Functional Analysis of Metagenomes by Likelihood Inference (FAMLI) Successfully Compensates for Multi-Mapping Short Reads from Metagenomic Samples

Jonathan L Golob[1,2¶] and Samuel S Minot[1*¶]

Affiliations

(1) Vaccine and Infectious Disease Division, Fred Hutch, Seattle, WA

(2) Division of Infectious Diseases, Department of Medicine University of Washington, Seattle, WA

* Corresponding Author:

Samuel S. Minot

1100 Eastlake Ave E. E4-100

Seattle, WA 98109

¶ These authors contributed equally to this work.

Abstract

Short-read metagenomics use high-throughput sequencing to reveal the functional capability of microbial communities. The computational analysis of the raw short-read data is complicated by the shared functional domains in peptides, and the complex evolutionary origins of many peptides (including recombination and truncation), resulting in a given short read aligning equally well to many possible peptides. We find that a short read known to be from one peptide will on average align equally well to about 160 other peptide sequences not present in the sample.

Here we describe an iterative algorithmic approach to successfully map reads to their true origin peptides and introduce a software package FAMLI that implements this algorithm. We demonstrate that FAMLI is able to identify peptides from a wide variety of metagenomes with a consistent precision of about 0.8, and recall of about 0.6, while retaining O(n) runtimes. This compares favorably to alternative approaches, including *de novo* assembly (that results in superior precision, but much inferior recall and much larger computational resource needs as compared to FAMLI), or hybrid taxonomic-identification approaches (that results in less consistent performance for extremely novel communities, as compared to FAMLI).

Addressing the problem of short reads aligning equally well to hundreds of more peptides than are truly present in a sample is a key challenge any successful short-read-metagenomics software must address. We present an effective approach to mitigate this problem and improve the accuracy of functional metagenomic analysis.

Author Summary

Proteins are typically made up of modules that are shared with other proteins; a given short piece of protein can be shared with hundreds of other identified proteins. As a result, it is difficult to try to identify which proteins are present in a collection of microbes using short pieces of proteins. If this problem is not addressed, studies that rely on short stretches of peptides will falsely report how many proteins are present in a microbial community. We developed an algorithm to address this problem and wrote an open source software package called FAMLI that can successfully identify which proteins are present, while ignoring most of those not present.

Introduction

High throughput (or "next-generation") sequencing has uncovered the hidden complexity of microbial communities living within and upon the human body, as well as the link between the human microbiome and health [1–4]. The taxonomic composition of a microbial community can be inferred by sequencing PCR amplicons spanning variable regions of a taxonomically informative gene (i.e. the 16S rRNA gene or the CPN60 gene) [5,6]. Alternatively, DNA recovered from a sample can be put through Whole-Genome Sequencing (WGS), which samples the complete genomic content of a sample via random fragmentation and sequencing [7]. While WGS is generally thought to cost more than 16S sequencing, it provides the advantages of (1) providing genomic data from all organisms in a sample—not limited to any single domain of life; (2) enabling a high degree of taxonomic resolution which identifies the subspecies and strains present in a sample; and (3) generating a "functional" metagenomic profile of the *genes* that are present in a sample in addition to the organisms which contain those genes [8].

There are three broad computational approaches used to generate an estimate of the functional metagenome (gene content) of a microbial community from WGS reads: (1) The inferred taxonomic composition can be used to construct a custom database of protein-coding genes from the set of reference organisms detected in the sample (e.g. HUMAnN2, MIDAS) [9,10]. (2) *De novo* assembly, in which the WGS reads are combined into contigs, which can be further used to identify open reading frames (e.g. SPAdes, IDBA-UD) [11,12]. (3) The WGS reads can be directly aligned

against a closed reference of protein coding sequences (which is included as a component of HUMAnN2).

Inferring gene content from taxonomic composition is complicated by frequent divergence between taxonomy and genome content, due to horizontal gene transfer and other non-clonal events that redistribute genomic content between taxa. *De novo* assembly of sequences into contigs is computationally slow and relatively inefficient insofar as it provides poor recall (or high false-negative rate) when determining the peptide content of a sample.

Mapping of reads to a closed reference of peptides is challenged by reads that align to multiple references: "multi-mapping" reads. We cannot assume that there is an underlying directed acyclic graph structure (e.g. a phylogeny) for peptides that can be used to resolve these multi-mapping reads. Peptides can evolve by duplication events, truncation, homologous recombination, and other means that result in the sharing of highly conserved domains between otherwise distinct peptide sequences.

Here we demonstrate the magnitude of the problem of multiple-mapping of short-reads to peptides, revealing a large number of equally-scored alignments; if one simply includes all peptides for which there is at least one short read that aligns equally as well as to any other peptide, the false positives outnumber true positives by an average of about 160:1. We describe an algorithmic approach to correctly assign these multiply aligned WGS reads to the proper reference sequence,

implemented as the open source software package FAMLI (Functional Analysis of Metagenomes by Likelihood Inference). With FAMLI, we are able to improve our precision (number of true positives divided by the sum of false and true positives) to about 80%; this performance is consistent over a range of community types. FAMLI is more efficient than *de novo* assembly at identifying protein-coding sequences present in a community with regards to both read depth and computational time. While FAMLI can be used as a standalone tool to identify protein-coding genes, it could also easily be used to enhance the precision of existing bioinformatics tools (e.g. HUMAnN2, MIDAS). We present FAMLI as an efficient and precise method to determine the metagenomic protein-coding gene content of a microbial community.

Results

Short reads from microbial metagenomics samples align equally well to multiple possible references

Three hundred and sixty simulated reads were generated from three peptides. These simulated reads were aligned against the UniRef100 database, and all alignments of equal quality to the best match were retained. This resulted in a total of 57,983 equally-well-scored-by-read alignments three true peptides, an average of 161 (median of 165, minimum of 1 and maximum of 384) equally-well aligning references for each read. More often, to account for sequencing errors and poor representation in the reference database, we accept alignments within a certain percentage of the best alignment for a given read. When we accept all alignments within about 10% identity ('top-10') of the best alignment for a read, this balloons

to 100,468 alignments for the 360 reads, an average of 279 (median of 268, minimum of 77 and maximum of 537) references recruited per read.

