JOHNS HOPKINS
UNIVERSITY

Zanvyl Krieger School of Arts and Sciences
Advanced Academic Programs

410.734.81 and 410.734.82
Practical Introduction to Metagenomics

Topic: Community profiling

Instructor: Joshua Orvis

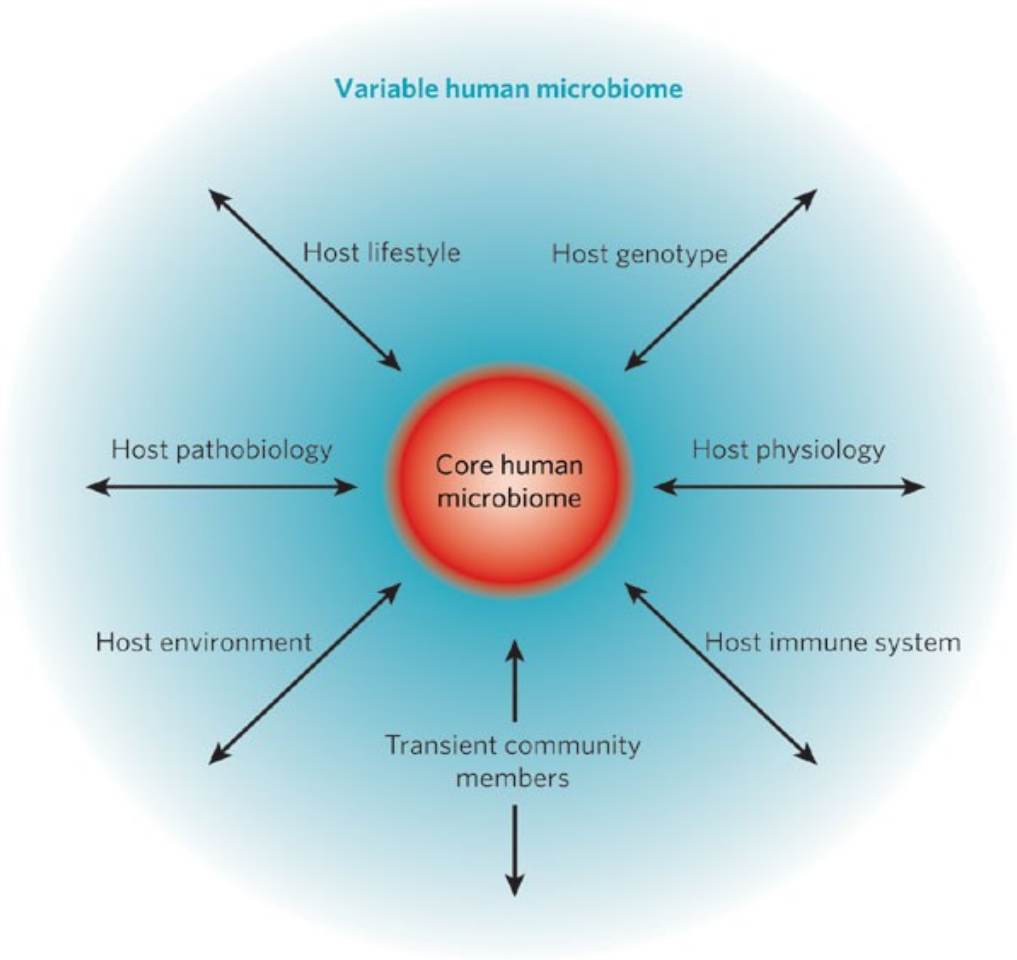
What is community profiling and why do it?

Community profiling is the determination of the abundance of each kind of microbe in a sample. We do it both to see what organisms are present in a sample and to see how changes in that profile affect the habitat. In the case of human microbiome studies, there is a lot of interest in how changes in that community profile affects our health.

The NIH, EU and HMPs question whether there is a **core human microbiome** of genes or species that we all share.

Are changes in the relative abundances of members of human-associated microbial communities important?

One possibility is that everyone shares the same microbial species but that the abundance of individual species varies by orders of magnitude in different people in ways that affect health and disease.

