NOTES

Lecture 9: Genome Assembly

Overview

- next generation sequencing
- assembly problems and solutions
- methods and metrics

Clones

$$n = \frac{\ln(1-P_o)}{\ln(1-f)}$$

n = number of clones in a gene library
P_o = desired probability of gene in library
f = fraction of genome in one insert

- for a probability of 0.99 (99% chance of a desired gene in the library) and an average insert size of 40Kb, the number of clones required varies as a function of genome size

$$n_{E.coli} = \frac{\ln(1-0.99)}{\ln\left(1-\frac{4\times10^4}{4.6\times10^6}\right)} = 5.3\times10^2 \text{ clones}$$

$$n_{H.sapiens} = \frac{\ln(1-0.99)}{\ln\left(1-\frac{4\times10^4}{3\times10^9}\right)} = 35\times10^5 \text{ clones}$$

$$384 \frac{\text{clones}}{\text{plate}} \times 66 \frac{\text{plates}}{\text{rack}} \times 5 \frac{\text{racks}}{\text{shelf}} \times 5 \frac{\text{shelves}}{\text{freezer}} = 633{,}600 \frac{\text{clones}}{\text{freezer}}$$

Module 9.1 Genome Assembly: Platforms for "Now Generation Sequencing"

Platform	# reads per run	Read Length	Cost per Megabase	Throughput per run	Run time	Limitations	Detection Method	Sequencing Method	Paired Ends
Sanger	384	1,000 bp	$1000	100 kb	<1 day	Low throughput, expensive	Fluorescence	Chain termination method	Yes
454 Titanium	500,000	1,000 bp	$25	500 Mb	1 day	Expensive, homopolymer runs	Fluorescence	Sequencing by synthesis	No
Illumina HiSeq 4000	10 billion	2X 100 bp	$0.0018	1,500 Gb	3.5 days	Short reads, lots of data, run time	Fluorescence	Sequencing by synthesis	Yes
Illumina MiSeq	30 Million	2X 300 bp	$0.133	15 Gb	2.5 days	Short reads	Fluorescence	Sequencing by synthesis	Yes
PacBio Sequel SMRT Cell 1M	500,000	~10,000 bp	$0.05	20 Gb	20 hours	Need a lot of intact DNA	Fluorescence	Single molecule, real-time (SMRT)	No
Oxford Nanopore Minion MK1B	3 Million	~10,000 bp	$0.03	30 Gb	Real-time (<48 hours)	High sequencing error (85%), Need a lot of intact DNA	Amperage	Nanopore-based sensing	No

Notes:

Lessons Learned

- The Human Genome Project unit cost $1 per base pair with a total cost of $3,000,000,000 ($3B) tax payer dollars at $146,000 per gene (based on 20,500 genes predicted)
- A relatively low throughput method base on large insert clones was used to order and orient contigs making for more accurate genome assembly
  - human DNA (multiple individuals) → ligated to vector → transformed yeast to clone → sanger sequencing
- a competing private effort by Celera Genomics (Craig Venter) use a more scalable small insert clone approach that cost $100,000,000 ($100M) at $4,878 per gene with many more gaps in assembly

Why De Novo Assembly?

- non-model or uncultivated organism → no reference genome exists
- interested in novel genomic elements not present in standard reference,
  - e.g. effective for organisms with high intraspecific genome plasticity such as fungal pathogens
- often used to update older references when newer sequencing methods are used such as PacBio or Oxford Nanopore
- *de novo* reconstruction of a genome reduces dependence of using a reference as prior information

What is an Assembly?

- An assembly is a hierarchal data structure that maps the sequence data to a putative reconstruction of the target
  - it groups reads into contigs, and contigs into scaffolds
- contigs provide a multiple sequence alignment of reads plus the consensus sequence
  - the scaffolds ("supercontigs" or "metacontigs") define the contig order and orientation and the size of the gaps between contigs
  - scaffold topology may be a simple path or network
- assemblies are measured by the size and accuracy of their contigs and scaffolds

Reference-Guided De Novo Assembly

- NOT the same as reference based alignment
- Can use a related reference to aid in the assembly process
- these methods still involve a de novo step and can help when there is an available related reference genome

