LLNL-TR-487552

FY11 Report on Metagenome Analysis using Pathogen Marker Libraries Lawrence Livermore National Laboratory, Pathogen Informatics Group

Shea N. Gardner, Jonathan E. Allen, Kevin S. McLoughlin, Tom Slezak

May 31, 2011

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Abstract

A method, sequence library, and software suite was invented to rapidly assess whether any member of a pre-specified list of threat organisms or their near neighbors is present in a metagenome. The system was designed to handle mega- to giga-bases of FASTA-formatted raw sequence reads from short or long read next generation sequencing platforms. The approach is to pre-calculate a viral and a bacterial "Pathogen Marker Library" (PML) containing sub-sequences specific to pathogens or their near neighbors. A list of expected matches comparing every bacterial or viral genome against the PML sequences is also pre-calculated. To analyze a metagenome, reads are compared to the PML, and observed PML-metagenome matches are compared to the expected PML-genome matches, and the ratio of observed relative to expected matches is reported. In other words, a 3-way comparison among the PML, metagenome, and existing genome sequences is used to quickly assess which (if any) species included in the PML is likely to be present in the metagenome, based on available sequence data.

Our tests showed that the species with the most PML matches correctly indicated the organism sequenced for empirical metagenomes consisting of a cultured, relatively pure isolate. These runs completed in 1 minute to 3 hours on 12 CPU (1 thread/CPU), depending on the metagenome and PML. Using more threads on the same number of CPU resulted in speed improvements roughly proportional to the number of threads. Simulations indicated that detection sensitivity depends on both sequencing coverage levels for a species and the size of the PML: species were correctly detected even at ~0.003x coverage by the large PMLs, and at ~0.03x coverage by the smaller PMLs. Matches to true positive species were 3-4 orders of magnitude higher than to false positives. Simulations with short reads (36 nt and ~260 nt) showed that species were usually detected for metagenome coverage above 0.005x and coverage in the PML above 0.05x, and detection probability appears to be a function of both coverages. Multiple species could be detected simultaneously in a simulated low-coverage, complex metagenome, and the largest PML gave no false negative species and no false positive genera. The presence of multiple species was predicted in a complex metagenome from a human gut microbiome with 1.9 GB of short reads (75 nt); the species predicted were reasonable gut flora and no biothreat agents were detected, showing the feasibility of PML analysis of empirical complex metagenomes.

Background

The current approach to labeling metagenomic data is to use sequence alignment tools that align each read to a part of each reference sequence and report a summary of the top reference matches. The limitations of this approach are scalability and computational cost. Applying this approach to the Sargasso Sea metagenome data of 1.8 million reads (of length 700) would take approximately 42 days on a 64 CPU cluster (Huson et al. 2007). Using sequence alignment to label data sets produced by the newest high throughput sequencer (Illumina HiSeq 2000), would require 4,630 days of compute time, indicating that a 1,000-node cluster would need 5 days to process one data set. Biosecurity applications require accurate and complete information in a matter of hours, and would benefit from a system that could handle output from a large number of sequencers distributed around the world generating data on a continuous basis (Franz and Lehman 2009).

An alternative to sequence alignment are the composition based approaches, which match the frequency of occurrence for short (k=6) kmers (oligos of length k) in the query read with

frequency of occurrence for kmers found in broad classes of organisms (Teeling et al 2004, Chatterji et al. 2008, and McHardy et al. 2007). This approach is more scalable then sequence alignment but lacks the ability to provide detailed discrimination of the sample contents.

We advocate a hybrid approach, replacing costly pair-wise sequence alignment with faster kmer searches to a reference database using larger values for k (k=18+). We built Pathogen Marker Libraries (PMLs) with kmers selected to be informative for a set of bacteria and viruses of interest. By informative, we mean that the kmers are family specific, genus specific, species specific, etc. Widely occurring kmers in human or non-target microbial sequence databases are excluded from the PMLs, so that metagenome matches to the PMLs suggest the presence of particular microbes of interest. Here we present 1) 3 design approaches for alternative PMLs, 2) methods for building a pre-computed database of expectations relative to a large microbial sequence database, 3) methods and software for rapidly comparing a metagenome to a PML, 4) results of testing on actual metagenomes (raw, unprocessed short reads), and 5) results of testing on simulated metagenomes where ground truth is known.

Methods

PML construction

PMLs were designed to cover the pathogenic species and near neighbors in 5 viral families and 8 bacterial families listed in Table 1.

Table 1

size (# bases)	# sequences	Family
1.45E+09	696	Enterobacteriaceae
6.44E+08	296	Bacillaceae
5.95E+08	187	Burkholderiaceae
4.03E+08	126	Clostridiaceae
1.62E+08	63	Brucellaceae
1.25E+08	71	Francisellaceae
5.74E+07	54	Rickettsiaceae
2.90E+07	154	Poxviridae
1.83E+07	16	Coxiellaceae
4.15E+06	1359	Bunyaviridae
3.07E+06	266	Togaviridae
1.10E+06	222	Arenaviridae
1.00E+06	53	Filoviridae
3.50E+09	3563	Total

Three alternative methods were used to construct candidate PMLs. The first method was based on the KPATH conserved/unique sequences (Slezak et al. 2003). The second approach was based on family-specific kmers, clustering sequences in a family by shared kmers and selecting a subset of those kmers for inclusion in a PML based “heavyweight clusters” covering all levels of conservation in the family whether or not clusters map to NCBI taxonomy (any conserved subgroups, including but not limited to strains, species, genus, etc.). The third method

pulls out all family-specific regions relative to a large database of non-target sequence, regardless of conservation within the family. Additional files for all of the libraries are available with annotations as to the genomes matching each marker sequence and the genes they land on (gene annotations are slow due to limitations on hitting the NCBI website, so may not be complete by May 31, 2011 for the largest libraries. Smaller libraries have completed, and updated gene annotations for the larger libraries will be sent upon completion.)

First Approach: KPATH conserved/unique signatures

KPATH is LLNL's software system for designing pathogen identification signatures and has been described in detail (Slezak et al. 2003). KPATH signature design runs for a target set to identify the conserved and unique sequences. These conserved/unique sequences represent for the members of the target set conserved regions of at least 18 nt that are not found in any other sequenced bacterial or viral genome outside of the target set. Different chromosomes, viral segments, or plasmids must be run separately, since all the members of a target set must share the selected regions. Species-level and genus-level target sets must also be run separately. We note that while KPATH is roughly equivalent to the Insignia software recently put out by University of Maryland, as both determine regions of relative uniqueness, KPATH was designed to be able to create signatures at varying levels of resolution (genus/species/strain). KPATH PMLs were the lists of conserved/unique regions for the pathogenic bacteria and viruses and some of their near neighbors from the families in Table 1. 126 KPATH runs were performed on 12 and 26 bacterial genera and species, and 6 and 44 viral genera and species, respectively. A complete list is provided in an excel spreadsheet accompanying this document (KPATH_pathogen_reference_marker_101015.xlsx). Many of the species (e.g. VEE, WEE, EEE, CCHF) were so divergent that no conservation exists for regions >=18 nucleotides, so those viruses had no representation in the KPATH viral PML. Therefore, we pursued an alternate approach to enable better pathogen target representation in the PMLs than was possible from the strict consensus calculations by KPATH.

Second approach: Kmer clusters

The kmer method described below is more flexible with regards to finding markers at different taxonomy levels simultaneously, i.e. genus, species, and strain-specific subsequences can be found in the same set of calculations for a family, as can markers for different chromosomes, viral segments, and plasmids. This is preferred over many separate KPATH runs, some of which yield no markers (e.g. divergent viral species). The following steps are used to generate "Kmer" PMLs.

1. Enumerate kmers in a family, k=16 or 17. Smaller k is higher uniqueness stringency.
2. Delete any kmers that match assembled human genome or bacteria or virus not in family
3. Cluster sequences in the family that share kmers. All possible sequence combinations that share 1 or more kmers are calculated. Each sequence combination is a cluster. A sequence may be found in more than one cluster, but a kmer maps to only one cluster of sequences that contain that kmer. There can be an extremely large number of clusters, so splitting the kmers into smaller subsets facilitates fitting all clusters in memory until those supported by too few kmers can be eliminated (step 4 below). Some of the clusters map to the established taxonomy (all members of a species, all members of a genus) and some do not (some members of a species, some members of multiple species). Incorrect

taxonomic labels (mislabeled or unlabeled species or strains) are surprisingly common: 430,848 of 2,755,253 sequences in the KPATH database, 16%, have no species label in the NCBI taxonomy table. We have encountered mislabeled strains or species of Burkholderia, Francisella, and Bacillus that we discovered during SNP or other phylogenetic analyses. When possible, we confirmed and corrected the labeling errors with subject matter experts, but suspect that other cases exist in the data, particularly for less intensively examined genomes.

4. Weight clusters by # of shared kmers. "Heavyweight clusters" that share a lot of kmers usually mean there is some meaningful phylogenetic similarity (or could also mean that a gene has been horizontally transferred), and kmers from these "heavyweight" clusters are what are included in the PML. Note that PML sequences contain both the kmers that map to nodes of the kmer tree as well as homoplastic kmers from heavyweight clusters. In contrast, those kmers forming clusters of sequences sharing only a few kmers are phylogenetically noisy and represent highly variable sequence which should be omitted from a PML. The threshold for the number of kmers in a cluster required to include that cluster's kmers in the PML was ~100 for viruses and ~800 for bacteria. These thresholds are approximate because the calculations were divided in parallel across CPUs with a lower threshold for each CPU. IO limitations prevented storing all of the small intermediate files that did not pass the threshold for a single node, so exact threshold counts are not feasible. In addition, no markers that are present in only a single sequence were included, since isolate and strain identification was not a goal of the project. Alternative libraries could be designed that did include strain specific kmers.
5. Using positional information from a sequence in the cluster, extend overlapping kmers from the heavyweight clusters into maximum-length unique substrings (MLUS). MLUS of minimum=18, 19, or 20 bases long make up the PML. Thus, when k=16 for the uniqueness calculations, there must be 3 kmers shifted by 1 base in a heavyweight cluster to make up an 18-mer, the minimum length of a marker sequence.
6. Build 12 Kmer libraries with different k (step 1) and minimum MLUS (step 6) for all bacteria and viruses in the families in Table 1: Unique k=16, 17 X and minimum length marker=18, 19, 20 with markers for each of 2 kingdoms (virus, bacteria) [2 x 3 x 2= 12]. These are named KmerUniq[16|17]Extend[18|19|20][Virus|Bacteria]
7. Also built smaller libraries by manually inspecting and selecting clusters that correspond to some bacterial genera and species of interest. This was labor intensive, error prone (unlabeled or mislabeled species), and unscalable for large numbers of sequences or large numbers of clusters, so was not pursued further. Tests on a number of bacterial species libraries computed manually showed lower sensitivity than the automated, rule-based libraries described in the steps above (data not shown).