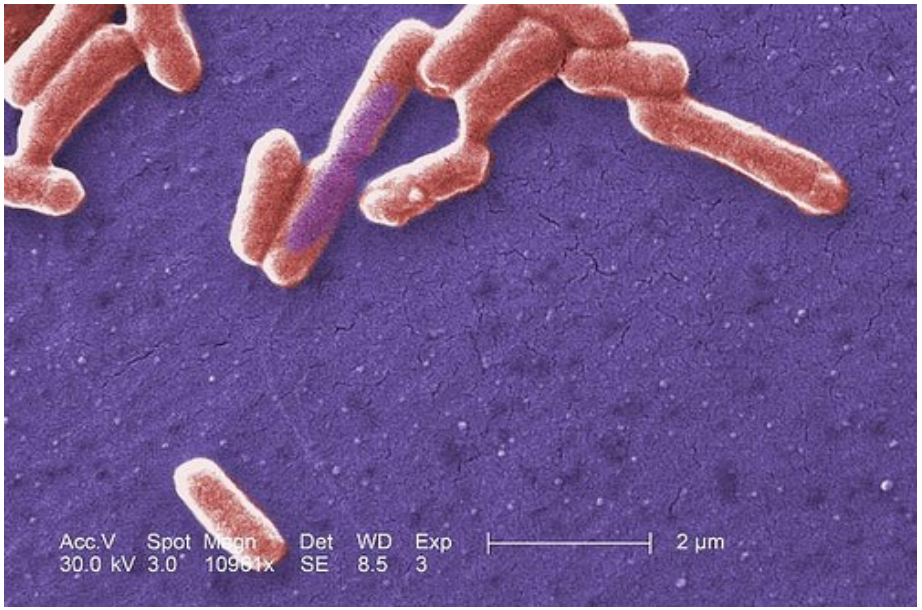


Hamady & Knight, 2009

Turnbaugh, et al. *Nature*. 2007

Big and little changes can matter

For example, the proportional representation of the bacterial phyla Fermicutes, Actinobacteria, and Bacteroidetes in the gut is associated with obesity in both humans and mice (Ley et al. 2005, 2006c; Turnbaugh et al, 2006, 2008).



For example, the proportional representation of the bacterial phyla Fermicutes, Actinobacteria, and Bacteroidetes in the gut is associated with obesity in both humans and mice (Ley et al. 2005, 2006c; Turnbaugh et al, 2006, 2008).

These establish the observed affects of broad bacterial groups, but the organisms with relatively smaller counts (such as pathogens) can be transformative as well.

For example, we think of *E. coli* as a classic gut bacterium, but the entire Gamma-proteobacteria phylum that contains it typically comprises much less than 1% of gut bacteria – rather, *E. coli* just grows well in culture and can thus be detected a low abundance.

If rare species are generally important, much deeper characterization of the microbiome may be required.

Hamady & Knight, 2009

Community profiling methods

In any given metagenomics project two of the most common questions are, “What organisms are there, and what can they do?”

Community profiling attempts to answer the first of these, and there are several different general approaches for doing this, among them are:

- Read/contig alignment to reference genomes
- Ribosomal targeted sequencing
- Alignment to marker genomic elements (non-ribosomal)

This week we're going to discuss the first two of these approahces.

What is the difference between aligning to reference genomes and ribosomal collections?

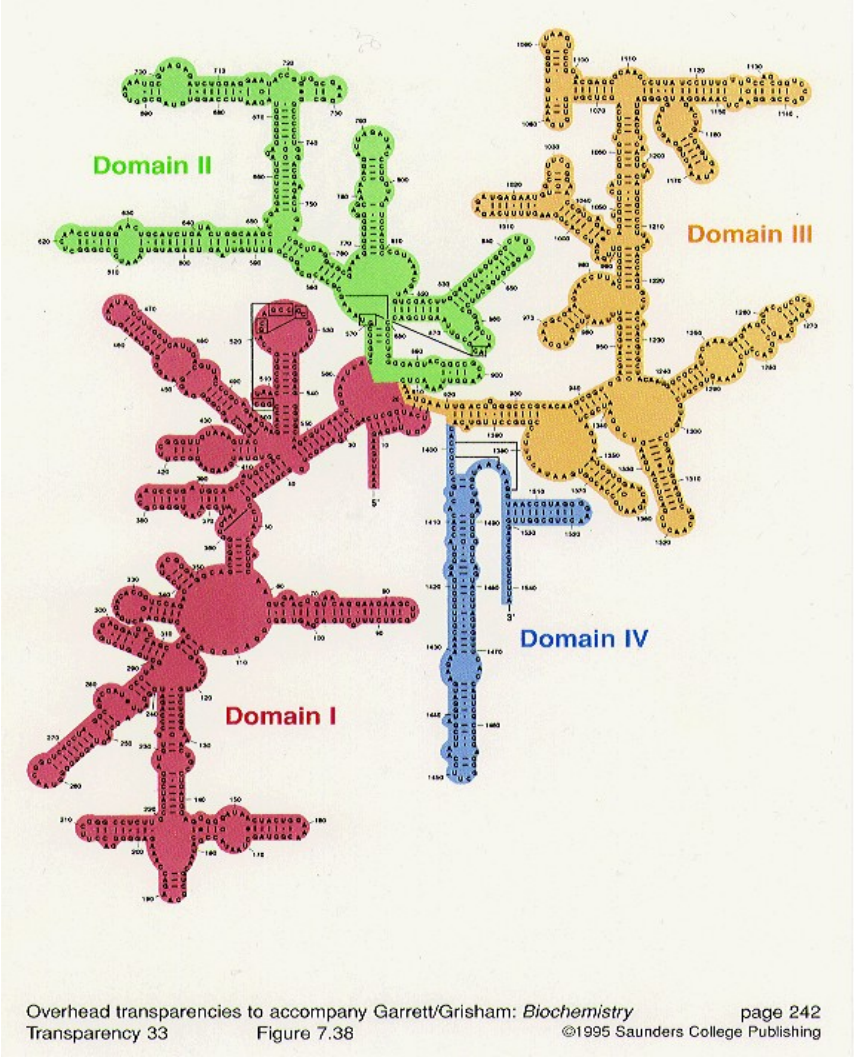
What can you do with the results of each?