Read Coverage

$$R_N = \frac{CT}{rL(P_f)}$$

R_N = # of reads needed to complete target sequence
C = depth of genome coverage
T = length of target DNA sequence in bases
rL = average length of a read (i.e. $Q > 20$)
P_f = pass rate, fraction of reads above Q threshold

- to sequence the E. coli genome using dye-termination sequencing methods at 10X coverage:

$$R_{E.coli} = \frac{(10)(4.6 \times 10^6)}{(700)(0.8)} = 82{,}143 \text{ reads}$$

Probability of Nucleotide Recovery

$$P_0 = e^{-c}$$

P_0 = probability that a base is not sequenced
e = numerical constant equal to 2.71828
c = fold sequence coverage = $\frac{LN}{G}$
L = read length
N = number of reads sequenced
G = length of target DNA sequence in bases

- Lander-Waterman model assumes sequences are randomly distributed over genome (Poisson distribution)
- at 10-fold sequence coverage the probability that any base is not sequenced is 0.000045 → 0.99995 $(1-P_0)$ of the target has been sequenced

Assembly Problems

- DNA sequencing technologies share the fundamental limitation that read lengths are much shorter than even the smallest genomes
- whole shotgun sequencing (WGS) overcomes this limitation by oversampling the target genome with short reads from random positions
  - assembly software reconstructs the target sequence
- assembly software challenged by repeat sequences, sequencing error, non-uniform coverage, and computational complexity

Poisson Distribution

- distributions are not always accurate
- biology is unpredictable

- e.g. **TTAGGGTTAGGGTTAGGG**
  - How many times should we expect to find this 18bp sequence by chance in the human genome?

$$(6.5 \times 10^9)(4^{-18}) = 0.08 \text{ or } < 1 \text{ match}$$

(6.5×10^9: human genome in bases; 4^{-18}: 4 bases)

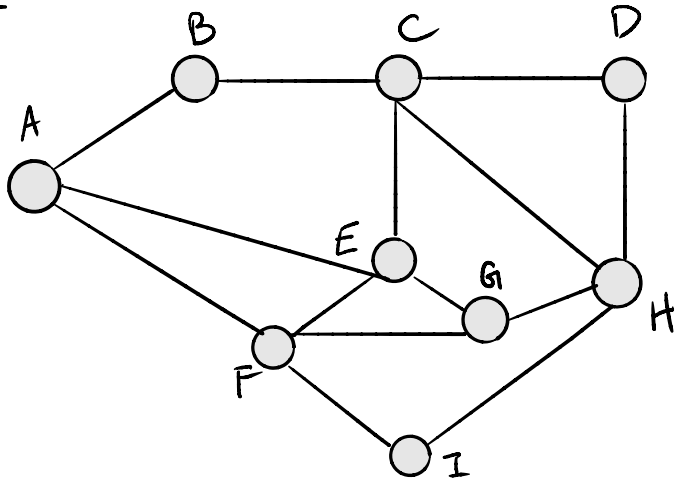
  - however, if you search the human genome for sequence, you will find 100s of exact matches (repeats)
  - if there were no repeats, 17bp would be enough to specify a unique position in the human genome

$$\frac{\log(6.5 \times 10^9)}{\log(4)} = 16.24$$

Reality Check

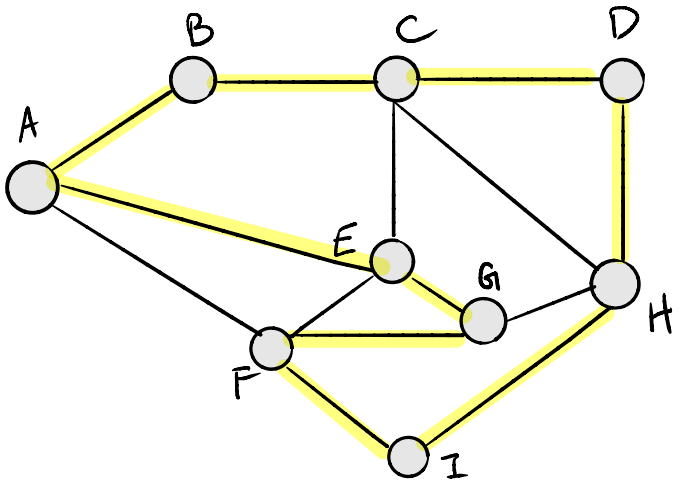
- high depth of <u>coverage</u> is needed to produce high-quality assembly
- sequencing depth needed for completion will depend on genome size, GC composition, repeat content, sequencing platform
- genome completion can also be estimated by counting the number of conserved single copy marker genes in genome
- read length matters

