



Using populations of human and microbial genomes for organism detection in metagenomes

Sasha K Ames, Shea N Gardner, Jose Manuel Marti, et al.

Genome Res. published online April 29, 2015

Access the most recent version at doi:[10.1101/gr.184879.114](https://doi.org/10.1101/gr.184879.114)

P<P	Published online April 29, 2015 in advance of the print journal.
Accepted Manuscript	Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.
Open Access	Freely available online through the <i>Genome Research</i> Open Access option.
Creative Commons License	This manuscript is Open Access. This article, published in <i>Genome Research</i> , is available under a Creative Commons License (Attribution 4.0 International license), as described at http://creativecommons.org/licenses/by/4.0/ .
Email Alerting Service	Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here .

Comprehensive immune receptor profiling.
Discover the **DriverMap™ AIR Assay** difference.



To subscribe to *Genome Research* go to:
<https://genome.cshlp.org/subscriptions>

Published by Cold Spring Harbor Laboratory Press

Using populations of human and microbial genomes for organism detection in metagenomes

Sasha K Ames^{1,3}, Shea N Gardner^{2,3}, Jose Manuel Marti⁴, Tom R Slezak^{2,3}, Maya B Gokhale^{1,3}, Jonathan E Allen^{2,3}

¹Center for Applied Scientific Computing

²Global Security Computer Applications Division

³Lawrence Livermore National Laboratory, Livermore CA

⁴Instituto de Física Corpuscular, CSIC-UVEG, Valencia, Spain

Abstract

Identifying causative disease agents in human patients from shotgun metagenomic sequencing (SMS) presents a powerful tool to apply when other targeted diagnostics fail. Numerous technical challenges remain, however, before SMS can move beyond the role of research tool. Accurately separating the known and unknown organism content remains difficult particularly when SMS is applied as a last resort. The true amount of human DNA that remains in a sample after screening against the human reference genome and filtering non-biological components left from library preparation has previously been underreported. In this study we create the most comprehensive collection of microbial and reference-free human genetic variation available in a database optimized for efficient metagenomic search by extracting sequences from GenBank and the 1000 Genomes Project. The results reveal new human sequences found in individual Human Microbiome Project (HMP) samples. Individual samples contain up to 95% human sequence and 4% of the individual HMP samples contain 10% or more human reads. Left unidentified, human reads can complicate and slow down further analysis and lead to inaccurately labeled microbial taxa and ultimately lead to privacy concerns as more human genome data is collected.

Introduction

Metagenomic analysis of human clinical samples is often confounded by the abundance of host genome present in the sample. This can be particularly challenging when trying to uncover a poorly characterized etiologic agent that may or may not be recoverable from a diagnostic next-generation sequencing run (Wilson et al., 2014; Yozwiak et al., 2012). Ideally each sequencer read should be examined and its relationship relative to all previously sequenced organisms accurately reported. Unfortunately, the most sensitive search needed to detect highly divergent organisms does not scale when naively examining an entire sequencing run that may total many gigabases (Zhao et al., 2012). A complementary approach is to use a metagenomic assembler; however, low coverage of individual microbial genomes can limit the potential for good quality assemblies in complex samples (Howe et al.,

2014) or samples dominated by host background. Marker based approaches present an important option for fast estimation of microbial content but only tag a small fraction of the input leaving potentially interesting fragments hidden in a large collection of unlabeled reads (Tu et al., 2014; Segata et al., 2012; Sunagawa et al., 2013; Liu et al., 2011; Minot et al., 2014; Berendzen et al., 2012). Recent efforts have demonstrated substantial progress in labeling all reads by applying ordered searches that attempt to reserve the most computationally expensive analysis for the fewest reads (Zhao et al., 2013; Cotten et al., 2014; Takeuchi et al., 2014; Nakamura et al., 2009; Byrd et al., 2014). A recent example is SURPI (Naccache et al., 2014), which maps reads to the human reference genome to subtract host reads prior to search of the GenBank NT database. DeconSeq is another tool that identifies human reads as contaminants by aligning reads to seven assembled human genomes (Schmieder and Edwards, 2011) to detect human variants beyond the single reference genome. A limitation of these approaches, however, is the use of read mapping tools that organize their respective database search against each reference gene/genome independently. This presents a fundamental scaling problem as more organisms in the population are added to the database. For example, the reference database used to build taxonomic profiles for Human Microbiome Project (HMP) samples (Martin et al., 2012) used a microbial reference database totaling 7.3 gigabases (Gb), which represents only a small fraction of the sequenced microbial strains available. In the case of SURPI although a 42 Gb GenBank NT database is searched, a best first match approach is taken, which requires careful down stream analysis to avoid overly specific calls that fail to recognize highly conserved elements.

Non-redundant search of a population of human genomes was recently shown to improve detection sensitivity using a new data structure to map reads to multiple human genomes simultaneously (Huang et al., 2013). This approach used assembled human genomes for reference but excluded microbial genomes. A related approach uses a large k -mer index as implemented in software tools Kraken (Wood and Salzberg, 2014) and Livermore Metagenomics Analysis Toolkit (LMAT) (Ames et al., 2013). Kraken supports efficient search by mapping k -mers to each source genome and storing the lowest common ancestor. A highly conserved genetic region found in hundreds of genomes is efficiently identified with the lowest common ancestor (LCA) in a single search step. Recent improvements to LMAT present a similar approach but place a greater emphasis on storing more data in two ways. First, rather than store the LCA for each k -mer, the list of source genomes and the taxonomic hierarchy are stored - up to a user specified limited number of taxonomic identifiers (set to 200) to improve discrimination at lower taxonomic ranks while placing an upper bound on overall runtime. When a k -mer maps to more genomes than allowed by the threshold, a pruning procedure retains only the higher rank taxonomy identifiers. Second, LMAT uses a larger microbial genome database, which totals over 115 gigabases of genomic sequence – assembled draft and complete genomes for virus, bacteria, fungi, protozoa and mitochondrial genomes of eukaryotes. The original database required at least 620 gigabytes (GB) of DRAM, which limited its use to researchers with access to large memory machines. To improve accessibility, the software was recently tuned to exploit a combination of DRAM and NVRAM (e.g. flash drives) (Ames et al., 2014) and in addition reduced the database size to 458 GB for the full and 17 GB for the marker databases, thus allowing the software to run on substantially lower cost machines. These recent optimizations additionally enable the use of a larger cluster where NVRAM is used as a supplemental memory resource. Scaling is supported in part by the use of NVRAM, which allows use of a much larger local single address space than is possible with DRAM and at a lower cost. For example, the 800 GB of NVRAM available per compute node in this study is commercially available at \$5 per GB compared to \$10 per GB of DRAM (assuming 1 terabyte of DRAM). A key benefit of NVRAM over traditional high performance disks is the ability to randomly access any part of

flash memory similarly to traditional memory at orders of magnitude higher input output operations per second (IOPS) than is possible with traditional single high performance disks. RAID disk configurations and parallel file systems with multiple storage servers increase IOPS, but also increase cost and complexity. Additionally, parallel file systems based on storage area networks (SAN) greatly increase latency compared to node local NVRAM and are not suitable for our extended memory use case. NVRAM represents an important middle memory tier that allows the NVRAM to act as a supplemental memory resource in a way that is not possible with traditional spinning disk technology.

Results

Taking advantage of a large NVRAM resource, this paper uses LMAT to present the first evaluation of a vastly expanded database of searchable reference-free human genetic variants for metagenomic search. Over 90 terabases of raw sequence data comprising 2,626 human individuals from the 1000 Genomes Project (HGP) (The 1000 Genomes Project Consortium) were evaluated using 1.05 million CPU hours to identify novel human genetic variants. Table 1 summarizes new databases that were created and compared, including a new database termed “LMAT-Grand”, which includes novel human variants and an expanded collection of synthetic sequence contaminants. The LMAT-Grand database is used to annotate the 18 terabases of Human Microbiome Project (HMP) data (The Human Microbiome Project Consortium 2012a) and report microbial contents of the HGP data. The number of newly discovered human reads in HMP samples were compared with using the standard human reference genome alone (LMAT-Ref) and using all assembled human genetic data available through GenBank (Benson et al., 2013) (LMAT-GenBank). While the vast majority of human genetic variants found outside of the human reference genome were recovered through the use of assembled GenBank genes, novel variants from the HGP were responsible for identifying as much as 40 times more human reads in the HMP dataset depending on the body site. The increased sensitivity in detecting human reads confirms that the vast majority (84.58%) of the samples contained 1% or fewer human reads. Nevertheless, if looking for low abundance organisms or investigating the smaller number of cases containing large amounts of human sequence, misidentification of these human reads can lead to costly and inaccurate downstream reporting of a sample’s microbial contents. New detailed genus, species and strain level taxonomic profiles, plasmid, and gene profiles for all HMP samples are made available along with the newly discovered human sequence and other novel sequences. Recent work presented searches of small subsets of HGP data that look for a small number of specific microbes (Langdon, 2014; Laurence et al., 2014). A search of all available sequenced microbes across the complete phase 3 HGP dataset is given for the first time to better quantify potential sources of microbial contamination. Finally, the expanded collection of searchable human sequence data is used to report 38 million new human reads found in the recently sequenced La Braña genome (Olalde et al., 2014). The new searchable databases and software are made freely available both for large memory and smaller memory machines. The results provide a new resource for more detailed and accurate assessments of human metagenomic sequencing data.