Figure 1 shows the FAMLI algorithm applied to the top-10 alignments. Figure 1A shows the coverage (or read depth by base pair) for the three true positive peptides, as well as the number of references-per-read at the start. After filtering for coverage evenness, figure 1B shows the read-depth of some successfully filtered away references, as well as some references not present in the simulated sample that pass this first test. By the conclusion of the first evenness filtering, 908 references remain (for the true three); the 360 reads remain with an average of 271 (median of 267, minimum of 77, and maximum of 398) equally-well aligning references.

Figure 1C depicts the iterative pruning of alignments by likelihood, showing the candidate references for one query being successfully filtered down over ten iterations to a single reference for the read (the true origin reference for this read). By the conclusion of ten iterations, all reads are successfully assigned now to their true origin reference (one reference per read).

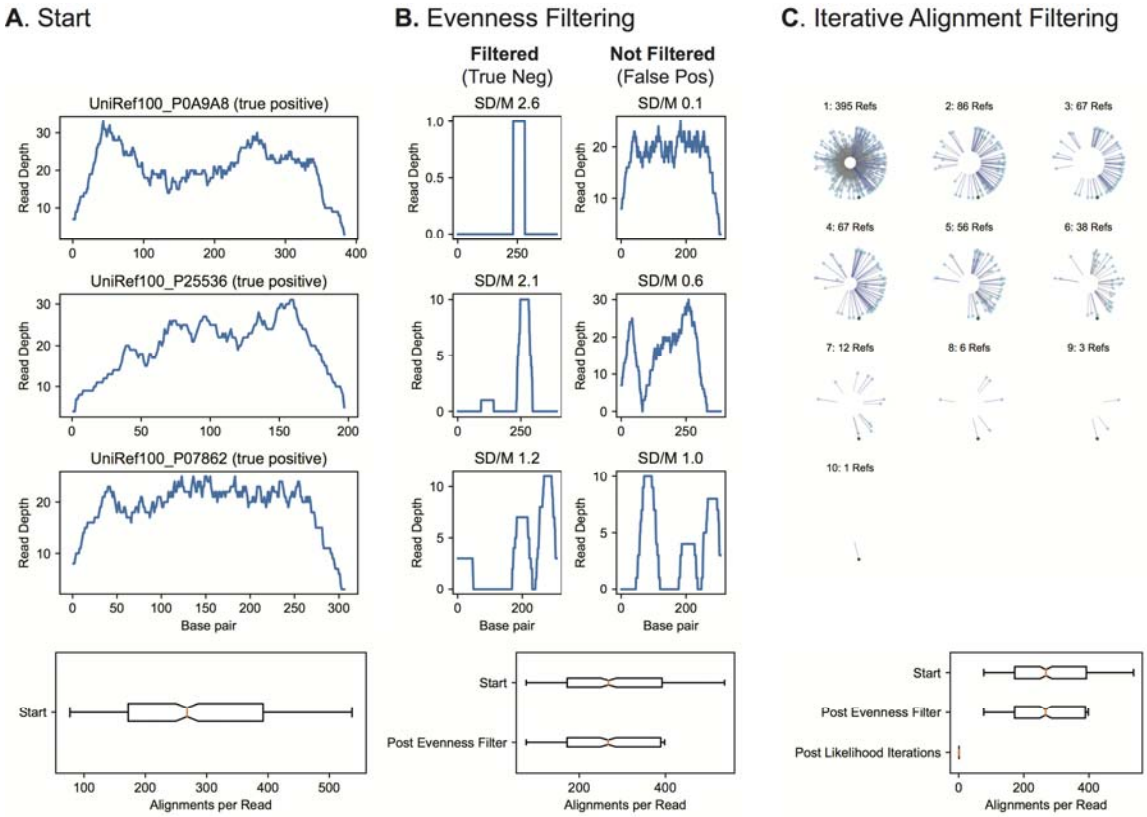

Figure 1: The problem of multiply-aligning short-reads, and the FAMLI algorithm schematized. Three hundred and sixty simulated reads were generated from three peptides. These simulated reads were aligned against the UniRef100 database, and all alignments of equal quality to the best match were retained. A) The read-depth coverage of the three true peptides. A given read has an average of 277 equally well-scoring alignments. B) Evenness filtering is used to remove the least likely to be present references from being considered. The left column is three randomly selected references that are successfully filtered at this step, the right three false references that are not filtered. C) The iterative likelihood-based filtering of one randomly selected reference's alignments. Each circle represents one remaining aligned reference for this read; the true positive origin reference is in

dark green. The length of each line is proportional to the calculated likelihood at this iteration. After the tenth iteration, only one reference (the correct) remains for this read.

FAMLI has consistent precision and recall for novel and reference organisms

We generated two simulated communities, each with twelve organisms. The organisms in the first simulated community are all well-described reference organisms, each with multiple available genomes (sim12-ref). The second simulated community is comprised of novel organisms (sim12-novel) each with no more than one complete genome available publicly in NCBI. Reads were generated *in-silico* via the ART software package. As our gold standard for protein-coding genes, we used Prokka to extract possible open reading frames (ORFs) from the available genomes; the resultant ORFs were aligned against UniRef90, and for each ORF the best hits (references that align to the ORF with the same quality as the best alignment score) are included as truly represented in the genome. Thus, we had a gold standard for both precision (true positives divided by the sum of true positives and false positives) and recall (true positives divided by the sum of true positives and false negatives). We also calculated the F-measure, which is the harmonic mean of precision and recall and quantifies overall accuracy.

