UCDAVIS Bioinformatics Core

Metagenomics Metatranscriptomics

Matthew L. Settles
Genome Center Bioinformatics Core
University of California, Davis
settles@ucdavis.edu; bioinformatics.core@ucdavis.edu

Sequencing Depth

- The first and most basic question is how many base pairs of sequence data will I get
  Factors to consider are:
  - 1. Number of reads being sequenced
  - 2. Read length (if paired consider then as individuals)
  - 3. Number of samples being sequenced
  - 4. Expected percentage of usable data

$$bpPerSample = \frac{readLength * readCount}{sampleCount} * 0.8$$

- The number of reads and read length data are best obtained from the manufacturer's website (search for specifications) and always use the lower end of the estimate.

Genomic Coverage

Once you have the number of base pairs per sample you can then determine expected coverage

Factors to consider then are:

1. Length of the genome
2. Any extra-genomic sequence (ie mitochondria, virus, plasmids, etc.). For bacteria in particular, these can become a significant percentage

$$\frac{ExpectedCoverage}{sample} = \frac{(readLength * numReads) * 0.8}{numSamples} * \text{num.lanes}}{TotalGenomicContent}$$

Metagenomics Sequencing

Considerations (when a literature search turns up nothing)

- Proportion that is host (non-microbial genomic content)
- Proportion that is microbial (genomic content of interest)
- Number of species
- Genome size of each species
- Relative abundance of each species

The back of the envelope calculation

$$\frac{numReads}{sample} = \frac{Coverage * (AverageGenomeSize)}{ReadLen * DilutionFactor * (1 - hostProportion)} * \frac{1}{0.8}$$

Metagenomics Sequencing

Considerations (when a literature search turns up nothing)

- Proportion that is host (non-microbial genomic content)
- Proportion that is microbial (genomic content of interest)
- Number of species
- Genome size of each species
- Relative abundance of each species

The back of the envelope calculation

$$\frac{numReads}{sample} = \frac{Coverage * (AverageGenomeSize)}{ReadLen * DilutionFactor * (1 - hostProportion)} * \frac{1}{0.8}$$

Amplicons vs. Metagenomics

- Metagenomics
  - Shotgun libraries intended to sequence random genomic sequences from the entire bacterial community.
  - Can be costly per sample ($500 to multi thousands $$ per sample)
  - Better resolution and sensitivity to characterize the sample
  - Due to cost, can only do relatively few samples
- Amplicon community profiling
  - Sequence only one regions of one gene (e.g. 16s, ITS, LSU)
  - Cheap per sample (at scale, down to $20/sample)
  - Due to cost, can do many hundreds of samples make more global inferences

Community Sequencing Designs

- Taxonomic Identification
  - Amplicon based (e.g. 16s variable regions)
  - Shotgun Metagenomics
- Functional Characterization
  - Shotgun Metagenomics
  - Shotgun Metatranscriptomics (active)
- Genome Assembly, Function and Variation
  - Shotgun Metagenomics
  - Shotgun Metatranscriptomics

UCDAVIS Bioinformatics Core

Cost Estimation

- DNA/RNA extraction and QA/QC (Bioanalyzer/Gels)
- Metatranscriptomes: Enrichment of RNA of interest and RNA library preparation
  - Library QA/QC (Bioanalyzer and Qubit)
  - Pooling ($10/library)
- Metagenomes: DNA library preparation
  - Library QA/QC (Bioanalyzer and Qubit)
  - Pooling ($10/library)
- Community Profiling: PCR reactions
  - Library QA/QC (Bioanalyzer and Qubit/microplate reader)
  - Pooling
- Sequencing (Number of Lanes / runs)
- Bioinformatics (General rule is to estimate the same amount as data generation, i.e. double your budget)

http://dnatech.genomecenter.ucdavis.edu/prices/

Cost Estimation

- Amplicons
  - 384 Samples
    - Amplicon generation ($20/sample)= 384`*20 = $7,680
  - Sequencing PE300, target 30K reads per sample
  - Bioinformatics
- Metagenome
  - 12 samples (DNA)
  - Expectations: Host Proportion 40%, use average genome size of eColi, Target the 1% and coverage of 20
  - Sequencing PE150
  - Bioinformatics

METAGENOMICS Software

- Remember this is “Data Science”, focus on the questions you wish to answer, the path to answering those questions and not on what is the ‘best’ software to use.
  - Software is rarely generated for the ’generic’ situation
  - Often untied to type of sequencing data, though often dependent
  - Often one and done, look for software that is currently being supported.
- Work in a comparative framework, consistency of results across samples indicates comparability, biases are at least consistent and/or unassociated with the project factors.

