Supplemental Material

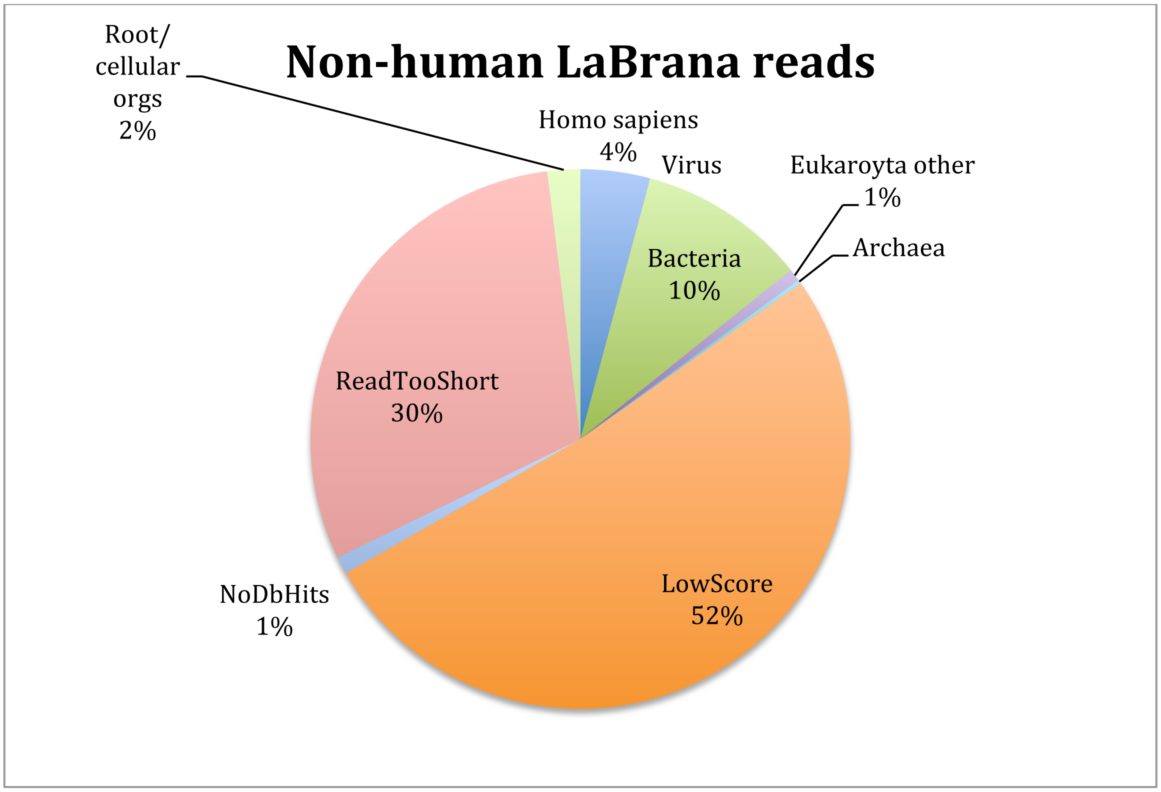
Using populations of human and microbial genomes for organism detection in metagenomes

Sasha K Ames, Shea N Gardner, Jose Manuel Martí, Tom R Slezak, Maya B Gokhale, Jonathan E Allen