What sort of sequence data is required for each?

What are the costs involved (both financially and computationally)?

Targeted rRNA vs. metagenomic sequencing

The 16S (for prokaryotes and archaea) and 18S (for eukaryotes) rRNA subunits are often targeted for sequencing and are used as stable phylogenetic markers to define which lineages are present in a sample.



It is much cheaper to use rRNA because only one gene out of each genome is examined, but metagenomic profiles are essential for understanding the functions encoded in those genomes.

16S rRNA sequencing studies are especially useful for characterizing which kinds of organisms are present in a wide range of samples, especially when differences at or above the genus level distinguish the samples.

Metagenomic studies are especially useful for characterizing microbial assemblages at a functional level.

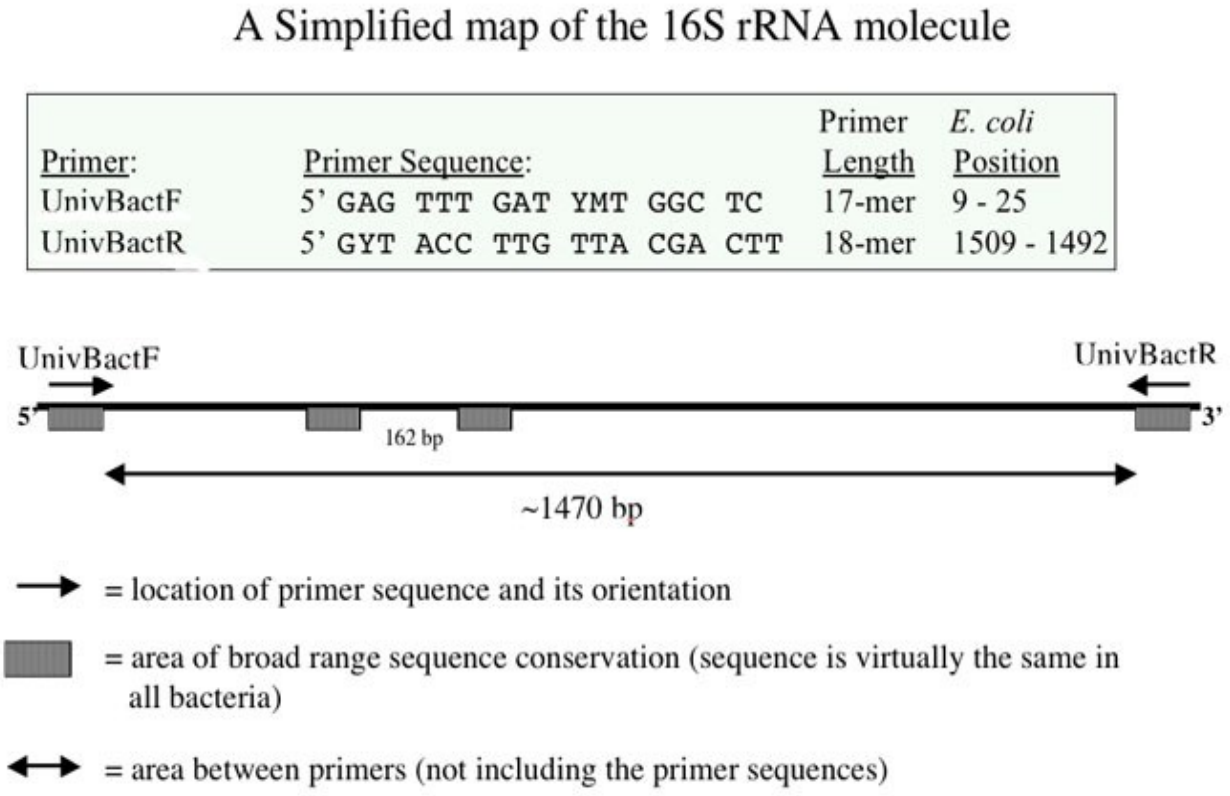
Hamady & Knight, 2009

16S rRNA sequencing

The ribosomal subunits (such as 16S) are so critical to the function of a cell that they remain highly conserved relative to most other genes and are universally present in organisms of their type (prok, euk, arch.)

The diagram below shows the conventional PCR strategy for targeted RNA sequencing. Primers are designed for the ultra-conserved regions on each end and pointed inward. With sufficient read length (Sanger), the entire rRNA subunit can be sequenced.

