Ready for a break in programming and grid computing after last week's assembly and gene prediction work?  :)

The major topics of this week are reference genomes and the resources that provide and utilize them.  From fragment recruitment plots (which we cover even more next week) to studies of taxonomic diversity and rates of genetic divergence, several papers we've covered so far illustrate how reference genomes are used in microbiome studies.

For this homework assignment you're going to start building your own experimental metagenomic dataset with reference genome collection.  Choose wisely, as you'll be building on these datasets as the course goes along, ultimately using them in both your mid-term and final projects.

You've read a lot about metagenomic projects at this point, seen several detailed examples, and this week are introduced to both general resources (GOLD, IMG/M, etc.) as well as sites made for specific habitats (HMP, and initially CAMERA.)  Your tasks for this homework assignment then are:

1. Choose a habitat of interest to you which has publicly-available metagenomic sequence data.  You can browse literature or use any of the resources from this week's lecture notes, such as IMG/M's Microbiome Projects Map.  It needs to have FASTQ data publicly available for download, but you don't need to collect it yet.

2. Briefly describe your chosen habitat (a paragraph or two) along with why and how you chose it.  Summarize the data available for it including format, sequencing technology used, and size.

3. Collect between 10 and 20 (annotated) reference genomes you believe should be present in the sample.  For each, you should have Genbank, protein fasta, and nucleotide fasta files.  List these out in a table with the following fields:

    - Organism name

    - Taxon ID

    - Genomic molecule count (contig/scaffold count)

    - Gene count

 Organize these under your home area in directories like this:

 ~/project/refgenomes/$tax\_id/genomic.fna

 ~/project/refgenomes/$tax\_id/genomic.gbk

 ~/project/refgenomes/$tax\_id/protein.faa

 Your write-up should include the statistics table described just above and a summary of why you chose each of these reference genomes for your particular habitat.

We will be using these in many of the following lessons as the semester progresses, so be sure to pick a habitat that's interesting to you and for which sufficient data are available.

**Metagenomic analysis of permafrost microbial community response to thaw**

<<http://www.osti.gov/bridge/purl.cover.jsp?purl=/1051660/1051660.PDF>>

I’ve always been interested in global warming and amazed by individuals who Because ocean temperatures increase slower than land temperatures, the northern hemisphere warms up faster due to the presence of more land, and since global warming is primarily caused by increasing levels of greenhouse gases, I thought what kind of effect the thawing of permafrost would have on global warming. With the help of GOLD, I chose the metagenomic dataset collected from the permafrost microbial communities from Central Alaska <<http://img.jgi.doe.gov/cgi-bin/m/main.cgi?section=TaxonDetail&page=taxonDetail&taxon_oid=2067725009>>.

Permafrost is soil that has been frozen for thousands of years and contains organic matter which microbes can process into methane and thus add to the levels of greenhouse houses. It has been estimated that the arctic permafrost has potential to have more than 250 times the greenhouse gas emissions from the United States in 2009.

    - Organism name: Exiguobacterium sibiricum

    - Taxon ID: 262543

    - Genomic molecule count (contig/scaffold count):

    - Gene count: 3156

    - Organism name: Exiguobacterium anarcticum B7

    - Taxon ID: 1087448

    - Genomic molecule count (contig/scaffold count):

    - Gene count:

    - Organism name: Octadecabacter antarticus 307

    - Taxon ID: 391626

    - Genomic molecule count (contig/scaffold count)

    - Gene count

    - Organism name: Octadecabacter antarticus 238

    - Taxon ID: 391616

    - Genomic molecule count (contig/scaffold count)

    - Gene count

    - Organism name: Rickettsia sibirica 246

    - Taxon ID: 272951

    - Genomic molecule count (contig/scaffold count)

    - Gene count

    - Organism name: Janthinobacterium sp. Marseille

    - Taxon ID: 375286

    - Genomic molecule count (contig/scaffold count)

    - Gene count:

    - Organism name: Moritella sp. PE36

    - Taxon ID: 58051

    - Genomic molecule count (contig/scaffold count)

    - Gene count

    - Organism name: Glaciecola punicea DSM 14233

    - Taxon ID: 1121923

    - Genomic molecule count (contig/scaffold count)

    - Gene count

    - Organism name: Pseudoalteromonas sp. SM9913

    - Taxon ID: 234831

    - Genomic molecule count (contig/scaffold count)

    - Gene count

    - Organism name: Pseudoalteromonas atlantica T6c

    - Taxon ID: 342610

    - Genomic molecule count (contig/scaffold count)

    - Gene count

    - Organism name: Glaciecola sp. HTCC2999

    - Taxon ID: 455436

    - Genomic molecule count (contig/scaffold count)

    - Gene count

    - Organism name: Polaromonas sp. JS666

    - Taxon ID: 296591

    - Genomic molecule count (contig/scaffold count)

    - Gene count

    - Organism name: Psychrobacter cryohalolentis

    - Taxon ID: 335284

    - Genomic molecule count (contig/scaffold count)

    - Gene count

    - Organism name

    - Taxon ID

    - Genomic molecule count (contig/scaffold count)

    - Gene count

    - Organism name: Pseudomonas sp. PAMC 25886

    - Taxon ID: 1125977

    - Genomic molecule count (contig/scaffold count)

    - Gene count

    - Organism name: Psychrobacter articius 273-4

    - Taxon ID: 259536

    - Genomic molecule count (contig/scaffold count):

    - Gene count: 2212

    - Organism name: Planococcus anarcticus DSM 14505

    - Taxon ID: 1185653

    - Genomic molecule count (contig/scaffold count)

    - Gene count