As a convenient way to visualize some of the major clusters captured by kmer analyses, a kmer-based tree can be built from a pairwise distance matrix between all pairs of sequences in the family. Distance between sequences i and j is calculated as _____

Neighbor-joining (the "neighbor" program in the PHYLIP software; Felsenstein 2005) was used to build the tree from the distance matrix.

For well-curated sequences (finished genomes, separate components like plasmids and different chromosomes) the kmer based tree accurately estimates phylogeny (Figure 1).

However, if the sequences in a family are a mixture of finished sequence that contains only a single chromosome or plasmid and draft sequences that contain plasmid and all chromosomes (Figure 2), or have experienced substantial horizontal gene transfer or convergent evolution, the tree may diverge from phylogeny. The tree enables a visualization of just a small number of the clusters which map to the nodes of that particular tree. A large number of additional clusters may exist that do not map to a node, for example, clusters that contain sequences branching from multiple nodes or only a subset of the sequences under a node. These are "homoplastic" kmers relative to that kmer tree. Homoplastic kmers can make up the bulk of kmers, and therefore are an important component of the Kmer PMLs.

Kmer based trees were built for all the families in Table 1, and are provided in slides accompanying this document (Kmer tree plots.pptx). One is shown based on k=16 for Coxiellaceae (Figure1) and Bacillus anthracis genomes and plasmids (Figure 2). The numbers of kmers specific to each node are plotted, and the numbers of genome-specific kmers are given in brackets by each genome. Figure 2 clearly shows that when draft data mixing sequence elements like plasmids and chromosomes into a single sequence entry is combined with finished data, a kmer tree cannot capture the actual phylogenetic pattern. The BA[L|R|V|S], BA[N|I] and other draft genomes with plasmid and chromosome data combined into a single sequence entry (contigs glued together with intervening N's) do not cluster with their closest near neighbors, and instead all stack up above the finished genomes but along different branches than the plasmids, since they share most of their kmers with the finished genomes but also thousands of kmers with the plasmids. For example, SNP analyses clearly show that BAI_A0442_IsolateB clusters with A0442 and the KrugerB genomes (Figure 3), but on the kmer tree BAI branches next to the other B** draft genomes. The kmer counts for the various nodes for the plasmids have 0 kmers mapping to them because those kmers are also contained in the BA[L|R|V|S], BA[N|I] draft sequences branching off the chromosome part of the tree. The plasmid kmers map to clusters containing both draft "genomes" and plasmids that are captured by the cluster data but are not visually evident in the kmer tree. *This illustrates why it is important to use kmers in the PML from all heavyweight clusters, not just those that map to nodes of a tree, as described in step 4 above.* In summary, a) clustering sequences based on shared kmers is an automated and scalable method to classify subsequences across the entire range of conservation and divergence, b) draft sequence data with combined elements like plasmids and chromosomes will give a kmer based tree that differs from a phylogenetic tree, and c) visualizing the data in a tree illustrates only a single pattern from the much richer and more informative set of patterns captured in the cluster data. A phylogenetic network representation could enable a richer visual representation of the complex relationships among sequences.

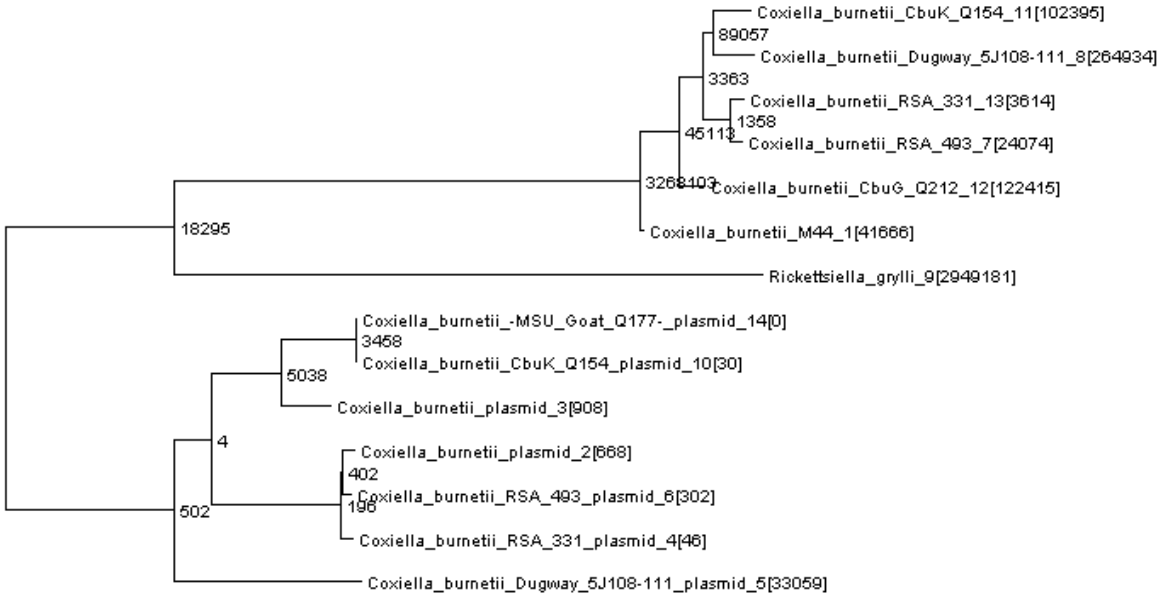

Figure 1: Kmer tree for Coxiellaceae sequences.

Figure 2: Kmer tree based on k=16 for Bacillus anthracis genomes and plasmids.

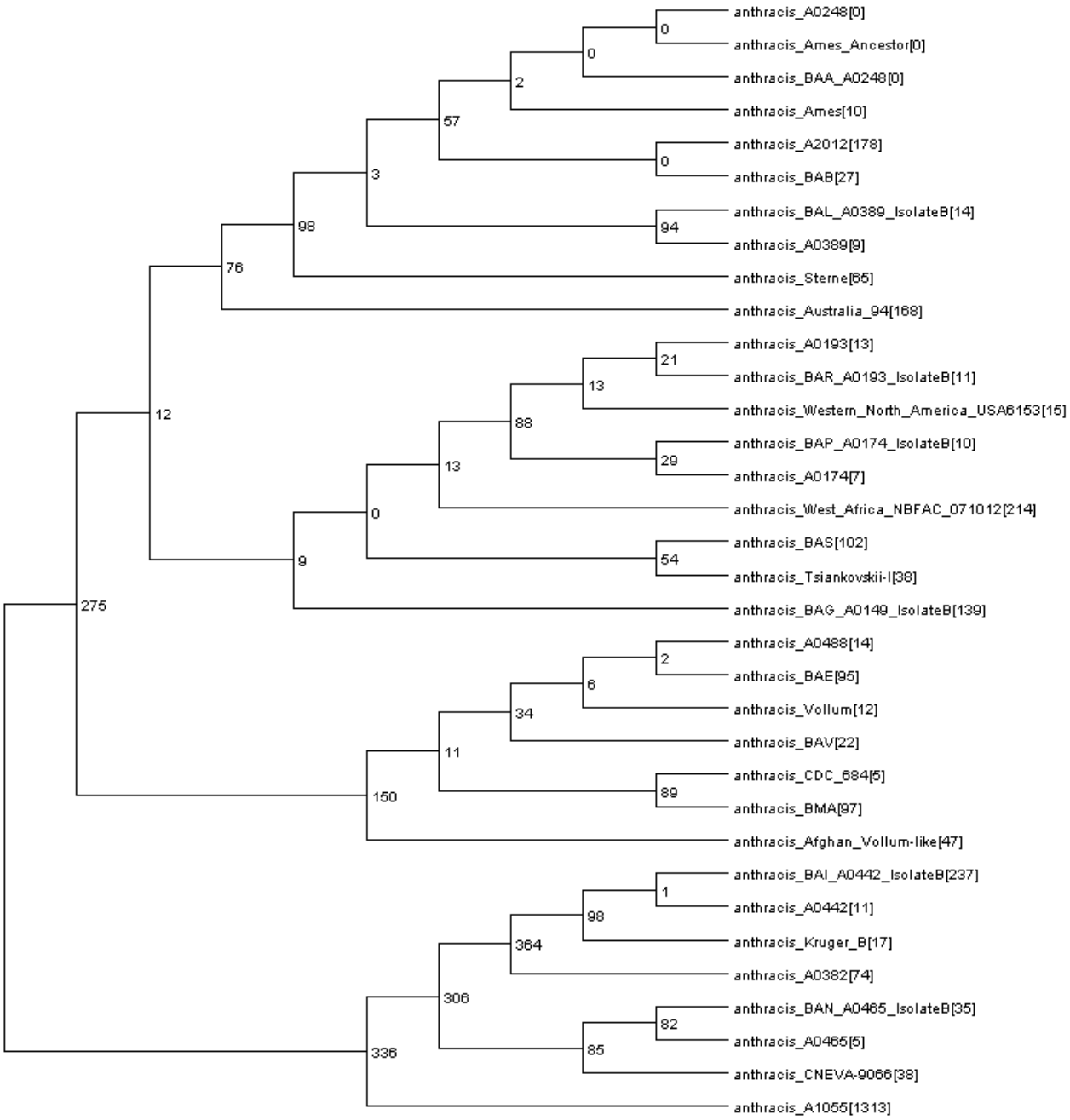

Figure 3: SNP-based tree for Bacillus anthracis finished and draft genomes, calculated using an updated version of the kSNP software (Gardner and Slezak 2010). Counts of the alleles shared by the genomes under a node are plotted at each node, and counts of genome-specific SNPs (or sequencing errors) are given in brackets after the strain name.