LMAT-Grand identifies new human sequence in HMP

The first objective was to determine the ability to detect new human sequences in metagenomic samples that were previously screened against the human reference genome using new human genomic data. In addition to the human reference genome, two sources of human genomic data were considered: human sequence from GenBank and genomic data from HGP. Adding all human genetic sequence from GenBank to the LMAT database is a non-trivial undertaking, which required downloading 72 gigabases of sequence; however, it is computationally tractable. By contrast, adding the genetic variation from 2,626 human genomes was a major computational undertaking, requiring nearly half a petabyte of temporary storage, and 1.05 million CPU hours on a new cluster (Catalyst) designed for data intensive computing (see Methods).

Table 2 shows a count of distinct human labeled 20-mers added from each source of human genomic DNA: the reference assembly (LMAT-Ref), GenBank (LMAT-GenBank), and HGP (LMAT-Grand). The vast majority of human 20-mers not found in the human reference assembly were found in GenBank with fewer 20-mers unique to HGP. This gives a conservative report of HGP specific 20-mers given the strict selection criteria used to extract 20-mers from the HGP data (see Methods). Despite the smaller number of HGP unique 20-mers and their minimal impact on database size (see column 3 of Table 1), their inclusion led to identifying substantially more human reads in select body sites in the HMP.

Figure 1 shows the difference in human labeled reads across the body types sampled in the Human Microbiome Project (HMP) using 131 representative samples. All HMP samples were previously screened against the human reference genome to remove human reads using BMTagger¹. The results show that for many cases, the largest improvement is obtained through the addition of GenBank, however, in every sample improvement in identified human reads is observed when the HGP database is included. Overall, 1.7 times more human reads were detected using HGP data over GenBank alone and 2.4 times more than from the human reference genome alone. Additionally, a collection of 1000 reads taken from world regional distinct populations were spiked into HMP mock community samples to compare detection sensitivity between LMAT-Grand and LMAT-Ref, which showed up to 1.4 times improvement in detection sensitivity (see Supplemental Material).

Figure 2 shows a histogram for the abundance of human reads across the 9,025 HMP individual sequencer runs examined using the HGP derived database. The average amount of human DNA in each sequencer run is 2%, however, 4% of the runs contained 10% or more human reads with 75% of these samples coming from anterior nares (nostrils). Sequencer runs with an abundance of human reads ($\geq 10\%$) came from different individuals and two different sequencing centers, reducing the likelihood that these newly identified human sequences come from a single contaminating source.

To obtain evidence to confirm the newly identified human reads detected in HMP samples, blastn (Altschul et al., 1990), with default search settings was used to compare the reads against the NT database. Averaged over the 131 representative samples, support was found for 32% of the human reads showing eukaryote sequence (primates, other mammals, plants, insects, etc.) as the top match, including 28.5% with top match to human. Less than 2% had their top BLAST match to bacteria indicating a small false positive rate. These matches were examined and determined to originate from subsequences from a small number of partial bacterial genomes. Most of the reads tagged as human by LMAT-Grand had no

¹ http://hmpdacc.org/doc/HumanSequenceRemoval_SOP.pdf

BLAST matches using the default settings. Additional analysis that follows indicates that LMAT was more sensitive than default BLAST.

A more sensitive BLAST search (using a word size of 8 instead of 28 and max e-value of 0.01) for human reads from the sequencer run with the highest percentage of human reads was then examined (SRR059474 from an Anterior nares body site). Figure 3 shows a breakdown of the reads as a percentage of the human read counts (e.g. estimated abundance) and as a percentage of read clusters after clustering reads with CD-HIT (Fu et al., 2012). As a percentage of distinct read clusters, the vast majority (86%) have a top match to a primate sequence, giving strong support for the accuracy of the newly identified human read calls. As a percentage of distinct reads, the vast majority (71%) of reads are not recognized with BLAST matches. The most abundant sequence element shows a top match to *Solanales*, primarily originating from a single highly redundant cluster. This could represent a true biological contaminant, however, the relatively high e-value of 0.001 indicates high sequence divergence making it difficult to confidently assert the taxonomic identity. The 4 most abundant read clusters (comprising 57% of the human reads) with no initial BLAST match were checked for matches in the NT database with e-values up to 10. One of the 4 clusters showed exclusive best matches to human with an e-value of 0.75 giving the strongest support for a human origin. In the other cases the results were inconclusive with poor scoring matches (e-value ≥ 2.7) to bacteria and eukaryotes. The reads were checked for chimera to explain the lack of BLAST matches to human sequence, but insufficient numbers of chimera were found to explain the paucity of BLAST support. The unidentifiable human reads were also checked for their abundance across individuals in the 1000 Genomes Project collection, which show that approximately 25% of the reads were associated with at least 100 individuals (see Supplemental Material).

To further determine the identity of the new human sequences, human reads with BLAST hits to RepBase (Jurka et al., 2005) were compared with comparable NT and NR BLAST hits to identify known repetitive sequences with no evidence of human sequence in GenBank, but match a primate near-neighbor and show no conflicting protein matches. There were 33 distinct read clusters with a best match to a non-human primate sequence spanning 16 distinct repeat elements including SINE1 elements and others (see available data). Errors in the NR database were detected in the process of comparing human reads that matched to bacterial proteins in NR. A collection of human sequences was found to be misidentified as *Leuconostoc* sp. *DORA_2* proteins (GI: 566236135, 566236136, 566236137, 566238011, 566240723). While the blastx search showed 2,038 human reads matched to the five different *Leuconostoc* sp. *DORA_2* protein identifiers, 2,033 of the reads were found to have significant blastn matches to human sequences in NT. The sequences originate from premature infant guts and highlight the need to correctly recognize human reads to avoid microbial misattribution.

To further validate novel human read identification, LMAT-Grand was searched against HMP Mock Community samples as a negative control using both the staggered and even abundance Illumina samples (SRR172903 and SRR172902 respectively). Results from the staggered sample are reported as representative results. LMAT-Grand reported 4,585 human labeled reads and 669 of these could be validated with supporting BLAST alignments, indicating a small amount of human contamination even in these carefully constructed controls designed to contain only microbial DNA. However, 2,135 reads were found to be likely false positives given their lower BLAST e-value to a microbial genome than a human genomic sequence. Although this reflects a tiny fraction of the approximately 7.6 million reads, to confirm these reads were not having a disproportionate impact on over aggressive human read calling,

the collection of putative false positive human reads identified in the previous set of 131 HMP samples and the mock community samples were used to remove 20-mers from LMAT-Grand that were mislabeled human (see Supplemental Material). The revised LMAT-Grand was used to re-tag all human identified sequence in the HMP samples, which reduced the number of human reads by only 1.5%. In the examined high human abundance case (SRR059474), the number of human reads was reduced by only 0.3%. Therefore, while the corrected LMAT-Grand database is made available along with the corrected collection of human reads, other output was retained.

HMP samples contain limited world region and personal identifiable data

To consider whether newly labeled human reads could present potential privacy concerns, the reads were examined for short tandem repeat (STR) markers and world region identifiable markers. The STR detection software lobSTR (Gymrek et al., 2012) was run on the reads LMAT classified as human from the HMP data, grouping reads by subject ID across samples (combined from multiple sequencer runs) and body sites for the runs available for download, for a total of 178 subjects. An average of 25 short tandem repeats (STRs) per subject were identified, with a maximum of 54 in reads from one individual (see Supplemental Material). LobSTR reported the vast majority on Chromosome 5, one on Chromosome X, and none on Chromosome Y. Although none of these markers corresponded to 16 CODIS markers used for forensic attribution (based on positional information from <http://lobstr.teamerlich.org/ystr-codis.html>), it is possible that a database containing more STR markers could facilitate some level of individual identification using human reads remaining in the HMP samples after the standard procedure of read mapping to a single reference to remove human reads. The STRs were also checked against a library of newly catalogued HGP derived STRs to determine the amount of overlap (Willems et al., 2014). There are 78 of the 178 subjects with at least 20 of the hg19 reference loci with periods of 2-6, and a few subjects have up to 33 of the reference loci. 89 of the approximately 700,000 reference loci were present in the human reads identified by LMAT from the HMP samples, predominantly on Chromosome 5, but with at least one on every chromosome except Chromosomes 18, 21, X and Y. While the STRs do not currently provide personal identifying information it is conceivable that over time a sufficiently large database of human genomes labeled with personal identifying information could pose a privacy concern.

