Numerous uncharacterized and highly divergent microbes which colonize humans are revealed by circulating cell-free DNA

Mark Kowarsky[a], Joan Camunas-Soler[b], Michael Kertesz[b,1], Iwijn De Vlaminck[b], Winston Koh[b], Wenying Pan[b], Lance Martin[b], Norma F. Neff[b,c], Jennifer Okamoto[b,c], Ronald J. Wong[d], Sandhya Kharbanda[e], Yasser El-Sayed[f], Yair Blumenfeld[f], David K. Stevenson[d], Gary M. Shaw[d], Nathan D. Wolfe[g,h], and Stephen R. Quake[b,c,i,2]

[a]Department of Physics, Stanford University, Stanford, CA 94305; [b]Department of Bioengineering, Stanford University, Stanford, CA 94305; [c]Chan Zuckerberg Biohub, San Francisco, CA 94158; [d]Department of Pediatrics, Stanford University School of Medicine, Stanford University, Stanford, CA 94305; [e]Pediatric Stem Cell Transplantation, Lucille Packard Children's Hospital, Stanford University, Stanford, CA 94305; [f]Division of Maternal–Fetal Medicine, Department of Obstetrics and Gynecology, Stanford University School of Medicine, Stanford University, Stanford, CA 94305; [g]Metabiota, San Francisco, CA 94104; [h]Global Viral, San Francisco, CA 94104; and [i]Department of Applied Physics, Stanford University, Stanford, CA 94305

Contributed by Stephen R. Quake, July 12, 2017 (sent for review April 28, 2017; reviewed by Søren Brunak and Eran Segal)

Blood circulates throughout the human body and contains molecules drawn from virtually every tissue, including the microbes and viruses which colonize the body. Through massive shotgun sequencing of circulating cell-free DNA from the blood, we identified hundreds of new bacteria and viruses which represent previously unidentified members of the human microbiome. Analyzing cumulative sequence data from 1,351 blood samples collected from 188 patients enabled us to assemble 7,190 contiguous regions (contigs) larger than 1 kbp, of which 3,761 are novel with little or no sequence homology in any existing databases. The vast majority of these novel contigs possess coding sequences, and we have validated their existence both by finding their presence in independent experiments and by performing direct PCR amplification. When their nearest neighbors are located in the tree of life, many of the organisms represent entirely novel taxa, showing that microbial diversity within the human body is substantially broader than previously appreciated.

cell-free DNA | microbiome | metagenomics | biological dark matter

The advent of high-throughput DNA sequencing has led to powerful new approaches to studying the diversity of life on Earth, ranging from single-cell genome sequencing (1, 2) to large-scale metagenomic analysis of bulk DNA from various microbial ecosystems (3–5). Applying this approach to a variety of environmental samples has led to the discovery of many new phyla, expanding knowledge of the diversity of the tree of life (6), while large human microbiome studies, such as the Human Microbiome Project (HMP) (7) and MetaHIT (8, 9), have characterized many previously unknown taxa at easily accessible body sites. However, those projects targeted specific niches, such as the gut or skin, and therefore do not detect organisms residing in other body sites or those possessing very low abundances. Here, we take advantage of the fact that blood is a medium that samples virtually the entire body and collects molecules—including DNA—released by the organisms which colonize humans in all body sites.

The existence of circulating nucleic acids in blood has been known since the mid-20th century (10), but only in the last few years has the advent of high-throughput sequencing led to clinical diagnostics based on these nucleic acids [also known as cell-free DNA (cfDNA) or RNA], including detecting fetal abnormalities (11), transplanted organ rejection events (12, 13), and signatures of cancers (14). It is not only human cells that shed their nucleic acids into the blood: DNA from plant-based foods has been detected (15), and other life forms such as viruses, bacteria, and fungi release their DNA and RNA into the blood, a phenomenon which has been exploited to determine the presence of infectious disease (12, 16) and to measure alterations of the virome due to pharmacological immunosuppression (17). There are roughly an order of magnitude more nonhuman cells than nucleated human cells in the body (18, 19); combining this observation with the average genome sizes of a human, bacterium, and virus (Gb, Mb, and kb, respectively) suggests that approximately 1% of DNA by mass in a human is derived from nonhost origins. Previous studies by us and others have shown that indeed approximately 1% of cfDNA sequences appear to be of nonhuman origin, but only a small fraction of these map to existing databases of microbial and viral genomes (16). This suggests that there is a vast diversity of as yet uncharacterized microbial diversity within the human microbiome and that this diversity can be analyzed through "unmappable" sequencing reads.

We analyzed the cfDNA-derived microbiomes of 1,351 samples from 188 patients in four longitudinally sampled cohorts—heart transplant (HT), 610 samples (76 patients); lung transplant (LT), 460 samples (59 patients); bone marrow transplant (BMT), 161 samples (21 patients); and pregnancy (PR), 120 samples (32 patients)—and discovered that the majority of assembled

Significance

Through massive shotgun sequencing of circulating cell-free DNA from the blood of more than 1,000 independent samples, we identified hundreds of new bacteria and viruses which represent previously unidentified members of the human microbiome. Previous studies targeted specific niches such as feces, skin, or the oral cavity, whereas our approach of using blood effectively enables sampling of the entire body and reveals the colonization of niches which have been previously inaccessible. We were thus able to discover that the human body contains a vast and unexpected diversity of microbes, many of which have highly divergent relationships to the known tree of life.