Third approach: Unique_k19 libraries

The third approach was to find the family specific regions for every sequence in the family, regardless of conservation. This process is much like the KPATH library construction without the manual target set selection for every species or genus of interest (instead all

sequences/all species in a family were included) and without the problematic conservation requirements that eliminated too many regions for divergent species.

These calculations were done most recently, so a larger set of sequences was used to eliminate non-unique regions. In addition to screening against the bacteria and viruses not in the target family (as in the KPATH and Kmer PMLs) and the assembled chromosomal and mitochondrial human genome (Kmer PMLs only), the sequences were also screened for uniqueness against fungal and archaeal complete sequences, the human version hg19 sequences including repeat regions and unassembled pieces that have not been localized on a chromosome yet, sequences from the SILVA ribosomal RNA database outside the target kingdom (i.e. bacteria were not screened against bacterial rRNA sequences), and the RepBase16.01 database of repeat elements (Jurka et al 2005). Any match of 19 bases or longer to non-target sequence was eliminated. This length is slightly longer than the matches of 16-18 for the Kmer and KPATH libraries, and was necessary because of the expanded set of sequences for uniqueness checks in order to have sufficient unique sequence for a marker library. These uniqueness calculations required thousands of cpu-hours.

In each family, unique regions of at least 18 bases were gathered from all genomes and clustered using cd-hit (Li and Godzick 2006) at a 99% similarity level to remove redundant sequences. That is, many genomes contained nearly identical unique substrings, so these were clustered and the longest sequence in each cluster was taken as the representative of that cluster. Clustering at 95% yielded libraries only slightly smaller than the 99% clustering, so we used the 99% level. These non-redundant unique substrings were gathered for all the families in Table 1, and these comprised the "Unique_k19" viral and bacterial PMLs.

Advantages and disadvantages of each PML approach

KPATH PML

Pros

- simple to compute using an established method that has been used to predict thousands of successful PCR-based signatures.
- libraries are small, so they are fast to run against a metagenome
- there should be little or no marker overlap, as a consequence of building a single consensus sequence from multiple target sequences

Cons

- conservation criterion is overly strict, resulting in few or no signature regions for viruses at the species and genus levels
- libraries are small, so while they have sufficient sensitivity for pure samples (cultured isolates) they have low sensitivity against complex mixtures (e.g. soil).
- manual selection of target sets and initiation of runs for every species and chromosome, segment, or plasmid component is unscalable
- currently no uniqueness checks against human genome
- expressly excludes virulence or antibiotic resistance genes that are not conserved and unique to a single species

Kmer PML

Pros

- library size is tunable based on uniqueness, minimum marker size, and the minimum number of shared kmers in those clusters used for a library. Large libraries are very

sensitive even for complex samples, while a small library can be used for pure/simple samples
- automated to compute
- markers for all species in a family are included, both pathogenic and near neighbors
- markers for both highly conserved and species or strain specific regions are included, to detect both novel unsequenced organisms in a family and characterize known organisms to the species level.

Cons
- Substantial software development with memory and speed optimization was required to design these libraries
- Currently markers may partially overlap one another by lengths shorter than the value of k used for clustering, so markers are not independent in a strict sense and the library is larger than it needs to be if all overlaps can be eliminated (future version will address this to facilitate statistics that assume marker independence)

Unique_k19

Pros
- Conceptually simple to compute. Essentially the KPATH process, without the overly strict conservation requirements and manual target set selection.
- Most stringent uniqueness filters employed in terms of number and types of sequences checked, although the length of allowable non-unique regions is 1-3 bases longer than in the other libraries
- Since no checks for conservation, strain specific markers are included in the library so strain identification may be possible

Cons
- Regions are unique at 19-mer level, so there may be more 18-mer non-specific hits
- One of the larger libraries, more sensitive than smaller libraries but possibly more false positive matches
- The downside of keeping strain-specific markers is that there may be marker overlap, and thus markers are not strictly independent. Clustering sequences at 99% similarity reduced marker overlap to some extent, but not entirely.

In some of the tables and figures, abbreviated library names have KU=KmerUniq and E=Extend, and PML_bacteria_unique_k19 is Unique_k19.

Table 2: Summary information about the size of each PML

PML	Number Bases	Number Markers	Avg Marker Length	Minimum Marker Length	Maximum Marker Length
KPATHbacteria	1,566,740	62,806	24.9	18	1183
KmerUniq16Extend18Bacteria	22,274,728	1,182,918	18.8	18	39
KmerUniq16Extend19Bacteria	10,153,474	509,515	19.9	19	39
KmerUniq16Extend20Bacteria	4,967,139	236,550	21.0	20	39

KmerUniq17Extend18Bacteria	3.16E+08	16,131,587	19.6	18	67
KmerUniq17Extend19Bacteria	1.86E+08	8,957,361	20.8	19	67
KmerUniq17Extend20Bacteria	1.19E+08	5,421,907	22.0	20	67
PML_bacteria_unique_k19	1.39E+08	3,983,469	34.9	19	759
KPATHvirus	516,165	11,718	44.0	18	2131
KmerUniq16Extend18Virus	421,979	22,259	19.0	18	35
KmerUniq16Extend19Virus	206,519	10,289	20.1	19	35
KmerUniq16Extend20Virus	108,498	5,130	21.1	20	35
KmerUniq17Extend18Virus	3,143,718	158,636	19.8	18	71
KmerUniq17Extend19Virus	1,962,684	93,023	21.1	19	71
KmerUniq17Extend20Virus	1,316,057	58,990	22.3	20	71
PML_virus_unique_k19	1,189,237	32,835	36.2	19	247

Calculating Expected and Observed Matches to PML

Expected matches

Bacterial PMLs were compared to the complete bacterial sequences in the KPATH database, and virus PMLs to viral sequences in KPATH to calculate a list of the expected matches. KPATH is LLNL's database of only complete finished and draft sequences, including chromosomes, plasmids, genomes, and viral segments. Single genes or other sequence fragments are not included. We used Mummer to perform rapid calculations of maximal exact matches of at least 18 nt, which were stored in separate files for every sequence, with separate directories for each PML. Redundant matches that were exact match subsequences of longer matches were removed using cd-hit. Sequences with fewer than 100 nt of PML matches summed across all ≥18-mer matches for a given PML were removed from consideration as possible matching sequences, since coverage was too low in the PML for reliable detection. To avoid unnecessary bookkeeping that would slow down the calculations, only the matching sequences were stored; the identity of the marker sequence in the PML containing that match was not stored. The process of computing the expected matches is only done once per PML, and stored for all future tests against new metagenomes.

Observed matches

Given a metagenome sequence (raw short or long reads in fasta format), run_match_library finds the matches of at least 18 nt between reads and a PML. Neither the entire read nor the entire marker are required to match, so long as there is an exact match of at least 18 bases (found with Mummer). This allows flexibility to handle both short and long reads and PMLs with both short and long marker sequences. Redundant subsequence matches captured by longer matches are ignored. Then the observed matches of the PML and metagenome versus the expected matches of the PML and every KPATH database sequence are compared.

Since many of the more conserved markers (e.g. genus-level markers) match multiple sequences, the sequences with matches were ranked by number of observed matches. Sequences for which ALL matches could be entirely explained by a higher (with more matches) or equal ranking sequence were not reported. So unless a different strain or near neighbor explained

additional matches, and not just a subset of already-explained matches, it was not reported. Thus, hits should be interpreted as a species hit rather than a strain match, since strain specific markers were not included in most PMLs. However, if at least 2 matches spanning more than 30 nt could not be explained by a higher ranking sequence, the sequence was reported, with both the number of matches not explained by a higher ranked sequence and the total matches including those already explained. Thus, sequences are reported only if they provide additional information beyond others already reported. The entire set of matching sequences (including duplicating matched sequences) is also reported in the Matches subdirectory, to maintain a complete reporting system. Sequence variation between existing genomes and unsequenced isolates, e.g. "missing links" between available genomes, or mixtures of multiple species sequenced at very low coverage, or chimeric genomes could account for these lower ranked matches. Extending this process to provide strain-level information is a possible future enhancement. While this would increase library size, improvements in hardware and algorithms are possible that could dramatically speed up the calculations. We are pursuing funding to develop novel algorithms for rapid, high resolution metagenome analysis taking advantage of recent improvements in large, fast, random-access, persistent memory technology. Since plasmids were combined with chromosomal sequence in many of the draft genomes, matches to these were scored twice, once ranking them with all the sequences, and a second time ranking them only with plasmids. Otherwise, plasmid hits could be omitted from the list of top hits if all the matches to a plasmid were also captured by a draft genome ranking higher by match count.

Output Files

A report file "metagenomeName.PML" in the output directory specified in the run_match_library script gives some summary information about which PML and metagenome were compared, total sizes, number of markers and reads, and average marker and read length. The output file listing the hits is "metagenome.PML.DBtype.TOP_HITS.names.taxonomy" in the output directory specified in the run_match_library script. DBtype is "bacteria" or "virus" depending on which database of expected matches was used. Since the information in taxonomy tables at NCBI and KPATH is imperfect/incomplete, the files with the full header information (with gi and kpath ID's, usually with some strain or plasmid name) are provided in "metagenome.PML.DBtype.TOP_HITS.names". Pruning to report only the sequences with Obs/Expected number of matches ≥3% is reported in another file: "metagenome.PML. TOP_HITS.3percentOvsE". This file is provided for convenience only when examining organisms with sequencing coverage greater than approximately 0.5x, and all of the analyses presented below use the "metagenome.PML.DBtype.TOP_HITS.names.taxonomy" results, rather than the pruned *TOP_HITS.3percentOvsE, unless explicitly stated otherwise.