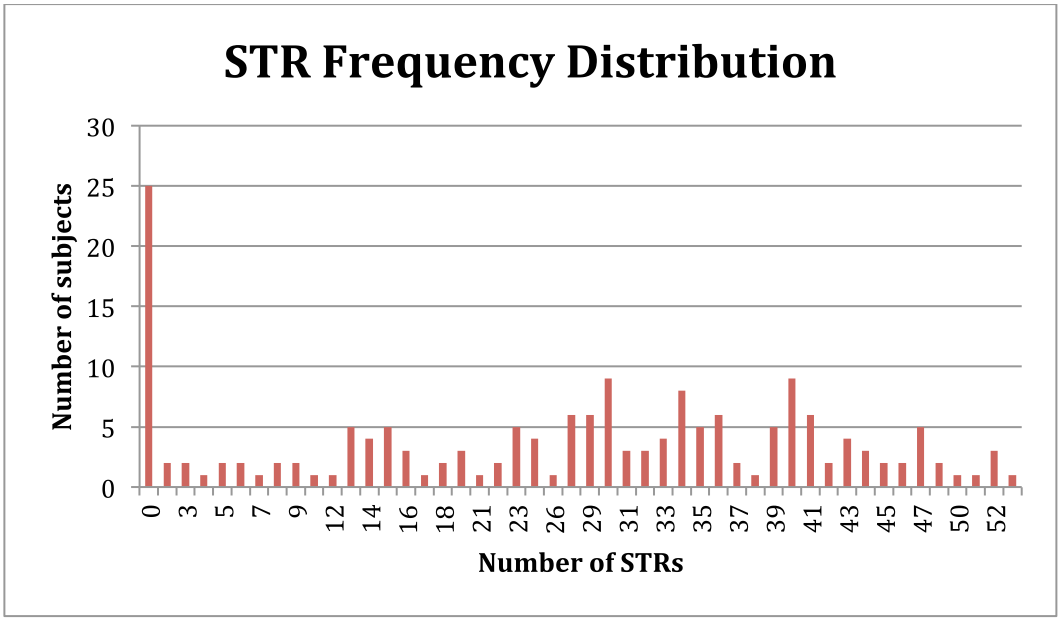
Microbial content of HGP

Epstein Barr virus (*human herpesvirus 4*) was used as part of the cell culturing process and is found in 1,360 samples with an abundance of up to 2.6 x 10^{-5}. While many additional human herpesvirus types could be found in 1,371 samples (type 2, 6A, 6B, 7, and 8) only type 6A and 6B could be found in a small (7) number of samples with an abundance above the frequency at which another retrovirus associated with cell culture contamination was detected (1.8 x 10^{-7}, NA12763). The retrovirus associated with cell culture contamination – *Squirrel monkey retrovirus* (NCBI taxonomy ID 11856) was originally discovered and isolated from Lymphoblastoid cell line (Oda et al., 1988) and is found in 10 human samples and is likely to be a product of the immortalized cell culture, also a Lymphoblastoid cell line used to preserve the HGP DNA. While presumably the low abundance Human herpesvirus is associated with some human infection, given the presence of other viral contaminants from cell culture it seems impossible to exclude the possibility that the herpesvirus were contaminants, inadvertently propagated through cell culture.

Also of interest was the persistence of the fungal pathogen *Cryptococcus neoformans*. Reads were labeled *Cryptococcus neoformans* in 159 individuals total, with 52 individuals having at least 100 fungal reads and up to 1.9 x 10^{-5} of the sample (5,816 reads) contaminated. Although the LMAT reference library contains three different neofomans strains (*grubii H99*, *B-3501A* and *JEC21*) only the *grubii H99* strain is reported.



Supplementary Figure S1: LaBrana reads that did not map to the human reference genome. LowScore refer to reads assigned a taxonomic label with a match score below the default threshold (0). ReadTooShort were reads with too few *k*-mers (minimum 35) to compare. NoDbHits refer to reads that are above the minimum *k*-mer cutoff but could not be assigned any taxonomic label due to lack of matching *k*-mers in the database. Taxonomy labels without an assigned abundance (Virus and Archaea) reflect values that are less than 1%.