Author contributions: M. Kowarsky, M. Kertesz, I.D.V., and S.R.Q. designed research; M. Kowarsky, J.C.-S., I.D.V., W.K., W.P., L.M., N.F.N., J.O., R.J.W., S.K., Y.E.-S., Y.B., D.K.S., G.M.S., and S.R.Q. performed research; M. Kowarsky contributed new reagents/analytic tools; R.J.W., S.K., Y.E.-S., Y.B., D.K.S., and G.M.S. recruited patients and collected samples; M. Kowarsky and S.R.Q. analyzed data; and M. Kowarsky, N.D.W., and S.R.Q. wrote the paper.

Reviewers: S.B., Novo Nordisk Foundation Center for Protein Research; and E.S., Weizmann Institute of Science.

Conflict of interest statement: N.D.W. is an employee of Metabiota and founder of Global Viral; S.R.Q. is a founder of Karius. M. Kertesz is an employee and founder of Karius, but all work was performed while at Stanford and before he joined the company. All other authors declare no conflict of interest.

Freely available online through the PNAS open access option.

Data deposition: Sequencing data are accessible on the NCBI sequence read archive (accession nos. PRJNA263522, PRJNA222186, PRJNA385009, and PRJNA385180).

[1]Present address: Karius, Redwood City, CA 94065.

[2]To whom correspondence should be addressed. Email: quake@stanford.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1707009114/-/DCSupplemental.

sequences are derived from previously unidentified organisms. For example, we found numerous novel anelloviruses in immunocompromised patients, which represent a doubling of identified members in that viral family. Over two thirds of the sequences are bacterial, and the majority are most similar to proteobacteria; however, many large contigs can only be classified at the phylum or superkingdom level. We also found numerous novel phages throughout the population. Multiple independent analyses confirm the existence of these novel sequences.

Unmappable Reads from the Human Microbiome Can Be Assembled and Annotated

We sequenced a total of 37 billion molecules from the 1,351 samples of cfDNA, of which 95% of reads passed quality control. Of these, an average of 0.45% did not align to the reference human genome (GRCh38) (Fig. 1*A*, *Left*), in line with our expectations of the nonhuman DNA sources in the body. Of these putatively nonhuman reads, only approximately 1% could be identified in a curated microbiome database of almost 8,000 species of known bacteria, viruses, fungi, and eukaryotic pathogens (Fig. 1*B*, *Left*). This miniscule fraction of reads encompasses the known microbiome. Less than 1,800 known species (800 known genera) are observed across all samples. The rarefaction curve of species prevalence quickly plateaus, and the species abundance distribution has only a slight positive skew (*SI Appendix*, Fig. S1). These both indicate that the number of known species we measure has saturated (20), and deeper or broader sequencing of cfDNA from humans is unlikely to substantially increase the richness of known species.

We performed de novo assembly on the remaining nonhuman reads to uncover new species in the dark matter of cfDNA. The construction of assemblies used an iterative approach (Fig. 1*E*, *Top*). Nonhuman reads were assembled on a per-sample basis, and reads that aligned to low-complexity or human-derived contigs were removed. This process of assembly and cleaning was repeated for remaining reads, pooled first by patient and then by cohort, and resulted in a total assembly of 40 Mbp. Over 25 megabases of low complexity or residual human-derived contigs were removed (15% of reads) (Fig. 1*A*, *Right*), many of which were identified as human microsatellites or primate BAC/FOSMID clones (Dataset S1). The iterative assembly process captures more reads in each stage as the number of reads pooled together increases (*SI Appendix*, Fig. S2). The cohort assemblies constructed 7,190 contigs larger than 1 kbp and 131 larger than 10 kbp. Compared with the proportion of reads that map to known organisms, an order-of-magnitude greater fraction of reads map to the cohort assemblies (Fig. 1*B*, *Right*).

To select for "novel" contigs likely to originate from uncharacterized genomes, a series of filtering steps were applied (Fig. 1*E*, *Bottom*). The first two filters enrich for contigs that have a high gene content (i.e., predicted genes span at least 60% of bases) and low homology at both the nucleotide and protein level to any previously known sequence (*SI Appendix*, Figs. S3 and S4; i.e., BLAST alignments span less than 20% of bases and an average gene identity of less than 60%). Application of these filters results in the selection of 4,354 contigs over 1 kbp. Two additional steps removed contigs that have homologies in expanded microbiome sequence databases. First, we aligned all assemblies from phases II and III of the HMP and blacklisted the 47 contigs that had matches. As many of the HMP assemblies are of previously known organisms or had been deposited in the National Center for Biotechnology Information (NCBI) nucleotide (nt) database, the BLAST coverage filter encompasses most of the HMP-filtered contigs (*SI Appendix*, Fig. S5). The second filter used the web service Onecodex (21) to remove contigs with homologies in their database.