The 9 columns in the *TOP_HITS* files report the following data:

1) The first column reports both the total number of matches and in parentheses the subset of matches not already explained by a better-matching sequence. More specifically: the number ≥18 nt matches in the metagenome for PML markers from that database sequence, and (Number of these matches not already explained by a sequence above).
2) Number of expected ≥18 nt matches of database sequence to the PML
3) Ratio of column 1/column 2: Obs/Exp number of matches
4) Number of matching bases spanned by the matches in column 1, and in parentheses the number of matching bases not already explained by a sequence above. This corrects for

differences in the length of matches, information which is not captured by the simple counts in column 1.

5) Number of expected matching bases covered by the expected matches in column 2
6) Ratio of column 4/column 5: Obs/Exp number of matching bases
7) Number of bases in the database sequence
8) Coverage in the PML for this sequence, computed as column 5/column 7. This indicates sensitivity of the PML for this sequence.
9) Species or name of sequence. The ".taxonomy" in the file name indicates that species names have been pulled from a taxonomy table (NCBI and KPATH taxonomy database). Many of the sequences are missing species information in the taxonomy database, so sequence names were automatically parsed as best as possible for species and chromosome, segment or plasmid identifiers. Full sequence names are given in the file ending ".names" (no ".taxonomy" at the end).

A number of additional intermediate output files are in the Mummer and Matches subdirectories. This data can be removed if it is not of interest, but it may be useful for debugging or more in depth investigation of reads containing matches.

Although no formal statistical model has been developed to assess likelihoods and p-values, the ratio of observed/expected matches represents information about 1) whether a species is present, and 2) coverage in the metagenome for that species. We are seeking funding to devise statistical models to estimate p-values for presence/absence and confidence intervals, as that is beyond the scope of the currently funded work.

Data

Empirical Metagenomes (unassembled short read collections)

We downloaded datasets from the Short Read Archive (SRA) of NCBI. The datasets we tested are listed in Tables 3 and 4. All are unassembled reads; 8 datasets are of cultured organisms and one is a true metagenomic sample.

Table 3

Metagenome	Organism	Platform
DRR000002	Bacillus subtilis subsp. subtilis str. 168	Illumina Genome Analyzer II
DRR000184	B. anthracis BA104	Illumina Genome Analyzer II
ERR011207	A human gut microbial gene catalog	Illumina Genome Analyzer II
ERR015579	Yersinia enterocolitica biotypes	Illumina
SRR000340	Francisella tularensis B-SA	454
SRR004172	Brucella abortus bv. 5 str. B3196	454
SRR005754	Brucella melitensis bv. 1 str. F1/01	Illumina Genome Analyzer II
SRR039956	Francisella tularensis subsp. tularensis FSC043	Illumina
SRR133640	Yersinia pestis KIM D27	Illumina Genome Analyzer II

Table 4

Metagenome	Number bases	Number reads	Average read length	Minimum read length	Maximum read length
DRR000002	5.9E+08	16354270	36	36	36
DRR000184	3.8E+08	7631281	50	50	50
ERR011207	1.9E+09	25274238	75	75	75
ERR015579	8E+08	21535250	37	37	37
SRR000340	4.2E+07	401300	105.417	34	250
SRR004172	9.4E+07	359502	262.452	35	375
SRR005754	3.5E+08	6869470	51	51	51
SRR039956	2.2E+08	5743234	38	38	38
SRR133640	2.3E+08	2311784	100	100	100

We did not report on the testing of any viral PMLs since we found no metagenomes for the viruses in Table 1. We plan to expand the taxonomic coverage (e.g. Caliciviridae, Adenoviridae, etc) of more clinically common viruses in the PMLs for testing next year on available metagenomes and simulated reads.

Simulated Metagenomes

Single species simulations

Reads were simulated according to the error models provided in Metasim (Richter et al. 2008). 454 reads averaged ~260 bases and Illumina reads exactly 36 bases with the default "Empirical" model. Simulations were done for 15 bacterial genomes, listed in Table 5. For each species/error model combination, we simulated "metagenomes" of 50, 500, and 5000 reads for each species, to mimic various coverage levels (the exact coverage depends on the length of each genome). Although these single species simulated datasets are not metagenomes in the usual sense of a complex sample of multiple species, for ease of presentation we refer to them as metagenomes in the sense that they are short, unassembled raw reads. Then we ran the PMLs against each metagenome and counted the number of instances that the top hit was the correct species and the number of matching bases in correct top hits, correct genus hits, and other (incorrect genus) hits.

Complex Mixture Simulations

To simulate a complex mixture, simulated reads from the 15 bacteria in Table 5, plus two Escherichia coli genomes (K12_DH10B and O157H7_EDL933) and 13 viruses were mixed into one complex metagenome composed of reads from 30 genomes (29 species). There were 500 reads per genome, each 36 nt long, simulated with the Empirical error model in Metasim, or 0.002x-0.016x coverage per genome (interpolating from Table 5). This complex metagenome was run against each bacterial PML.

Table 5: Genomes and coverage levels for simulated 454 and Illumina reads.

	Coverage			
Genome	50 reads of 36 nt	5000 reads of 36 nt	50 reads of ~260 nt	5000 reads of ~260 nt
Bacillus_anthracis_AmesAncestor	.003	0.034	0.002	0.239
Bacillus_subtilis_spizizenii	0.0005	0.045	0.003	0.313
Brucella_melitensis_ATCC23457	0.0005	0.054	0.004	0.374
Brucella_suis_ATCC_23445	0.0005	0.054	0.004	0.377
Burkholderia_cenocepacia_AU1054	0.0003	0.025	0.002	0.171
Burkholderia_mallei_SAVP1	0.0003	0.033	0.002	0.228
Clostridium_botulinum_A_ATCC19397	0.0005	0.047	0.003	0.324
Clostridium_perfringens_SM101	0.0006	0.062	0.004	0.431
Coxiella_burnetti_Dugway	0.0008	0.083	0.006	0.579
Francisella_philomiragia_ATCC25017	0.0009	0.09	0.006	0.624
Francisella_tularensis_SCHU_S4	0.001	0.095	0.007	0.66
Rickettsia_akari_Hartford	0.0015	0.146	0.01	1.015
Rickettsia_prowazekii_Madrid_E	0.0016	0.162	0.011	1.125
Yersinia_pestis_CO92	0.0004	0.039	0.003	0.269
Yersinia_pseudotuberculosis_IP31758	0.0004	0.038	0.003	0.265

Results

Empirical Metagenomes

The results in the *3percentOvsE files for each PML vs each metagenome are provided as supplementary data in the file "PML_metagenome_hits.xlsx". In summary, every PML clearly reported the top hit as the correct species that was sequenced. The B.anthracis (DRR000184) and Y.pestis (SRR133640) metagenomes also had strong hits to multiple plasmids. The others had no or only weak plasmid hits. From the "metagenome.PML.DBtype.TOP_HITS.names.taxonomy" files (including all top hits, not only those with Obs/Exp≥3%), the number of bases matching the best sequence (the correct species in the metagenomes tested), sequences in the same genus, or sequences in other genera are plotted in Figure 4. For these high-coverage actual metagenomes from cultured isolates, Obs/Exp ratios are close to 1 (see the "PML_metagenome_hits.xlsx" file), and there are ~1-3 orders of magnitude more matching bases to the correct species than to near neighbors (NNs) or more distant species.

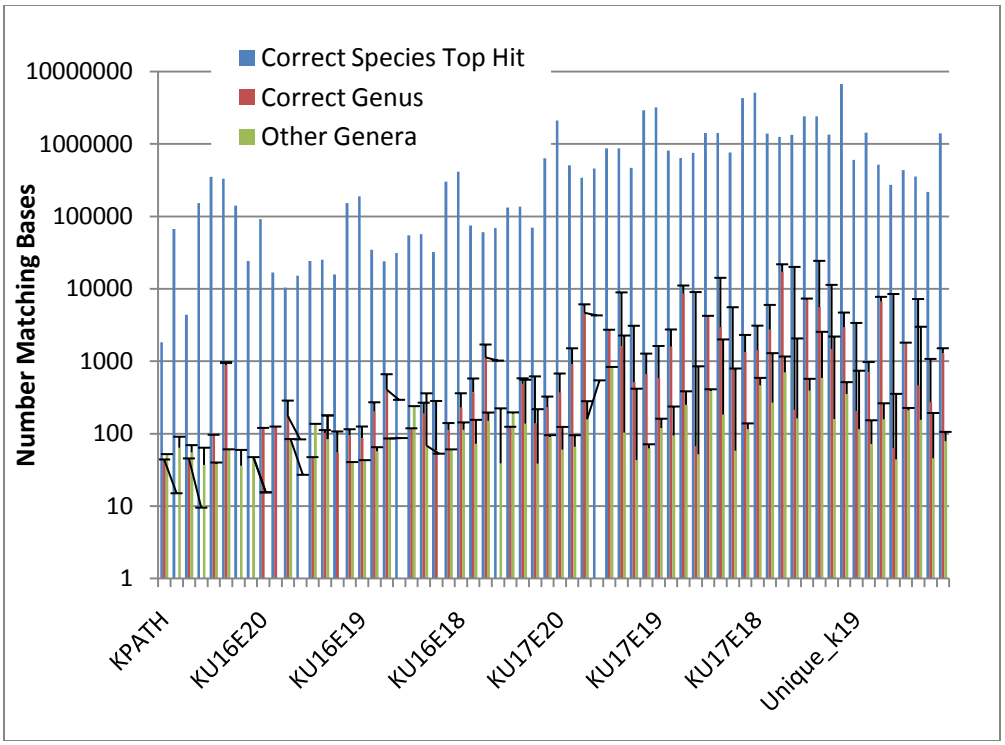

Figure 4: The number of bases matching the PML markers for the correct species, the average matching other species in the correct genus, or other genera. Error bars are across different sequences in that group (e.g. matching genomes from the correct genus, or matching genomes from other genera). Base counts were of the number of matching bases not already explained by a better-matching sequence, and each bar represents a different metagenome.