UCDAVIS Bioinformatics Core

Taxonomic Classification and abundance

Taxonomic Assignment

KrakEN

- A taxonomic classifier using k-mers, current db contains > 75Gb of microbial/viral genome data (unique kmers).
- Assigns each read to its lowest common ancestor in the tree in a taxonomic tree based on the set of kmers in a read
- Can build your own database
- Requires a large server, 128Gb to 256Gb of memory
- https://ccb.jhu.edu/software/kraken/

Bracken (Bayesian Re-estimation of Abundance with KrakEN) –

- computes the abundance of species in DNA sequences from a metagenomics sample
- https://ccb.jhu.edu/software/bracken/

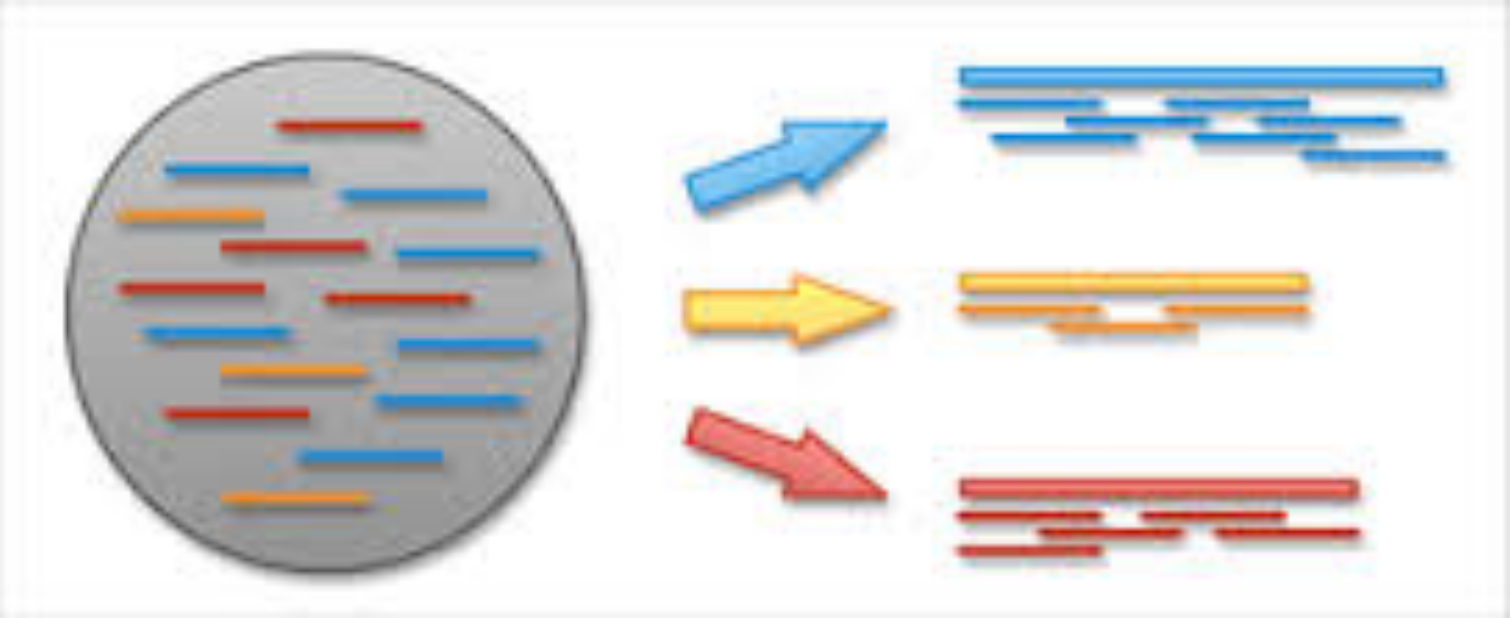
Taxonomic Assignment

Taxonomer

- Assigns taxonomy to sequencing reads from both clinical and environmental samples.
- http://taxonomer.iobio.io/info.html
- http://taxonomer.iobio.io/