Graphs



- a graph or network is a representation used to draw inferences between related nodes (people, reads, k-mers, species, etc.) and the links (edges) between them

Hamiltonian Cycle



- a hamiltonian circuit or cycle: visit each node (city, sequence, ...) exactly once, returning to the start.
  - this is an NP-hard (non-deterministic polynomial-time hard) problem in computer science

Overlap Layout Consensus

- overlap between sequence reads is used to create a link between them resulting in a directed graph based on all versus all alignment that completes a Hamiltonian cycle.
  - this identifies reads that can be merged to generate a consensus contig sequence
- the genome sequence is then assembled by aligning sequences of adjacent contigs and calculating a path through these alignments that will produce a non-redundant sequence often called a Tiling Path.
- some short-read assemblers that are based on this method include Celera, SSAKE, VCAKE, SHARCGS and Roche's proprietary 454 assembler, Newbler
- the nodes are whole reads → useful for maintaining long distance sequence information (but computationally expensive)

Celera Workflow

- remove repeats, mask low complexity regions based on a priori knowledge and sequence quality scores → repeat masker
- identify overlaps between reads at user defined length and identity thresholds
- assemble high-confidence sequences from overlapping reads
- order and orient the contigs using mate pair information
  - fill gaps with strings of 'N' ambiguity characters
- attempt to resolve sequencing errors

OLC Challenges

- polymorphisms, errors, repeats and other ambiguities result in forking paths that reduce contig lengths and increase graph complexity
- OLC is useful on data sets containing fewer than billions of reads > 300 bp in length
  - however, very short reads (< 300 bp per read) are not well suited to OLC
  - the sheer number of reads makes the overlap graph, with one node per read, extremely large and lengthy to compute
- computational complexity relates to both the number of pairwise comparisons and the number of edges in the graph
  - this information is difficult to parallelize because the entire graph must be stored in RAM

De Bruijn Graphs

- De Bruijn Graph: A Euler path that crosses every edge exactly once without repeating, if it ends at the initial vertex then it is a Euler cycle

De Bruijn Assembly

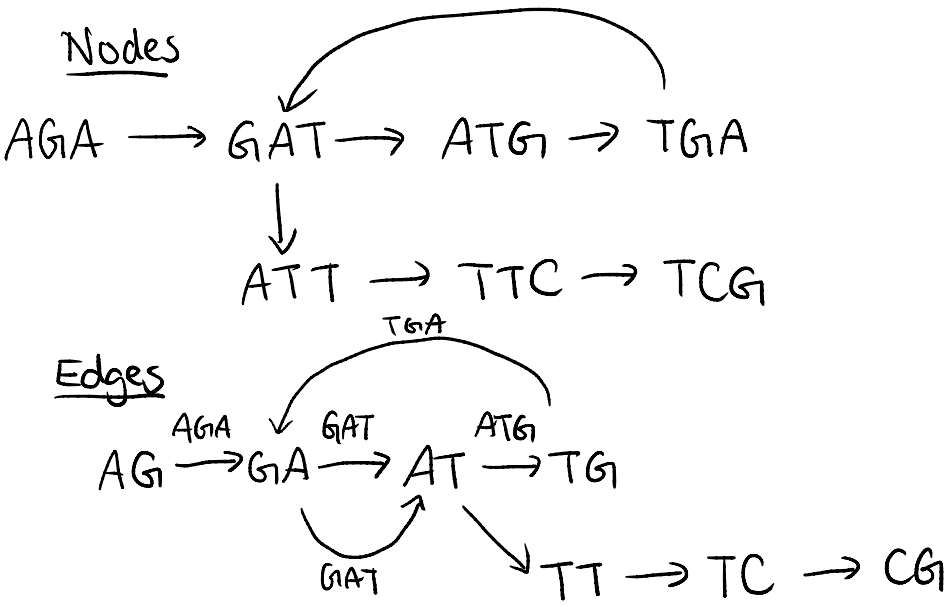
- reads are decomposed into k-mers that can be represented as nodes or edges
  - Eulerian graph: k-mers are edges
  - Hamiltonian graph: k-mers are nodes