We compared five approaches to identify which protein-coding sequences were present in the community: (i) *de novo* assembly via SPAdes, extraction of ORFs from contigs via Prokka, and then recruitment of UniRef90 groups aligning the closest to

each ORF; ii) a "Top 20" approach in which the WGS reads are aligned directly to UniRef90, and the top 20 hits are included for each read iii) a "Top Hit" approach in which the WGS reads are aligned directly to UniRef90, with all top-scoring protein-coding sequences included for each WGS read; iv) HUMAnN2 analysis on the WGS reads directly, using the "protein family" output table; and v) FAMLI output using the UniRef90 database. The accuracy for these five approaches is presented for both the sim12-ref (Figure 2) and sim12-novel (Figure 3) datasets, including the F-measure, recall, precision, and total false positive CDS ("coding sequence") for the entire community, as well as the recall rate for each individual organism.

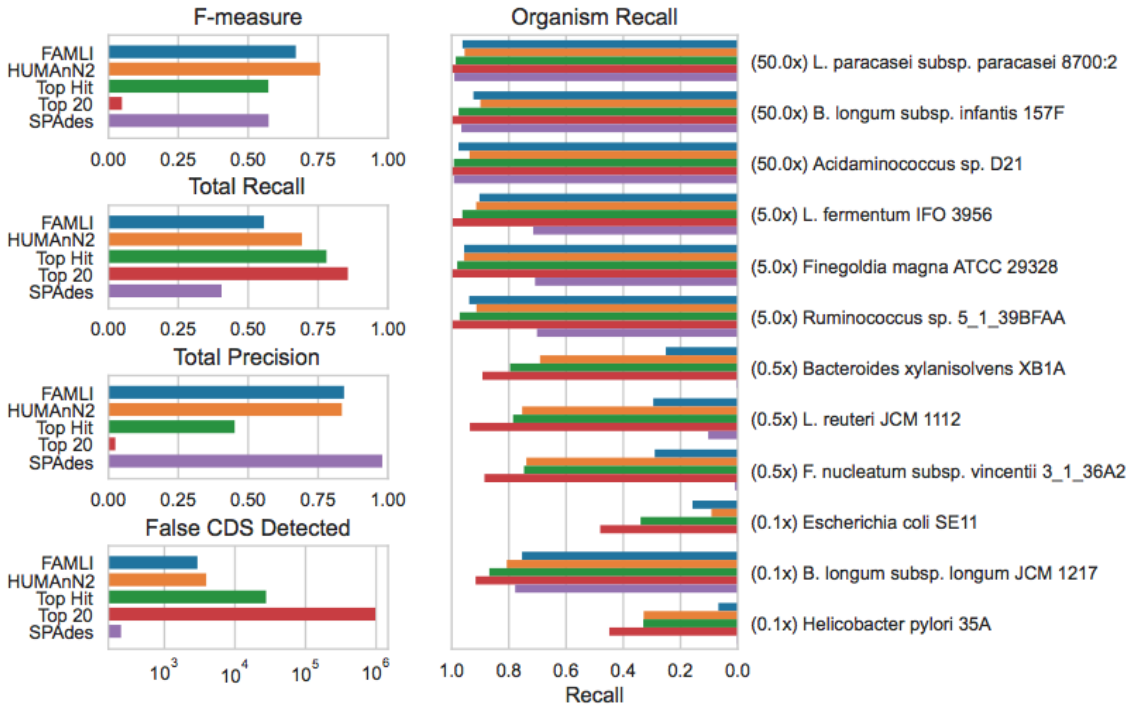

Figure 2. Accuracy of protein-coding sequence detection in the sim12-ref dataset. Accuracy metrics for the total community (F-measure, recall, precision, and number of false positive CDS detected) are shown on the left, with organism-specific

recall rates shown on the right. The same color and order is used to distinguish each computational method across all subplots.

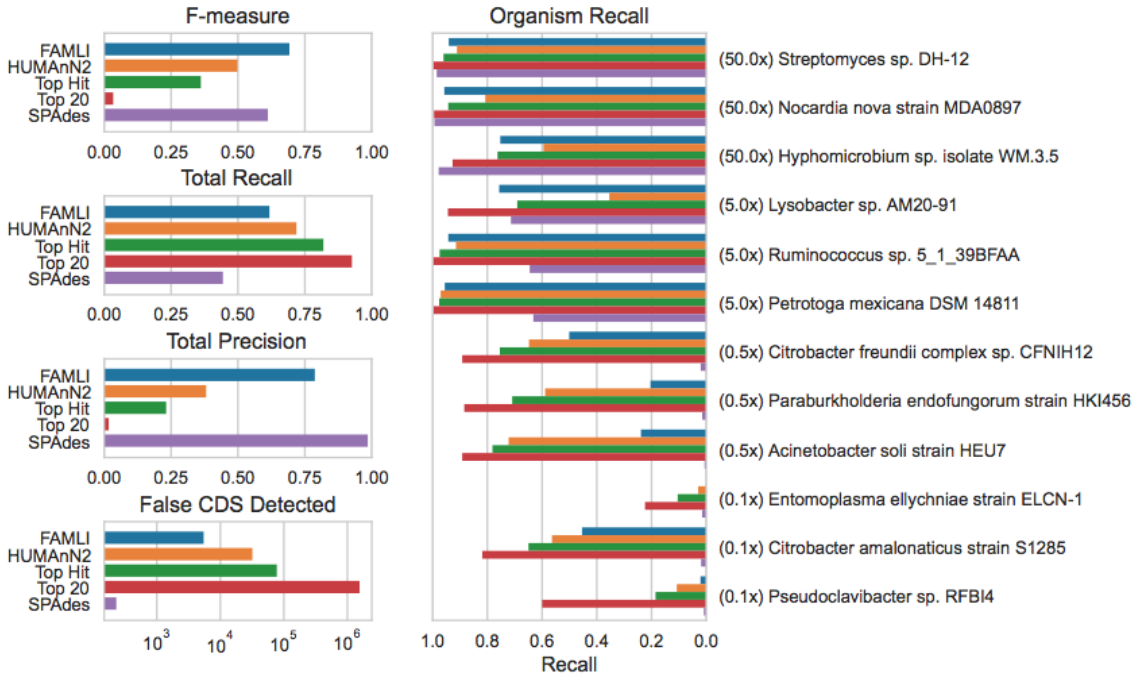

Figure 3. Accuracy of protein-coding sequence detection in the sim12-novel dataset. Accuracy metrics for the total community (F-measure, recall, precision, and number of false positive CDS detected) are shown on the left, with organism-specific recall rates shown on the right. The same color and order is used to distinguish each computational method across all subplots.