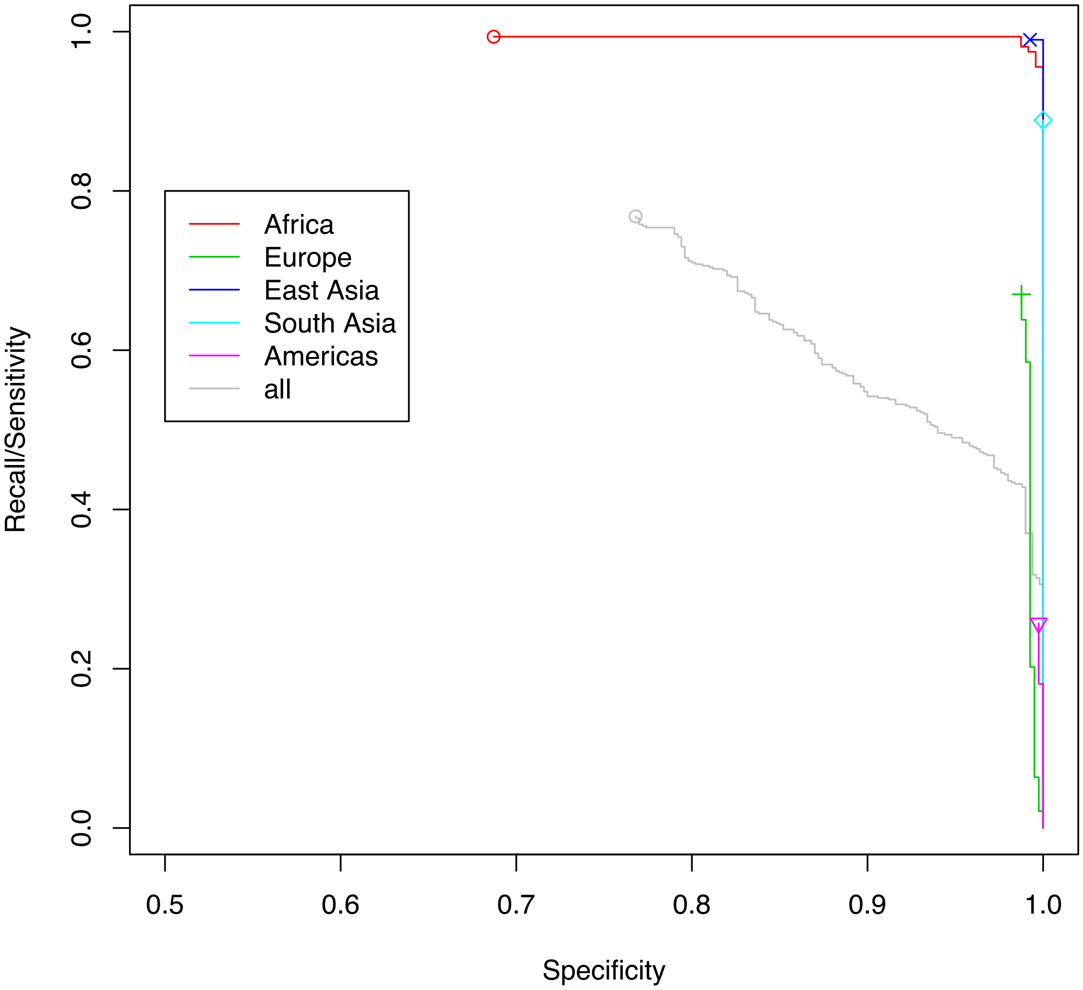
Supplementary Figure S2: Frequency distribution of the number of STRs identified by LobSTR per subject ID from the reads LMAT classifies as human in the HMP data.

Number of STRs	Chromosome
3	chr1
15	chr2
6	chr3
15	chr4
4056	chr5
2	chr6
44	chr7
3	chr8
2	chr9
2	chr10
7	chr11
77	chr12
1	chr13
1	chr14
2	chr15
1	chr16
1	chr17
0	chr18
2	chr19
2	chr20
0	chr21
237	chr22