Considerable progress has been made in recent years using targeted genetic loci from the human genome to recognize shared genealogies (Elhaik et al., 2014). To investigate the potential for ethnic host identification in HMP samples, a search database was created (LMAT-Region in Table 1) to conduct world region identification from HGP genomes and ethnic codes. The LMAT-Region database is similar in structure to other LMAT databases except for two custom levels added to the taxonomy tree under the human species node. Taxonomy tree leaves are the HGP ethnic codes and their parents are the five world regions (Africa, Europe, East Asia, South Asia and Americas). The LMAT-Region contains human sequence only and is meant to regionally classify previously identified human reads. Supplementary Figure S3 shows ROC curves measuring sensitivity and specificity on HGP individuals, which reflect an optimistic assessment of accuracy – assuming future test individuals are drawn from the HGP population and there is relatively high coverage of the human genome present in the sample. Supplementary Figure S4 reports the ratio of total 20-mers to distinct 20-mers as a function of distinct 20-mers present in an individual human sample and provides the reference point to assess whether human region identifying reads are present in the HMP samples. The human reads for the 178 participating HMP individuals were searched against LMAT-Region and found to yield very small numbers of region world identifiable reads (at most 76). The small number of region identifiable reads

combined with the much larger number of reads used to accurately identify regions in the HGP data suggest that region identifying information from the host in HMP is not readily accessible.

Inclusion of all draft genomes in search database improves read bin sensitivity

Including draft genomes in the reference database presents an opportunity to recover more population level genetic variation using the rapidly growing number of draft assembled genomes available through GenBank. Including draft genomic sequence must be weighed with the potential for misclassification from included contaminants in draft assemblies. This is particularly problematic for human pathogens like *Toxoplasma gondii* where substantial amounts of un-recognized human sequence are included in the draft assemblies. The HGP data provide a valuable negative control where evidence of microbial content indicates potential false microbial identification in human samples. In response to uncovering falsely identified genomes, the “synthetic construct” taxonomy category was expanded to include contaminant sequence that would otherwise preclude the use of error prone draft assemblies. The added sequences include an updated collection of synthetic vectors (such as cloning vectors) (Allen et al., 2008), phiX sequence (used by Illumina as a sequencing control) and Illumina adaptors. Across the representative subset of 131 HMP sequencer runs, an average of 1.7% reads were tagged as synthetic. In addition, the use of GenBank and HGP human sequence reduced the percentage of *Toxoplasma* reads from 0.37% to 0.0003%. Although the average may seem relatively low across 131 samples using only the human reference, in individual cases the percentage of *Toxoplasma* reads were as high as 11.6% and use of HGP data reduced the percentage to 0.0087% demonstrating a powerful tool for recognizing human sequence inadvertently included in assembled microbial genomes.

An important motivation for using microbial population data is to reduce the number of unknown sequences, which require interrogation with additional computationally time-consuming methods. The hypothesis is that searching a read against all genetic variants of a strain or species will increase the number of reads that can be identified. Two metrics were used to measure the impact of adding populations of genomes – percentage of reads with no *k*-mer matches in the database (NoDbHits) and reads that have too few matches to the reference database to confidently use the assigned taxonomic label (termed LowScore). Use of the new human reference data (LMAT-Grand) did not increase the NoDbHits number but decreased LowScore reads by 2.8% essentially allowing more human reads to be confidently labeled. The bigger impact was observed when excluding microbial draft sequence. Table 3 compares results from three databases, LMAT-Grand, LMAT-RefSeq (a subset of Grand consisting of only complete RefSeq genomes) and the previously published mapping rate for the existing read mapping based microbial profile method available through the HMP Data Analysis Center (HMP DACC) at the web portal for accessing HMP related analysis and data (The Human Microbiome Project Consortium 2012b). The table shows a dramatic reduction in unlabeled reads from the previously reported 12% (Martin et al., 2012) down to 0.05% reads with no putative label. Thus, the number of reads with no database match is dramatically reduced by 240-fold to 139 megabases total for the HMP dataset.

A strikingly different microbial profile is given when relying on the RefSeq database alone as shown in Figure 4. The figure measures the amount of agreement between different databases at the species and genus level using different minimum abundance thresholds to decide organism presence. Given the larger database sizes required to store the complete collection of genome sequence, two forms of a marker library are made available to support applications on smaller memory machines. The marker

library containing only microbial sequences (ML in Figure 4) uses the LMAT-ML database described in Table 1. The marker library containing both the microbial sequences and a collection of human *k*-mers to support tagging human reads (ML+humanNoprune in Figure 4) is represented by the LMAT-ML-Human database described in Table 1. The marker libraries are a subset of the LMAT-Grand database, retaining *k*-mers that are unique to different taxonomic ranks (Ames et al., 2013). Despite an order of magnitude smaller size (17 gigabytes versus 458 gigabytes), the marker libraries showed closer agreement to the full database compared with the RefSeq methods and the profiles available from the HMP DACC. MetaPhlAn taxonomic profiles are also available through the HMP DACC and an updated version of MetaPhlAn profiles are included in Figure 4 as an additional profile reference point.

The species most commonly identified by HMP DACC mappings that were not detected in the same sequencer runs by LMAT-Grand were examined in detail, to determine if they are HMP DACC false positive calls or LMAT false negative calls. Up to 100 reads that HMP DACC mapped to the indicated species were extracted from each HMP DACC bam file, and BLAST searched against NT, the bacterial and viral subset of the full sequence database used by LMAT, and a database of the synthetic constructs. The fraction of BLAST hits to the indicated species was calculated for the NT and full database as well as the fraction of reads with hits to the vector database, given in Table 4. BLAST matches to NT and LMAT-Grand showed very few matches to the HMP DACC species call. Up to 31% of the reads matched synthetic constructs. The HMP DACC-mapped reads to each species were run through LMAT and the fraction of classified reads at the genus level or higher is indicated. LMAT classified up to 84% of these reads as conserved at the genus level or above. For those few remaining reads that LMAT assigned at the species level, the species with the most reads is indicated with the number of reads in parentheses. None of these species are in the reference database used by HMP DACC. The last column shows the fraction of these LMAT species-classified reads for which the LMAT species call is among the top BLAST hits. Few or none of the reads have the HMP DACC call in the BLAST matches except for candidate division TM7, which has nearly identical BLAST matches as *Saccharimonas aalborgensis*, the call made by LMAT. Therefore, the reads that LMAT does classify to species have high BLAST support for the LMAT classification. These results support the conclusion that these are false positive species classifications by the HMP DACC profiles due to a combination of overly specific calls for reads that are in fact much more widely conserved, matches to synthetic constructs that contaminate many of the draft genomes in reference sequence databases, and failure to detect the correct organism due to a lack of representation in the HMP DACC reference database.

For the first time, profiles of the plasmid and gene content are provided for all the HMP sequencer runs. Horizontal transfer and replicate sequencing complicate interpretation of plasmid and gene profiles, particularly when the same plasmid or gene is sequenced multiple times but differently named. Since LMAT takes a “best first match” approach, the reported plasmid or gene may have equal match scores to multiple plasmids or genes. Nonetheless taxonomic identification can be linked with gene calls to identify mobile genes of particular interest. For example, the 131 representative set was searched for plasmids known to carry the important drug resistance factor NDM-1 (*Klebsiella pneumonia* plasmid pNDM-US, *Acinetobacter baumannii* plasmid pAbNDM-1, or *Escherichia coli* plasmid p271A) (Hu et al., 2012) and then reads used to identify these plasmids were checked for their gene assignments. In total, 33 of the 131 sequencer runs indicate the presence of a NDM-1 related gene (using a minimum of 10 reads to support the NDM-1 related gene call) and in 7 sequencer runs the genes were linked with the p271A plasmid. Five of the 7 were taken from oral samples (Throat, Palatine Tonsils and Keratinized Gingiva). The number of reads used to support the calls is relatively small ranging from 10 to 94

indicating very low abundance in each sample. As a second example application, the 131 representative set was checked for human identified reads with gene assignments. Only 6% of the human reads were assigned gene calls and 1% of these reads were associated with non-protein coding genes suggesting that the vast majority of detected human sequences are non-protein coding.

Microbial content in HGP

Although HGP samples do not reflect clinical metagenomic sequencing protocols the data provides an important “negative control” for documenting microbial content in human samples, which should have little to no microbial content with clinical relevance. Only five individuals displayed microbial content for a single organism with greater than 10^{-4} ($1.5\text{--}3.4 \times 10^{-3}$) read abundance all for an organism identified as *Acidovorax* sp. KKS102, which likely reflect an environmental contaminant. Abundance is calculated using the number of reads assigned to a taxonomic label divided by the total number of raw sequencer read pairs (i.e. not quality control filtered read counts). Individual microbial contaminants could be detected in 1,783 of the samples at levels below 10^{-4} . Total in sample contamination abundance was found to range from between 0.1% and 1% of the reads in 1,415 of the samples. The majority of these contaminants are tagged as either “synthetic construct”, which reflect known sources of contamination such as cloning vectors and adaptors, and two high rank categories “root” and “cellular organism”, which reflect reads that match to genomes found across the taxonomy tree. While some of these reads reflect true biological conservation (such as phage represented both as a virus and integration in a bacterial chromosome) many of these reads are expected to reflect errors in the reference database. The relatively high abundance of contaminating sequences in the HGP indicate the need to accurately tag contamination so that true low frequency microbes of interest can be clearly separated and reported. While filtering out known contaminants appears to greatly reduce the remaining number of microbial reads, clearly additional work must be done to determine whether the identified microbe is truly associated with the biological sample being examined or a persistent environmental contaminant. Although a detailed analysis of the microbial content of the HGP data is beyond the scope of this report, the data is made available for further analysis and a brief summary of potentially interesting microbial content is included in the Supplemental Material, such as evidence of sample contamination and detection of retrovirus contaminants that were likely contributed by the immortalized cell culture used to store the HGP DNA.

