Metagenomics and Bioinformatics approaches

A. Dereeper, G. Sarah, F. Sabot

What is metagenomics?

Metagenomics (Environmental Genomics or Community Genomics) is the study of genomes recovered from environmental samples without the need for culturing them

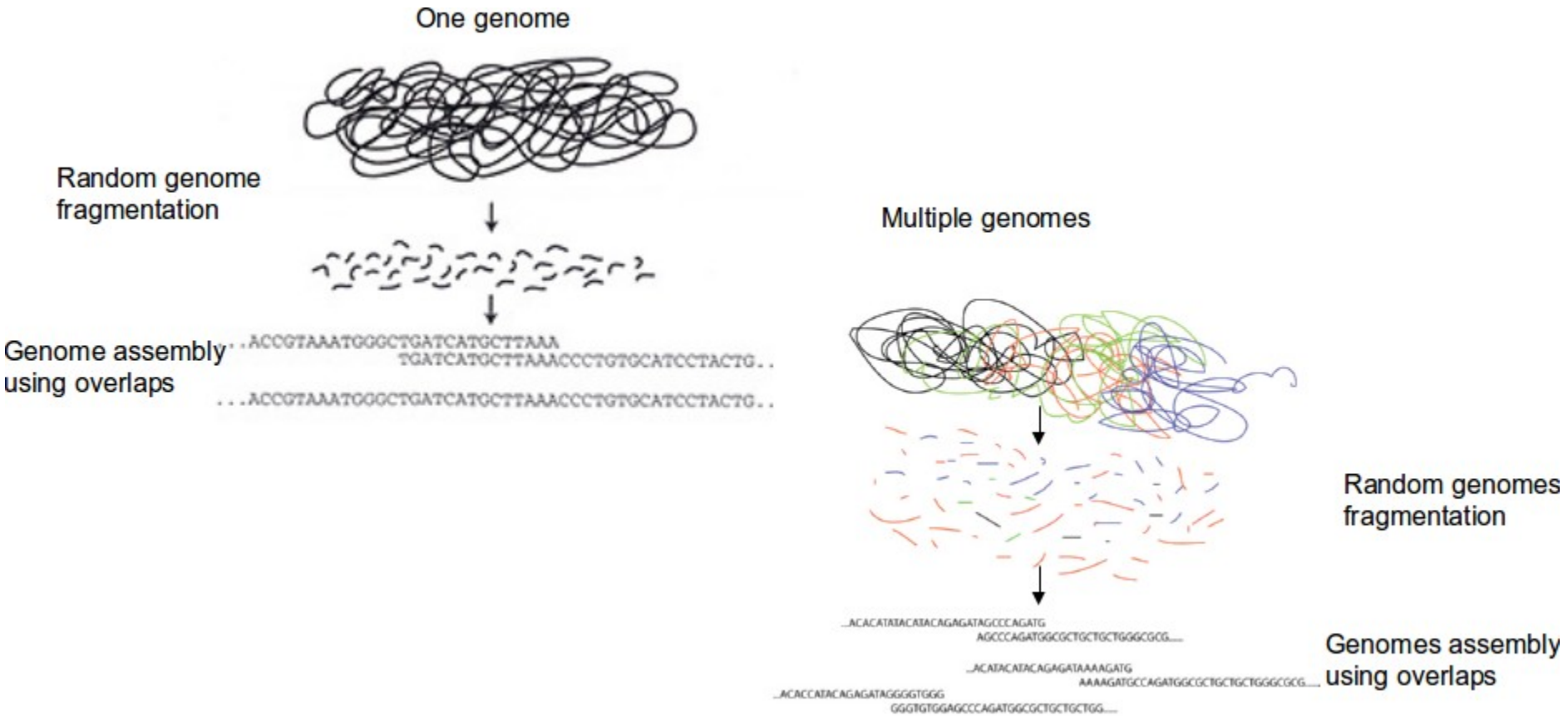
Metagenomics processes data using bioinformatics tools

=> Advantages:

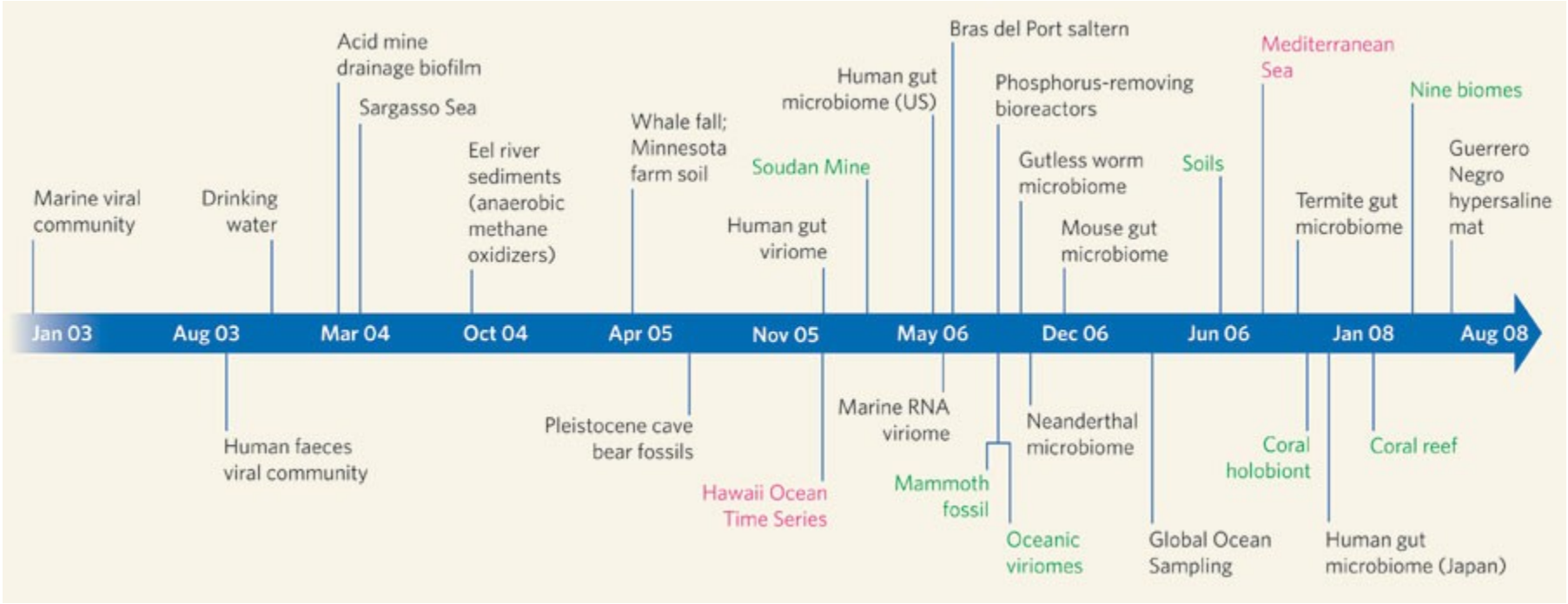
- Organisms can be studied directly in their environments bypassing the need to isolate each species

- There are significant advantages for viral metagenomics, because of difficulties cultivating the appropriate host

Whole genome shotgun sequencing for metagenomics



Examples of projects



TARA OCEANS

Bioinformatics approaches for metagenomics

- Binning based
  Attempts to “bin” reads into the genome from which they originated
  Compare reads to large reference database using BLAST
  => *Megan*, *Megablast*

- Marker based

  - Single gene
  ex: 16S, 18S, ITS…
  Qiime, *Mothur*

  - Multiple genes
  MetaPhlAn

Marker gene based analysis

- 16S rRNA most commonly used
  - Ribosomal RNAs are present in all living organisms
  - rRNAs play critical roles in protein translation
  - rRNAs are relatively conserved and rarely acquired horizontally
  - Behave like a molecular clock
  - Useful for phylogenetic analysis
  - Used to build tree-of-life (placing organisms in a single phylogenetic tree)

Marker gene based analysis

- Other marker genes used
  - Eukaryotic Organisms (protists, fungi)
    - 18S (http://www.arb-silva.de)
    - ITS (http://www.mothur.org/wiki/UNITE_ITS_database)
  - Bacteria
    - CPN60 (http://www.cpndb.ca/cpnDB/home.php)
    - ITS (Martiny, Env Micro 2009)
    - RecA gene
  - Viruses
    - Gp23 for T4-like bacteriophage

Marker genes vs Shotgun metagenomics

Marker Gene Profiling	Shotgun Metagenomics Profiling
Less expensive (~$100 per sample)	Still very expensive (~$1000 per sample)
Computational needs can be met by desktop / small server computers	Usually requires huge computational resources (cluster of computers)
Provides mainly taxonomic profiling	Provides both taxonomic and functional profiling
For 16S, majority of genes can be assigned at least to phylum level	Many more unassigned gene fragments (“wasted” data)
Relatively free of host DNA contamination	Prone to host DNA contamination

Overall bioinformatics workflow