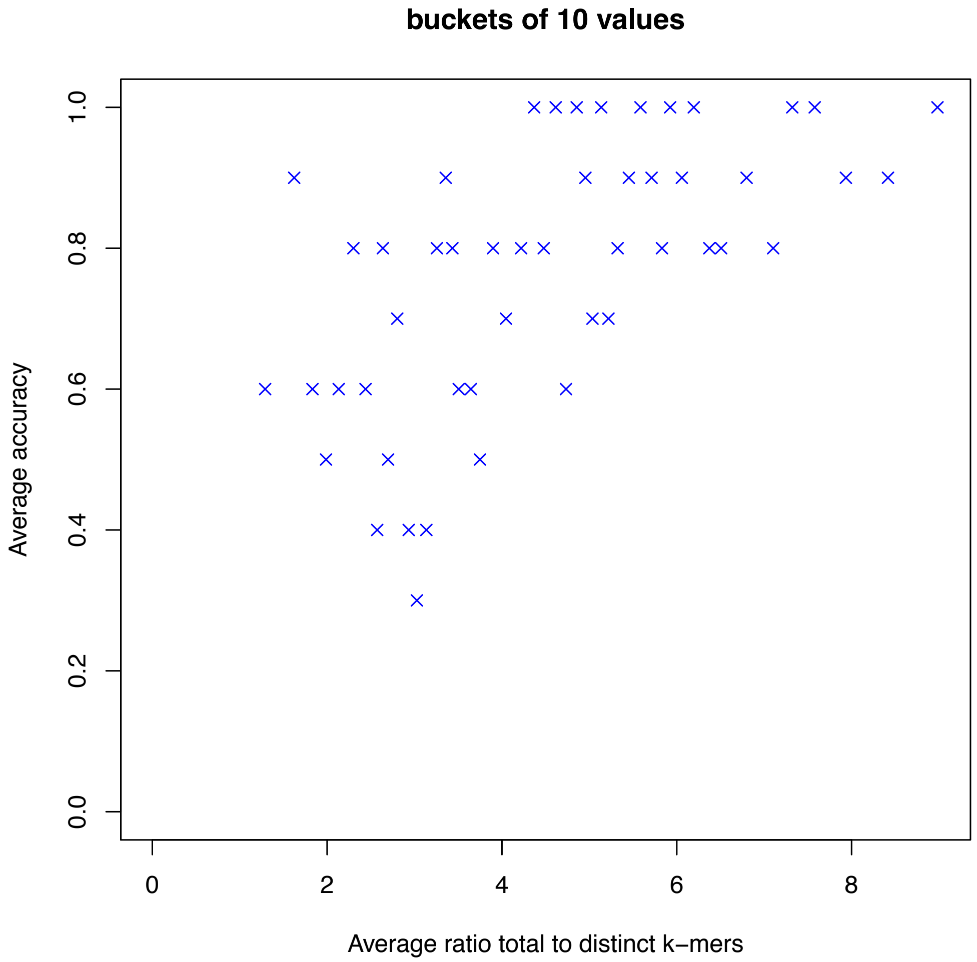
0	chrX
0	chrY

Table S1: Number of STRs per chromosome identified by LobSTR in the human-classified reads from the HMP data.