- a directed edge between nodes indicates that k-mers on those nodes occur consecutively in one or more reads with k-1 overlap
- the locally constructed graph reveals the global sequence structure of a genome
  - overlaps between nodes are implicitly captured by the graph, rather than explicitly computed, saving a substantial amount of computing time over other methods
- some short-read assemblers that are based on this method include Velvet, MEGAHIT, SPAdes, ABySS, ALLPATHS, and SOAPdenovo

AGATGATTCG

k-mer = 3

AGA
GAT
ATG
TGA
GAT
ATT
TTC
TCG



The Key is "K"

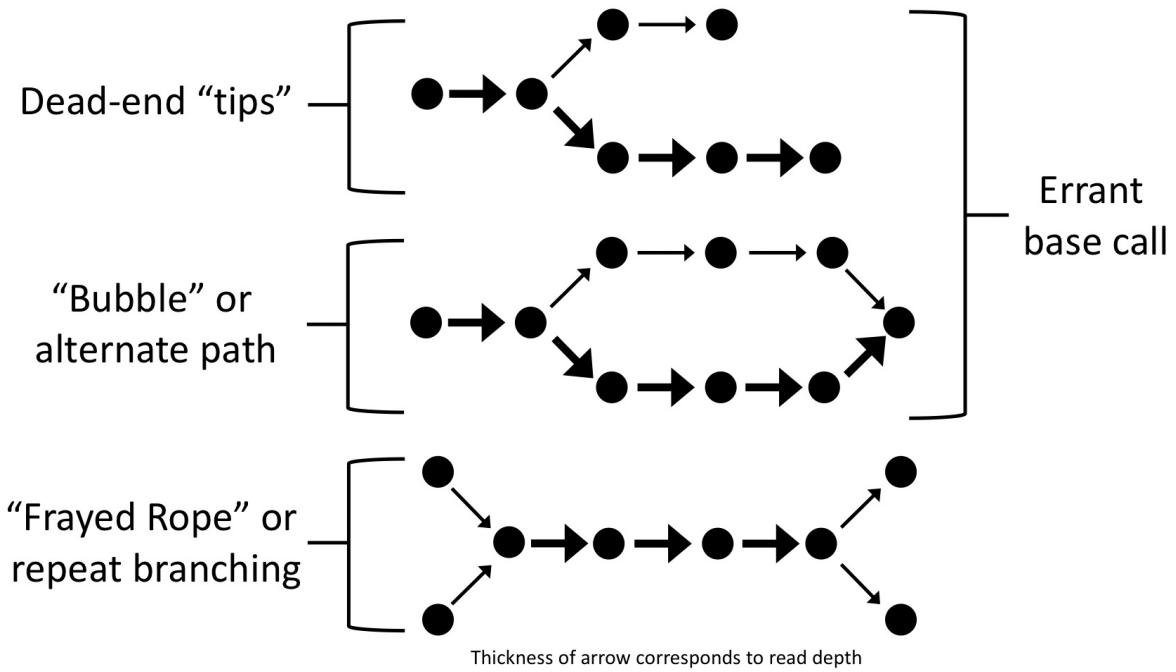
- the size of K used to generate the K-mers is critical to a contiguous genome assembly
- if K is too small, the graph becomes tangled
- if K is too large, the graph becomes fragmented
- the best way to find this value is to sample a range of K-mer values resulting in different assembly outcomes that can be empirically assessed using metrics

Velvet Workflow

- a set of algorithms collectively named Velvet, manipulates de Bruijn graphs efficiently to both eliminate errors and resolve repeats
  - VelvetH: Create a k-mer hash table with sequence coverage information
    - multiplicity = number of times you find a given k-mer
  - VelvetG: Construct the de Bruijn Graph (DBG)
  - Compress the DBG based on unambiguous edges
  - simplify the DBG including tip removal, bulge removal and removal of erroneous graph connections
  - resolve repeats
  - extract the contigs