Plotting the ratio of incorrect/correct species matches (Figure 5) shows that the smallest library, the KPATH PML, had only 3% matching sequence to false positives compared to true positives, and this percentage declines as the PML size increases and more correct matches occur; the larger the library, the higher the specificity, as hits to incorrect species and genera diminish relative to hits to the correct species. However, since these are actual metagenomes, we do not actually know ground truth: possible low level contamination may mean that some hits could be real rather than false positives.

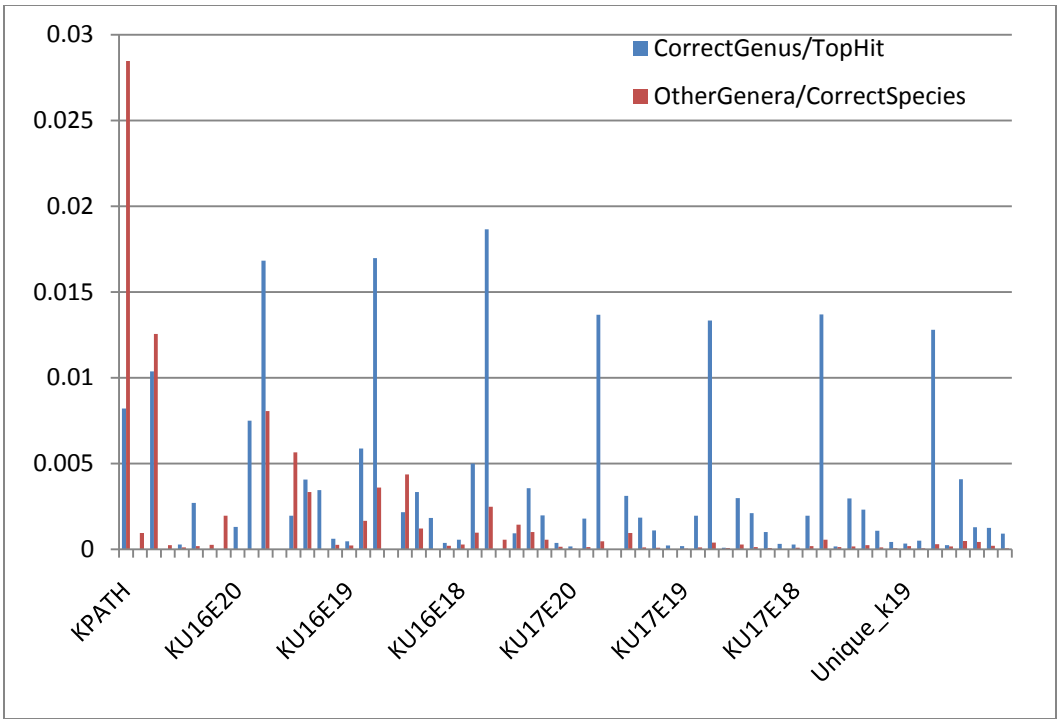

Figure 5: Ratio of bases matching NNs or distantly related species relative to the best (correct) match species. Base counts were of the number of matching bases not already explained by a better-matching sequence. Each bar represents a different metagenome. The Brucella abortus sequence had the highest bars, due to the similarity of B.abortus with other Brucella species, considered different subspecies by some taxonomists. Even for B.abortus, NN matches were less than 2% of those to the correct target species. The plot shows the number of bases not explained by a better match. For "correct genus" and "other genera" the average with standard error bars is across sequences with matches.

Calculation Times

The average times required to run the empirical metagenomes against the PMLs is shown in Figure 6 and 7, on the 12 CPU 48 GB server we had available for testing using 12 threads (1 per CPU). Run times averaged across PMLs for each metagenome were less than an hour (Figure 6). Run times varied more by PML, with the smallest PMLs running in less than 5 minutes on average across metagenomes, and the largest PML averaging 1.75 hours (Figure 7). Full timing information for every metagenome/PML comparison is provided in the Appendix. The fastest comparison finished in just over a minute, and the slowest required 3 hrs 17 minutes. However, we recently tried running with 36 threads and saw a significant speedup of 70% on average. For example, the longest running comparison dropped from 3 hrs, 17 minutes with 12 threads down to 37 minutes with 36 threads, an 81% speedup. Of course, all results were the same regardless of the number of threads. This is a variable that should be tested further, to find the threading level that minimizes run time.

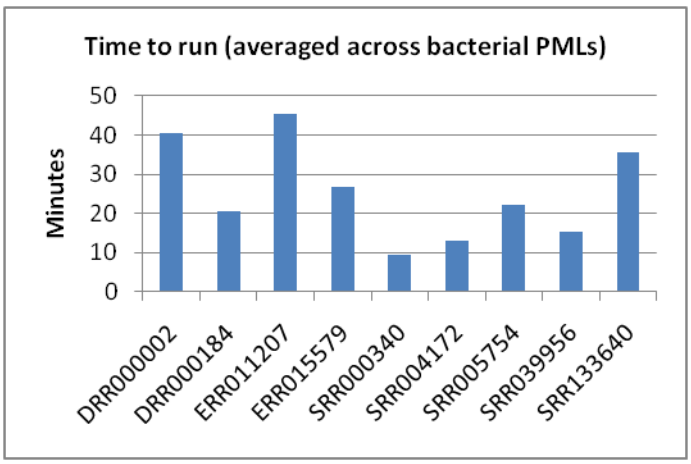

Figure 6: Time required to run each metagenome against the various bacterial PMLs, averaged across PMLs.

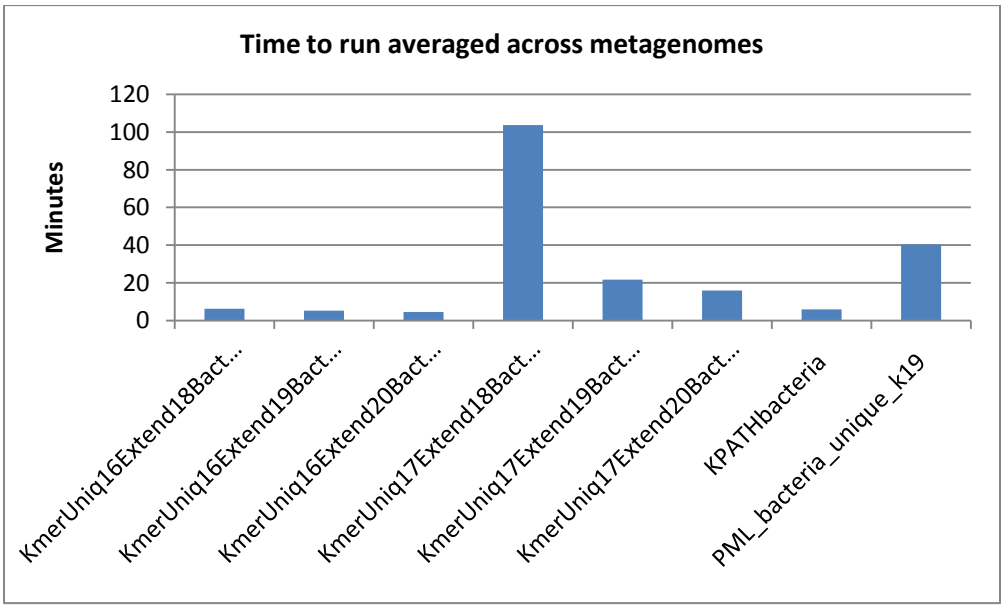

Figure 7: Time required to run each metagenome against the various bacterial PMLs, averaged across metagenomes.

Complex Empirical Metagenome

Results of each PML against the complex "human gut microbial gene catalog" are in the supplementary file "ERR011207_3percentOvsE.xlsx". This was the largest and most complex metagenome examined, with 1.9 GB of sequence as 75 nt Illumina reads. There were dozens of species with hits, including Escherichia coli, Clostridium sp., Bacteriodes sp., Odoribacter sp., Sutterella, Shigella, Faecalibacterium, Phascolarctobacterium, and others, many of which were seen across all PMLs. Reassuringly, there were no hits to biothreat agents, suggesting a low false positive rate for agents of concern, even in highly complex metagenomes predicted to contain NNs to those agents (e.g. Clostridium sp., Bacillus plasmids, Burkholderiales bacterium).

Simulated Metagenomes

Single Species Simulations

The KPATH library failed to identify the correct species in 50-90% of cases for 50-500 Illumina reads and 50 454 reads (Figure 8). The KmerUniq17Extend[18|19] libraries consistently had the best rates of species identification, even with very low coverage levels provided by 50 Illumina or 454 reads (~0.0003x-0.003x). For higher coverage levels ~0.03x-0.3x (5000 Illumina reads, 500-5000 454 reads), all the libraries except the KPATH PML had similar success at identifying the correct species. In most cases B.mallei was mis-identified as a B.pseudomallei, even for the largest libraries, since B. mallei is very similar to B.pseudomallei but with a large deletion. Looking at the ratio of Obs/Exp matches it is clear that B.mallei is a better match, although since the hits are sorted by number of matches rather than the Obs/Exp ratio, B.pseudomallei is listed as the top hit. The heuristics employed to reduce the redundant reporting of multiple strains sometimes causes omission of the correct species from the TOP_HITS list if ALL the matches are also explained by the NN, especially for low read coverage and small PMLs. Nevertheless, we still opt to sort by number of matches rather than the Obs/Exp ratio because genomes with low coverage in a PML, including non-pathogenic species for which markers were not specifically designed, can have a high Obs/Exp ratio due to a small denominator. Since intermediate files with all the matching sequences are provided in the Matches subdirectory of the output directory, one can check those for species that one suspects might be pruned out in the TOP_HITS file, as well as all the strains that have approximately equivalent matches.