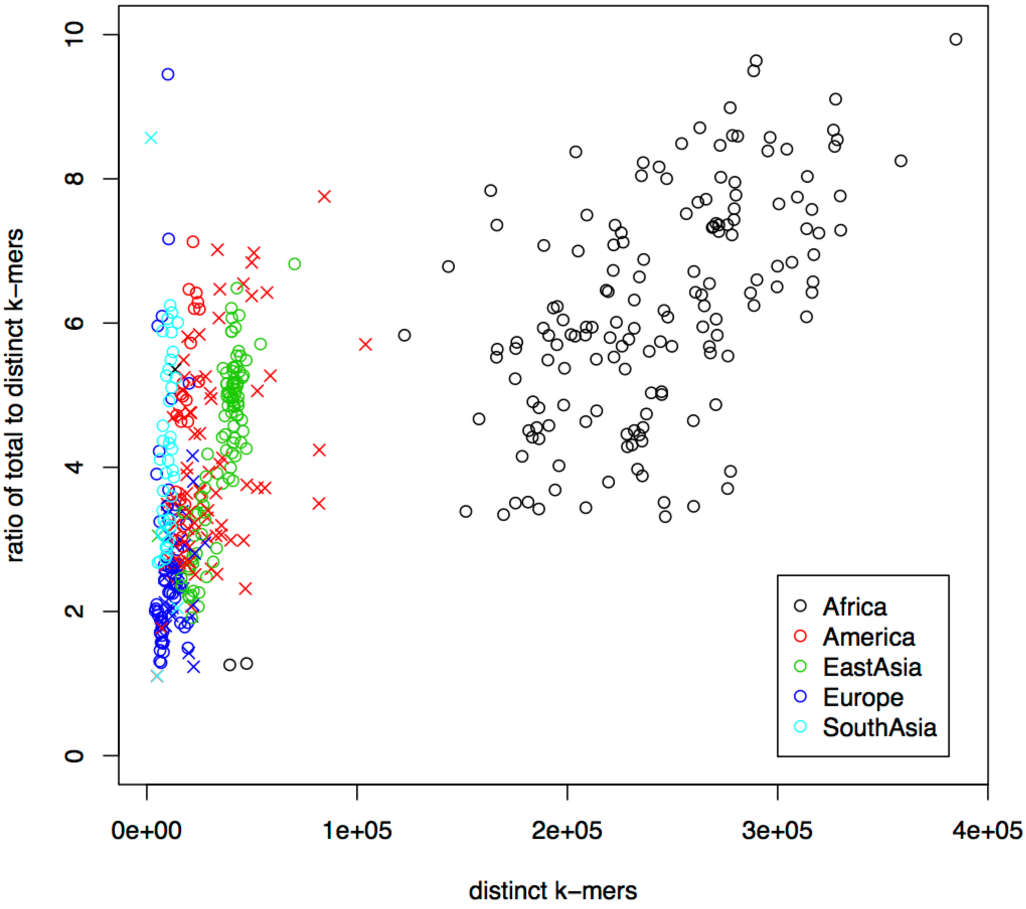
World Regional Classification



Supplementary Figure S3: ROC curves of world Regional LMAT classification. Each point shown with a symbol on the left side of each curve shows the recall (fraction total samples of the specified group identified) and specificity (total not false positives for identifying using a particular group label). We increase specificity through requiring a larger threshold of reads for the called region versus the next region listed in results by read count. Samples labeled "Africa" make up the vast majority of the false positives, while almost perfect recall. Samples from individuals categorized under "Americas" had the lowest recall, and also "Europe" had below average recall, yet relatively few false positives. We can understand the challenge in correct classification for samples from individuals of Americas in light of the diversity for indigenous peoples of Mexico as suggested in [Moreno-Estrada 2014].



Supplementary Figure S4: accuracy of LMAT Regional classification given total to distinct 20-mers. The x-axis measures the average ratio of total 20-mers to distinct 20-mers over 10 samples of close proximity by the ratio value. The y-axis measures the average accuracy --- samples correctly identified --- over the same ten samples. The pattern shown here suggests a trend of increased potential accuracy given the increase in additional 20-mers (as a result of increased coverage).



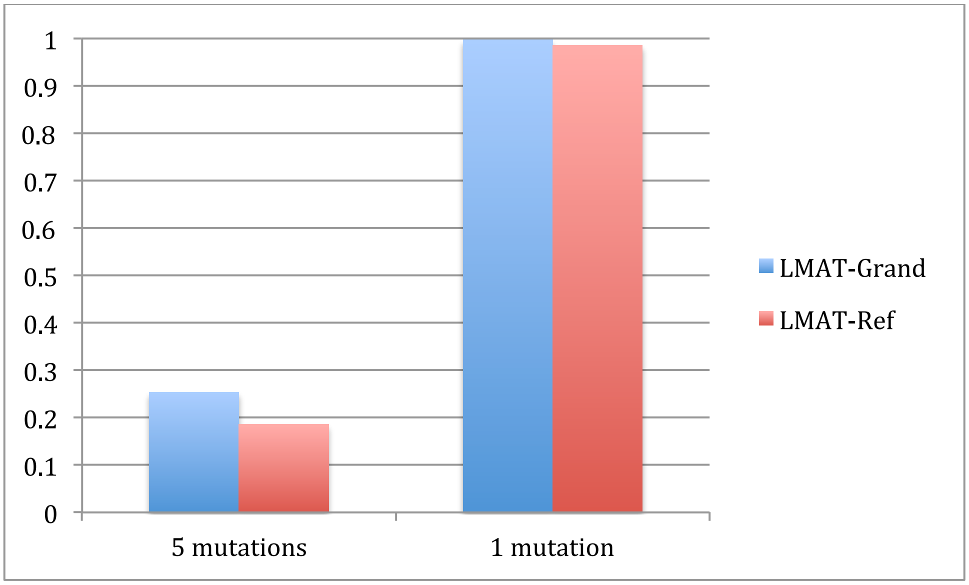
Supplementary Figure S5: individual regional call sample metrics. Each sample point shown measures the distinct 20-mers found to make the classification and the ratio of total to distinct 20-mers used. Each point is colored (see legend) to indicate the actual region of the sample. The point symbol differs depending on a correct "O" or incorrect "X" classification. Notably, most of the African samples had a larger distinct 20-mer count than the other regions, and this group had a relatively higher rate of recall than the others.

Removal of mislabeled human 20-mers from LMAT-Grand

Mislabeled human *k*-mers were taken from LMAT labeled human reads identified using the BLAST from the false positive mock community control data and the 131 HMP sequencer runs. To avoid overaggressive human *k*-mer removal (e.g. true human reads that may be incorrectly included into microbial genes from the NT database), candidate false human 20-mers that are also found in the human reference genome were retained (on the presumption that the reference has already been carefully assembled and screened).