To control for potential contaminants from the extraction columns, we prepared six sequencing libraries using the same protocol used in the plasma-based cfDNA samples, but instead of plasma, we used either water or DNA extracted from a human cell line. Filtered reads were aligned to the assembled contigs, resulting in blacklisting an extra 114 contigs due to their presence in control samples. An additional step to check for contaminants was performed using the nonhost reads from 300 cfDNA samples obtained from nonhuman primate plasma. No contigs were observed at a level above 1 read per kilobase in more than 158 samples, with 75% of contigs observed at this level in less than 27 samples. The highly variable and nonubiquitous expression of these contigs in primate samples indicates that these are not common contaminants from the laboratory or kits.

After all of the filters, a total of 4,187 contigs remain, which can be further reduced to 3,761 novel candidates after merging contigs with significant overlaps. For later comparisons, a further 773 contigs are classified as "known" (>80% BLAST coverage and >1 kbp) and a further 598 as "divergent" (>1 kbp and neither known nor novel). The majority of assembled bases are

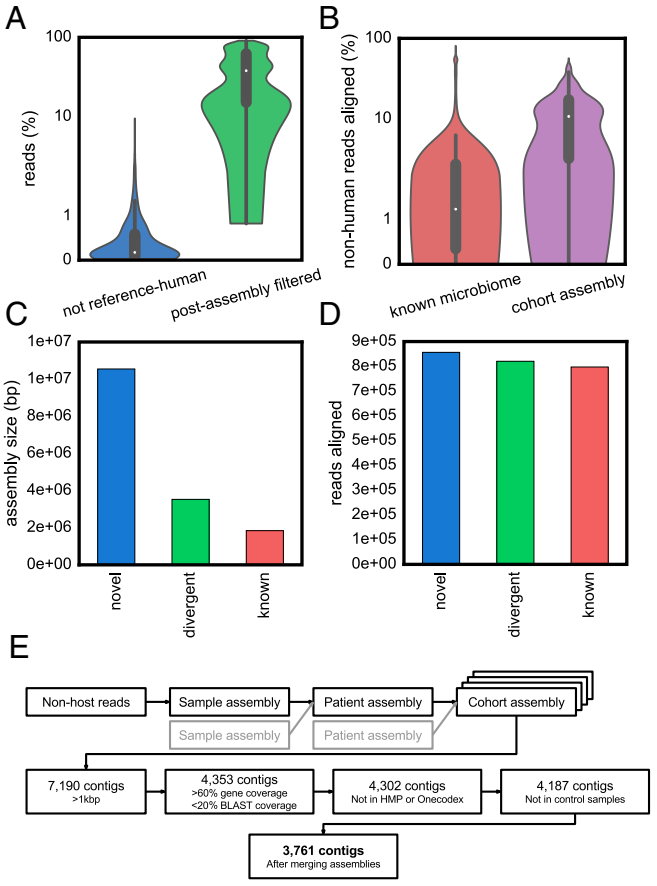

Fig. 1. (*A* and *B*) Violin plots showing the distributions of the percentage of reads associated with various stages of the assembly pipeline. From left: (*A*) reads that did not align to the reference human sequence; reads that were not removed during postassembly cleanups of additional human-like or low-complexity sequences; and (*B*) nonhuman reads that aligned to a curated microbiome database; nonhuman reads that aligned to the cohort assemblies. The white dot is the median value, and the thicker black bar is the interquartile range. (*C*) The total size of novel, divergent, and known contigs in the cohort assemblies. (*D*) Number of reads aligning to novel, divergent, and known contigs in the cohort assemblies. (*E*) Schematic of the iterative assembly process and novel contig selection. Between each step of assembly, further human and low-complexity contigs are blacklisted, with reads pooled from multiple samples/patients used in the next assembly. Multiple filters are performed on the long (>1 kbp) contigs to select for sequences that have not been previously observed and are unlikely to be contaminants.

novel, with 11 Mbp assembled compared with 3.5 Mbp of known or divergent contigs (Fig. 1*C*). The number of reads aligning is similar for all classes of contig (Fig. 1*D*). The lower coverage of novel contigs may have kept them hidden from previous analyses, and only by pooling many samples together were these sequences found (22). Biases in read coverage across the bacterial contigs may also be used to estimate their growth dynamics (23).

Novel Contigs Are Not Artifacts or Contaminants

As validation that the novel contigs are not merely contaminants or assembly artifacts, we performed a series of additional tests of their existence. The fragmented nature of cfDNA prevents direct PCR amplification of large regions of the contigs, so instead we (*i*) screened external datasets and samples for novel sequences at the read level, (*ii*) used bioinformatic approaches to assess assembly quality, and (*iii*) independently measured the presence of novel contig sequences by PCR in additional samples.