Thankfully for short-read sequencers, 250bp reads can be essentially as good as full-length sequences for many microbial community comparisons and can even be useful for taxonomy assignment, provided that the region of the 16S rRNA is carefully chosen, for example the V2 or V4 regions (Liu et al. 2007, 2008; Wang et al. 2007)

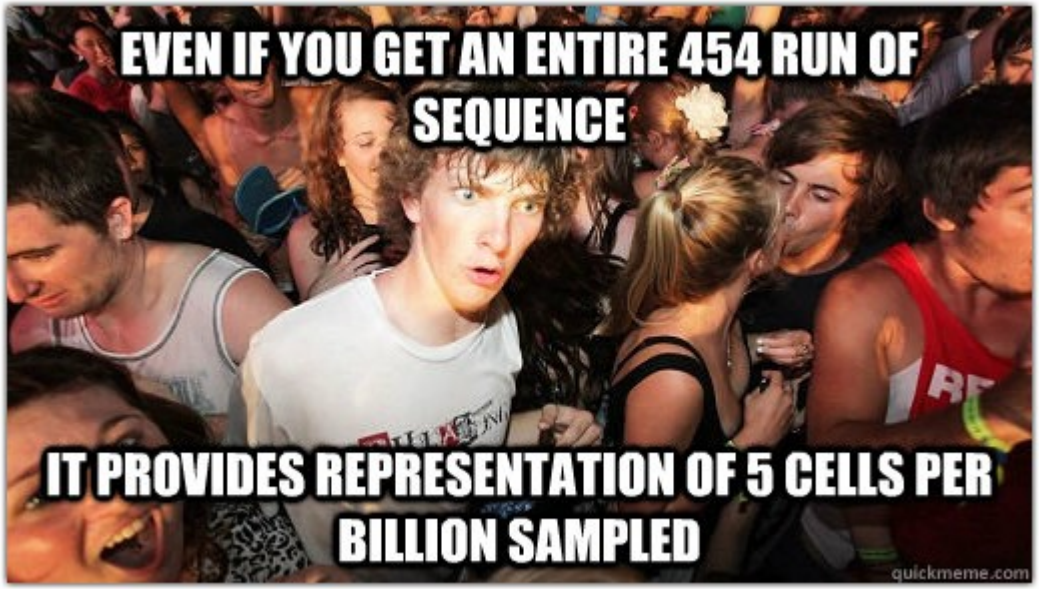


Some disappointing math

If the goal is complete characterization of all sequences in a sample, vast numbers of sequences may be required if many species are rare of if the diversity is high, such as in seawater or soils. Such a tiny fraction of the total number of cells is sampled that *characterizing the full, absolute diversity is not a reasonable goal*.

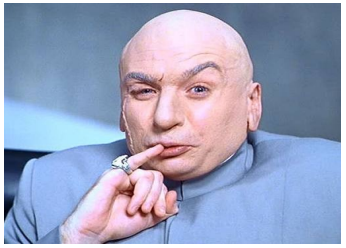
Nonsense you say? The math:

A full 454 run currently recovers ~$5x10^5$ sequences, so if there are ~10^{14} microbes in the gut, that means:



At an estimated $10k per 454 run, the 200 million runs it would take (minimum) to get at coverage of at least 1 16S rRNA molecule per cell would cost

$2,000,000,000,000



The point here is just to keep in mind the scale of what your sequences represent and know that low-abundance organisms can be problematic.

Hamady & Knight, 2009

Metagenomic sequencing

Both 16S sequencing and metagenomic sequencing have their specific applications, benefits and drawbacks. Indeed, larger studies like the HMP employ both as part of their analysis. See the Hamady and Knight paper in the “Further Reading” slide at the end of this presentation more comparisons and descriptions of the two.

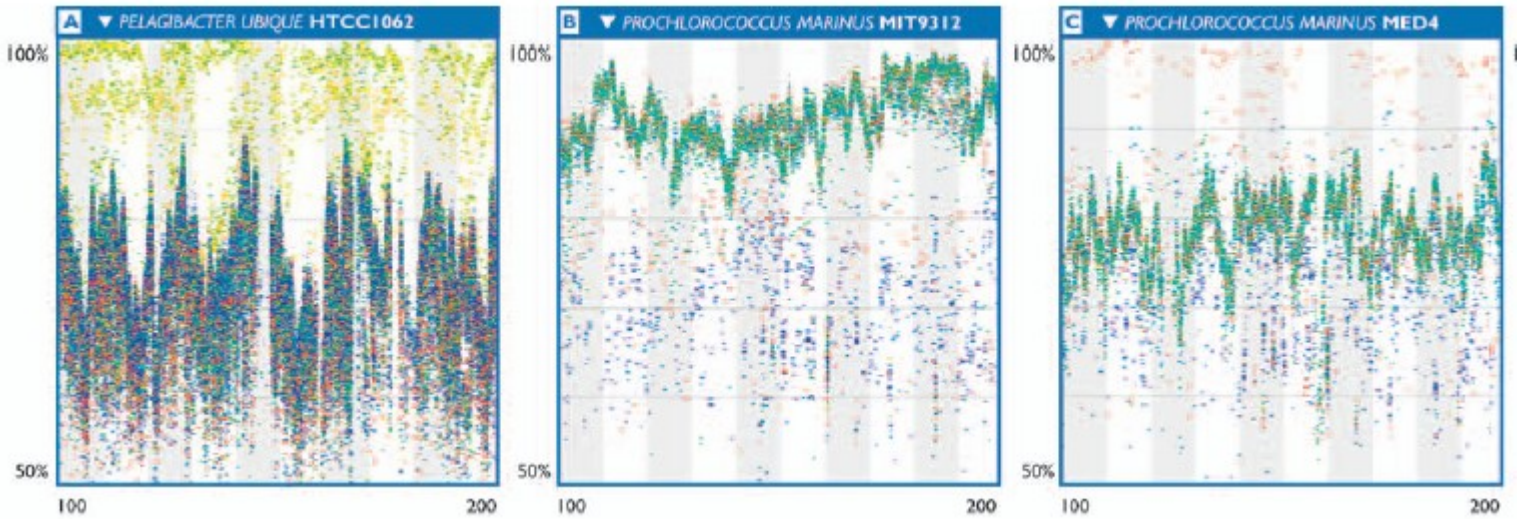
For this lesson I'm going to focus more on how to use metagenomic sequence data for community profiling. (This shouldn't be a huge surprise in a metagenomics course.)