We then repeated the analysis of mock samples to confirm the removal of the false positive human calls and tested the database on a subset of HMP samples (including those with highest human content). The number of human reads dropped from 4,585 to 1,009 with 662 human confirmed and 168 putative false positive reads. False positives can be eliminated by raising the minimum score threshold from 0 to 1. For the total complement of human reads previously identified, we identified and removed 1.5% of the human reads, which also reflected the average human read reduction in the 131 HMP sequencer runs tested. The "cleaned" version of the LMAT-Grand database is released with the corrected newly detected human reads. Since the reduction in human read calls does not appear to change other reported findings, we retain the remaining reported results.

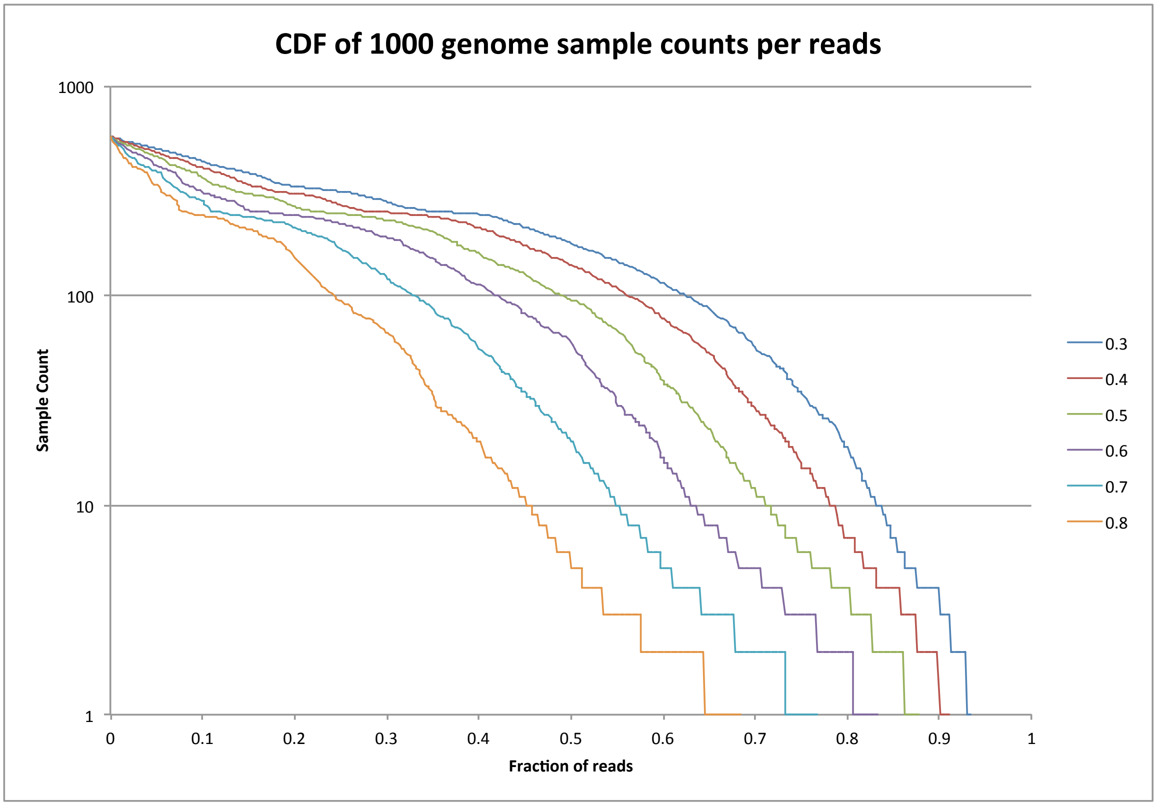
Sensitivity tests with human spike-ins of mock community samples



Supplementary Figure S6. Fraction of human reads correctly labeled with reads containing 1 mutation or 5 mutations.

Using the LMAT world region database, reads from 1000 HGP, which were classified as unique to one of the five world-region classifications, were selected to reflect regional genetic diversity. Two hundred reads from each world region were chosen and only reads with Q30 or higher base calls were included. The software utility *msbar* (Rice et al., 2000) was used to introduce either 1 or 5 mutations (SNPs or other small variations) into each read to simulate varying levels of divergence from the database. Results comparing LMAT-Grand and LMAT-Ref are shown in Supplementary Figure S6. The results show comparable performance with the presence of 1 mutation between the two databases but LMAT-Grand shows a 1.4 fold improvement in number of reads detected with greater simulated divergence introduced to the query set further demonstrating that use of the expanded collection of human sequence supports more robust human read detection in the presence of host genetic diversity.

