**METAGENOMICS**

## LESSON ONE – WHAT IS METAGENOMICS

### WHAT IS METAGENOMICS?

* What it is not
  + In this house we don’t use “metagenomics” to mean “amplicon sequencing”
  + 16S “metagenomics” sequencing
* Here, we only mean “shotgun metagenomics”
  + “amplicon sequencing” or “metabarcoding” is sequencing a specific target region from many genomes (eg 16S rRNA gene)
  + “shotgun metagenomics” is (incomplete) sequencing of a mixture of genomes using an untargeted approach
* Why is shotgun metagenomics incomplete?
  + Environmental samples are information dense
  + One mL of seawater has about a million bacteria, tens of thousands of eukaryotes
    - THAT’S SO MANY BP
  + Even more genomic information in soil, and even more in stool
  + Reading truly all of that information would cost so much money... cannot be reasonably done

### WHAT KINDS OF QUESTIONS CAN METAGENOMICS BE USED TO ANSWER?

* Who is there?
  + Marker gene analysis
    - Taxonomic diversity
    - Phylogenic diversity
  + Binning
    - Taxonomic diversity
    - Phylogenic diversity
    - Novel taxa
  + Assembly
    - Genome diversity
    - Novel genomes
  + Questions
    - Is this gene present in this sample?
    - How many homologs of this gene appear in this sample?
    - Which genomes encode this gene?
    - Is this pathogen present in this environment
    - How closely related is this uncultured strain to this cultured representative
    - How many ecotypes of this bacterium appear in this environment
    - How many prokaryotic phyla are there in the world?
* What are they doing
  + Gene prediction
    - Gene diversity
    - Novel genes
  + Functional annotation
    - Protein family diversity
    - Functional diversity
  + Questions
    - What proteins do symbionts encode to mediate relationships with their hosts?
    - Which genes/pathways/genomes co-occur in this environment?
    - What antibiotic resistance genes does this community encode?
    - How many different carbon fixation pathways exist in hydrothermal vents?
    - Are there novel CRISPER-Cas systems yet to be discovered?
* Comparative metagenomics
  + Combination of who is there and what they are doing
  + Intercommunity similarity
  + Metadata correlations
  + Biomarker detection

### IS METAGENOMICS RIGHT FOR ME?

* You might tray amplicon sequencing if:
  + You need to detect rare genes of species
  + You’re working with eukaryotes
  + You have many (1000s) of samples to run
* You might try quantitative PCR if:
  + You only care about presence/absence
  + You want to quantify how many copies of a gene/species is present in a sample
* You might try isolate genomics if:
  + You can isolate your organism of interest
  + You’re working with eukaryotes
* You might try single-cell genomics if
  + Your interest is population genetics
  + Your interest is in novel taxa
  + Your interest is horizontal gene transfer and pangenomics
* You might try meta/transcriptomics if:
  + You want levels of gene expression
  + You’re working with eukaryotes