New La Braña human reads

Using LMAT-Grand, the impact of detecting new human variants was examined for a completely different type of human sample - the recently published La Braña metagenome (Olalde et al., 2014) taken from a 7,000 year old human sample where more than half of the reads failed to map to the human reference. With its expanded human genetic variation content, LMAT-Grand database was expected to detect substantially more human reads. All 889,666,990 unmapped reads were searched and compared (Supplemental pie chart in Figure S2) with the Extended Data Figure 9 in Olalde et al. The previously published analysis used a random sampling of 1 million unmapped reads, which were searched against NT using BLAST and all reads were compared against a smaller viral database to search for viruses. The new LMAT pie chart is surprisingly consistent with the predicted version previously reported. Analysis of all unmapped reads (not just a random subset) found 3% more reads could be assigned taxonomic labels than previously determined. Among the unmapped reads, 4% were newly detected to be human, which while comprising a relatively small percentage of the data constituted 37,290,232 reads. Additional analysis of the newly identified human reads will be needed to rule out the possibility of

contemporary contamination. An exceptionally small percentage of the reads (0.00014%) were identified as virus and appear to consist of phage.

Discussion

The increasing use of deep sequencing presents both a challenge and an opportunity for expanding the searchable knowledge base of circulating microbial life while characterizing microbes to support clinical applications (Fricke and Rasko, 2014; Klymiuk et al., 2014). On the one hand it is important to develop tools and databases that extract information from the new genetic variation recovered on an ongoing basis from sequencing, particularly as methods for assembly of new genomes from metagenomic sequencing advance (Albertsen et al., 2013; Nielsen et al., 2014). However, robust and high-throughput methods for error checking become critical to fully exploit the growing collection of observational data. This becomes apparent with a more exhaustive search of human genetic variants and synthetic sequences, which reveal draft microbial assemblies with contamination that can confound reference based microbial identification. This is currently handled with LMAT by expanding the collection of tagged reference sequences. As contaminants are encountered, if they are not explicitly matched to a synthetic construct or host genome, the read is tagged as “root” or “cellular organism” to denote a taxonomically conserved element along with the list of top matches. As future work, an explicit metagenomic profile of every assembled genome is envisioned, which can be used to identify and exclude problematic assemblies. A key limitation of this approach is the need for extensive sequencing of the target environment. The current LMAT database is enriched for genomic data associated with the human target environment. Metagenomic communities isolated from other environments will continue to present additional challenges until more of the host and microbial genomes are better represented in reference databases.

The identification of human STRs in HMP samples indicate that human identification may one day be possible but only with a vastly expanded database of human genomes, which is not currently available. The absence of clear ethnic region identification markers indicate that considerable additional work would be needed to tease out any possibly important host markers in the existing HMP samples. Nonetheless, the LMAT-Region database indicates the potential to identify informative genetic host markers when there is sufficient coverage of the host genome, which could become an increasingly common feature of human metagenomic samples as depth of sequencing coverage grows.

The first published analysis of HMP data reported the use of 8.8 terabases of shotgun metagenomic sequence from 681 biological samples (The Human Microbiome Project Consortium, 2012b). Since that time many more samples have been sequenced and the current collection of 18 terabases from 1,410 samples spread across 9,025 sequencer runs represents a significant expansion of collected data. Comprehensive search of host and microbial populations now present scaling challenges that limit the use of traditional search tools, which rely on examining each reference genome independently. The indexing strategy employed by LMAT tracks conserved sequence elements across the taxonomy tree and presents an important option for fast search but its application has been limited by large memory requirements. Recent advances using extended memory hierarchies that use lower cost flash drives show how exceptional performance can be achieved at dramatically lower per compute node costs and can support genomic search on a much larger scale than previously possible. For example, the taxonomic and gene content profiles for the 18 terabases of HMP data were completed in 37.7 hours on

a data intensive cluster (see Methods) and represent to our knowledge the first analysis for this large collection of HMP data.

There is a growing recognition of the challenges in using metagenomic sequencing for detecting low abundance pathogens (Naccache et al., 2013; Salter, 2014). The abundance of microbial content found in the HGP data further demonstrates the ongoing challenge to differentiate clinically relevant low abundance microbial content from other sources of biological contamination. Thus, the advances in methods to efficiently and accurately identify the complete profile of microbial contents within a sample must be accompanied with new strategies to avoid misattribution from unexpected sources.

Methods

Computational resource

The computational resource for processing the HGP and HMP data sets using LMAT is the Catalyst cluster within the Livermore Computing (LC) center. This resource provided 304 compute nodes with 24 Intel Ivy Bridge cores with 128 GB of DRAM and 800 GB PCIe attached local flash storage (NVRAM) per node. A complete copy of an LMAT database is stored on the flash drive of each node and accessed directly as if it were stored in DRAM with database contents being cached in DRAM. The open source Data Intensive Memory Map (DI-MMAP) (Van Essen et al., 2012; Van Essen et al., 2013) linux kernel module was used to support efficient DRAM caching of NVRAM pages. Pre-processing of the HMP data for quality control did not require access to a larger memory resource and therefore, was done using a more standard Linux cluster Aztec, a 96 node (Intel Xeon 5660, with 12 cores) and 48 GB DRAM per node.

Search database pruning

To maintain the ability to scale as more reference genome sequences are considered, the reference database uses a previously unpublished feature of LMAT that reduces the number of taxonomic identifiers pertaining to particular k -mers in the database. Earlier versions of the software assigned the taxonomy root to k -mers that have a larger number of taxonomy identifiers (IDs) than a cutoff parameter (originally set to 50). While this approach reduces the size of the database and improves runtime, as long lists of IDs produce slow LMAT runtimes, it sacrifices accuracy by removing the pertinent taxonomic information for the k -mers. The new approach, referred to as "pruning", removes all identifiers for the lowest rank first (e.g. strain), and then increasing up the taxonomy hierarchy (e.g. next species, then genus) as needed. We apply this approach at database creation, assuming that the identifiers have been preprocessed to include the identifiers for all ancestor taxons up to the LCA. Every k -mer that refers to more taxonomy identifiers than a specified cutoff parameter will be subjected to the pruning procedure. The taxonomy identifiers for the k -mer are placed in a priority queue data structure ordered by decreasing rank specificity (depth in the taxonomy tree). The data structure enables removal of all the taxonomy IDs of a particular rank until the total number of identifiers remaining in the queue is below the cutoff parameter, which becomes the reduced list of taxonomy identifiers for the k -mer in the database. The current database uses a cutoff value of 200 to limit the overall size of the database.

Assembled human genome data processing

From assembled human reference genome version 38, we extracted canonical *k*-mers using Jellyfish 2.0 (Marçais and Kingsford, 2011). The utility to create a searchable LMAT database merges a stream of *k*-mers in ASCII format. *k*-mers output from Jellyfish are sorted alphabetically and taken uniquely (unix uniq utility). These *k*-mers were merged with the LMAT microbial database to form the LMAT-Ref.

To create the LMAT-GenBank database, the GenBank human identified nucleotide sequences were retrieved using the NCBI get utilities. The 72 GB of sequence was run through the same Jellyfish *k*-mer extraction procedure and merged with LMAT-Ref. The LMAT-GenBank database then was used as the starting point for creating the LMAT-Grand database as outlined below.

Processing HGP data

To reduce the risk of errors, only “Phase 3” data as provided here: <ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/analysis.sequence.index> was used for analysis, after being downloaded via ftp over the course of several weeks. Nonetheless, a major impediment to incorporating unbiased (e.g. non reference based) variation from raw reads is the risk of mislabeling non-human contaminants as novel human variants. Although extensive QC was undertaken during human genome (HG) sequencing, the potential for non-human contamination cannot be ruled out. Therefore, reads were initially compared against an LMAT database to identify reads with 20-mer matches to the existing LMAT human genome reference. Reads that shared no 20-mers with the existing HG reference were identified as possible non-human reads and were explicitly classified with LMAT using its full microbial database. Reads assigned a minimum score of 1 (higher confidence) for a bacterial, viral or archaea taxa were excluded. Reads assigned a score of < 1 were included as human candidates with all other reads that had 1 or more human 20-mers present. (Reads with every 20-mer already matched to the existing HG reference were set aside and ignored.) Illumina guidelines recommend the use of minimum 30x coverage of minimum Q30 bases to infer a human SNP². We adopted this strategy by applying a Q30 mask to all reads (e.g. all bases with Q-value <30 were set to N), thus only 20-mers comprised exclusively of Q30 or greater were considered. This is much more restrictive than requiring that the single query base adhere to the Q-value threshold. This is an inherent limitation of operating in *k*-mer space. The computational efficiency afforded by the use of *k*-mers nonetheless expands the candidate reference set to compensate for the more restrictive quality filtering. In addition, a novel candidate 20-mer must occur at least 30 times in a human sample to be considered valid. While this biases toward identification of repetitive sequence and reduces the variants recovered from low coverage samples, we opted to err on the side of conservative selection criteria. Finally, to further reduce possible low level contaminants and false positives a novel 20-mer was required to occur in at least 4 human samples.