Dataset	Method	Precision	Recall	F-measure	False CDS
sim12-ref	FAMLI	0.84	0.56	0.67	2946
	HUMAnN2	0.84	0.69	0.76	3914
	Top Hit	0.45	0.78	0.57	27596

	Top 20	0.025	0.86	0.049	975275
	SPAdes	0.98	0.41	0.57	243
sim12-novel	FAMLI	0.79	0.62	0.69	5482
	HUMAnN2	0.38	0.72	0.50	31941
	Top Hit	0.23	0.82	0.36	77442
	Top 20	0.017	0.93	0.034	1557148
	SPAdes	0.99	0.44	0.61	230

Table 1. Accuracy of protein-coding sequence detection in the sim12-ref and sim12-novel datasets with five different computational methods.

As expected, *de novo* assembly has a nearly perfect precision (0.98 and 0.99 for the reference and novel communities respectively). The "Top 20" approach performed the most poorly, with a precision of 0.025 for reference organisms and 0.017 for novel organisms—reflecting the challenge of multi-mapping reads. HUMAnN2 had a huge variation in precision; on the community of reference organisms (sim12-ref, Fig. 2) precision was 0.83 versus only 0.38 on the novel organism simulated community (sim12-novel, Fig. 3). FAMLI was able to achieve a more consistent precision of 0.84 and 0.79 for the reference and novel organism communities respectively.

Recall was the worst for the *de novo* assembly, 0.40 for sim12-ref and 0.44 for sim12-novel, with *de novo* assembly struggling for most organisms with less than 5x coverage. FAMLI was able to recall 0.56 from sim12-ref and 0.62 from sim12-novel,

with most of the reduced recall concentrated in organisms with very low coverage. HUMAnN2 recovery rates were 0.77 for sim12-ref and 0.71 for sim12-novel. The Top Hit approach had recovery rates of 0.78 and 0.81 for the reference and novel communities respectively.

The F-measure score shows that FAMLI has the best combined performance of precision and recall across the two simulated datasets, with 0.67 in the sim12-ref dataset and 0.69 in the sim12-novel dataset. While HUMAnN2 provides a higher F-measure in the sim12-ref dataset (0.76), the result is lower in the sim12-novel dataset (0.50).

FAMLI is sequencing read efficient when identifying peptides from a sample

Next, we analyzed a set of stool samples analyzed as part of the Human Microbiome Project (HMP). On the basis of our simulated communities, in which peptides identified via *de novo* assembly were very likely to truly be present in the sample, we used peptides identified by assembly using complete set of reads as our target. The WGS reads for each biological specimen were subsampled at a range of levels, from 50,000 to 10,000,000. Figure 4 shows the detection rate of protein-coding sequences confirmed by assembly (vertical axis) versus subsampling (horizontal axis), as well as the rate of peptides recovered that were not recovered by *de novo* assembly.