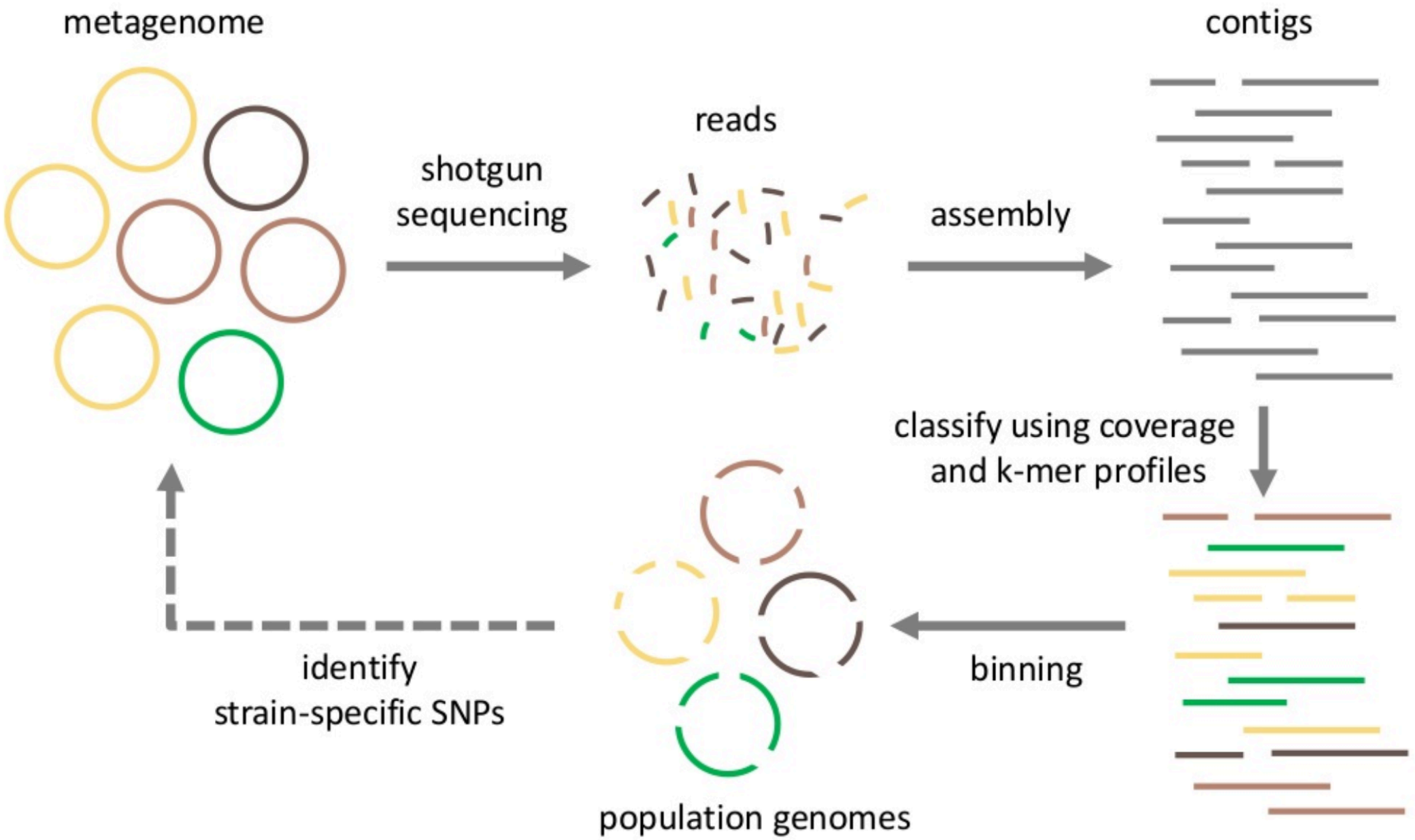
Sourmash

- Uses Minhash (another kmer technique) but low memory and low cpu needs
- http://sourmash.readthedocs.io/en/latest/