Graph Compression

- after graph construction, many edges are unambiguous
  - these edges are subsequently merged together to compress nodes
  - e.g. node A only has one outgoing arc to node B, and if node B has only one ingoing arc → merge
- following graph compression different graph structures need to be resolved to remove assembly errors and artifacts including "dead-end" tips, mutant bubbles and low coverage chimeric edges



St. Petersburg Genome Assembler

- SPAdes utilizes multi-sized de Bruijn graph which allows employing different values of K
- the program uses a paired de Bruijn graph (i.e. double layered)
  - the k-mers from individual reads build the inner graph which is used to construct contigs
  - the outer graph composed of "paired k-mers" with large insert size is used to resolve repeats and construct scaffolds
- SPAdes uses paired-end reads library as input
  - single-end library will result in an error
  - user sets one basic parameter for the input DNA type: isolate DNA, single cell, or plasmid

Evaluation of an Assembly

- often not straight forward, especially for organisms with little representation in public databases
- truth is not known but can be approximated
- always best to use a combination of evaluation metrics
  - not just contiguity but also accuracy of an assembly
- Quality Assessment Tool for Genome Assemblies (QUAST) provides an easy-to-use tool for evaluation and comparison of assembly outcomes

Assembly Metrics

- Nx: the largest contig length, L, such all contigs of greater than or equal to L sum up to at least x% of the bases of the assembly
  - e.g. N50 is the contig length such that contigs of that size and greater account for 50% of the genome
  - a measure of contiguity (NOT completeness)
- Mapping genomic or transcriptomic reads back to the assembly to identify suspicious regions
- Identification and counting of core conserved genes expected to be present based in the genome using Benchmarking Universal Single-Copy Orthologs (BUSCO)

BUSCOs

- provides a quantitative measure of genomic data completeness in terms of expected gene content
- evolutionary expectation these genes should be present, and present only once, in a complete assembly
- Completeness is quantified in terms of this expected gene content by assessing the orthology status of predicted genes
- allows comparisons across different assemblers or versions of the genome in terms of completeness

N50 Statistic

- given a set of contigs, defined as the shortest contig length such that all contigs of that length and greater sum to $\geq 50\%$ of the total genome assembly
  - e.g. contig lengths: 50kb, 30kb, 20kb = 100kb
    - 50% of 100kb = 50kb → N50 = 50kb
  - e.g. contig lengths: 31kb, 30kb, 20kb, 19kb = 100kb
    - 31kb + 30kb > 50kb → N50 = 30kb
- the larger the better

Nx Curve Generation

- Nx is calculated for all values of x for 1 to 100% of the genome assembly
- the "Nx" curve illustrates the contiguity of the assembly
- values are plotted where:
  - x-axis are values 1-100
  - y-axis are contig lengths

Summary

- there are two major assembly paradigms, OLC and DBG currently in play
- OLC is optimal when you have smaller data sets with longer reads and DBG are optimal for larger data sets with shorter reads
- Nx curves are a proper visualization for sequence contiguity (e.g. genome completion)