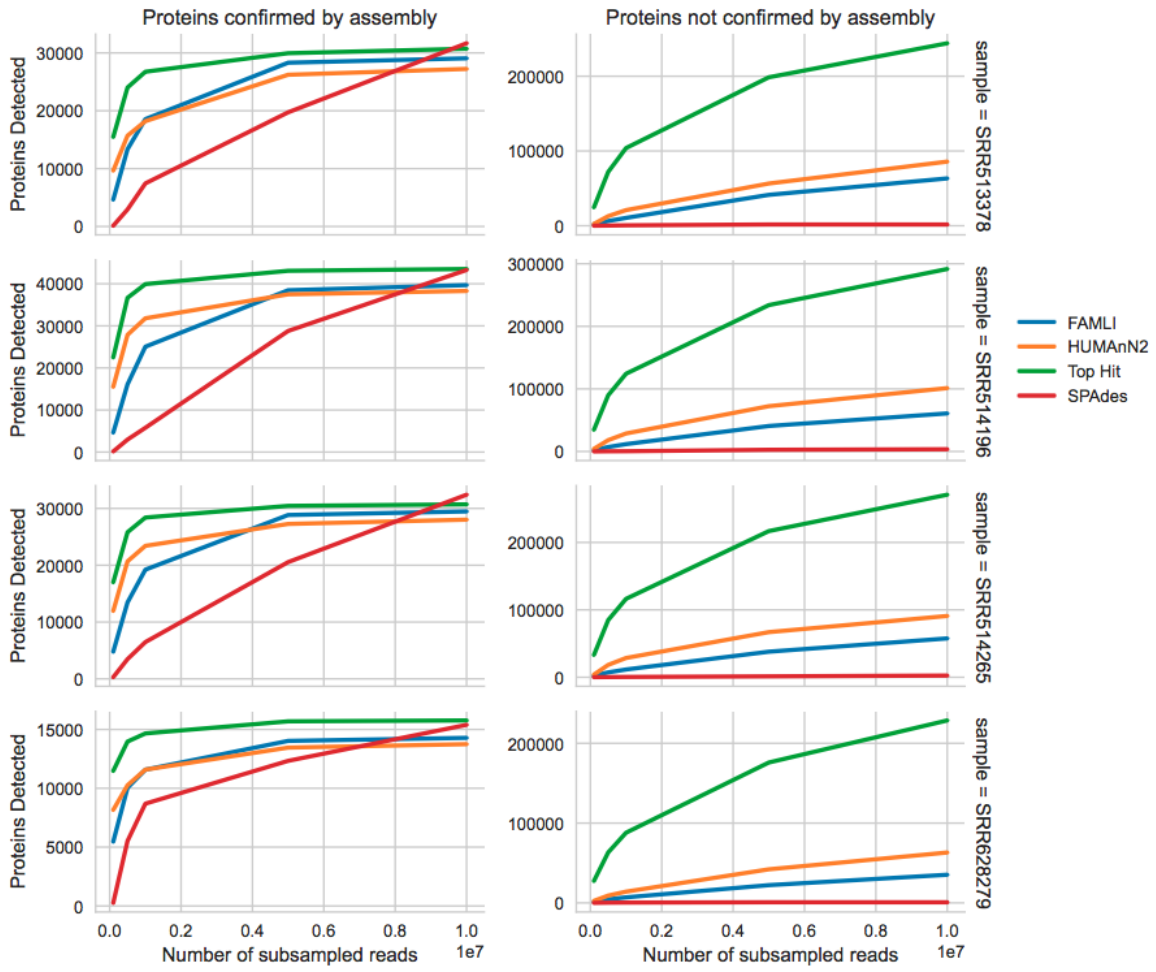

Figure 4. Efficiency and accuracy of protein-coding sequence detection in computationally subsampled stool microbiome samples. The number of WGS reads subsampled in indicated with the horizontal axis, while the number of protein-coding sequences detected by each method is indicated on the vertical axis. The left-hand panels indicate the set of protein-coding sequences that were confirmed by assembly with SPAdes, while the right-hand panel indicates the set of protein-coding sequences that were not confirmed by assembly. Each of the three rows of panels represents a different biological sample, which is identified by its SRR accession on the right hand side of the plot.

FAMLI and HUMAnN2 each identified a similar number of protein-coding sequences that were confirmed by assembly across this range of sequencing depths. Both of those methods were able to identify a greater number of confirmed sequences than *de novo* assembly at a given depth of sequencing. For the peptides identified that are not found in assembly, it is difficult to discern if these are true positives missed by assembly or false positives. Nevertheless, the Top Hit method identified the largest number of proteins which were not confirmed to be present via *de novo* assembly, while HUMAnN2 identified the second-most, and FAMLI detected the least.

FAMLI can be used to detect protein-coding sequences associated with disease from human microbiome samples

To test the application of FAMLI to an existing microbiome experiment, we reanalyzed the WGS data generated by Byrd, et al. (2017) from skin swabs of pediatric participants before, during, and after atopic dermatitis flares. Runtime was O(N) per number of reads (Figure 5). Out of the 745,534 UniRef90 groups identified across all 11 participants, only 724 were significantly enriched (GEE p-value < 6.7e-8, using the Bonferroni correction) and detected in at least 3 participants (Figure 6).

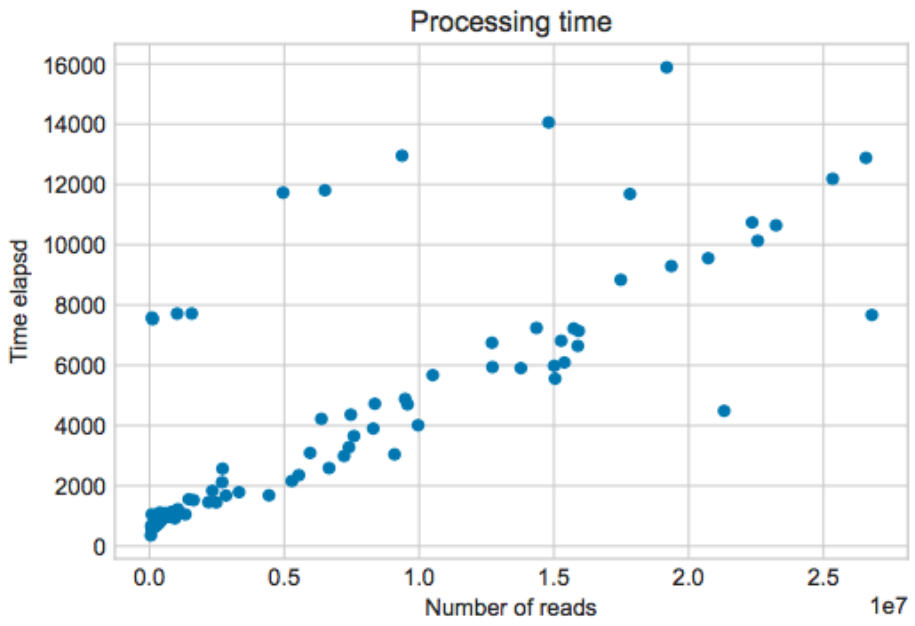

Figure 5. Processing time for FAMLI analysis of Byrd, et al. 2017 dataset. The number of WGS reads in each dataset is indicated on the horizontal axis, while the processing time required to (1) download the WGS data, (2) align against UniRef90 with DIAMOND, and (3) filter all resulting alignments with FAMLI is displayed on the vertical axis. Linear regression indicates that the processing time on average is 2,443 WGS reads per second, as well as an additional 1,286 second fixed cost for every analysis (database loading, download from SRA, etc.).