We implemented a workflow to handle preprocessing and post-processing steps around LMAT using shell scripts, which would operationally integrate into a batch job scheduling environment. This workflow processed all sequence files for a particular sample as input, and produced a set of ASCII *k*-

² http://res.illumina.com/documents/products/technotes/technote_snp_caller_sequencing.pdf

mers as output. Below we describe the steps taken within the workflow. The workflow was designed to run as a single "compute" job within the batch scheduling environment.

There were two passes of LMAT analysis on the HGP data. The first pass of LMAT ran concurrently with the preprocessing steps (via shell script pipe) where data was decompressed (unzip) and filtered for quality using seqtk with -Q 30. In this step, LMAT ran against a precursor to the LMAT-Ref database (an earlier version of LMAT-Ref using genomic data from 9/4/2013). To support identification of potentially new human *k*-mers, we made a simple modification to LMAT to report the number of *k*-mers retrieved from the database. Any reads that had *k*-mers not matching those in the database would be considered new for the purpose of integration.

Three post-processing scripts written in awk were used to scan through output from the first pass of LMAT. In the first post-processing task, reads that contain valid *k*-mers and one or more *k*-mers that do not match the database were pooled by sample identifier and set aside to be considered as potentially novel human content. Second, reads that do not contain *k*-mers matching human in our database but show strong enough matches to other organisms were set aside for future analysis. An additional awk script counts reads that produce 100% matches to human for all valid *k*-mers.

One of the challenges facing the processing of the samples is the diversity in size. 10 samples were larger than 100 GB in total data size so partitioning of the sample was necessary for a compute job to complete the workflow processing within a reasonable timeframe. Some human samples consist of several hundred constituent sequencer runs (input files). Our approach to partitioning each sample was to split the list of input files assuming that the distribution of file sizes would produce reasonable partitioning for each sample. A lesson learned from this approach is that many of the individual file sizes were even too large, so a better approach (employed for LMAT large runs using cluster resources) is to split the fasta (or fastq) on a per-read basis. The average time per sample was 16 hours but the longest jobs took 96 hours and some remained running on the compute nodes after most jobs completed. Because our workflow specific for this process started with compressed files, partitioning raw sequence data files does not readily apply; thus a future implementation would need to revise the workflow.

Jellyfish was used to enumerate all the canonicalized *k*-mers found in the reads determined in the post-processing to contain potentially novel human content. Following the 30x coverage requirement, for each sample, each *k*-mers must be counted at least 30 times. Distinct *k*-mers from Jellyfish are output as ASCII text, and the unix sort utility is used to produce a set for the individual sample that can be merged with other samples.

In a process that took several hours, the 2,646 input files containing sorted ASCII *k*-mers were merged (sort -m) in batches of up to 10 files per merge process. The distribution of number of *k*-mers found to occur in differing numbers of samples was examined and a minimum of 4 individuals was chosen as a cutoff to balance capturing a large number of *k*-mers while avoiding erroneous *k*-mers that may arise in a small number of samples. We used the unix uniq tool with -c to count occurrences in the merged output and with a single-line awk script, produce such output. This set of *k*-mers generated from the 1000 Genomes Project samples are merged with the previously extracted sets of human *k*-mers and added to the searchable LMAT reference database as done in previous steps for the GenBank human gene sequence collection.

As a second pass, microbial content profiles for the HGP data were generated using the LMAT-Grand database using all candidate new human reads identified in the first pass screening of the HGP dataset. The total processing time for the HGP dataset was 43,600 node hours (or 1,046,400 CPU hours), which could be completed in 6 days on the Catalyst cluster.

HMP data processing

All available HMP files were downloaded from the NCBI sequence read archive (<http://www.ncbi.nlm.nih.gov/bioproject/43017>) for Nov. 27 2013 using the Aspera download utility (SRA Handbook 2014), downloaded over the course of 66 hours. 9,113 SRA files were converted to fastq using the SRA SDK 2.3.2-4 and options `fastq-dump –split-files –split-spot –skip-technical –dumpbase`. `fastq-mcf` revision 488 (Aronesty) was used to trim reads using a list of Illumina sequence adaptors and discard reads of length less than 35. Reads were then N masked to ignore base calls with quality scores below 10 and reads with less than 35 valid bases were discarded. Paired reads were then merged to form a single contiguous read for LMAT processing. 9,025 sequencer runs were available for LMAT processing after quality filtering. All steps prior to running LMAT were run on the Aztec cluster described above and took 2,851 node hours, 820 of these hours were spent extracting the fastq format from the SRA files. A total of 10,046 node hours were used to generate the HMP taxonomic profiles, which could be completed with the Catalyst cluster in 33 hours. The gene content profiles were generated in 1,430 node hours and completed in 4.7 hours on the Catalyst cluster.

The average runtime for LMAT-Grand, LMAT-ML and LMAT-Gene was 17 kilobases per second per core (Kbp/s), 198 kbp/s and 63 kbp/s respectively. This translates to approximately 4.1 hours to analyze a 10 gigabase dataset on a 40-core machine for LMAT-Grand, 21 minutes for LMAT-ML and 1.1 hours for LMAT-Gene.

To compare output from different databases on a smaller subset of HMP samples, a collection of 131 sequencer runs were selected by randomly choosing up to 6 runs from the 19 body sites evenly divided by sex where appropriate. To examine accuracy more closely, the subset (6) of samples with existing taxonomic reports available through HMP DACC that could be matched exactly with samples in the 131 HMP set were compared (SRS052620, SRS022719, SRS053917, SRS052668, SRS057083, SRS022713). Due to the small sample size (6), a second collection of 73 samples non-overlapping with the 131 representative set were compared and showed similar overlap, thus the original collection was taken as a representative sample. The HMP DACC derived profiles used a custom read mapping and taxonomy calling pipeline and the following summary is based on the supplemental text provided by Martin et al., 2012. Briefly, a microbial database was constructed from archaeal, bacterial, viral and lower eukaryote organisms (including draft genomes). A custom procedure was developed to remove possibly redundant genomes, based on removing genomes with 90% similarity. Reads were mapped to genomes in the database using CLC bio aligner requiring 80% identity over 75% of the read. The minimum threshold for reporting a strain present is a coverage cutoff of 0.01x depth across 1% of the strain's genome.

The most recent available version of MetaPhlAn (MetaPhlAn2 version 2.0.0 beta3) was used to generate MetaPhlAn profiles reported in Figure 4 using default parameter settings. For clustering, version 4.5.4 was used with default settings. NCBI BLAST version 2.2.27+ was used for all BLAST based analysis.

For comparing differing taxonomic calls between different software, blastn was run with settings: -evalue 0.0001 -max_target_seqs 5.

Searchable gene database

In addition to the genome databases used for organism identification, a searchable gene database was created (LMAT-Gene in Table 1) using a 20-mer index and using sequences obtained from April 22 2014, consisting of GenBank NT genes and annotated genes from microbial genomes. Gene sequences were extracted from genomes and labeled with gene ID by combining information from the genome gff files in the NCBI genomes database (<ftp://ftp.ncbi.nih.gov/genomes/>), start and stop positions enumerated in the gene2accession file at NCBI (<ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2accession.gz>), and sequences in .ffn and .frn files at the NCBI genomes subdirectories for Bacteria, Viruses, Plasmids, Fungi, and Protozoa using custom perl and unix scripts. Single gene entries in NT were extracted using the BLAST blastdbcmd with a list of gene gi's obtained from the gene2accession file. The gene database includes 23,604,714,457 nucleotides for 14,020,775 genes and the downloadable search index is 120 gigabytes. Instead of assigning a taxonomic label to each read, an NCBI gene identifier is assigned. Genes are reported on a best first match basis and output with a minimum of 10% of the 20-mers matching a reference gene. In addition to providing gene content profiles that are independent of taxonomic assignments, the subset of reads that are assigned taxonomic identifiers with match scores above the minimum score cutoff (0) are summarized for their gene content to cross link the gene content profiles with the taxonomy content profiles.

Plasmid identification

Evolutionary characterization of plasmids is complicated by the potential for a plasmid to be found in multiple organisms. In addition, a plasmid may be isolated under different experimental conditions, sequenced and assigned different labels. A further complication is that draft genome sequencing projects may not identify the plasmid but include it as a segment along with other chromosomal segments. Thus, LMAT provides "rudimentary" plasmid identification, by explicitly tagging all known plasmid sequences during its database creation phase. At runtime when a taxonomic assignment is made, the list of top scoring taxonomic labels is checked for the presence of plasmids. If a plasmid is found to be among the top scoring candidates, the first encountered top scoring plasmid is chosen to replace the default taxonomic assignment. It should be noted that multiple plasmids may be assigned the same score but only one is reported. Therefore, use of plasmids for organism identification should be considered carefully with evidence for the organism from chromosomal information. Nonetheless, it is expected that the utility to explicitly identify genetic segments that are known to be associated with plasmids will be useful for downstream analysis.