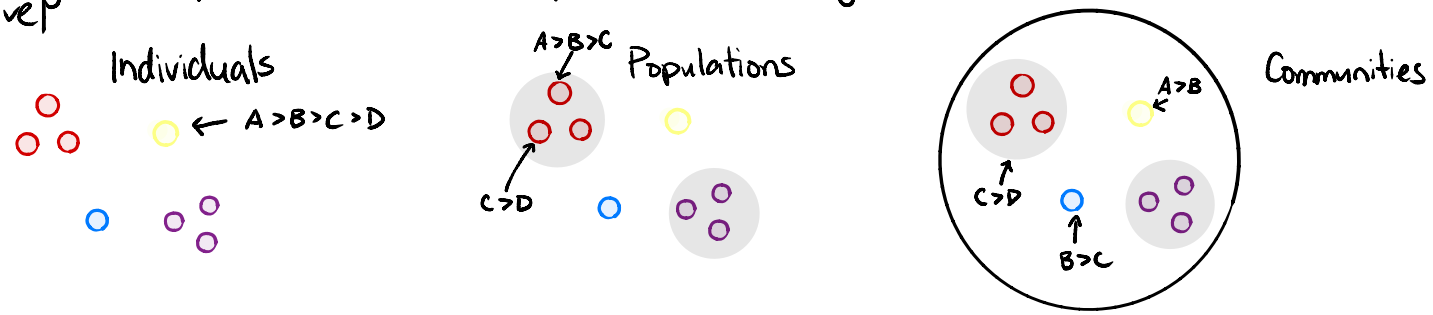
Lecture 10: Metagenomic Analysis

Overview

- living in a microbial world
- through the looking glass
- standards and assessment

Ecological Design

- Chisholm, 2000: The regulation of the pools and fluxes in biogeochemical cycles have their origins in the genetic inventory of individual microbes, and the regulation of these genes within the organism is determined by the environment. As such, one can look at the microbial food web as a collection of genomes whose expression and replication is coordinated through complex feedback loops at the organismal, population, and ecosystem level



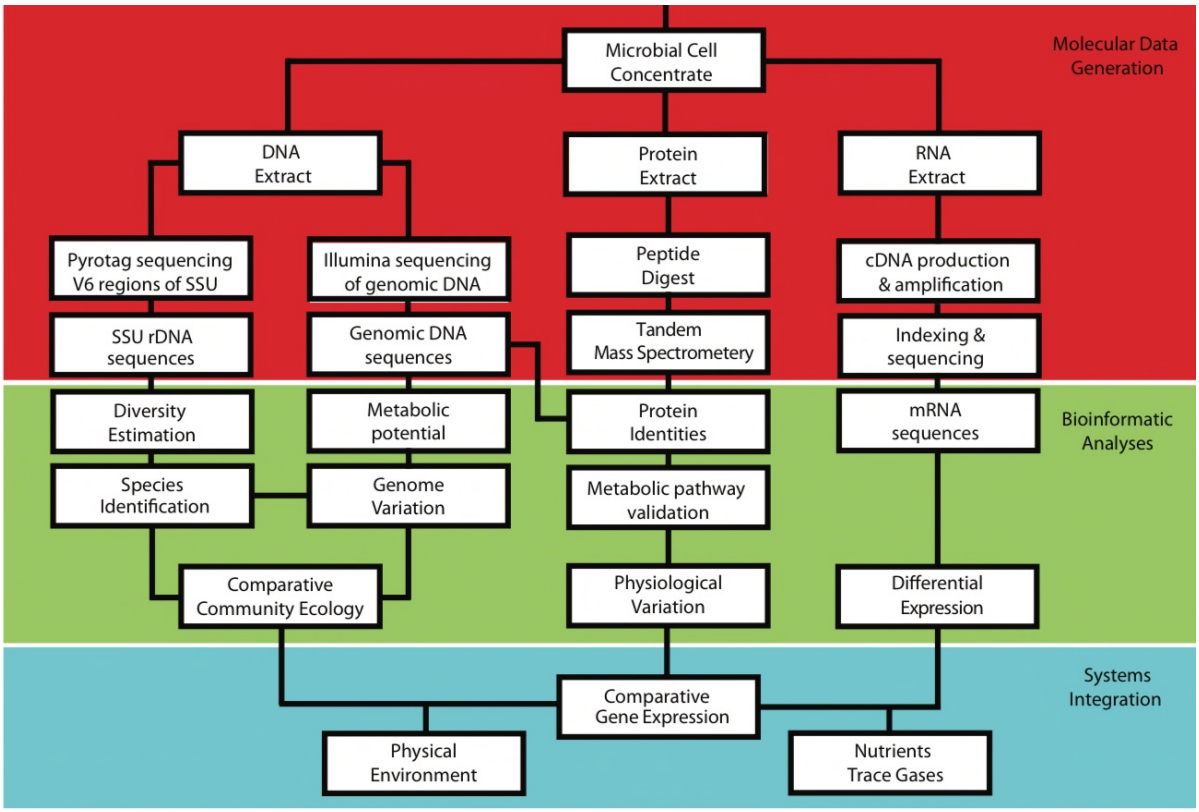
- increased connectivity as oxygen concentrations lower (Pearson correlation)
  - decrease in energy available increases cooperation and connectivity
- microbial connectivity } environmental sequences + engineered strains → biorefining ecosystems ← biomass → energy and materials

Foundational Questions

- what is the taxonomic and functional structure of the ecosystem?
- how does this structure change in response to environmental perturbation?
- what are the ecological and biogeochemical consequences of this change?
- what are relevant units of selection, conservation, or utilization within microbial communities?