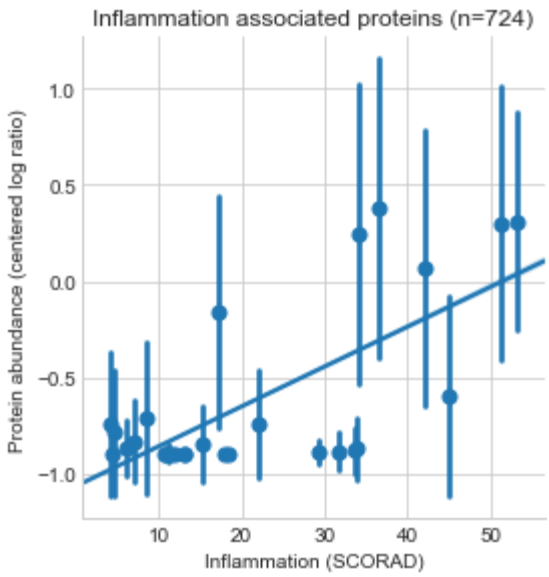

Figure 6. Detection of disease-associated protein coding sequences in the skin microbiome during atopic dermatitis. FAMLI used to reanalyze a study of the short read metagenomes of skin microbiomes in pediatric patients before, during, and after an atopic dermatitis flare. As inflammation increased, microbial peptides previously associated with inflammation increased in abundance.

The vast majority of inflammation-associated UniRef90 groups with a taxonomic annotation were annotated consistently with *Staphylococcus aureus* at the species, genus, and family levels (646 out of 695) (Table S1), confirming the previous hypothesis that atopic dermatitis is associated with outgrowths of *S. aureus*. A smaller set of proteins were annotated taxonomically with other skin-associated bacteria, including *Streptococcus pneumoniae* (n=7), *Acinetobacter baumannii* (n=2), and *Clostridioides difficile* (n=2). Moreover, virulence-associated functions were enriched in the set of inflammation-associated genes, including ABC transporters, adhesins, and siderophores (Table S1). In addition, we found a large number of

inflammation-associated genes with no annotated biological function (252 out of 724). With no prior assumptions about the organisms or metabolic pathways that may have been associated with disease, we were able to use FAMLI to identify the primary disease causing organism, while also identifying the protein coding genes (many lacking any prior annotation) with the highest degree of correlation with disease.

Discussion

The problem of short-reads from metagenomes aligning equally well to large numbers of distinct peptide sequences is a critical one that must be addressed in any successful analysis of short-read metagenomics data that involves alignment to a reference. As evident in our simulated communities, the ratio of true to false positive alignments can be in the hundreds to one, resulting in dismal precision unless the alignments are culled or filtered. We suspect some of the limitations experienced by software attempting to use short reads to identify the functional genes encoded by microbial communities, described by [14], may be due to this multi-mapping read problem.

Here we establish FAMLI as a precise and efficient (both in read depth and computation time) manner to overcome the multi-mapping read problem and precisely discerning the peptide content of a complex microbial community from randomly fragmented WGS reads. With FAMLI we were able to reproduce and

extend the findings of an existing study, pointing the ways by which FAMLI can facilitate future research.

As we considered the computational methods that are used to analyze metagenomic WGS datasets, our goal was to generate a biologically meaningful representation of functional genomic content while minimizing the number of *a priori* assumptions and maximizing the efficiency of comparison across large numbers of samples.

One of the most common *a priori* assumptions that is made by metagenomic analysis tools is the pre-estimation of species abundance, which is then used to subset the database of potential genes in the sample (e.g. HUMAnN2, MIDAS). While that approach has merit, we felt strongly that the introduction of that additional analytical step ran the risk of introducing analytical bias towards organisms with larger collections of reference genomes. The bias towards well-sampled organisms is observed when comparing the precision of HUMAnN2 in the sim12-ref dataset, compared to the sim12-novel dataset (Fig. 2). The consistent performance of FAMLI across these two datasets is likely due to the complete lack of any species-genome model used for that analysis.

Thinking about the relative merits of reference-based (e.g. UniRef90) or reference-free (e.g. *de novo* assembly) analysis methods, one of our primary considerations was the efficiency of comparing results across large numbers of samples. While reference-free approaches are free by definition from the bias inherent in reference

databases, that lack of common reference makes it extremely challenging to compare results between samples. For example, comparing a set of genes between N samples is an O(N^2) problem that scales exponentially with the number of samples. In contrast, by identifying proteins from a reference database (UniRef90), all results are inherently comparable without any additional computation (e.g. sequence alignment), in other words the complexity is O(1). Put simply, with *de novo* assembly (SPAdes) it is *much* more difficult to compare the results for 1,000 samples in contrast to just 10 samples, while for FAMLI or HUMAnN2 it is about the same.

Having formulated a computational approach that provides a high degree of accuracy on control samples (mock communities), we presented results from a clinical microbiome study with the goal of demonstrating how this analytical approach can be used by other researchers in the field. Without any prior hypotheses about the organisms or metabolic pathways implicated in disease, we were able to quickly identify a set of disease-associated protein-coding sequences. The set of proteins we identified are consistent with the outgrowth of pathogenic *S. aureus*, as is recognized for atopic dermatitis [15]. The specific protein-coding sequences identified in this analysis represent the subset of the *S. aureus* pan-genome which are most tightly linked to pathogenesis in this set of patients. In addition to a set of known pathogenesis mechanisms (iron siderophores, adhesion, and ABC transporters) we present a set of disease-associated protein-coding

sequences with no previously annotated function that may present opportunities to discover novel virulence mechanisms in atopic dermatitis.

The first application we would like to stress is that the FAMLI approach could be easily applied as part of larger bioinformatics pipelines that incorporate alignment against reference databases (e.g. HUMAnN2, MIDAS). Those pipelines produce a large amount of data over and above this implementation of FAMLI, and we have no goal of trying to replicate those useful tools. Instead, we would be happy to help contribute this algorithmic approach to their future enhancement and development.

The second application we hope to demonstrate is the detection of disease-associated microbial genes in the absence of any *a priori* information about the organisms that encode them. Microbial genomes are highly plastic, and microbiome-associated diseases may well be linked to functional genes that have not been sequenced in culturable reference organisms. By first identifying a set of genes of interest, additional analytical tools (e.g. *de novo* assembly) can be used in a targeted manner to reconstruct the organisms that encode them. Above all we see the FAMLI approach as a means of accelerating the *hypothesis-generating* process of microbiome research.