Metagenomics assembly

To determine if you've sequenced 'enough' to re-assemble 'most' of the community member's genetic content, look to what is left over - proportionally


metagenome
shotgun sequencing
reads
assembly
contigs
classify using coverage and k-mer profiles
binning
identify strain-specific SNPs
population genomes

Metagenome Assembly

- Many assemblers to choose from and more each day
  - Recent tutorial using cloud computing

  http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4496567/
- Most metagenomics assemblers use kmers
  - Either normalize reads by kmers (remove what appears to be redundant information)
  - Or bin by kmers (each bin is assumed to be a unique species), assemble each bin (first normalizing by kmers), or sort post assembly by bin.
- Map reads back to assembly to estimate coverage/ kmer count
- BUT then you have to do some with ambiguous contigs
  - Identify ORFs, marker genes, etc. to characterize gene/taxon content
  - IT IS all about the databases!!

MG-RAST

The MG-RAST system provides answers to the following scientific questions:

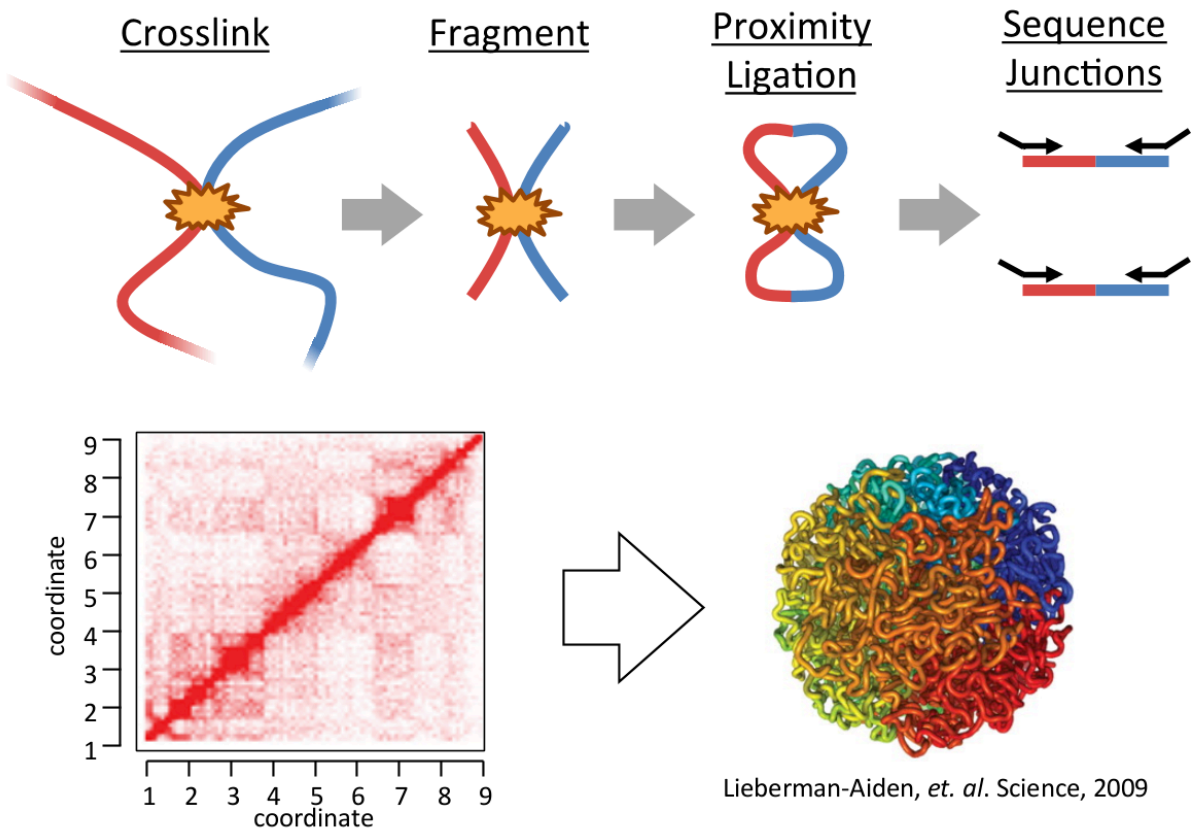
- Who is out there? Identifying the composition of a microbial community either by using amplicon data for single genes or by deriving community composition from shotgun metagenomic data using sequence similarities.
- What are they doing? Using shotgun data (or metatranscriptomic data) to derive the functional complement of a microbial community using similarity searches against a number of databases.
- Who is doing what? Based on sequence similarity searches, identifying the organisms encoding specific functions.
- Finally compare samples to each other

