Kaneshiro K, Suzuki K, Kawasaki Y, Shibamura N, Hashiramoto A: **Circulating cell free DNA: a marker to predict the therapeutic response for biological DMARDs in rheumatoid arthritis.** *Int J Rheum Dis* 2017, **20**:722–730.

236. De Mattos-Arruda L, Mayor R, Ng CKY, Weigelt B, Martínez-Ricarte F, Torrejon D, Oliveira M, Arias A, Raventos C, Tang J, Guerini-Rocco E, Martínez-Sáez E, Lois S, Marín O, de la Cruz X, Piscuoglio S, Towers R, Vivancos A, Peg V, Cajal SR y, Carles J, Rodon J, González-Cao M, Tabernero J, Felip E, Sahuquillo J, Berger MF, Cortes J, Reis-Filho JS, Seoane J: **Cerebrospinal fluid-derived circulating tumour DNA better represents the genomic alterations of brain tumours than plasma.** *Nat Commun* 2015, **6**:8839.

237. Venkatesan BM, Bashir R: **Nanopore sensors for nucleic acid analysis.** *Nat Nanotechnol* 2011, **6**:615.

238. Jain M, Olsen HE, Paten B, Akeson M: **The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community.** *Genome Biol* 2016, **17**:239.

239. Quick J, Ashton P, Calus S, Chatt C, Gossain S, Hawker J, Nair S, Neal K, Nye K, Peters T, De Pinna E, Robinson E, Struthers K, Webber M, Catto A, Dallman TJ, Hawkey P, Loman NJ: **Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of Salmonella.** *Genome Biol* 2015, **16**:114.

240. Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, Bore JA, Koundouno R, Dudas G, Mikhail A, Ouédraogo N, Afrough B, Bah A, Baum JHJ, Becker-Ziaja B, Boettcher JP, Cabeza-Cabrero M, Camino-Sánchez Á, Carter LL, Doerrbecker J, Enkirch T, Dorival IG-, Hetzelt N, Hinzmann J, Holm T, Kafetzopoulou LE, Koropogui M, Kosgey A, Kuisma E, Logue CH, et al.: **Real-time, portable genome sequencing for Ebola surveillance.** *Nature* 2016, **530**:228.

241. Castro-Wallace SL, Chiu CY, John KK, Stahl SE, Rubins KH, McIntyre ABR, Dworkin JP, Lupisella ML, Smith DJ, Botkin DJ, Stephenson TA, Juul S, Turner DJ, Izquierdo F, Federman S, Stryke D, Somasekar S, Alexander N, Yu G, Mason CE, Burton AS: **Nanopore DNA Sequencing and Genome Assembly on the International Space Station.** *Sci Rep* 2017, **7**:18022.

242. Wilson BD, Eisenstein M, Soh HT: **High-Fidelity Nanopore Sequencing of Ultra-Short DNA Sequences.** *bioRxiv* 2019:552224.

243. Tzimagiorgis G, Michailidou EZ, Kritis A, Markopoulos AK, Koidou S: **Recovering circulating extracellular or cell-free RNA from bodily fluids.** *Cancer Epidemiology* 2011:580–589.

244. Schwarzenbach H, Nishida N, Calin GA, Pantel K: **Clinical relevance of circulating cell-free microRNAs in cancer.** *Nat Rev Clin Oncol* 2014, **11**:145.

245. Tsang JCH, Vong JSL, Ji L, Poon LCY, Jiang P, Lui KO, Ni Y-B, To KF, Cheng YKY, Chiu RWK, Lo YMD: **Integrative single-cell and cell-free plasma RNA transcriptomics elucidates placental cellular dynamics.** *Proc Natl Acad Sci* 2017:201710470.

246. Bar-On YM, Phillips R, Milo R: **The biomass distribution on Earth.** *Proc Natl Acad Sci* 2018, **115**:6506 LP-6511.