Plasma aliquots from a lung transplant recipient in our cohort were independently prepared and sequenced in another laboratory using an alternative library preparation protocol (24). We downloaded the reads from the NCBI sequence read archive and processed them identically in the pipeline followed by aligning them to the database of novel contigs. The two sets of different preparations exhibit concordant rankings of the high-abundance contigs (*SI Appendix*, Fig. S6). We also downloaded data generated by another group (25) from pregnancy samples that were completely independently collected, extracted, and sequenced and compared the reads to our pregnancy cohort (*SI Appendix*, Fig. S7). Although these samples had only slightly more than a tenth as many nonhuman reads as our samples, there were still reads which mapped to the novel contigs discovered in our experiments. These external datasets provide additional evidence that the novel contigs are not local laboratory contaminants and that their presence is robust to different sample preparation protocols.

The quality of assembly was assessed using information from the iterative assembly approach. Alignments between contigs from different levels created a graph structure, allowing us to investigate the additive structure derived from pooling more samples together. Under ideal circumstances, contigs assembled at lower stages should be wholly contained at higher stages and they should not end up orphaned (i.e., having no homology to a later stage's contig). Only 3,301 of the 25,765 sample contigs (over 500 bp, nonhuman, not low-complexity) are orphaned, compared with the patient contigs. Over 3,000 of these are from single-end read samples and are ignored by the assembler at later stages, as the longer paired-end reads tend to be more informative. The patient-orphaned contigs mostly had high BLAST coverages (10,768 of 15,321 had coverage >20%) and encompassed the following genera most frequently: *Lactobacillus* (5,975), *Streptococcus* (1,449), *Saccharomyces* (1,315), *Leuconostoc* (1,051), and *Lactococcus* (610). Likewise to the sample-to-patient case, the vast majority of the remaining contigs (>85%) derive from patients who only have single-end sequencing. If we look in the other direction, i.e., at the lower contigs contained within the novel contigs, the majority of them derive from a single patient's assembly (with some reads recruited from others), and only 387 were not assembled at any lower stages. The increased complexity of assembling due to more pooling does not appear to reduce the power to assemble what is present or to produce strong evidence of chimeric sequences. The consistency of the novel microbiome was further analyzed by calculating the similarity between longitudinal samples from the same patients. Times within approximately 3 mo tend to be highly correlated before the observed contigs start to diverge.

An additional check of the pipeline with a synthetic control dataset was performed to test if relevant sequences are prematurely removed. This was comprised of 8,068 curated genomes from the NCBI of viruses, bacteria, and fungi from which high-quality synthetic 100-bp reads at 25-bp intervals were created. Only 28 genomes (Dataset S2) had more than 50% of their reads removed in the pipeline steps leading up to assembly. These include various plasmids and enterobacteria phages (expected to be filtered in UniVec), simian virus 40, a few candidate bacterial species, and two variants of the human papillomavirus. Less than 10% of genomes had more than 10% of their reads removed, and those that were removed are sequences known to be in the cleaning or subtraction databases, indicating that the subtraction steps are unlikely to remove relevant nonhuman sequences.

We performed PCR testing to verify the presence of novel contigs in 16 previously unsequenced samples derived from the plasma of eight heart transplant recipients. As the fragmented nature of cfDNA prevents the amplification of large fragments, we designed primers against numerous short (60–100 bp) templates. Sequencing libraries known to express the templates targeted were used as positive controls (fifth column in *SI Appendix*, Fig. S8). Eight primers for a novel TTV (LT_node_191) and nine primers from unrooted contigs (HT220, HT552, HT76, PR78) were tested (Dataset S3). One of the eight patients had evidence at both time points for four templates from the TTVs, and three additional patients had evidence for one TTV-derived template (Fig. 2). None of the unrooted templates were seen in these samples. This was not surprising, as the median of the maximum number of reads of the unrooted contigs in the samples is 21; in contrast, the TTVs have a median value of 180 reads observed. Only through sequencing many foreign molecules from hundreds of samples were the unrooted contigs (and other novel ones) able to be found.

Provisional Taxonomic Assignment of Novel Contigs

Among the 3,761 novel contigs we found 14 with ribosomal proteins, none of which resembled the 16S ribosomal subunit. This, coupled with the fact that many of the sequences are viral-like, led us to conclude that a taxonomic classification based on 16S rRNA sequences is unfeasible. Thus, the taxonomic position was estimated based on choosing the lowest common ancestor of the