MG-RAST

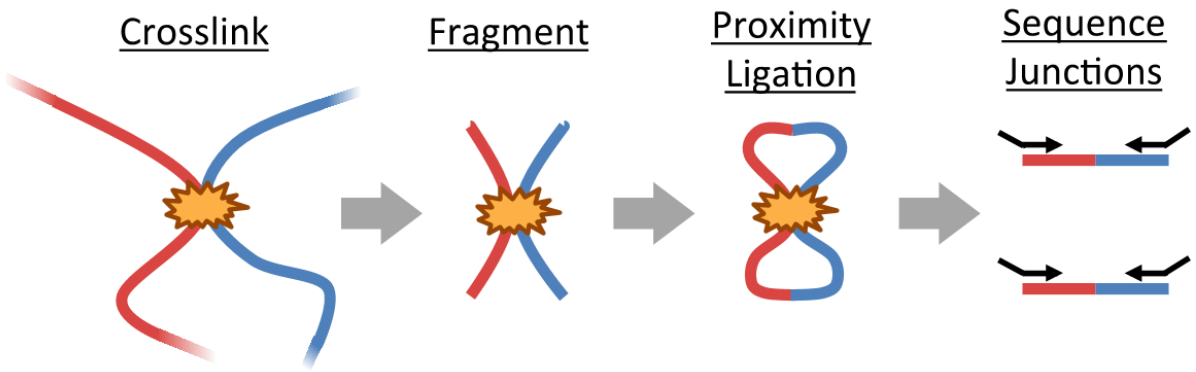
- Can upload
  - 16s amplicons
  - Metagenomes/Metatranscriptomes
  - Assembled contigs
  - Raw reads
- Use their resources for analysis, don't have to have your own computational resources
- More of a black box, but can download many of output data options
- Subjected to their philosophy for analysis of metagenomic/transcriptomic data

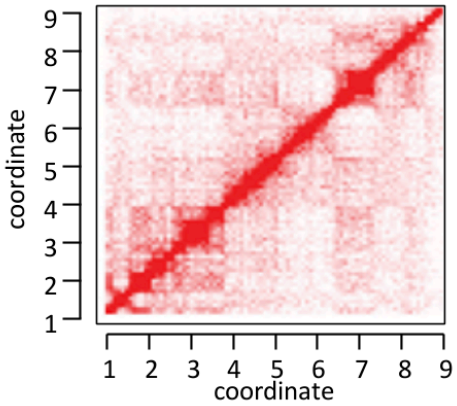
A downloadable alternative to MG-RAST is MEGAN5