### YOU’VE GOT DATA! NOW WHAT?

* Some examples
  + metaWRAP
  + YAMP
  + ATLAS
  + SqueezeMeta

## LESSON ONE – INTERACTIVE MATERIAL

### FASTQC

* Useage
  + fastqc file -o output.directory

### MULTIQC

* Useage
  + multiqc directory

## LESSON TWO – TAXONOMIC CLASSIFICATION USING K-MERS

### THE POWER OF K-MERS

* What is a k-mer
  + A word of DNA that is k bases long
  + 41 k-mers: A, T, C, G
  + 42 k-mers: AA, AT, AC, AG, TA, TT, TC, TG, CA, CT, CC, CG, GA, GT, GC, GG
  + 43 k-mers: ordered combos of three
    - All codons are 3-mers but not all 3-mers are codons
* Extracting k-mers
  + Typically k-mers are extracted by running a k-length window across all of the reads and sequences
* Complementarity
  + K-mer analysis is often simplified by storing (“hashing” or “indexing”) only the lexicographically lower reverse compliment (the “canonical” string)
  + For example, you can store a particular 4mer at CGAT or ATCG, and these are equivalent
  + Standard one to choose would be whichever one would come first in the dictionary
* Palindromes
  + In molecular biology, a palindrome is a DNA sequence that is its own reverse compliment
  + Restriction enzyme cut sites are common examples of palindromes
  + Some tools (eg assemblers) require odd-length k-mers to avoid palindromes, which induce self-loops in B Brujin graphs
  + An odd-length k-mer can never be its own compliment
* Choosing a k-mer length
  + The optimal k-mer length will be determined by the goal of your analysis
  + 1-mers: GC-content analysis
    - The distribution of k-mers in the genomes is not uniform
    - Even at the level of 1-mers, bacterial genomes have different, but highly conserves proportions
    - For a given species/group, the GC content tends to be similar across the whole genome
      * This may be different than what it is for other species though
    - Could be useful for detecting HGT!
  + 3-mers: codon usage
    - At the level of 3-mers, bacterial genomes have unique, but highly conserved useage of codons
    - Extends to untranslated regions and may be primarily related to GC content
    - Some codons are preferred over other
  + 4- or 5-mers: metagenomic binning
    - 4-mers (tetranucleotide) frequency
      * At the level of 4-mers, bacterial genomes have conserved “fingerprints) which are used in metagenomic binning tools and to detect horizontal gene transfer events
      * Tend to be relatively conserved across the genome, except perhaps where HGT has occurred and in ribosomal DNA areas (because this is so highly conserved)
      * Dimensional reduction
        + Some tools: PCA, t-SNE, DBSCAN, UMAP
        + Embed data with many, many, many dimensions into a 2-D space for visualization
  + 7- to 35-mers: taxonomic classification
    - Taxonomic classification
      * As k-mer length increases, the probability of a random match between two distantly related taxa decreases exponentially
      * This can be used to classify DNA sequences based on their k-mer profiles by comparing the k-mers to a pre-computed database of taxonomically-known k-mers
      * This technique is also used to identify adapters and other contaminants in raw reads
      * Some of these tools understand that some k-mers are more useful in determining taxonomic classification than others and will analyze accordingly
  + 21- to 127-mers: genome assembly
    - Genome assembly
      * Many use Brujin graphs based on k-mer overlap
      * If the k-mer length is too short, many unrelated regions are assembled together
      * By k-mers cannot be longer then sequence lengths, thus the benefit of long-read technologies

### CHOOSING A CLASSIFIER

* Sequence format
  + Nucleotide k-mers grow at 4x
  + Amino acid k-mers grow at 20x
  + So a nucleotide 31-mer=amino acid 12-mer
* Database
  + Whole genomes
  + Marker genes
  + Ribosomal RNA genes
* Search algorithm
  + K-mer indexing
  + MinHash
  + BWT
  + BLAST-like
* Examples of tools
  + K-mers plus genomes
    - Kraken (don’t use anymore)
    - Kraken2 (nuc or aa)
    - Krakenuniq (nuc)
    - CLARK (nuc)
    - MOTHUR classify seqs (16S rRNA)
  + BWT + genomes
    - Centrifuge (nuc)
    - Kaiju (aa)
  + BWT + marker genes
    - MetaPhlAn2
    - mOTU
    - AMPHORAnet
  + Reviewed vy Breitwieser et al. (a review of methods and databases for metagenomic classification and assembly

### DEMO: ADAPTER TRIMMING

* Trimmomatic
  + Many methods to trim by, including:
    - Crop (trim to a specific length by removing base reads from the end)
    - Sliding window (once the quality slips below a threshold, cut the window)
  + Use help for full list