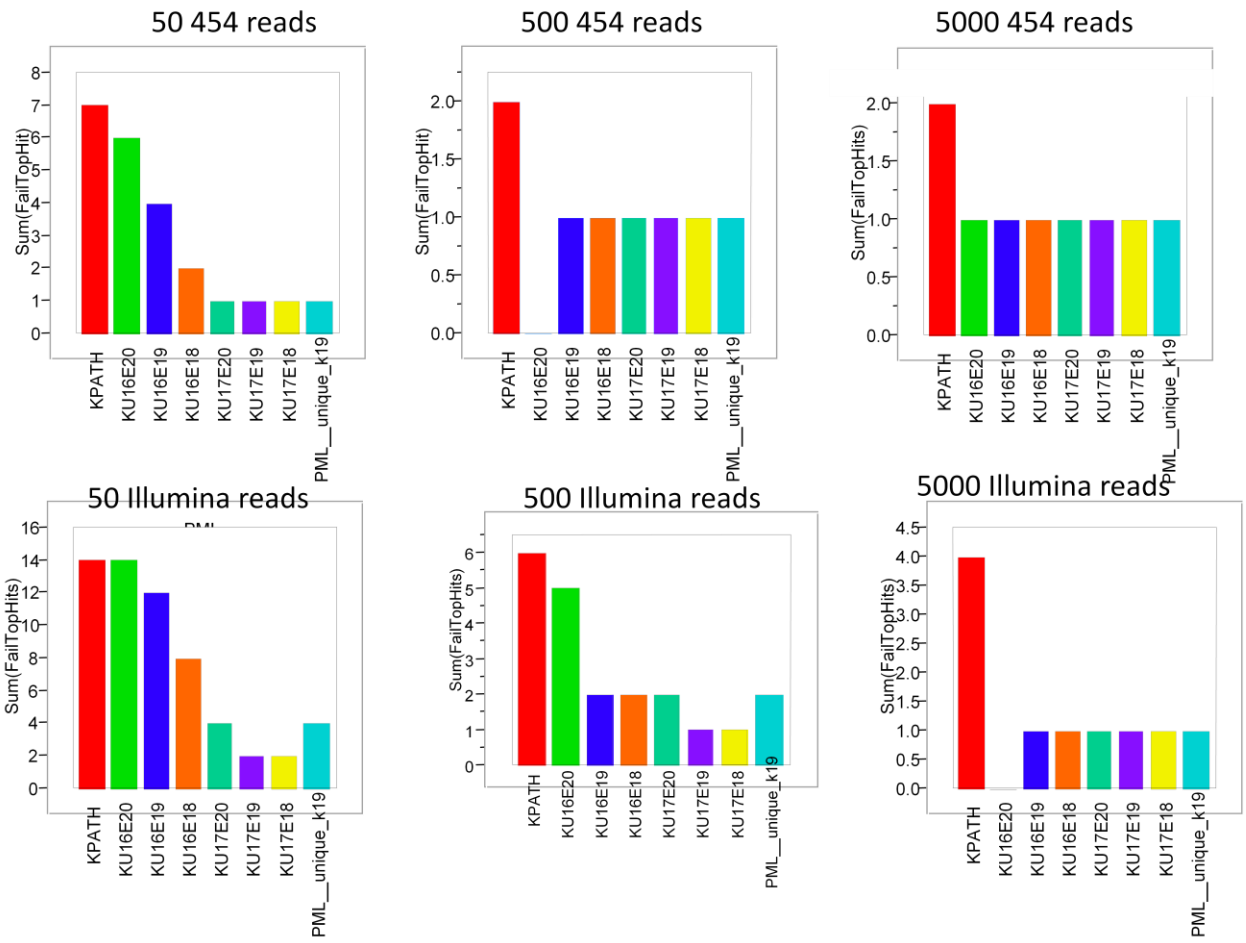

Figure 8: Number of instances that the top hit was not the correct species for simulated 454 and Illumina reads. A total of 15 species were tested. Smaller libraries and shorter or fewer reads resulted in more failures to pick the correct species as the top hit. Abbreviated library names have KU=KmerUniq and E=Extend.

Figure 9 shows the average number of bases matching the correct species as the top hit (red bars) and the number of matching bases not already explained by a better matching genome which match either another species in the correct genus (green bars) or another genus (blue bars). These are taken from the "metagenome.PML.DBtype.TOP_HITS.names.taxonomy" files (including all top hits, not only those with Obs/Exp≥3%). All libraries on average have more bases matching the correct top hit species than lower-scoring matches to other members of the correct genus or other genera, with one exception: KPATH PML for the extremely low coverage in the Empirical 50 read simulations, for which no hits (73% of the tests) or the incorrect species in the genus (20% of tests) was selected as the top hit. The larger libraries have more false positive hits to incorrect genera for the higher coverage simulations, although the false positive predictions were dwarfed by the true positive predictions by ~4 orders of magnitude, so at worst would be considered very minor constituents. Note that the plots in Figure 9 are on a log scale, so the ratio of matching bases of the correct species is 1000-10,000+ times higher than other species in the genus or other genera (Figure 10), and the ratio of correct/incorrect matches is 1-2 orders of magnitude higher for the larger than for the smaller PMLs, showing that the confidence in predicting the correct species goes up with library size. The Unique_k19 library is slightly less

specific than expected based on library size, however: the KmerUniq17Extend20 library is smaller than the Unique_k19 library but it results in relatively more matches to the correct species than to incorrect species.

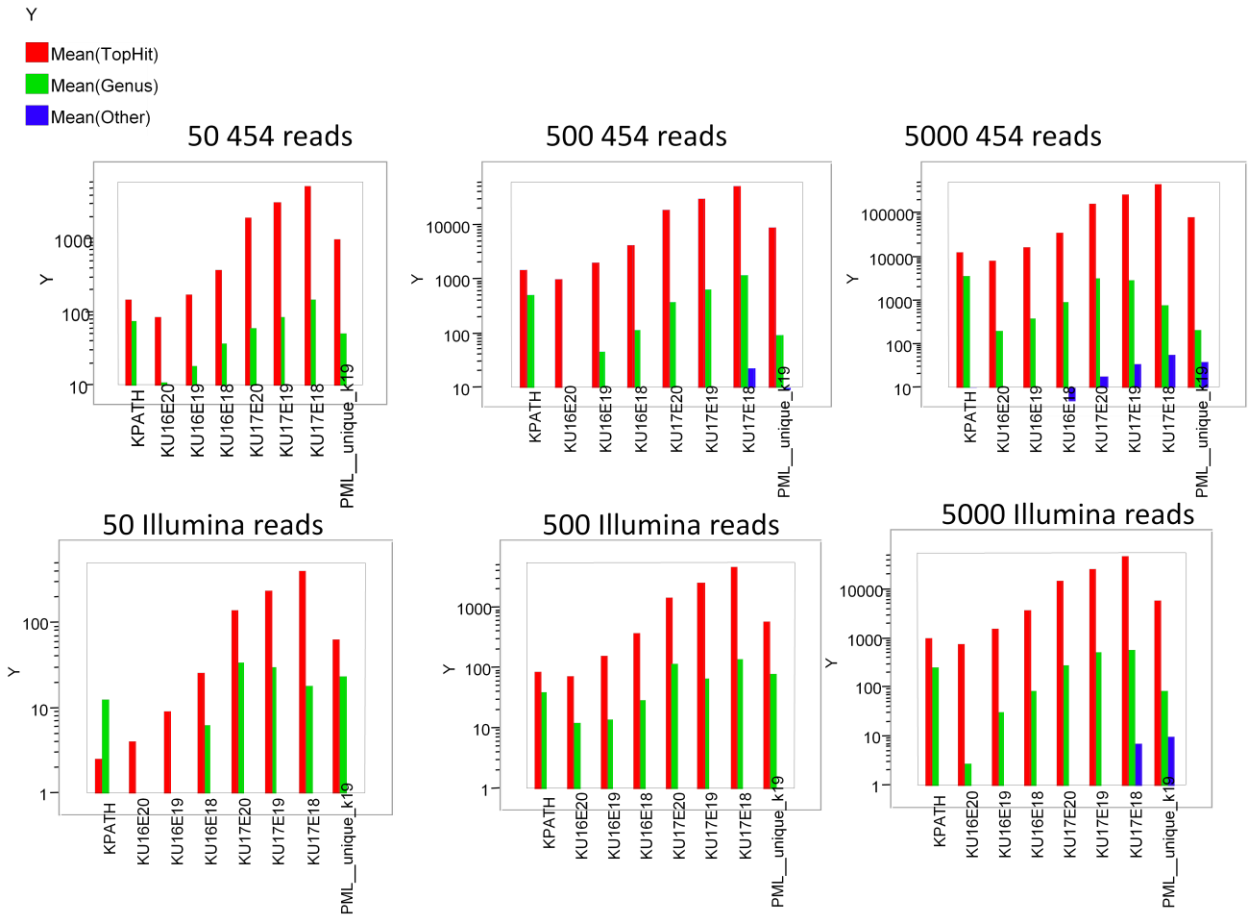

Figure 9: Number of bases matching the correct species (top hit), other species in the correct genus (Genus), or other genera (Other) were averaged across the 15 simulated metagenomes.

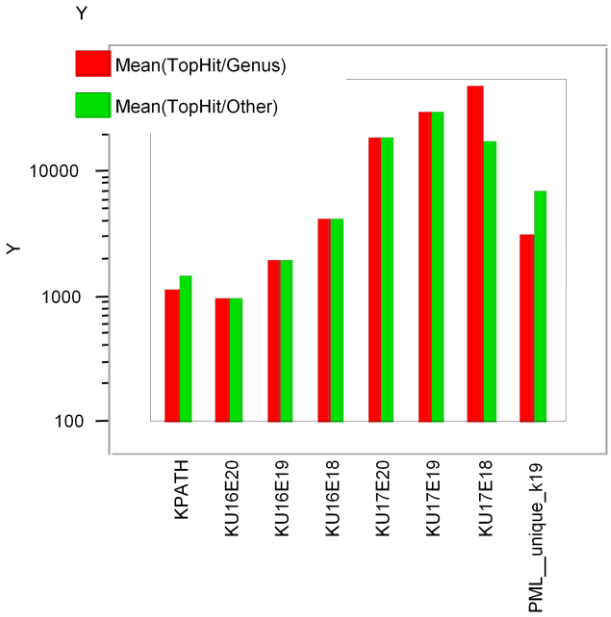

Figure 10: Ratio of the number of matching bases in matches to (correct species/other species in genus) and (correct species/other genera), for the Empirical 500 read simulations.