Biological Information

- DNA: genomes last years
- Transcripts: transcriptomes are time sensitive, and responsive to change, lasting seconds to minutes
- Proteins: lingering signatures of the environment (metabolite production) lasting minutes to days



SSU rRNA Gold Standard

- reverse transcribe rRNA to build a "ribosomal tree of life"
  - 16S rRNA gene is ubiquitous and conserved in bacteria
  - 9 variable regions, 10 conserved stems
- tag each rRNA to characterize isolates

NCBI Taxonomy

- ~360,000 Taxa
  - 25,000 Prokaryotes
  - 84,000 Animals
  - 65,000 Plants
  - 17,000 Viruses

Domain → Kingdom → Phylum → Class → Order → Family → Genus → Species

Cultivation-Independent Gene Surveys

- microbial community → environmental DNA → PCR Amplification → Amplicon Sequencing → Sequence Clustering
- put barcodes in amplicons so that same OTUs share barcodes
  ↳ nodes in network

Environmental Monitoring

- three distinct microbial communities or "functional groups" are observed based on molecular barcode analysis of the small subunit of the rRNA gene

Biological Data Structures

Tier-1: Highly Curated
↓
Tier-2: Moderately Curated
↓
Tier-3: Automatically Curated
↓
Tier-4: MAGs

} SAGs

Generic Workflow

- Sample Processing
  - Microbes
    - sample conditions of interest
  - Extract biological information
    - DNA, RNA, Protein, Metabolites
  - Sequence
    - Genome, Transcriptome, Proteome, Metabolome
- Data Processing
  - Assembly
    - construct and bin contigs from short read data
  - open reading frame (ORF) prediction
    - identify functional parts of genome from sequence motifs

- Annotation
  - search databases for subjects with taxonomic or functional identities
- Analysis
  - who is there?
  - what are they doing?
  - how do they respond to change?

Challenges

- Analysis is a Big Data problem
  - Volume
    - billions of sequences across many samples
  - Variety
    - environmental conditions, taxonomy and function
  - Complexity
    - multiple hierarchal levels of organization
  - Dynamism
    - lateral gene transfer and viral reprogramming
- Computation is heterogeneous
  - Software
    - many different software have to work together
  - Computation
    - local resources or high performance computing (HPC)
  - Parallelism
    - local (CPU) and cloud (distributed)
- need to maintain a data model throughout

Design Considerations

- modularity
  - rapid swap-in, swap-out and integration of new modules
- input/output
  - archive, retrieve raw and compare
- memory and performance issues
  - fast and cache friendly

Genome Sizes

- assuming that the number of unique genomes per gram of soil approximates 1 in 10^6 ...
  - ↳ How much sequence data would be needed to assemble these metagenomes assuming 100X coverage?

Assembly Problem

- complex such as soil metagenomes don't assemble very well
- need a lot more raw sequence data and computational power
- use unassembled short read data for gene finding and annotation

Gordian Knot

- multiple coexisting genotypes represented in one sample
- cross-section of naturally occurring heterogeneity
  - microdiverse clusters
- on average, no two clones derived from same genome
- sequence space is complex and interwoven
  - lateral gene transfer: phage, plasmid, integron...
  - symbiosis and syntrophy
  - ecological context and natural history

Metagenome Size

$$G_m = \sum_{i=1}^{\ell} n_i G_i$$

G_m = metagenome size in bases
ℓ = number of genomes in sample
n_i = number of copies of genome G_i
G_i = size of any given genome in sample of ℓ genomes