Materials and Methods

FAMLI Overall Approach

1. Align all input nucleotide reads in amino acid space (using the six-frame conceptual translation) against a reference database of peptides.
2. Calculate the coverage depth (CD) across the length of each reference, representing the number of reads aligning to each amino acid position of the reference.
3. Filter out any reference sequences with highly uneven coverage: $\frac{CD_{STD}}{CD_{Mean}} < 1.0$ (1)

   Where STD is standard deviation.
4. Calculate initial likelihoods for a given query coming from a subject using the alignment scores to weight the relative possibilities for a given query, normalizing the scores to total to 1 for a given query.
5. Iteratively, until no further references are pruned or a maximum number of iterations is reached:
   i. WEIGHTING and RENORMALIZING: The likelihoods of queries being from a subject from the prior iteration are weighted by the sum of likelihood for a given subject, and then renormalized to sum to 1 for each query.
   ii. PRUNING. Determine the maximum likelihood for each query. Prune away all other likelihoods for the query below a threshold.
6. Repeat filtering steps 2-3 using the set of deduplicated alignments resulting from step 4.

An interactive demonstration of our algorithm is available as a Jupyer notebook in the online supplemental only materials.

Here are some examples:

- For reference A and reference B that both have some aligning query reads, if **there is *uneven* depth for reference A** but relatively even depth across reference B, then **reference A is removed from the candidate list** while reference B is kept as a candidate.
- If query **read #1 aligns equally-well to reference A and reference C**, but **there is *2x more* query read depth for reference A as compared to reference C** across the entire sample, then **reference C's alignment is removed from the list of candidates for query read #1**.

Filtering subject reference peptides by coverage evenness

Coverage evenness is calculated for every reference peptide sequence (Step 2, above). On a per-amino-acid basis, alignment depth is calculated using an integer vector. It is expected that the 5' and 3' ends of the reference will have trail offs, thus the depth vector is trimmed on both the 5' and 3' ends. A mean coverage depth and the standard deviation of the mean are calculated. The standard deviation is divided by the mean. Both based on the Poisson distribution and some empirical efforts on our part, we set a threshold of 1.0 for this ratio as a cutoff of uneveness; **references with a coverage SD / MEAN ratio > 1.0 are filtered (Step 3, above)**.

Defining alignment likelihood

Let us consider the **likelihood that a given query *q* is truly from a given reference r** considering all of the evidence from all of the queries in a sample. For the terms of this discussion, we will describe this as the **likelihood** (L_{qr}) for a given assignment.

For this application we use the **bitscore**—an integrated consideration of the alignment length, number of mismatches, gaps, and overhangs—as a way of *initialization* for our likelihood calculations: Bitscore_{qr} is the quality of the alignment of query read *q* to reference peptide *r*.

We can use the bitscore of an alignment divided by the sum of bitscores for all the alignments for a given query sequence as a starting estimated **likelihood** of any individual query *q* is truly derived from a reference *r*: $\mathbf{L_{qr}}$.

$$L_{qr} = {Bitscore_{qr}}\Big/{\sum_{r=1}^{n_r} Bitscore_{qr}} \qquad (2)$$

Where n_r is the number of references.

We then begin an iterative process. In each iteration:

1) We calculate the **total weight** for every reference r, $\mathbf{TOT_r}$ using the **likelihood from the prior iteration, $\mathbf{LP_{qr}}$.**

   $$TOT_r = \sum_{r=1}^{n_r} LP_{qr} \qquad (3)$$

2) We calculate the new **likelihood** that any individual query *q* is truly derived from a reference *r*, $\mathbf{L_{qr}}$, by first calculating a weighted likelihood $\mathbf{W_{qr}}$

$$W_{qr} = LP_{qr} * TOT_r \qquad (4)$$

and determine the new likelihoods by renormalizing the likelihoods for each query to 1.0.

$$S_q = \sum_{q=1}^{n_q} W_{qr} \qquad (5)$$

$$L_{qr} = {W_{qr}}/{S_q} \qquad (6)$$

3) The **maximum likelihood for query q, $\mathbf{Lmax_q}$** is determined using the newly calculated likelihoods.

$$Lmax_q = \max\,(L_{qr}\ for\ all\ r) \qquad (7)$$

4) If the L_{qr} falls below the scaled maximum likelihood for query q, the **alignment is removed from consideration**:

For all query q, if

$$L_{qr} < scale * Lmax_q \qquad (8)$$

then L_{qr} is set to zero.

By default the scale here is set to 0.9 (or 90% of the maximum likelihood for query q).

Iterations are repeated until no more alignments are culled or a maximum number of iterations are reached.

FAMLI Implementation

FAMLI is available as an open source software package on GitHub at https://github.com/FredHutch/FAMLI. In addition, Docker images are provided at

https://quay.io/repository/fhcrc-microbiome/famli to facilitate easy usage by the research community with a high degree of computational reproducibility. FAMLI can be run with the single executable "famli.py", which encompasses:

1. Downloading reference data and query FASTQ files (supporting SRA accessions, local paths, FTP, and Amazon Web Service (AWS) S3 buckets)
2. Aligning query FASTQ files with DIAMOND
3. Parsing the translated alignments
4. Running the FAMLI algorithm to filter unlikely reference peptides and assign multi-mapping query reads to a single unique reference.
5. Summarizing the results in a single output file
6. Copying the output file to a remote directory (supporting local paths, FTP, and AWS S3 buckets)

The help flag ("-h" or "—help") can be used to print a complete list of options.

Comparison of FAMLI to HUMAnN2, SPAdes, Top Hit, and Top 20

The version of FAMLI presented in this paper is "v1.0", which can be found at https://github.com/FredHutch/FAMLI/releases/tag/v1.0. The version of DIAMOND used for translated nucleotide alignments in FAMLI v1.0 is DIAMOND v0.9.10.