Fragment recruitment

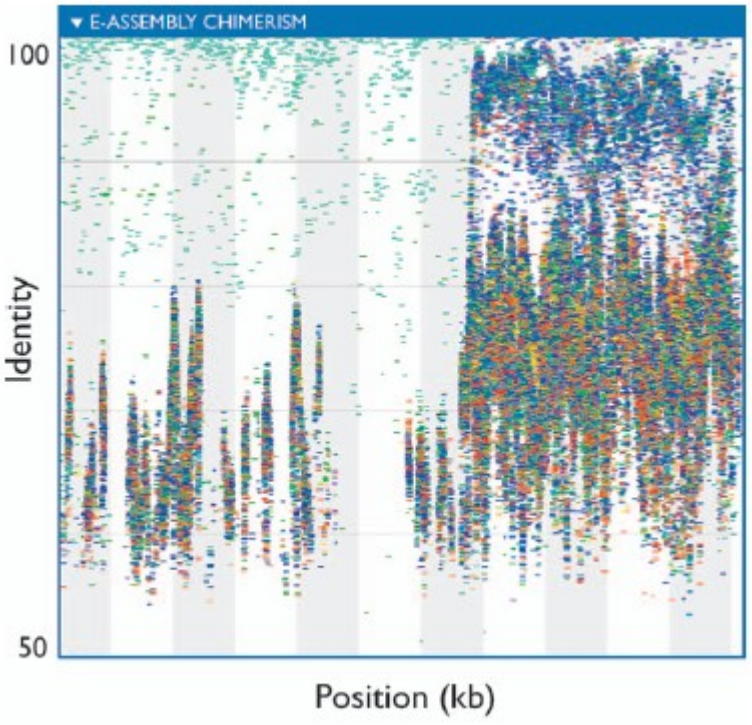
In a previous reading assignment, the Global Ocean Survey study, fragment recruitment plots were introduced.

The horizontal axis of each panel corresponds to a 100-kb segment of genomic sequence from the indicated reference microbial genome. The vertical axis indicates the sequence identity of an alignment between a GOS sequence and the reference genomic sequence. The identity ranges from 100% (top) to 50% (bottom). Individual GOS sequencing reads were colored to reflect the sample from which they were isolated. Geographically nearby samples have similar colors (see Poster S1 for key). Each organism shows a distinct pattern of recruitment reflecting its origin and relationship to the environmental data collected during the course of this study.

The following is a portion of Figure 2 from the paper, showing plots for 3 reference genomes.



Fragment recruitment



Fragment recruitment plots can also be used to find problems with assemblies.

Here, the contig's plot indicates the assembly is chimeric between two organisms, based on dramatic shifts in density of recruitment, level of conservation, and sample distribution.

FR-HIT

FR-HIT, a very fast program to recruit metagenomic reads to homologous reference genomes.
Niu, B., et al.
Bioinformatics. Vol 27 no 12 2011. p1704-1705. PMID: 21505035

FR-HIT first constructs a k-mer hash table for the reference genome sequences. Then for each query, it performs seeding, filtering and banded alignment to identify the alignments to reference sequences that meet user-defined cutoffs.

It is reported to be orders of magnitude faster than BLASTN and recruits 1-5 times more reads than (faster) SOAP2, BWA and BWA-SW.

BIOINFORMATICS APPLICATIONS NOTE
Vol. 27 no. 12 2011, pages 1704–1705
doi:10.1093/bioinformatics/btr252

Sequence analysis
Advance Access publication April 19, 2011

FR-HIT, a very fast program to recruit metagenomic reads to homologous reference genomes

Beifang Niu, Zhengwei Zhu, Limin Fu, Sitao Wu and Weizhong Li*
Center for Research in Biological Systems, University of California San Diego, La Jolla, CA, USA
Associate Editor: Martin Bishop

ABSTRACT

Summary: Fragment recruitment, a process of aligning sequencing reads to reference genomes, is a crucial step in metagenomic data analysis. The available sequence alignment programs are either slow or insufficient for recruiting metagenomic reads. We implemented an efficient algorithm, FR-HIT, for fragment recruitment. We applied FR-HIT and several other tools including BLASTN, MegaBLAST, BLAT, LAST, SSAHA2, SOAP2, BWA and BWA-SW to recruit four metagenomic datasets from different type of sequencers. On average, FR-HIT and BLASTN recruited significantly more reads than other programs, while FR-HIT is about two orders of magnitude faster than BLASTN. FR-HIT is slower than the fastest SOAP2, BWA and BWA-SW, but it recruited 1–5 times more reads.