True Positives (TP), False Negatives (FN), and False Positives (FP)

Figure 11 summarizes the results of the simulated metagenomes for each species described above in terms of TP, FN, and FP. Each point represents a combination of a simulated metagenome with a PML. TP and FN are a function of both the amount of sequence in the PML representing a genome ("coverage in PML") and the coverage by the reads in a metagenome ("Coverage" in the plots in Figures 11A,B,C). For a given combination of (coverage in PML, coverage in metagenome) there is only 1 data point, since each genome in the simulation is a different length (and thus has different coverage for a given number of reads) and has a different amount of conserved/unique sequence representing it in the PML. Rather than averaging across coverage levels and calculating sensitivity=TP/(TP+FN) across all metagenome/PML combinations, each TP and FN point is plotted separately (Figure 11A). Detection failures occur more often for species whose metagenome coverage is under 0.005x and for which the coverage in the PML is less than 0.05x, although this is a fuzzy boundary and a fair number of successes occur at lower coverage levels, and a few failures occur at higher coverage levels (Figure 11A).

Calculating specificity is problematic, since all species not detected in our database (with over 5000 bacterial genomes) would be considered true negatives, so the calculated specificity rates would be exceedingly high. False positive rates are more informative. There is a tradeoff compared to the sensitivity, with more false positives at higher metagenome and PML coverage levels. At about the same thresholds (metagenome coverage under 0.005x and coverage in the PML less than 0.05x) false positives in other genera are very uncommon (Figure 11B). Considering NN false positives in the same genus as the species used for the simulation in addition to false positives in other genera, there are more false positives below these thresholds, but usually less than a handful (≤5) in the same genus for any given simulation. The results presented above in Figures 9 and 10 clarify that on average the false positives have orders of

magnitude fewer bases matching the PML than the correct species (and fewer matches as well, data not shown). Therefore, simply counting false positives in the Figures 11B,C does not convey that true versus false positives are usually easily distinguished by comparing the number of matches or number of matching bases.

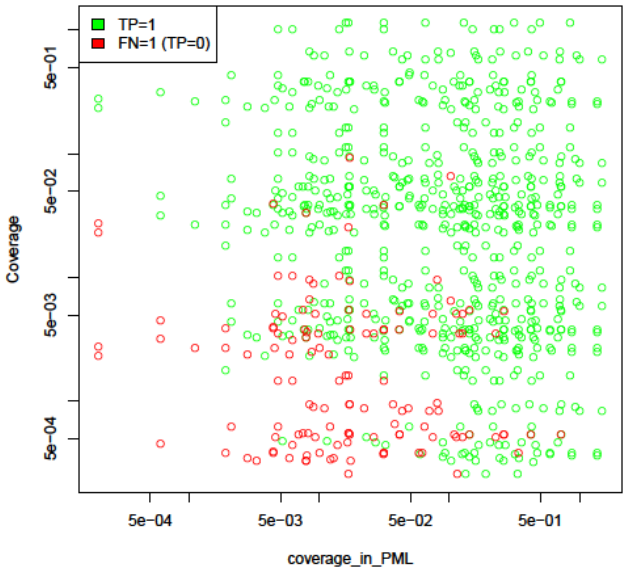

Figure 11A: Detection success and failure events for the simulated 454 and Illumina metagenomes, plotted as a function of coverage in the metagenome (Coverage) and coverage of the genome in the PML.

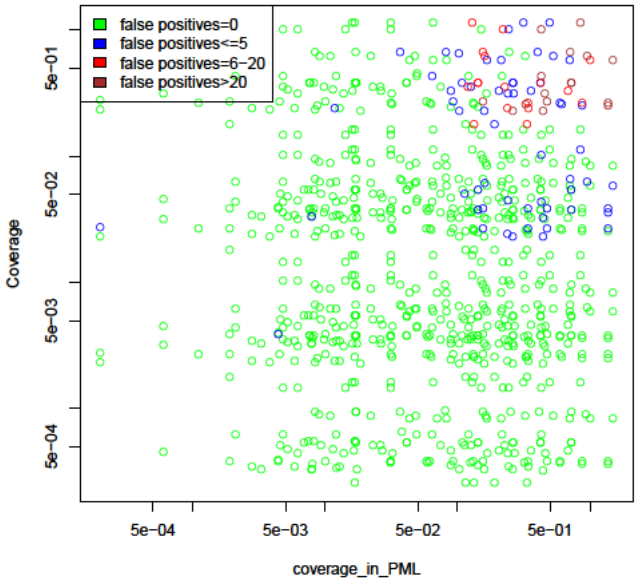

Figure 11B: False positives from other genera only, for the simulated 454 and Illumina metagenomes, plotted as a function of coverage in the metagenome (Coverage) and coverage of the genome in the PML.

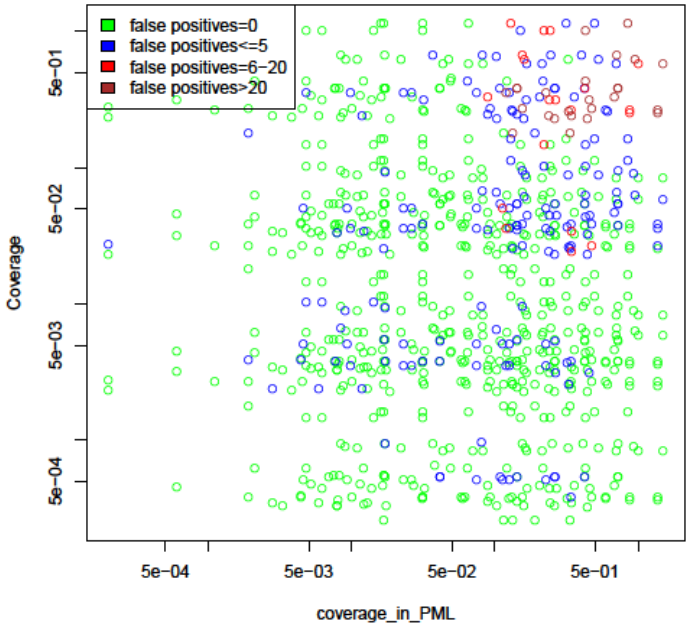

Figure 11C: Total False positives, same genus and other genera for the simulated 454 and Illumina metagenomes, plotted as a function of coverage in the metagenome (Coverage) and coverage of the genome in the PML.

Complex Mixture Simulation

For the simulated metagenome composed of a complex mixture of 30 genomes, the numbers of true positive, false positive, and false negative species reported in the "metagenome.PML.DBtype.TOP_HITS.names.taxonomy" files are shown in Figure 12. The results sort out by library size: KPATH had the most false negatives, followed by KmerUniq16Extend20, and so on. Many PMLs detected either Brucella suis or Brucella melitensis (usually suis) but not both, and all incorrectly detected Burkholderia pseudomallei (a false positive for every PML in Figure 12B) but not Burkholderia mallei, except for the largest PML (KmerUniq17Extend18) which detected them both. Only the Unique_k19 library detected organisms in the wrong genus (2 18nt matches to each of Colwellia and Neisseria), supporting the prediction in the list of "Cons" above that this library might be less specific (more FPs) than the others after controlling for PML size. At the low sequencing coverage levels of these simulations (most are <0.01x), 99% of all hits had Obs/Exp of less than 1%.

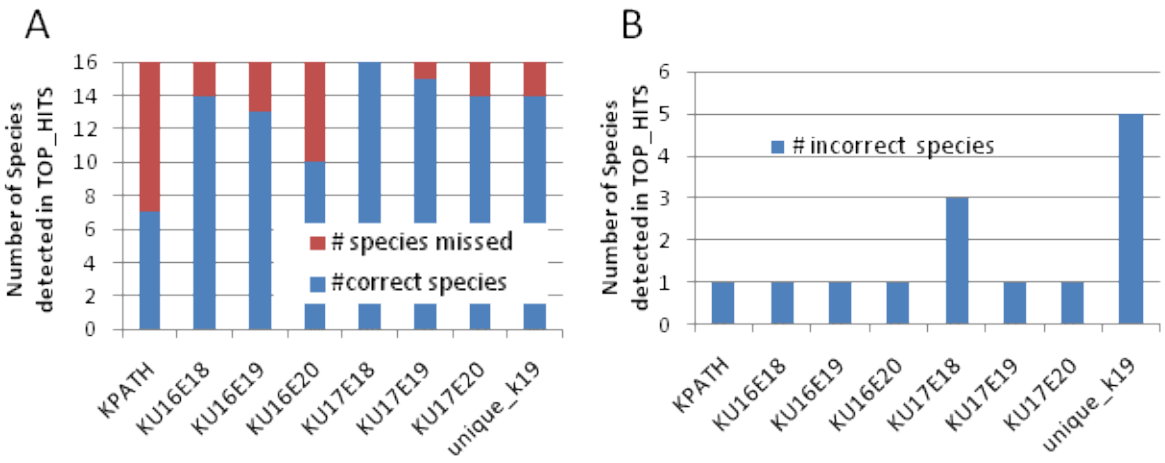

Figure 12: Counts of A) true positive, false negative and B) false positive bacterial species reported in the *TOP_HITS.names.taxonomy files for a simulated complex metagenome with 16 bacterial species and 13 viral species.

Discussion and Conclusions

Sequencing errors and novel unsequenced organisms may result in unanticipated matches to a PML. Distinguishing true positives for organisms sequences with extremely low coverage (e.g. <<0.0001x) from false positives will be very difficult, whether one uses the PML approach, BLASTs to a reference database, or read mapping to a set of reference genomes. Results presented here suggest that the PML approach can rapidly narrow the database of possible target species. A PML analysis to filter the reference database in minutes makes slower techniques like BLAST or read mapping, which align the entire sequence of a read and allow mismatches and gaps, more feasible for second line analysis. BLAST or read mapping for a large metagenome is not currently feasible against a large reference database with thousands of genomes, so a prefiltering step is necessary.