### DEMO: KRAKEN2

## LESSON THREE – TAXONOMIC CLASSIFICATION USING MINHASH

### MINHASH ALGORITHM

* Set similarity: jaccard index
  + We have two genomes that we’d like to caompare based on their gene content
  + A “set” is composed of a collection of genes in each genome
  + The jaccard similarity is a measure of set similarities given by the intersection of the sets divided by their union
  + Count the number of shared genes (included in both genomes), divide that by the total number of unique genes that both genomes collectively encode
* Genomes as sets: k-mers
  + Instead of using “genes” as sets, we could use k-mers
  + Break every genome into its component kmers (say k=21) and end up with about 3 million k-mers for each (in bacteria)
  + Computationally expensive when you have many sets to compare and you use the full set of k-mers
* Minhash sketches
  + One way to simplify a massive computational job
    - By approximating the jaccard index instead of calculating it exactly
    - Minhash does exactly this
  + For DNA sequences:
    - 1.) extract all k-mers from a sequence
    - 2.) use hash function to reproducibly convert k-mers into numbers that are randomly distributed over some range
    - 3.) choose the numerically lowest s hashes (min hashes) as the bottom sketch
    - 4.) compare bottom sketches against each other using the jaccard index
  + A hash function maps and input range to a fixed output range
    - Good hash functions has high compressibility, high randomization, low collision rate (where different sequences end up with the same output from the hash function), and high speed
  + Hashing functions
    - Used most in minhash implementations is murmurhash3
    - Q and A
      * Why not take the lexicographically smallest s k-mers
        + Because that is not random selection (could have biases due to GC content, unusual genes, etc)
      * Why not randomly select s k-mers from each set and sort them lexicographically
        + Because it is unlikely that two large sets will overlap with a small s unless they are highly similar. Using the bottom sketch ensures that overlap will be found if it exists
      * How many signatures are needed to get a good approximation?
        + Using minhash, the expected error rate is e=1/sqrt(s)
        + E=0.05 at s=400
* Bloom filters
  + Common way to speed up k-mer computations in bioinformatics: burton howard bloom filter to test whether an element is a part of a set
    - Trades large memory footprint for speed on tasks like rejecting singletons or error-containing k-mers
  + A bloom filter is described by two parameters
    - m=length of filter (eg 231 bits) initialized to all 0s
    - n(<<m)=number of hash functions that each produce a hash in the range of the filter (eg a to 231)
  + procedure
    - 1.) to insert element: calculate all n hashes and insert them y flipping the corresponding bits to 1
    - 2.) test whether an element is part of set: calculated all n hashes and check the bits
      * If all bits are set, then “maybe”
      * If at least 1 bit is unset, then “definitely not”
  + False positives are possible. False negatives are not possible. The desired false positive rate can be tuned by changing m and n
* Minhash compared to ANI
  + Average nucleotide identity (ANI) is a method to compare the similarity between core genomes
    - Minhash does a good job approximating ANI
* Minhash metagenome comparisons
  + Minhash can be used to quickly compare metagenomes for related content, including containment of a particular genome in the metagenome
  + Sourmash has these databases available for use in metagenome classification:
    - 60k RefSeq microbial genomes
    - 100k GenBank microbial genomes
    - 87k GenBank microbial genomes LCA (for lowest common ancestor analyses)
* Minhash implementations
  + Mash [github.com/marbl/mash]
    - Can use a bloom filter to eclude singletons and other method for min k-mer count
    - Implemented online at PATRIC similar genome finder
  + Sourmash [github.com/dib-lab/sourmash]
    - Can use a bloom filter with counting for min k-mer count
    - Can use a “scaled” (proportional) s rather than static
    - Can generate figures
    - LCA taxonomic classification
  + Bbsketch [sourceforge.net/projects/bbmap]
    - Can use min k-mer count
    - Can use optional scaled s
    - Can use blacklist (block out certain hashes – for example those used for illumina adapters)
    - LCA taxonomic classification
  + Finch [github.com/onecodex/finch-rs]
    - Can use a bloom filter for min k-mer count

### DEMO: CLASSIFICATION USING SOURMASH

### DEMO: QUALITY TRIMMING USING BBDUK

### DEMO: CLASSIFICATION USING SENDSKETCH