Availability: http://weizhongli-lab.org/frhit.

Contact: liwz@sdsc.edu

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on February 19, 2011; revised on April 8, 2011; accepted on April 11, 2011

1 INTRODUCTION

Metagenomic data provide a more comprehensive picture for our understanding of the microbial world. An important step of such understanding is to compare the raw sequencing reads against the available microbial genomes to analyze the phylogenetic composition, genes and functions of the samples. Such a procedure, referred to as fragment recruitment, was introduced in the Global Ocean Sampling (GOS) metagenomics study (Rusch *et al.*, 2007).

Sequences from metagenomic samples exhibit great differences from the available genomes. Although there are thousands of available complete microbial genomes, they hardly cover the broad and diverse species in many metagenomic samples. A typical metagenomic dataset may have hundreds or thousands of species, and many of them are novel. Therefore, it is critical for fragment recruitment methods to align reads to homologous genomes.

In the GOS study, BLAST (Altschul *et al.*, 1997) was used for fragment recruitment. However, it is too slow to handle large datasets. The explosion of next-generation sequencing data stimulated the development of new mapping programs, such as SOAP (Li *et al.*, 2008), Bowtie (Langmead *et al.*, 2009), BWA (Li and Durbin, 2009) and many others. These programs are several orders of magnitude times faster than BLAST, but they can only identify very stringent similarities that tolerate only a few mismatches and gaps. So these mapping programs are insufficient for fragment recruitment. The slightly slower programs like BLAT (Kent, 2002), SSAHA2 (Ning *et al.*, 2001) and LAST (Kielbasa *et al.*, 2011) can recruit more reads than the mapping programs, but their fragment recruiting capacities are still limited. In this article, we present a new fragment recruitment method, FR-HIT. Given reference genomes, metagenomic reads and sequence identity and alignment length cutoffs, the goal of FR-HIT is to align the most reads to references with minimal computational time.

*To whom correspondence should be addressed.

2 METHODS AND IMPLEMENTATION

FR-HIT first constructs a k-mer hash table for the reference genome sequences. Then for each query, it performs seeding, filtering and banded alignment to identify the alignments to reference sequences that meet user-defined cutoffs.

2.1 Constructing k-mer hash table

The reference genome sequences are converted into a k-mer hash table. The default value of k is 11 and can be adjusted from 8 to 12. We include overlapping k-mers at an equidistant step from reference sequences. A reference sequence of length m contains $(m-k)/(k-p)+1$ k-mers with an overlap of p bases. Here, p is also a user-adjustable parameter. The hash table stores the indexes of reference sequences and the offset positions of k-mers on reference sequences.

2.2 Seeding

Seeding identifies candidate blocks, which are fragments of reference sequences that can be potentially aligned with the query. For each query, we count all its overlapping k-mers and scan the k-mer hash table to collect the k-mers shared by reference sequences.

We identify pieces of reference sequences that the query can be aligned to. These pieces are anchored by the shared k-mers. For a reference, any cluster of ≥ 2 pieces within b bases will derive a candidate block. This block covers all the pieces in that cluster and has extra b bases at each end. Here, b is the bandwidth to be introduced in Section 2.4. If two candidate blocks overlap, they are joined together into one candidate block. We repeat this until no overlapping blocks are observed.

2.3 Filtering

Filtering removes the candidate blocks that do not enclose qualified alignments. K-mer filtering was originally used in QUASAR (Burkhardt *et al.*, 1999). Two sequences of length n with Hamming distance ε share at least $n+1-(\varepsilon+1)k$ common k-mers (Jokinen and Ukkonen, 1991; Owolabi and Mcgregor, 1988). Here, ε is the number of mismatches in an alignment. Based on user-defined length and sequence identity cutoffs, we calculate the number of mismatches and reject the candidate blocks that do not have enough common k-mers. In this step, the length of a k-mer is 4.

Genometa

FR-HIT is a command-line utility with relatively simple input/output that can be incorporated into other tools and pipelines.

Some users prefer a GUI-based package that includes visualization of results.

Genometa enables identification of bacterial species and gene content from datasets generated by inexpensive high-throughput short read sequencing technologies.

It is a Java-based tool that runs on both Linux and Windows, and uses the bowtie algorithm as its alignment engine.

