METAGENOMICS

* fastqc fastq.file -o output.directory
  + run quality control on a fastq file
* multiqc directory
  + searches for fastqc outputs within directory, compares them
* kraken2 -db database fastqfile
  + run kraken2for classifying kmers
  + –confidence #set confidence level between 0 and 1
  + --report #gives a nice report format
* bbduk.sh in= out= ref=adapters ktrim= mink= hdist= minlen= ow=
  + for trimming sequences (removing adapters, etc)
  + see help
  + or to remove artifacts, ref=artifacts,phix
* krona
  + can be installed using conda
  + useful for visualizing kraken2 outputs
* trimmomatic
  + for trimming sequences for quality
  + use help
* sourmash compute -k kmer.length --scaled=scale.fract data
  + calculates signature (much smaller than original file)
  + creates a list of hashes
  + can also do
    - sourmash compare
    - sourmash gather
    - sourmash index
    - sourmash plot
    - sourmash search
    - sourmash lca
* clumpify.sh in1= in2= out= dedupe optical ow=
  + remove optical duplicates (typical to illumine sequencing)
  + will also reorder the sequences to maximize compression
  + use help
* filterbytile.sh in= out= ow=
  + averages the quality across each microtile, gets rid of low quality reads
* sendsketch.sh address=database
  + compares query sketches to others, prints their kmer identity
  + can use servers of reference sketches, don’t have to download your own
  + available databases:
    - nt
    - Refseq
    - Silva
* taxserver.sh
  + to set up your own private database to query
* bbmap.sh
  + map sequence to genome (use help)
  + some examples:
    - minid: default is 76% ID match with no indels
    - threads: default is all (don’t use all on a shared machine!)
    - ref: reference
    - in1, in2: use both for paired reads (fwd and back)
    - outm: only writes out the reads that are mapped (this saves some space I the memory)
* bbwrap.sh
  + map multiple sequences to a genome (use help)
  + in1 is two fwd seqs separated by a comma, in2 is two back reads in same order
  + mapper= (default is bbmap)
* bbsplit.sh
  + lets you decide what to do with reads that map to multiple places in the genome (toss them, pick only the best quality map, keep all the maps, write into separate file)