The difficulty in distinguishing B.mallei from B.pseudomallei illustrate that gene loss and small changes in common genes are much harder to detect with the reference marker approach than cases where novel genetic material readily defines the pathogenic species of interest. Thus, a second line analysis with BLAST or read mapping could include other species in the same genus as those detected by the PML in the filtered reference database, to ensure more accurate species identification, particularly for species differing by deletion or minor sequence variations.

Another alternative approach to the PML methods that we considered was metagenome analysis with the LLNL TriTool detection simulation software, which compares a sequence data set with the LLMDA (Lawrence Livermore Microbial Detection Array) probes designed by Shea Gardner. The advantage of this would be that we have already carefully designed dozens of family-specific probes for every species in our KPATH whole genome database, and our biostatistician (Kevin McLoughlin) has developed and our biologist (Crystal Jaing) has extensively tested in lab experiments a rigorous statistical methodology for predicting the presence of multiple species. However, the LLNL TriTool software does not currently run at the speed and memory footprint we need for metagenome analysis, although it may be possible to remedy this given time and funding. A member of our team attempted to run a small metagenome (285 MB) against the LLMDA version 2 probes with TriTool. Although the

MDAv2 probe set is much smaller than any of the PMLs, the TriTool calculations ran for two days, most of which was required to BLAST the probe sequences against the metagenome reads. In comparison, the PML software was able to analyze this metagenome in minutes. A few features that make the TriTools software difficult to scale for gigabase-sized metagenomes are the slow BLAST step and the statistical likelihood calculations that require large amounts of memory; we run all the TriTools calculations on a dedicated machine with 192 GB of memory. Additional optimization efforts are likely to lead to solutions for the speed and memory bottlenecks of the TriTools method.

In conclusion, PMLs were designed and a method to rapidly compare a metagenome to a genome database was invented. We developed a suite of software and tested it against real and simulated metagenomes. Rapid species identification (usually under 20 minutes, ranging from 1 minute to 3 hours with 12 threads) is possible for organisms represented in the PML, and substantial speedup is possible using more threads (3 hours dropped to 37 minutes with 36 threads). We found low false positive rates and high sensitivity for genomes in a metagenome at coverages of ~0.03x and below, depending on the PML used. Simulations with short reads (36 nt and ~260 nt) showed that the correct species were usually detected for metagenome coverage above 0.005x and coverage in the PML above 0.05x. We envision the PML approach as a first-pass rapid analysis to pre-filter a genome database prior to more in-depth read alignments against a subset of genomes from the original database.

References

Chatterji, S., I. Yamazaki, Z. Bai, and J. Eisen. (2008) CompostBin: A DNA composition-based algorithm for binning environmental shotgun reads. In RECOMB 2008.

Felsenstein, J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Franz, DR and R. Lehman. (2009) Global security engagement a new model for cooperative threat reduction. Technical report, National Academy of Sciences.

Gardner SN, Slezak T. Scalable SNP analyses of 100+ bacterial or viral genomes. (2010) J. Forensic Research, 1:107.

Huson, DH, A. F. Auch, J. Qi, and S. C. Schuster. (2007) MEGAN analysis of metagenomic data. Genome Research, 17(3):377–386.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., Walichiewicz, J. (2005) Repbase Update, a database of eukaryotic repetitive elements. Cytogentic and Genome Research 110:462-467.

Li, W, Adam Godzik, (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences", Bioinformatics, (2006) 22:1658-9.

McHardy, A.C, H. G. Martn, A. Tsirigos, P. Hugenholtz, and I. Rigoutsos. (2007) Accurate phylogenetic classification of variable-length dna fragments. Nat Methods, 4:63–72.

Richter DC, Ott F, Auch AF, Schmid R, Huson DH (2008): MetaSim—A Sequencing Simulator for Genomics and Metagenomics. PLoS ONE 3(10): e3373.

Slezak, T., Kuczmarski, T., Ott, L., Torres, C., Medeiros, D., Smith, J., Truitt, B., Mulakken, N., Lam, M., Vitalis, E., Zemla, A., Zhou, C. E., Gardner, S. N. (2003). Comparative genomics tools applied to bioterrorism defense. Briefings in Bioinformatics, June 2003, 4: 133-149.

Teeling, H., J. Waldmann, T. Lombardot, M. Bauer, and F. Glockner. (2004) TETRA: a web-service and a standalone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. BMC Bioinformatics, 5(1):163.

Appendix: Timings

Time to run PMLs against test metagenomes on a single node with 12 CPU and 48 GB RAM (on the aztec cluster at LLNL), without using multiple threads per CPU Using more threads significantly speeds up the calculations, as shown for the KmerUniq17Extend18Bacteria library.

PML	Metagenome	Time (h:mm:ss or m:ss on 12 CPU, 48GB RAM, 12 threads)	Time (h:mm:ss or m:ss on 12 CPU, 48GB RAM, 36 threads)	Ratio of compute time for 36 threads/ 12 threads
KmerUniq16Extend18Bacteria	DRR000002	07:37.2		
KmerUniq16Extend18Bacteria	DRR000184	04:34.0		
KmerUniq16Extend18Bacteria	ERR011207	15:45.4		
KmerUniq16Extend18Bacteria	ERR015579	08:29.5		
KmerUniq16Extend18Bacteria	SRR000340	02:23.9		
KmerUniq16Extend18Bacteria	SRR004172	03:26.6		
KmerUniq16Extend18Bacteria	SRR005754	04:36.7		
KmerUniq16Extend18Bacteria	SRR039956	04:18.4		
KmerUniq16Extend18Bacteria	SRR133640	05:06.7		
KmerUniq16Extend19Bacteria	DRR000002	06:25.6		
KmerUniq16Extend19Bacteria	DRR000184	03:34.4		
KmerUniq16Extend19Bacteria	ERR011207	14:25.9		
KmerUniq16Extend19Bacteria	ERR015579	07:02.9		
KmerUniq16Extend19Bacteria	SRR000340	01:44.0		
KmerUniq16Extend19Bacteria	SRR004172	02:40.8		
KmerUniq16Extend19Bacteria	SRR005754	03:48.9		
KmerUniq16Extend19Bacteria	SRR039956	02:59.0		
KmerUniq16Extend19Bacteria	SRR133640	03:47.8		

KmerUniq16Extend20Bacteria	DRR000002	05:36.9		
KmerUniq16Extend20Bacteria	DRR000184	03:03.0		
KmerUniq16Extend20Bacteria	ERR011207	13:12.5		
KmerUniq16Extend20Bacteria	ERR015579	06:13.7		
KmerUniq16Extend20Bacteria	SRR000340	01:22.3		
KmerUniq16Extend20Bacteria	SRR004172	02:06.1		
KmerUniq16Extend20Bacteria	SRR005754	03:13.7		
KmerUniq16Extend20Bacteria	SRR039956	02:44.2		
KmerUniq16Extend20Bacteria	SRR133640	02:52.5		
KmerUniq17Extend18Bacteria	DRR000002	3:17:20	37:16.9	0.19
KmerUniq17Extend18Bacteria	DRR000184	1:01:35	21:56.5	0.36
KmerUniq17Extend18Bacteria	ERR011207	2:29:15	47:02.0	0.32
KmerUniq17Extend18Bacteria	ERR015579	1:18:55	29:05.2	0.37
KmerUniq17Extend18Bacteria	SRR000340	39:55.6	17:11.7	0.43
KmerUniq17Extend18Bacteria	SRR004172	55:22.6	20:49.5	0.38
KmerUniq17Extend18Bacteria	SRR005754	1:38:57	26:17.6	0.27
KmerUniq17Extend18Bacteria	SRR039956	1:04:42	21:50.6	0.34
KmerUniq17Extend18Bacteria	SRR133640	3:07:59	35:49.0	0.19
KmerUniq17Extend19Bacteria	DRR000002	30:40.9		
KmerUniq17Extend19Bacteria	DRR000184	16:00.7		
KmerUniq17Extend19Bacteria	ERR011207	41:34.5		
KmerUniq17Extend19Bacteria	ERR015579	23:21.2		
KmerUniq17Extend19Bacteria	SRR000340	10:48.1		
KmerUniq17Extend19Bacteria	SRR004172	13:49.4		
KmerUniq17Extend19Bacteria	SRR005754	18:47.8		
KmerUniq17Extend19Bacteria	SRR039956	14:48.9		
KmerUniq17Extend19Bacteria	SRR133640	25:11.5		
KmerUniq17Extend20Bacteria	DRR000002	24:13.8		
KmerUniq17Extend20Bacteria	DRR000184	10:30.1		
KmerUniq17Extend20Bacteria	ERR011207	31:32.3		
KmerUniq17Extend20Bacteria	ERR015579	17:28.5		
KmerUniq17Extend20Bacteria	SRR000340	07:17.7		
KmerUniq17Extend20Bacteria	SRR004172	09:26.1		
KmerUniq17Extend20Bacteria	SRR005754	13:18.0		
KmerUniq17Extend20Bacteria	SRR039956	10:47.8		
KmerUniq17Extend20Bacteria	SRR133640	19:14.6		
KPATHbacteria	DRR000002	05:39.5		
KPATHbacteria	DRR000184	04:35.0		
KPATHbacteria	ERR011207	12:04.9		
KPATHbacteria	ERR015579	06:52.1		
KPATHbacteria	SRR000340	02:45.9		
KPATHbacteria	SRR004172	04:00.9		
KPATHbacteria	SRR005754	08:02.9		

KPATHbacteria	SRR039956	05:14.8
KPATHbacteria	SRR133640	03:35.1
PML_bacteria_unique_k19	DRR000002	46:27.1
PML_bacteria_unique_k19	DRR000184	1:01:30
PML_bacteria_unique_k19	ERR011207	1:26:41
PML_bacteria_unique_k19	ERR015579	1:05:04
PML_bacteria_unique_k19	SRR000340	10:04.9
PML_bacteria_unique_k19	SRR004172	13:16.2
PML_bacteria_unique_k19	SRR005754	27:24.3
PML_bacteria_unique_k19	SRR039956	16:40.4
PML_bacteria_unique_k19	SRR133640	36:36.4