OPEN ACCESS Freely available online

PLOS ONE

Genometa - A Fast and Accurate Classifier for Short Metagenomic Shotgun Reads

Colin F. Davenport[1]*, Jens Neugebauer[1], Nils Beckmann[2], Benedikt Friedrich[2], Burim Kameri[2], Svea Kokott[1], Malte Paetow[2], Björn Siekmann[2], Matthias Wieding-Drewes[2], Markus Wienhöfer[2], Stefan Wolf[2], Burkhard Tümmler[1], Volker Ahlers[2], Frauke Sprengel[2]

1 Pediatric Pneumology, Allergology and Neonatology, Hannover Medical School, Hannover, Lower Saxony, Germany, 2 Department of Computer Science, University of Applied Sciences and Arts, Hannover, Hannover, Lower Saxony, Germany

Abstract

Summary: Metagenomic studies use high-throughput sequence data to investigate microbial communities *in situ*. However, considerable challenges remain in the analysis of these data, particularly with regard to speed and reliable analysis of microbial species as opposed to higher level taxa such as phyla. We here present Genometa, a computationally undemanding graphical user interface program that enables identification of bacterial species and gene content from datasets generated by inexpensive high-throughput short read sequencing technologies. Our approach was first verified on two simulated metagenomic short read datasets, detecting 100% and 94% of the bacterial species included with few false positives or false negatives. Subsequent comparative benchmarking analysis against three popular metagenomic algorithms on an Illumina human gut dataset revealed Genometa to attribute the most reads to bacteria at species level (i.e. including all strains of that species) and demonstrate similar or better accuracy than the other programs. Lastly, speed was demonstrated to be many times that of BLAST due to the use of modern short read aligners. Our method is highly accurate if bacteria in the sample are represented by genomes in the reference sequence but cannot find species absent from the reference. This method is one of the most user-friendly and resource efficient approaches and is thus feasible for rapidly analysing millions of short reads on a personal computer.

Availability: The Genometa program, a step by step tutorial and Java source code are freely available from http://genomics1.mh-hannover.de/genometa/ and on http://code.google.com/p/genometa/. This program has been tested on Ubuntu Linux and Windows XP/7.

Citation: Davenport CF, Neugebauer J, Beckmann N, Friedrich B, Kameri B, et al. (2012) Genometa - A Fast and Accurate Classifier for Short Metagenomic Shotgun Reads. PLoS ONE 7(8): e41224. doi:10.1371/journal.pone.0041224

Editor: Niall James Haslam, University College Dublin, Ireland

Received February 16, 2012; **Accepted** June 19, 2012; **Published** August 21, 2012

Funding: CD performed much of this research while a member of the international research training group 'Pseudomonas Pathogenicity and Biotechnology' sponsored by the Deutsche Forschungsgemeinschaft grant GRK 653/3, and is now supported by the DFG SFB 900 (project Z1). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: davenport.colin@mh-hannover.de

Introduction

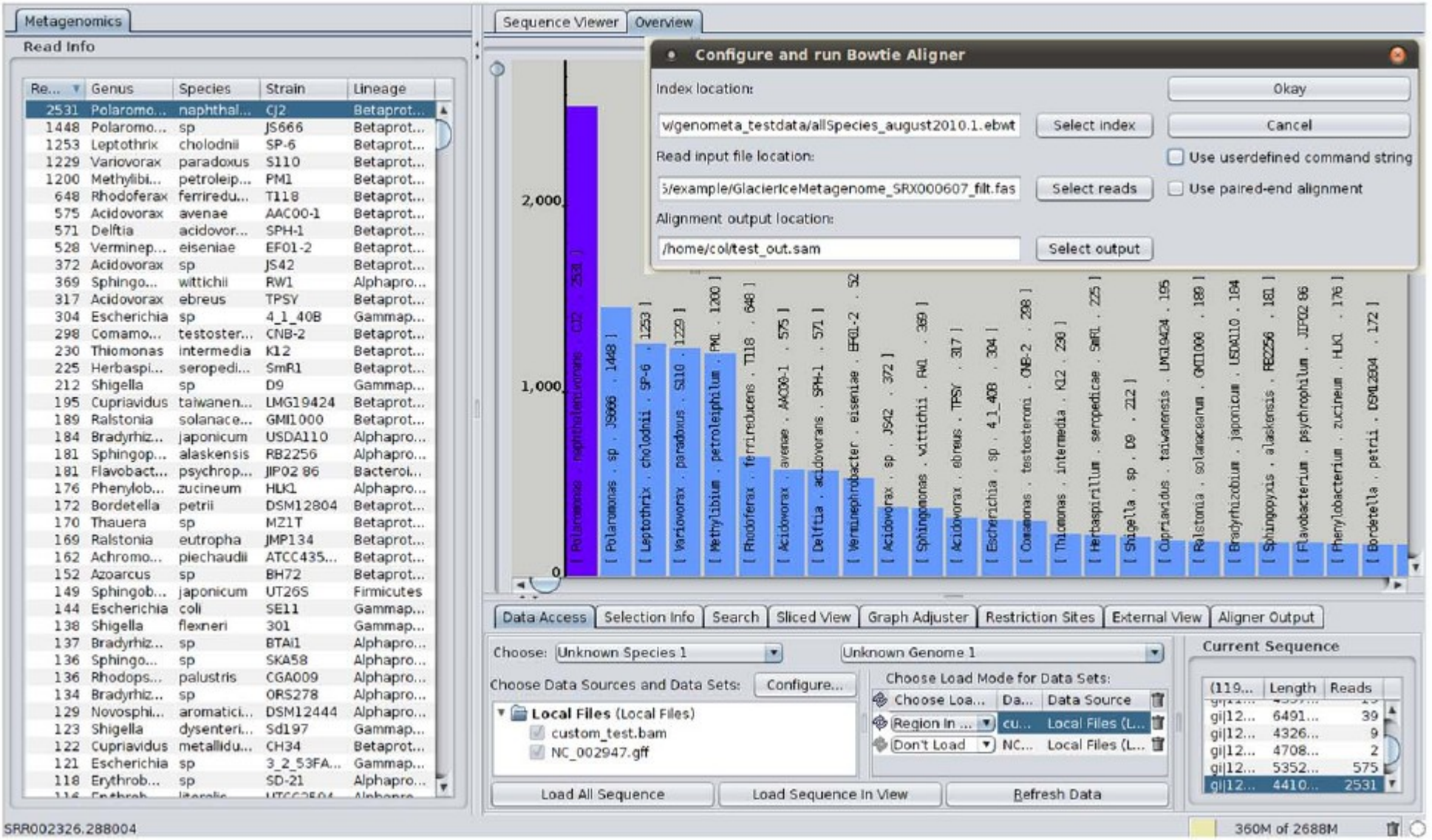
Metagenomics, the analysis of microbial communities directly within their natural environments, continues to gain traction in both the environment and in the clinic. In metagenomics, sequence reads can be used to predict both the abundance and functional capacity of the microbes present by molecular means. Sequence read data from high throughput sequencing platforms like Illumina and SOLiD are by far the most cost-effective per base pair sequenced [1], yet downstream analysis remains challenging, with algorithmic speed an issue. Despite this, extensive short read datasets are beginning to accumulate [2,3,4].