World regional classification

The LMAT-Region database used a modified NCBI taxonomy hierarchy that designated "species", "region", and "ethnic" as the named ranks. Homo sapiens was retained as the sole NCBI designated species, with the five world regions included as child nodes and the 27 ethnic groups (The 100 Genomes Project Consortium, 2010) included as children nodes of their respective parent region.

Given the ethnic identification of the 2,646 samples all *k*-mers from samples with same ethnicity that appear in 4 or more samples were selected as ethnic identifiers (N=1-4 was considered). These *k*-mer sets were combined into a single multi-fasta file to be input to LMAT's database creation pipeline (LMAT manual v. 1.2.4), in which each fasta header contains only the taxonomy identifier for the ethnicity. The "pruning" threshold was set to 1 to produce *k*-mers with single taxonomy identifiers

For validation of the database 500 individuals were chosen from the HGP selected to be a random sampling from the 27 ethnic groups. Region classification was configured to use only reads with 100% valid *k*-mer matches to a particular ethnicity or region. Region calls were made based on taking the ratio between the first and second most abundant region calls using read counts. Regional read counts are obtained by combining the individual ethnic specific read counts. Null models were considered but did not appear to improve results and were excluded from the analysis.

Data access

- 1) Genus (Supplemental File 6 and 10), species/strain (Supplemental File 5 and 8), and plasmid (Supplemental File 7 and 9) taxonomy profiles are available for the complete collection of HMP and HGP sequencer runs respectively.
- 2) Gene content profiles for all HMP sequencer runs with a cross-reference to taxonomy calls are available at ftp://gdo-bioinformatics.ucllnl.org/pub/lmat/hgp_hmp_2014/hmp.gene_cl_profiles.tar.gz.
- 3) All reads with no match to the database in the HMP collection are available as Supplemental File 4.
- 4) New human reads found in the La Braña genome are available through the NCBI Sequence Read Archive under accession SRX992275. La Braña metagenomic taxonomy profiles are provided in Supplemental File 11.
- 5) The LMAT software is freely available as open source at lmat.sourceforge.net and version 1.2.4 used in this study is included as Supplemental File 2.
- 6) Databases are available for download via anonymous ftp and are retrieved using the `get_db.sh` script included in the LMAT software distribution. The databases are listed with their actual file name, which are used as input to download a specific database. LMAT-Grand (`lmat-4-14.20mer.db`), LMAT-ML (`kML.v4-14.20.g10.db`), LMAT-ML-Human (`kML+Human.v4-14.20.g10.db`), LMAT-Genes (`lmat.genes.7-14.db`) and LMAT-Region (`lmat-world-region.db`).
- 7) Human repeat sequences and STRs are available as Supplemental File 3 and Supplemental File 12 respectively.
- 8) The HMP DACC staff is in the process of removing the human reads from the public HMP data, the reads will be made available through the HMP dbGAP study.

Acknowledgements

This work was performed under the auspices of the US Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. Laboratory Directed Research and Development (33-ER-2012 and 08-ER-2011); Livermore Computing is acknowledged for providing

early access and technical support to enable use of the Catalyst cluster for this study. We thank Iñigo Olalde for making the unmapped La Braña reads available for analysis.

Conflict of Interest: none declared.

Reference

- Albertsen, M., Hugenholtz, P., Skarshewski, A., Nielsen, K.L., Tyson, G.W., and Nielsen, P.H. (2013). Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat Biotech* 31, 533–538.
- Allen, J.E., Gardner, S.N. and Slezak, T.R. (2008). DNA signatures for detecting genetic engineering in bacteria. *Genome Biol* 9, R56.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.
- Ames S, Allen JE, Hysom DA, Lloyd GS, Gokhale MB. 2014. Design and Optimization of a Metagenomics Analysis Workflow for NVRAM. In *IEEE 28th International Parallel & Distributed Processing Symposium Workshops* pp. 556 - 565.
- Ames SK, Hysom DA, Gardner SN, Lloyd GS, Gokhale MB, Allen JE. 2013. Scalable metagenomic taxonomy classification using a reference genome database. *Bioinformatics* 29, 2253–2260.
- Aronesty E. 2013. Comparison of sequencing utility programs. *TOBioiJ*
DOI:10.2174/1875036201307010001
- Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2013. GenBank. *Nucleic Acids Res* 41, D36–D42.
- Berendzen J, Bruno W, Cohn J, Hengartner N, Kuske C, McMahon B, Wolinsky M, Xie, G. 2012. Rapid phylogenetic and functional classification of short genomic fragments with signature peptides. *BMC Res Notes* 5, 460.
- Byrd A, Perez-Rogers J, Manimaran S, Castro-Nallar E, Toma I, McCaffrey T, Siegel M, Benson G, Crandall K, Johnson W. 2014. Clinical PathoScope: rapid alignment and filtration for accurate pathogen identification in clinical samples using unassembled sequencing data. *BMC Bioinformatics* 15, 262.
- Cotten M, Oude Munnink B, Canuti M, Deijs M, Watson SJ, Kellam P, van der Hoek L. 2014. Full Genome Virus Detection in Fecal Samples Using Sensitive Nucleic Acid Preparation, Deep Sequencing, and a Novel Iterative Sequence Classification Algorithm. *PLoS ONE* 9, e93269.
- Elhaik E, Tatarinova T, Chebotarev D, Piras IS, Maria Calò C, De Montis A, Atzori M, Marini M, Tofanelli S, Francalacci P, et al. 2014. Geographic population structure analysis of worldwide human populations infers their biogeographical origins. *Nat Commun* 5:3513.

Van Essen B, Hsieh H, Ames S, Gokhale M. 2012. DI-MMAP: A High Performance Memory-Map Runtime for Data-Intensive Applications. *High Performance Computing, Networking, Storage and Analysis* pp. 731–735.

Van Essen B, Hsieh H, Ames S, Pearce R, Gokhale M. 2013. DI-MMAP—a scalable memory-map runtime for out-of-core data-intensive applications. *Clust Comput* 1–14.

Fricke WF, Rasko DA. 2014. Bacterial genome sequencing in the clinic: bioinformatic challenges and solutions. *Nat Rev Genet* 15, 49–55.

Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28, 3150–3152.

Gymrek M, Golan D, Rosset S, Erlich, Y. 2012. lobSTR: A short tandem repeat profiler for personal genomes. *Genome Res* 22, 1154–1162.

Howe AC, Jansson JK, Malfatti SA, Tringe SG, Tiedje JM, Brown CT. 2014. Tackling soil diversity with the assembly of large, complex metagenomes. *Proc Natl Acad Sci* 11 (13) 4904–4909

Hu H, Hu Y, Pan Y, Liang H, Wang H, Wang X, Hao Q, Yang X, Yang X, Xiao X, et al. 2012. Novel Plasmid and Its Variant Harboring both a blaNDM-1 Gene and Type IV Secretion System in Clinical Isolates of *Acinetobacter lwoffii*. *Antimicrob Agents Chemother* 56, 1698–1702.

Huang L, Popic V, Batzoglou S. 2013. Short read alignment with populations of genomes. *Bioinformatics* 29, i361–i370.

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res* 110, 462–467.

Klymiuk I, Högenauer C, Halwachs B, Thallinger GG, Fricke WF, Steininger C. 2014. A Physicians' Wish List for the Clinical Application of Intestinal Metagenomics. *PLoS Med* 11, e1001627.

Langdon W. 2014. Mycoplasma contamination in the 1000 Genomes Project. *BioData Min.* 7, 3.

Laurence M, Hatzis C, Brash DE. 2014. Common Contaminants in Next-Generation Sequencing That Hinder Discovery of Low-Abundance Microbes. *PLoS ONE* 9, e97876.

Liu B, Gibbons T, Ghodsi M, Treangen T, Pop M. 2011. Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. *BMC Genomics* 12, S4.

Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics* 27, 764–770.

Martin J, Sykes S, Young S, Kota K, Sanka R, Sheth N, Orvis J, Sodergren E, Wang Z, Weinstock GM, et al. 2012. Optimizing Read Mapping to Reference Genomes to Determine Composition and Species Prevalence in Microbial Communities. *PLoS ONE* 7, e36427.

Minot ST, Turner SD, Ternus KL, Kadavy DR. 2014. SIANN: Strain Identification by Alignment to Near Neighbors. *bioRxiv* doi:10.1101/001727.

Naccache SN, Greninger AL, Lee D, Coffey LL, Phan T, Rein-Weston A, Aronsohn A, Hackett J, Delwart EL, Chiu CY. 2013. The Perils of Pathogen Discovery: Origin of a Novel Parvovirus-Like Hybrid Genome Traced to Nucleic Acid Extraction Spin Columns. *J. Virol.*

Naccache SN, Federman S, Veeeraraghavan N, Zaharia M, Lee D, Samayoa E, Bouquet J, Greninger AL, Luk K.-C, Enge B, et al. 2014. A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. *Genome Res* Published in Advance June 4, 2014, doi:10.1101/gr.171934.113

Nakamura S, Yang C.-S, Sakon N, Ueda M, Tougan T, Yamashita A, Goto N, Takahashi K, Yasunaga T, Ikuta K, et al. 2009. Direct Metagenomic Detection of Viral Pathogens in Nasal and Fecal Specimens Using an Unbiased High-Throughput Sequencing Approach. *PLoS ONE* 4, e4219.

Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, Plichta DR, Gautier L, Pedersen AG, Le Chatelier E, et al. 2014. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat Biotech* 32(8):822-8.

Olalde I, Allentoft ME, Sanchez-Quinto F, Santpere G, Chiang CWK, DeGiorgio M, Prado-Martinez J, Rodriguez JA, Rasmussen S, Quilez J, et al. 2014. Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European. *Nature* 507, 225–228.

Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, Parkhill J, Loman NJ, Walker AW. 2014. Reagent contamination can critically impact sequence-based microbiome analyses. *BMC Biology* 12:87.

Schmieder R, Edwards R. 2011. Fast Identification and Removal of Sequence Contamination from Genomic and Metagenomic Datasets. *PLoS ONE* 6, e17288.

Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. 2012. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Meth* 9, 811–814.

SRA Handbook (2014). SRA Handbook: Aspera Transfer Guide (Bethesda (MD): National Center for Biotechnology Information).

Sunagawa S, Mende DR, Zeller G, Izquierdo-Carrasco F, Berger SA, Kultima JR, Coelho LP, Arumugam M, Tap J, Nielsen HB, et al. 2013. Metagenomic species profiling using universal phylogenetic marker genes. *Nat Meth* 10, 1196–1199.

Takeuchi F, Sekizuka T, Yamashita A, Ogasawara Y, Mizuta K, Kuroda M. 2014. MePIC, Metagenomic Pathogen Identification for Clinical Specimens. *Jpn J Infect Dis* 67, 62–65.

The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* 467, 1061 – 1073.

The 1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491, 56–65.

The Human Microbiome Project Consortium. 2012a. Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214.

The Human Microbiome Project Consortium. 2012b. A framework for human microbiome research. *Nature* 486, 215–221.

Tu Q, He Z, Zhou J. 2014. Strain/species identification in metagenomes using genome-specific markers. *Nucleic Acids Res* 42 (8): e67.

Willems TF, Gymrek M, Highnam G, The 1000 Genomes Project, Mittelman D, Erlich Y. 2014. The landscape of human STR variation. *Genome Res* Published in Advance August 18, 2014, doi:10.1101

Wilson MR, Naccache SN, Samayoa E, Biagtan M, Bashir H, Yu G, Salamat SM, Somasekar S, Federman S, Miller S, et al. 2014. Actionable Diagnosis of Neuroleptospirosis by Next-Generation Sequencing. *N Engl J Med* 370, 2408–2417.

Wood D, Salzberg, S. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* 15, R46.

Yozwiak NL, Skewes-Cox P, Stenglein MD, Balmaseda A, Harris E, DeRisi JL. 2012. Virus Identification in Unknown Tropical Febrile Illness Cases Using Deep Sequencing. *PLoS Negl Trop Dis* 6, e1485.

Zhao G, Krishnamurthy S, Cai Z, Popov VL, Travassos da Rosa, AP, Guzman H, Cao S, Virgin HW, Tesh RB, Wang D. 2013. Identification of Novel Viruses Using VirusHunter -- an Automated Data Analysis Pipeline. *PLoS ONE* 8, e78470.

Zhao Y, Tang H, Ye Y. 2012. RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics* 28, 125–126.

Tables

Table 1. Summary of examined LMAT databases. The full microbial genome database includes all draft and complete genomes from virus, bacteria, archaea, fungi, and protozoa from 04/11/2014. Vector sequence, Illumina adaptors, and a phiX genome are included and tagged as ‘synthetic construct’.

Database Name	Description	Size (GB)
LMAT-Ref	Microbial Genomes plus Human reference genome (hg19)	455.31
LMAT-GenBank	LMAT-Ref plus GenBank Human	457.57
LMAT-Grand	LMAT-GenBank plus HGP sequences	457.62
LMAT-RefSeq	LMAT-Grand minus draft microbial genomes	82.47

LMAT-ML	Microbial only marker library	16.70
LMAT-ML-Human	Microbial markers plus all human sequences	17.52
LMAT-Genes	All NT genes plus genes from microbial genomes	120.00
LMAT-Region	Human-only region specific markers	1.35

Table 2. Number of new human k -mers ($k=20$) added to searchable database and total k -mers in the database. Counts k -mers added from the human reference assembly (LMAT-Ref), from GenBank (LMAT-GenBank) and the 1000 Genomes Project (LMAT-Grand). Total k -mer count microbial and human sequences. Percent of database

Database	New human k -mers	Total Human k -mers	Total k -mers	Percent human	Percent increase
1 – LMAT-Ref	2,295,537,227	2,295,537,227	31,464,525,440	7.30	-
2 – LMAT-GenBank	209,632,954	2,505,170,181	31,742,908,834	7.89	0.88
3 – LMAT-Grand	7,249,146	2,512,419,327	31,749,680,687	7.91	0.90

Table 3. Percentage of unlabeled reads. See text for a description of the three databases, LMAT-Grand, LMAT-RefSeq and HMP DACC. LowScore are reads assigned a taxonomic label with a match score below the minimum threshold. NoDbHit reads have no match to the reference database. The size of the reference database is given in gigabases (Size Gb).

	LMAT-Grand	LMAT-RefSeq	HMP DACC
LowScore	14.81	42.34	-
NoDbHit	0.05	2.76	12
Size (Gb)	115	9.13	7.65

Table 4. BLAST based taxonomic assignments of reads with different labels assigned by LMAT and HMP DACC output. Values are reported as fraction of total reads.

Organism called by HMP DACC but not LMAT	NT BLAST hits	Bacterial /viral BLAST hits	BLAST matches to synthetic constructs	LMAT classified reads at or above genus	Most common species detected by LMAT (number of reads)	BLAST hits for LMAT species call
<i>Abiotrophia defectiva</i>	0.00	0.04	0.04	0.81	<i>Clostridium clostridioforme</i> (4)	1
<i>Actinobacillus minor</i>	0.00	0.02	0.14	0.44	<i>Haemophilus parainfluenzae</i> (128)	1 (combined reads)
<i>Actinobacillus pleuropneumoniae</i>	0.09	0.13	0.11	0.63	<i>Haemophilus parainfluenzae</i> (81)	
<i>Actinobacillus succinogenes</i>	0.04	0.00	0.11	0.43	<i>Haemophilus parainfluenzae</i> (114)	
<i>Haemophilus ducreyi</i>	0.02	0.00	0.31	0.62	<i>Haemophilus parainfluenzae</i> (62)	

<i>Mannheimia succiniciproducens</i>	0.06	0.02	0.13	0.46	<i>Haemophilus parainfluenzae</i> (107)	
<i>candidate division TM7 single-cell isolate TM7b</i>	0.00	0.19	0.10	0.62	<i>Candidatus Saccharimonas aalborgensis</i> (28)	0.89
<i>Catonella morbi</i>	0.00	0.06	0.08	0.84	<i>Phascolarctobacterium</i> sp. CAG:207 (10)	1