- in any given metagenome sample, genotypes of different sizes appear at different frequencies
  - therefore a metagenome of size G_m composed of genomes of sizes G_i through G_n can be viewed as a sum of fractions where each component genome constitutes a fraction of G_m

$$G_m = p_1 G_i + p_2 G_j + \dots p_n G_n$$

→ sequence until approaching saturation

Minimal Information

- Developed by Genome Standards Consortium for reporting bacterial and archaeal genome sequences from isolates and mixed communities
- Based on the Minimum Information about Any (x) Sequence (MIxS) standards that incorporate metadata including sampling location and environmental conditions
- focus on assembly quality, completion and contamination to establish community-wide standards for analysis of MAGs and SAGs

Genome Reporting Standards

- there are currently 12,728 MAGs and 2174 SAGs deposited in the Joint Genome Institute (JGI) Genomes OnLine Database (GOLD)

Criterion	Description
Finished	Single contiguous sequence without gas or ambiguities with consensus error rate $\geq$ Q 50
HQ MAG/SAG	Multiple fragments where gaps span repetitive regions. Presence of the 23S, 16S and 5S rRNA genes and at least 18 tRNAs. $>$ 90% completeness $<$ 5% contamination
MQ MAG/SAG	Many fragments with little to no review of assembly other than reporting of standard assembly statistics $\geq$ 50% completeness $<$ 10% contamination
LQ MAG/SAG	Many fragments with little to no review of assembly other than reporting of standard assembly statistics $<$ 50% completeness $<$ 10% contamination

Quality Control Metrics

- Assembly statistics include but are not limited to: N50, L50, largest contig, number of contigs, assembly size, percentage of reads that map back to the assembly, and number of predicted genes per genome
- Completion: ratio of observed single-copy marker genes to total single-copy marker genes in chosen marker gene set
- Contamination: ratio of observed single-copy marker genes in ≥ 2 copies to total single-copy marker genes in chosen marker geneset

Generating Sequence Bins

- contigs and scaffolds from a metagenome assembly can be binned using a variety of metrics including the identification of single copy marker genes
  - e.g. phylogenetic anchors or the identification of well-resolved functional gene anchors that represent a specific metabolic process
- in addition, Gt C content or k-mer profiling methods can be used to identify intrinsic nucleotide distribution patterns to statistically assign contigs to discrete sequence bins

Phylogenetic Anchors

- translated nucleotide sequences are mapped onto branches of corresponding reference tree based on sequence homology
- the size of the circles and bars represent the proportion of genes that can be probabilistically assigned to an internal node (circles) or phylotype (bars) through a maximum likelihood scoring assessment to reference tree
- at present, reference trees exist for 16S rRNA genes, 18S rRNA genes and 40 universal COGS
- Tree of Life (Archaea, Bacteria, Eukaryota)

  RuBisCo, nifH, nifD, MMO

- within each reference tree there are functional anchors

Critical Assessment

- lack of consensus regarding benchmarking genome assembly and binning methods complicates performance assessment
- the Critical Assessment of Metagenome Interpretation (CAMI) challenge engages developers to benchmark their programs on complex simulated data sets derived from 753 real world genomes
- CAMI compared 25 programs commonly used in metagenomic assembly and binning pipelines...

Genome binner (% contamination)		Recovered genomes (% completeness)		
		>50%	>70%	>90%
Gold standard		753	753	753
CONCOCT	<10%	275	272	262
	<5%	267	265	256
MetaWatt 3.5	<10%	500	475	405
	<5%	476	452	393
MetaBAT	<10%	247	228	195
	<5%	234	216	186
MyCC	<10%	250	240	197
	<5%	220	211	173
MaxBin 2.0	<10%	390	385	343
	<5%	356	352	316

- genomes were divided according to their average nucleotide identity (ANI) into "unique strains" (genomes with <95% ANI to any other genome) and "common strains" (genomes with an ANI $\geq$ 95% to another genome in the data set)
- genome binners performed well when no closely related strains were present
  - taxonomic binners reconstructed taxon bins of acceptable quality down to the family rank leaving a gap in species and genus-level reconstruction

SAG Extrapolation

- metagenomic contigs were binned with SAGs using a combination of BLAST, tetranucleotide frequency distribution, and manual curation
- 51 metagenome contigs > 5 kb with 95% similarity to MGA SAGs