Sequence reads in bacterial metagenomic analyses can be derived by whole genome shotgun sequencing, or targeted sequencing of 16S rRNA amplicons. These alternative techniques do lead to significant taxonomic differences in results, based upon the evaluation of 33 metagenomes [5]. In other words, the decision to select targeted 16S amplicon sequencing or untargeted whole genome sequencing will lead to different predictions of the taxonomy of a metagenome. Sequencing of 16S rRNA remains a popular approach [6] in metagenomics despite its well known limitations [7,8]. Estimates of taxon abundances can be biased by large differences in rRNA copy number between even closely related species [9], and the fact that not all rRNA genes amplify with PCR primers [10]. In fact, the number of copies of the rRNA gene in bacteria range from 1–15 [9], rendering rRNA-based approaches more suitable for qualitative than quantitative metagenomics. Because of these reasons, we anticipate whole genome shotgun metagenomes will be preferable to sequencing of rRNA amplicons in the future.

Ideally, researchers require programs which can perfectly assign reads to individual microbial strains. This goal is not realistically possible due to the very high sequence similarity between strains, the reads errors inherent to sequencing, and the lack of reference genome sequences for some phyla. However, the optimal result for a metagenome dataset must remain species level read assignments, and not unspecific matches to phyla such as Firmicutes or Proteobacteria. Attributions to higher taxonomic levels, while

Genometa

View of the graphical results of Genometa.



Bowtie

Bowtie advertises itself as an “ultrafast, memory-efficient short read aligner”, and it deserves that description.

It can align short DNA sequences (reads) to the human genome at a rate of over 25 million 35-bp reads per hour. Bowtie indexes the genome with a Burrows-Wheeler index to keep its memory footprint small: typically about 2.2 GB for the human genome (2.9 GB for paired-end).

It is already installed on DIAG, but you can also download it here:

http://bowtie-bio.sourceforge.net/

The input to bowtie is a 'reference' which can be a single genome such as human, or group of unrelated sequences pooled together, like bacterial reference genomes.

Running bowtie

For any bowtie run you first need to index your reference genome. This only needs to be done once. This can be a memory-intensive and time-consuming step. There are many pre-built indexes available for download from the bowtie site.

The *bowtie-build* command takes two arguments – the FASTA file of your reference and a label for the index to be created:

```
bowtie-build NC_002127.fna e_coli_O157_H7
```

This will create four files with that label that end with the ".ebwt" extension.

You can then run bowtie like this:

```
bowtie e_coli_O157_H7 reads/e_coli_1000.fq
```

You'll need to redirect the output to a file. More commonly, you'll want to write the results in SAM format, which is used by many downstream tools:

```
bowtie -S e_coli_O157_H7 reads/e_coli_1000.fq output.sam
```

For more information, see the official documentation:
http://bowtie-bio.sourceforge.net/tutorial.shtml

Further reading

"Microbial community profiling for human microbiome projects: Tools, techniques and challenges"

Micah Hamady and Rob Knight

Genome Res. 2009 Jul;19(7):1141-52. PMID: 19383763.

"Metagenomic microbial community profiling using unique clade-specific marker genes."

Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C.

Nat Methods. 2012 Jun 10;9(8):811-4. PMID: 22688413.

"Short pyrosequencing reads suffice for accurate microbial community analysis."

Liu Z, Lozupone C, Hamady M, Bushman FD, Knight R.

Nucleic Acids Res 35: e120. 2007. PMID: 17881377.