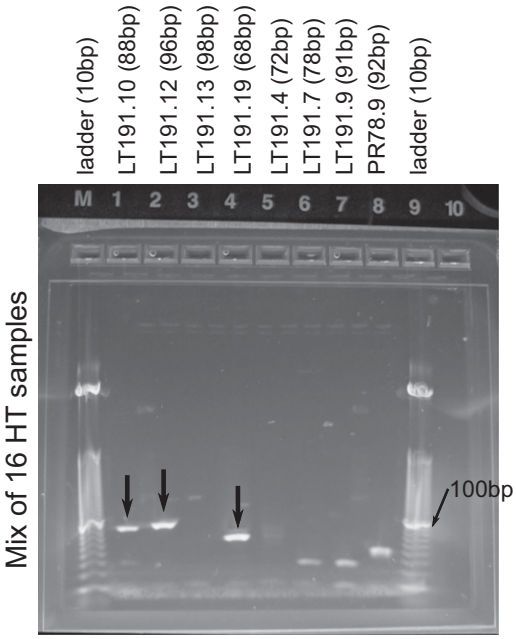

Fig. 2. Gel showing the presence of three segments of LT_node_191 (a novel torque teno virus) in the cfDNA deriving from the heart transplant recipients (eight patients, two samples per patient). Demultiplexed gels and no-template controls are in *SI Appendix*, Fig. S8.

majority of genes with homology, based on the gene's taxonomic assignment from alignment. For known contigs, this method agrees with the best nucleotide alignment of a contig at the species level over 85% of the time (94% accurate at genus level), indicating the suitability of this approach for taxonomic classification. Contigs with both bacterial and phage genes had their taxonomic placement refined by ignoring the phage genes if that made the assignment more specific. To visualize the taxonomic diversity, a "solar system" was constructed for all novel contigs (Fig. 3). Individual dots are contigs, with the rings representing the taxonomic levels; their radial position within the ring is proportional to the average gene identity (inner is high, outer is low), and the angular position is dependent on the superkingdom or phylum to which the contig was assigned. The 844 yellow contigs placed on the outermost ring cannot confidently be placed in any superkingdom based on gene homology because for almost all of them the predicted genes have no known homology. Over two thirds of sequences appear to be bacterial; the majority of these are most similar to proteobacteria. Most longer contigs (>5 kb) are bacterial or prophage-like and can only be placed at the phylum or superkingdom levels.

Many novel sequences cannot be classified based on their gene-level homologies. There is a clear trend in the reduction of both the proportion of genes identified and the degree of similarity to known genes as we compare the known, divergent, and novel contigs (Fig. 4 and *SI Appendix*, Fig. S9 *A–C*). This lack of homology is the reason there are 840 unrooted contigs unable to be placed into any existing superkingdom (the yellow points in the outer solar system of Fig. 3). The majority of these (n = 739) have four or fewer predicted genes. Only four of the unrooted contigs have any identified genes, and most of those that are identified are of unknown function (full table in Dataset S4).

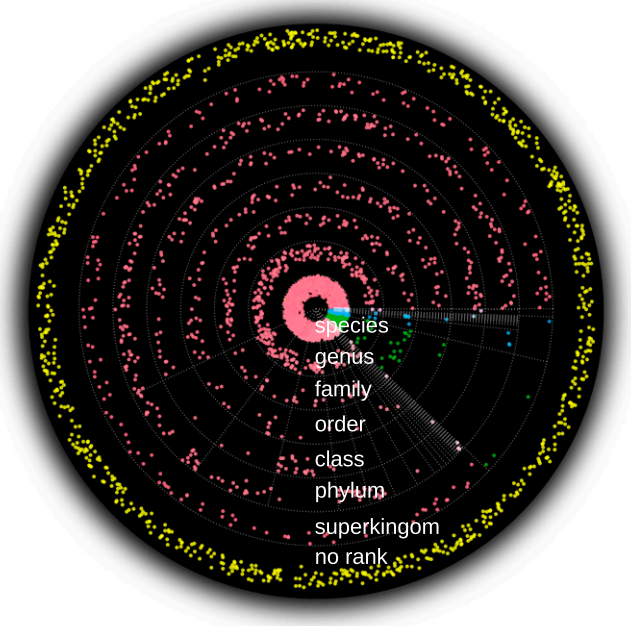

Fig. 3. Solar system plot of all novel contigs (>1 kbp). Each ring represents a different taxonomic level, with the intraring radius representing the average gene homology. Points are randomly assigned within sectors based on their superkingdom or phylum, as represented by the gray dotted spokes. The yellow contigs outside the last ring could not be assigned to any superkingdom; the radial scattering is to help illustrate the density of contigs. Colors correspond to superkingdom (counterclockwise from 3 o'clock: bacteria, viruses, eukaryota, archaea), saturation to the phylum.