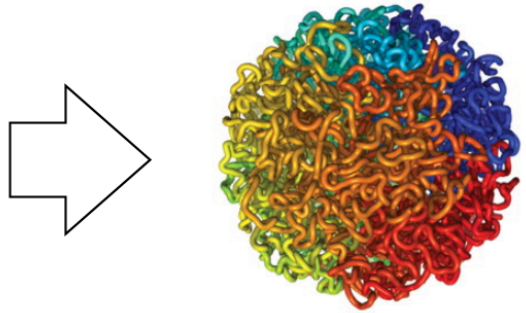
Metagenomes with Hi-C



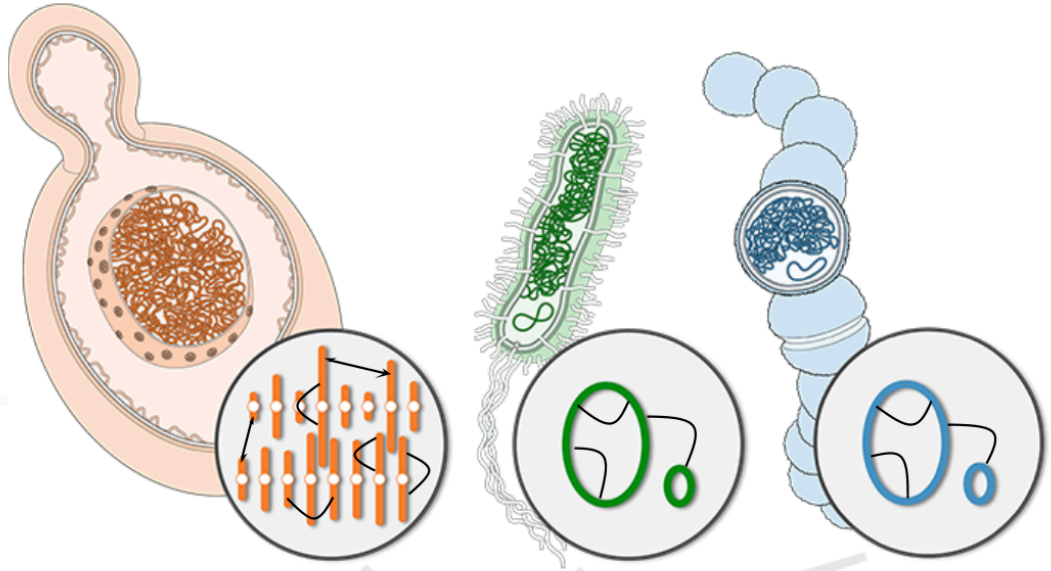
UCDAVIS Bioinformatics Core







Lieberman-Aiden, *et. al*. Science, 2009



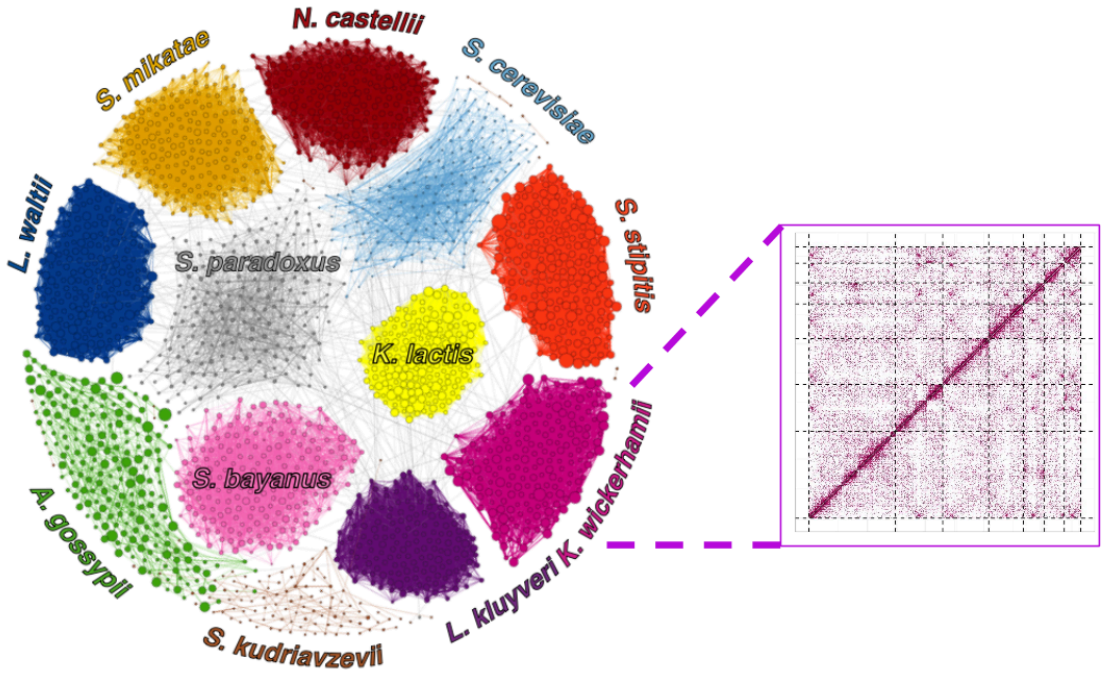
Shotgun sequencing

With Hi-C

UCDAVIS Bioinformatics Core

Phase Genomics

Reference-quality pro- and eu- karyotic genomes from mixed populations



ProxiMeta™ Hi-C: Genome-scale deconvolution & assembly of metagenomic samples