Novel human reads associated with individual HGP samples



Supplementary Figure S7. Cummulative Distribution Function (CDF) showing the fraction of reads (x-axis) that are associated with different numbers of human samples (y-axis) from the HGP. Reads are associated with a sample based on the minimum number of shared *k*-mers between the read and the sample for different cutoffs ranging from 80% (0.8) down to 30% (0.3).

Since mapping individual reads to the complete 90 terabase collection represents a substantial computational challenge, we created a new human LMAT database that tracks the *k*-mers present in individual samples. We then searched the unidentifiable reads against the newly created LMAT database and track the number of samples each read is found in by using different minimum percent identity cutoffs to consider different confidence levels for associating an individual read with an individual sample. Using the highest cutoff of at least 80% *k*-mer match to an individual sample (Supplemental Figure S7) shows approximately 45% of the reads could be associated with 10 or more samples and 25% of the reads could be associated with 100 or more samples, which gives greater confidence that these reads are reproducible in multiple individuals.

Checking novel human reads for their chimeric potential

To check for evidence of chimeras, the LMAT "raw output" showing all of the candidate matches were checked. Evidence of an identifiable chimera would come when a read shows two similar match scores to two different organisms. Most chimeras will be effectively screened out from being called human through LMAT's lowest common ancestor approach, where a lowest common ancestor taxa is reported only if its score is within one standard deviation of best match. Out of the 295,571 human reads SRR059474, 4418 reads had a second match with 1.5 standard deviations of the best match, which was used to check for possible chimera. 84% of the reads have second best matches to 9 different taxa (with the remaining reads matched to a wide variety of taxa in small numbers). 7 out of the 9 most common taxa are different draft genomes of *Toxoplasma gondii* strain variants (including the largest number of reads), which likely reflect matches to human genomic contamination erroneously included in the draft assemblies. The second abundant category is reads matched with synthetic sequence. Although reads go through a screening step to filter out barcodes, sequence errors can lead to some synthetic sequence remaining. Finally, some reads have a second best match to "root", which is a feature of the LMAT scoring process that indicates a "weak" match and reflects that these reads (903 of 4418) are lower scoring reads closer to the 0 default scoring cutoff used. Taken together, these results do not indicate that chimera represent a large portion of the novel human reads.

References

Oda, t, Ikeda S., Watanabe, Hatsushika M, Akiyama K and Mitsunobu F. "Molecular cloning, complete nucleotide sequence, and gene structure of the provirus genome of a tetrovirus produced in a human lymphoblastoid cell line." *Virology* 167 (2), 468-476 1988.

Andrés Moreno-Estrada, Christopher R. Gignoux, Juan Carlos Fernández-López, Fouad Zakharia, Martin Sikora, Alejandra V. Contreras, Victor Acuña-Alonzo, Karla Sandoval, Celeste Eng, Sandra Romero-Hidalgo, Patricia Ortiz-Tello, Victoria Robles, Eimear E. Kenny, Ismael Nuño-Arana, Rodrigo Barquera-Lozano, Gastón Macín-Pérez, Julio Granados-Arriola, Scott Huntsman, Joshua M. Galanter, Marc Via, Jean G. Ford, Rocío Chapela, William Rodriguez-Cintron, Jose R. Rodríguez-Santana, Isabelle Romieu, Juan José Sienra-Monge, Blanca del Rio Navarro, Stephanie J. London, Andrés Ruiz-Linares, Rodrigo Garcia-Herrera, Karol Estrada, Alfredo Hidalgo-Miranda, Gerardo Jimenez-Sanchez, Alessandra Carnevale, Xavier Soberón, Samuel Canizales-Quinteros, Héctor Rangel-Villalobos, Irma Silva-Zolezzi, Esteban Gonzalez Burchard, Carlos D. Bustamante. "The genetics of Mexico recapitulates Native

American substructure and affects biomedical traits." *Science* 344 (6189) 2014 pp. 1280-1285

Rice P, Longden I, Bleasby A. "EMBOSS: the European Molecular Biology Open Software Suite." Trends Genet. 2000 Jun;16(6):276-7.