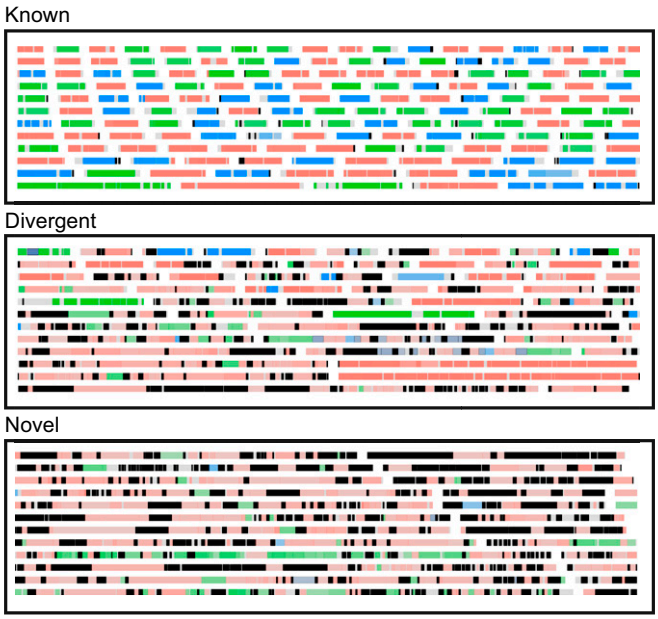

Fig. 4. Examples of contigs and their gene assignments for known, divergent, and novel contigs. Width of the box is equivalent to 60 kbp. Genes are colored by superkingdom (salmon, bacteria; lime green, viruses; dodger blue, eukaryota; fuchsia, archaea; black, no homology; gray, no gene); saturation is proportional to the gene identity.

As alignment-based methods failed to identify any genes from the remaining unrooted contigs, an estimated superkingdom assignment was made based on the codon use of genes (26). Comparing all genes with and without homology shows that these classes possess significantly different codon uses ($P <$ 1e-3 after Bonferroni correction, Mann–Whitney U test; *SI Appendix*, Fig. S10), which is not merely an artifact of differential GC bias (*SI Appendix*, Fig. S11). An adaptive boosting classifier trained on known genes allowed us to bin all except 16 of the unrooted contigs into single or multiple superkingdoms (*SI Appendix*, Fig. S12). The vast majority (n = 640) are bacteria, and, similarly to the rooted contigs, many others (n = 87) have viral and bacterial genes and are treated as phage or prophages. There are hints that horizontal gene transfer across superkingdoms may be prevalent in biological dark matter, with 106 unrooted contigs having diverse sets of genes. Despite this substantial taxonomic distance from known species, we were able to consistently assign the novel contigs to taxonomic locations on the tree of life.

Human Microbiome Has Been Substantially Undersampled in Previous Studies

The novel sequences indicate that the microbiome's species richness is much higher than is estimated from known species. If we count all of the taxa the contigs are assigned to, the number of taxa of known and divergent contigs is ~20% of the total number of species observed in the curated microbiome database. Conservatively counting contigs assigned to higher levels only once (e.g., only count "proteobacteria" once), we observe over 1,000 novel taxa. If the discovery rate is similar to known and divergent contigs, this gives a range of 1,000–5,000 new species (i.e., a 50–250% increase over known species).

The distributions of average gene identity within each rank are clearly distinct for novel, divergent, and known contigs (*SI Appendix*, Fig. S13), which is reflected in their solar system plots (*SI Appendix*, Figs. S14 and S15). Although the divergent and known contigs are fewer in number, they are much more precisely placed, both in their

taxonomic level and their average gene identity (reflected in the points being placed closer to the rank rings in the solar system). As more sequences are identified or classified, we expect many of the novel contigs in the outer orbitals to be attracted toward the center. The novel contigs are not restricted to individual samples or patients and most appear ubiquitously across the cohort populations. Each novel contig is observed in a median of 51 patients, and the median number of novel contigs observed in each patient is 924 (also see histograms in *SI Appendix*, Figs. S16 and S17). Even though elements of the human microbiome have been well-studied by deep sequencing from particular body sites, it is evident that other niches accessed by plasma contribute substantial novel diversity, which we report here for the first time.

Human Microbiome Contains an Unexpected Diversity of Novel Phages and Viruses

From the 2,917 placed novel contigs, 276 (9%) correspond to novel viral sequences, which are predominantly either phages or torque teno viruses (TTVs). Distinguishing between a phage and its bacterial host is difficult with short sequences, as they both are prone to incorporate each other's genes. Indeed, of the 523 contigs containing phage genes, 333 also have bacterial genes. Nonetheless, identifying these is important, as the contigs with the most predicted genes are all phage or prophage candidates. Half of the genes have no homology for the top 15 such contigs, with some (HT_node_2, HT_node_16, cluster_12) having over three quarters of genes without matches. Of the identified genes, their closest homologies are commonly hypothetical proteins. These two facts conspire to make functional annotation unfeasible at this time. The non-hypothetical genes present on the phage contigs include DNA primases, phage-tail proteins, terminases, capsid proteins, and DNA polymerases I and III. Many of their bacterial genes are from species within the proteobacteria phylum, with some from actinobacteria (HT_node_11, LT_node_6) or containing genes from a flavobacterium (bacteroides phylum) phage. Additional sequences were searched for in the recently published global virome from the Joint Genome Institute (JGI) (4). Only 21 of the unmerged contigs had matches over 50% of their lengths, and these derived from just six of JGI's scaffolds. Three of these are from Coloradan soil samples, one each from a wastewater bioreactor, a freshwater lake, and the upper troposphere. Our contigs matching to these are mostly best assigned as a cyanobacteria phage or unclassified Siphoviridae. The phage-like contigs were highly prevalent across patients and exhibited only minor clustering by cohort (*SI Appendix*, Fig. S18). These phages are likely to be associated with bacteria in the background microbial flora. Given that these contigs are observed ubiquitously in our data (77–287 samples) and seen in the JGI data only in a few exotic environments, we hypothesize that they are bona fide members of the human virome, and their presence in the JGI data is due to coincident discovery of related environmental species.