Figure Legends

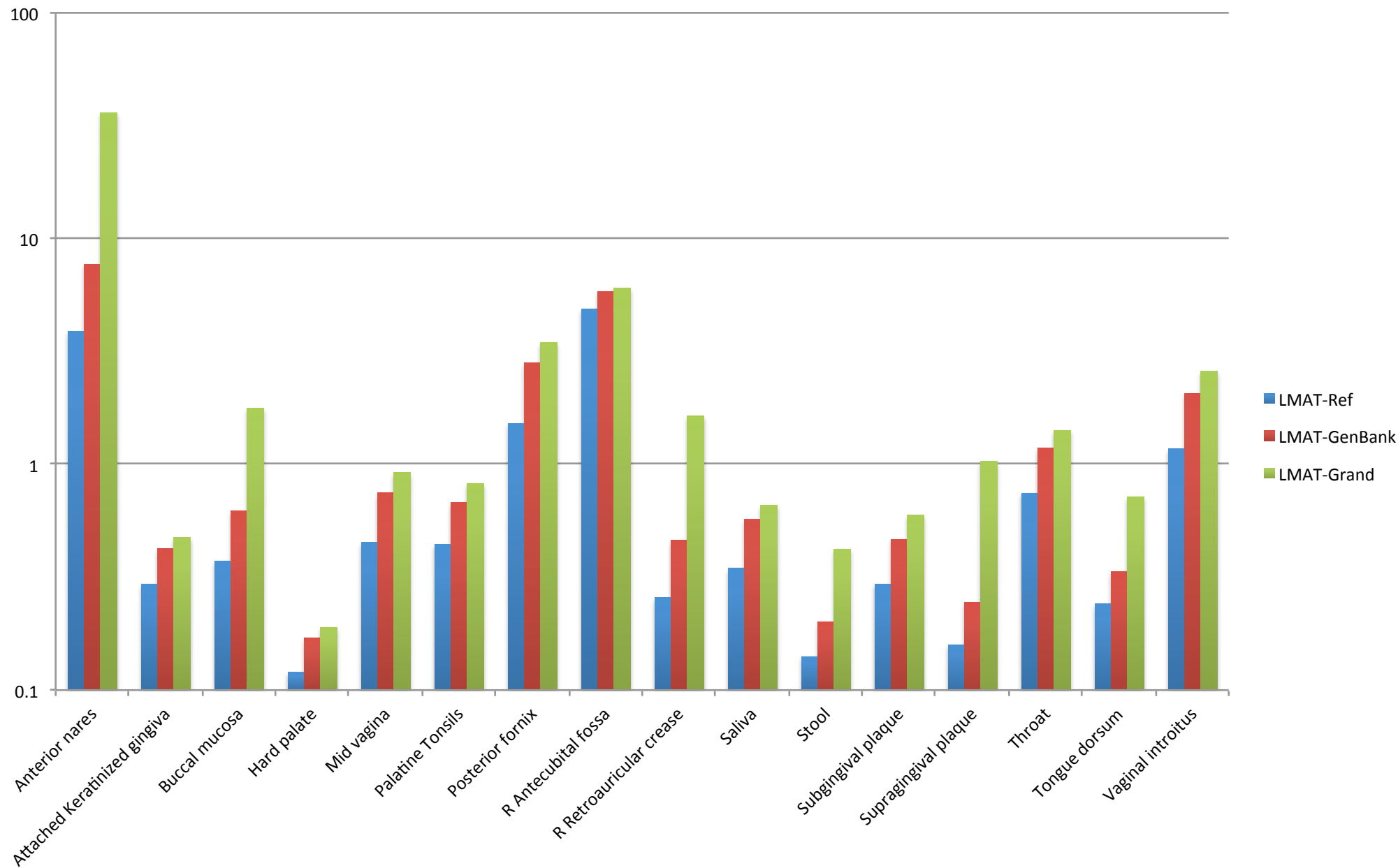
Figure 1. Average percentage of reads identified as human sequence in HMP samples, using LMAT-Ref, LMAT-GenBank or LMAT-Grand by body site.

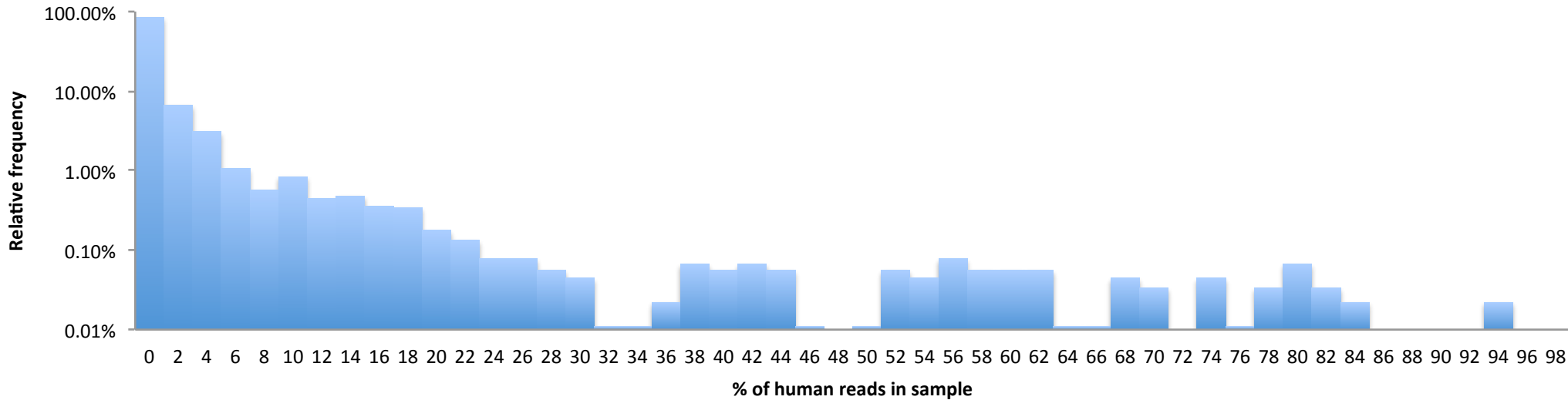
Figure 2. Histogram showing how often different amounts of human reads are found across the collection of sequencer runs. The x-axis reports human read abundance in sequencer runs in bins of two percent. The y-axis shows the percentage of sequencer runs with the amount of human reads specified on the x-axis using a log scale. The highest fraction of human reads in a sequencer run is 94 percent and found in 1 run.

Figure 3. Sensitive BLAST search based assignment of reads from an HMP sample reported to have a high abundance of newly labeled human reads. Left pane shows the distribution of taxonomic assignments after reads were binned into clusters of similar reads. The right pane shows the raw abundance based on read counts for each read assignment. Taxonomic assignments with a 0% abundance label reflect percentages below 1%.

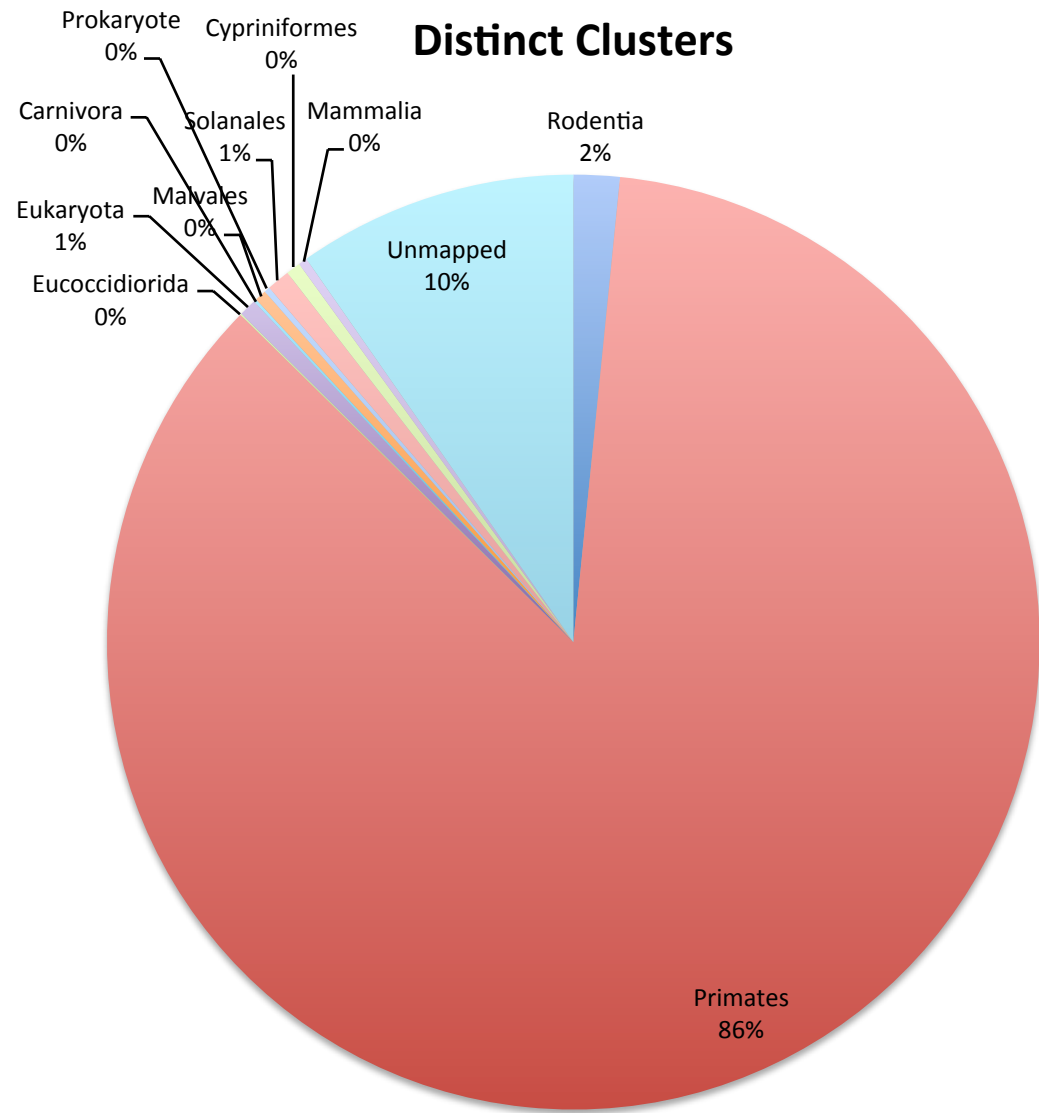
Figure 4. Fraction of shared genus (left panel) and species (right panel) calls. ROC curve shown using different minimum abundance thresholds to make organism calls. Different taxonomy calling methods are shown: HMP DACC, MetaPhlAn and LMAT taxonomy calls with different database types: LMAT-RefSeq (RefSeq), LMAT-ML (ML) and LMAT-ML-Human (ML-humanNoprune).

Average Percentage of Reads labeled Human per Body Site





Distinct Clusters



Abundance