Comparative analysis of the simulated communities used HUMAnN2 v0.11.1--py27_1, and all code used to run HUMAnN2 can be found in the GitHub repository https://github.com/FredHutch/docker-humann2 (v0.11.1--6), which is based on the BioBakery Docker image quay.io/biocontainers/humann2:0.11.1--py27_1.

The code used to run SPAdes can be found in the GitHub repository https://github.com/FredHutch/docker-SPAdes (v3.11.1--3), which uses SPAdes v3.11.1 and Prokka v1.11.

The "Top Hit" and "Top 20" results were generated by aligning raw reads with DIAMOND (v0.9.10) using the following settings:

- Top Hit: min-score=20, query-cover=90, top=0, id=90, query-gencode=11
- Top 20: max-target-seqs=20, query-gencode=11

The reference database used for the analysis was UniRef90 (www.uniprot.org/uniref/) [16], downloaded on January 30th, 2018.

Simulation of microbial communities

Microbial communities were simulated using ART v2.5.8, with default simulation settings for the Illumina HiSeq2500 instrumentation. Reference genome sequences were downloaded from NCBI and simulated at a range of sequencing depths (from 0.1x to 50x).

Analysis of the Byrd, et al. 2017 atopic dermatitis dataset

We used FAMLI to analyze a dataset of skin microbiome samples collected from pediatric participants experiencing atopic dermatitis [13]. The sample name, human participant ID, body side, and inflammation SCORAD was obtained from Table S3 of that publication, referencing WGS datasets submitted to NCBI as part of BioProject PRJNA46333. FAMLI was run using the Docker image hosted publicly on Quay

(quay.io/fhcrc-microbiome/famli:v1.0) and the UniRef90 database downloaded on January 4th, 2018. Following the methods from that study, we combined the results for the samples collected from the antecubital crease (Ac) and popliteal crease (Pc) for each participant at each timepoint. Detection of protein-coding sequence were correlated with inflammation was performed using GEE (from the sklearn package in Python), using a Gaussean variance model and independence by human participant. We applied a Bonferroni correction for an adjusted p-value threshold of 6.8e-8, and further filtered the protein-coding sequences to those found in at least 3 different participants.

Acknowledgements

We would like to acknowledge the Microbiome Research Initiative at the Fred Hutch, lead by David N Fredricks, for supporting this work and Dan Tenenbaum from the scientific computing group at the Fred Hutch for help with establishing computational resources for this manuscript.

References

1. Cho I, Blaser MJ. The human microbiome: at the interface of health and disease. Nat Rev Genet. Nature Publishing Group; 2012;13: 260–270. doi:10.1038/nrg3182
2. Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, et al. Structure, function and diversity of the healthy human microbiome. Nature.

Nature Publishing Group; 2012;486: 207–214. doi:10.1038/nature11234

3. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature. Nature Publishing Group; 2010;464: 59–65. doi:10.1038/nature08821

4. NIH HMP Working Group TNHW, Peterson J, Garges S, Giovanni M, McInnes P, Wang L, et al. The NIH Human Microbiome Project. Genome Res. Cold Spring Harbor Laboratory Press; 2009;19: 2317–23. doi:10.1101/gr.096651.109

5. Methé BA, Nelson KE, Pop M, Creasy HH, Giglio MG, Huttenhower C, et al. A framework for human microbiome research. Nature. Nature Publishing Group; 2012;486: 215–221. doi:10.1038/nature11209

6. Golob JL, Margolis E, Hoffman NG, Fredricks DN. Evaluating the accuracy of amplicon-based microbiome computational pipelines on simulated human gut microbial communities. BMC Bioinformatics. 2017;18. doi:10.1186/s12859-017-1690-0

7. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Shotgun metagenomics, from sampling to analysis. Nat Biotechnol. 2017;35: 833–844. doi:10.1038/nbt.3935

8. Scholz MB, Lo C-C, Chain PS. Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. Curr Opin Biotechnol. Elsevier Current Trends; 2012;23: 9–15. doi:10.1016/J.COPBIO.2011.11.013

9. Abubucker S, Segata N, Goll J, Schubert AM, Izard J, Cantarel BL, et al. Metabolic Reconstruction for Metagenomic Data and Its Application to the

Human Microbiome. Eisen JA, editor. PLoS Comput Biol. Public Library of Science; 2012;8: e1002358. doi:10.1371/journal.pcbi.1002358

10. Nayfach S, Rodriguez-Mueller B, Garud N, Pollard KS. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. Genome Res. 2016;26: 1612–1625. doi:10.1101/gr.201863.115

11. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol. Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA ; 2012;19: 455–477. doi:10.1089/cmb.2012.0021

12. Peng Y, Leung HCM, Yiu SM, Chin FYL. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics. Oxford University Press; 2012;28: 1420–1428. doi:10.1093/bioinformatics/bts174

13. Byrd AL, Deming C, Cassidy SKB, Harrison OJ, Ng W, Conlan S, et al. Staphylococcus aureus and Staphylococcus epidermidis strain diversity underlying pediatric atopic dermatitis. Sci Transl Med. 2017;9: 1–13. doi:10.1126/scitranslmed.aal4651

14. Carr R, Borenstein E. Comparative Analysis of Functional Metagenomic Annotation and the Mappability of Short Reads. Gibas C, editor. PLoS One. Public Library of Science; 2014;9: e105776. doi:10.1371/journal.pone.0105776

15. Geoghegan JA, Irvine AD, Foster TJ. Staphylococcus aureus and Atopic Dermatitis: A Complex and Evolving Relationship. Trends Microbiol. Elsevier; 2017;0. doi:10.1016/j.tim.2017.11.008

16. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics. Oxford University Press; 2015;31: 926–932. doi:10.1093/bioinformatics/btu739