In contrast, the novel TTVs show strong clustering among cohorts (*SI Appendix*, Fig. S19) on the basis of immune system status. TTVs encompass anellovirus families and are known to be enriched in immunocompromised patients (17). Their detection here represents another validation of the taxonomic assignment based on gene homology. A phylogenetic tree was built using these contigs along with reference sequences from previously characterized anelloviruses (Fig. 5). Both reference and de novo assembled contigs can be clustered into three distinct classes: a class of nonhuman infecting reference anelloviruses (green), a set of novel anelloviruses that have only 35–48% sequence similarity with any reference sequence (yellow), and the rest, which we identify as new species or subspecies of existing TTVs. This work has almost doubled the total number of anelloviruses found in humans, including a new potential genus of human-infecting anelloviruses. Therefore, assembly-based metagenomic methods can uncover

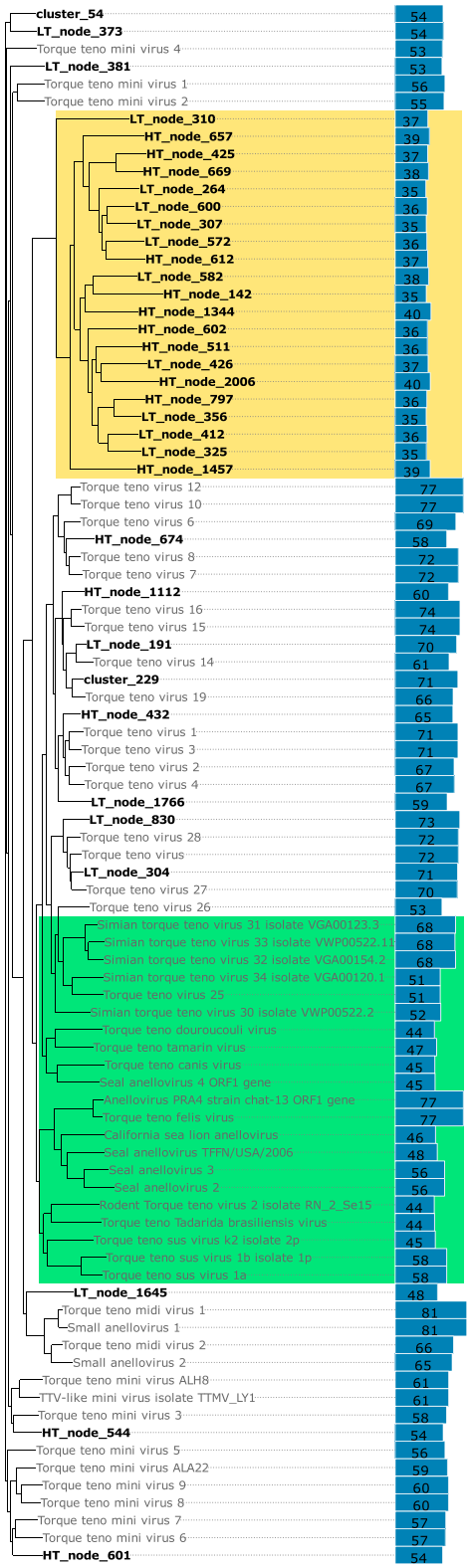

Fig. 5. Phylogenetic tree of reference and de novo assembled (bold) torque teno viruses (TTVs). Yellow sequences are a class of contigs divergent from known TTVs; green are known TTVs in animals. On the right are blue bars (with numbers) indicating the percentage identity of the nearest (nonself) reference sequence.

vast amounts of prevalent diversity even in known viral families with numerous reference sequences.

Conclusion

Deep sequencing of cfDNA from a large patient cohort revealed previously unknown and highly prevalent microbial and viral diversity in humans. This demonstrates the power of alternative assays for discovery and shows that interesting discoveries may lurk in the shadows of data acquired for other purposes. Many megabases of new sequences were assembled and placed in distant sectors of the tree of life. With deeper sequencing and targeted sample collection, we expect numerous new viral and bacterial species to be discovered in the circulating nucleic acids of organisms that will complement existing efforts to characterize the life within us. Novel taxa of microbes inhabiting humans, while of interest in their own right, also have potential consequences for human health. They may prove to be the cause of acute or chronic diseases that, to date, have unknown etiology and may have predictive associations that permit presymptomatic identification of disease. Assemblies and predicted genes are accessible in Datasets S5 and S6.

Materials and Methods

Plasma was extracted from whole-blood samples as previously described (11) with sequencing, preprocessing, and analysis of the known microbiome using our existing pipeline (16). The study was approved by the Stanford University Institutional Review Board. All patients provided written informed consent. Nonhuman reads were grouped by sample, patient, or cohort and assembled with SPADes (27), gene annotations provided by PRODIGAL (28), and homologies mostly determined using blastn or blastx against the NCBI nt and nonredundant protein (nr) databases. The solar system plot was constructed using the NCBI taxonomy and the ETE3 (29) python package. The anellovirus tree was constructed using FastTree (30) (GTR+CAT model and gamma option). Other cfDNA samples were downloaded from the Sequence Read Archive and processed identically to our samples before alignment to the novel contig database. For further information, see *SI Appendix, Materials and Methods*.

ACKNOWLEDGMENTS. We thank David Grimm and Helen Luikart for providing us with extra plasma samples for validation experiments. This work was supported in part by the Bill and Melinda Gates Foundation, the March of Dimes Prematurity Research Center at Stanford University, and the Stanford Child Health Research Institute. This work was supported by the John Templeton Foundation as part of the Boundaries of Life Initiative (Grant 51250). N.D.W. was supported in part by the US Agency for International Development (USAID) Emerging Pandemic Threats PREDICT program (Cooperative Agreement GHN-AOO-09-00010-00).

1. Marcy Y, et al. (2007) Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proc Natl Acad Sci USA* 104:11889–11894.
2. Gawad C, Koh W, Quake SR (2016) Single-cell genome sequencing: Current state of the science. *Nat Rev Genet* 17:175–188.
3. Tringe SG, Rubin EM (2005) Metagenomics: DNA sequencing of environmental samples. *Nat Rev Genet* 6:805–814.
4. Paez-Espino D, et al. (2016) Uncovering Earth's virome. *Nature* 536:425–430.
5. Sunagawa S, et al. (2015) Structure and function of the global ocean microbiome. *Science* 348:1261359.
6. Hug LA, et al. (2016) A new view of the tree of life. *Nat Microbiol* 1:16048.
7. Human Microbiome Project Consortium (2012) Structure, function and diversity of the healthy human microbiome. *Nature* 486:207–214.
8. Nielsen HB, et al.; MetaHIT Consortium; MetaHIT Consortium (2014) Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat Biotechnol* 32:822–828.
9. Qin J, et al.; MetaHIT Consortium (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464:59–65.
10. Mandel P, Metais P (1948) Les acides nucleiques du plasma sanguin chez l'homme. *CR Seances Soc Biol Fil* 142:241–243.
11. Fan HC, Blumenfeld YJ, Chitkara U, Hudgins L, Quake SR (2008) Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. *Proc Natl Acad Sci USA* 105:16266–16271.
12. De Vlaminck I, et al. (2014) Circulating cell-free DNA enables noninvasive diagnosis of heart transplant rejection. *Sci Transl Med* 6:241ra77.
13. Snyder TM, Khush KK, Valantine HA, Quake SR (2011) Universal noninvasive detection of solid organ transplant rejection. *Proc Natl Acad Sci USA* 108:6229–6234.
14. Schwarzenbach H, Hoon DSB, Pantel K (2011) Cell-free nucleic acids as biomarkers in cancer patients. *Nat Rev Cancer* 11:426–437.
15. Spisák S, et al. (2013) Complete genes may pass from food to human blood. *PLoS One* 8:e69805.
16. De Vlaminck I, et al. (2015) Noninvasive monitoring of infection and rejection after lung transplantation. *Proc Natl Acad Sci USA* 112:13336–13341.
17. De Vlaminck I, et al. (2013) Temporal response of the human virome to immunosuppression and antiviral therapy. *Cell* 155:1178–1187.
18. Sender R, Fuchs S, Milo R (2016) Revised estimates for the number of human and bacteria cells in the body. *PLoS Biol* 14:e1002533.
19. Luckey TD (1972) Introduction to intestinal microecology. *Am J Clin Nutr* 25: 1292–1294.
20. Preston FW (1948) The commonness, and rarity, of species. *Ecology* 29:254–283.
21. Minot SS, Krumm N, Greenfield NB (2015) One codex: A sensitive and accurate data platform for genomic microbial identification. *bioRxiv*. Available at: www.biorxiv.org/content/early/2015/09/25/027607. Accessed March 23, 2016.
22. Dutilh BE, et al. (2014) A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat Commun* 5:4498.
23. Korem T, et al. (2015) Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples. *Science* 349:1101–1106.
24. Burnham P, et al. (2016) Single-stranded DNA library preparation uncovers the origin and diversity of ultrashort cell-free DNA in plasma. *Sci Rep* 6:27859.
25. Karlsson K, et al. (2015) Amplification-free sequencing of cell-free DNA for prenatal non-invasive diagnosis of chromosomal aberrations. *Genomics* 105:150–158.
26. Dick GJ, et al. (2009) Community-wide analysis of microbial genome sequence signatures. *Genome Biol* 10:R85.
27. Bankevich A, et al. (2012) SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 19:455–477.
28. Hyatt D, et al. (2010) Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11:119.
29. Huerta-Cepas J, Serra F, Bork P (2016) ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. *Mol Biol Evol* 33:1635–1638.
30. Price MN, Dehal PS, Arkin AP (2010) FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One* 5:e9490.