

Genomics

Science of whole genomes
Part I: Sequencing a genome

Problem

- Sequence several billion nucleotides
- Assemble into chromosomes
- Identify genes, regulatory units, Tes
- Verify

Sequence a genome

- Genomes range from 10^6 to $>10^9$ bp
- Sequence read averages 500 bp
- Must break up genome
- Assemble short pieces into chromosomes
- Sequence each bp at least 10 times

Two complementary strategies

- Directed sequencing of large, ordered chromosome chunks
- Whole Genome Shotgun sequence--Randomly sequence small chromosome chunks

Cloning

- A piece of DNA is propagated in a microbe usually bacteria.
 - High fidelity replication
 - Large quantities easily obtainable.
- Usually use *E. coli*
- Need a vector
 - Allows maintenance in host
 - Convenient cloning sites

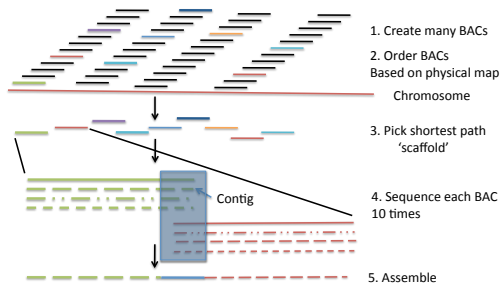
Types of vectors

- Plasmids
 - Small
 - Hold up to 15kb of DNA
- BAC
 - Large
 - Holds 100kb or more DNA

Directed Sequencing

- Genome broken into 100 kb pieces
- Each piece cloned into BAC
- Ends of the BAC are sequenced
- BACs are ordered into a tiling path or scaffold across a chromosome.

Directed Sequencing



Whole Genome Shotgun sequence

- BAC Library
- 2 Plasmid Libraries
 - 2kb library
 - 10 kb library
- Ends of every clone in each library is sequenced
- Computer assembles sequence

Most Sequencing Projects

- Most Sequencing projects use a combination of physical maps and WGS
- Computers assemble sequence into contig
- Contigs assembled into chromosomes when possible

Problems in assembly

- Repetitive sequences
 - Transposable elements
 - Di and tri nucleotide repeats
 - Ribosomal genes
- Gaps
 - “Unclonable” sequence
 - Problematic sequence
 - Heterochromatin

Finishing

- Humans attempt to close gaps
- PCR directed cloning

Annotation

- How to make sense of raw sequence
- Software predicts genes
 - Error prone
 - Cannot predict regulatory sequence
- TEs ignored (Our job!)

Draft vs Complete

- Genomes can be finished to draft or complete sequence
- Dependent on:
 - Cost
 - Scientific Need

Environmental sequencing

- Take a sample from the environment
- Clone all DNA in that sample
- Randomly sequence clones
- Deposit unassembled sequence
- Examples
 - Sargasso Sea Metagenome
 - Viral Metagenome of Yellowstone hot springs

Genomics

Science of whole genomes
Part II: Understanding the data

Bioinformatics

- Data mining, computational biology
- Use of software to make sense of Terabytes of data.
- GenBank
 - 240,000 named organisms (3000 new added/month)
 - 145 billion nucleotides (doubles every 18 months)

Bioinformatics

- Union of Computer Science and Biology
 - Mine genome sequence: genomics
 - Mine protein structure and function: proteomics
 - Analyze gene expression data
 - Analyze images of biological specimens
 - Assemble biochemical pathways
 - Visualize huge amounts of data
- Bioinformatics is a tool that is used to generate hypotheses.

This week: Genomics

- Learn some bioinformatics tools: PubMed, Blast, TATE.
- Use bioinformatics to search out TEs.
- Learn about sequence relatedness using multiple alignment and phylogenetic trees.
- Learn how bioinformatics predicted an active TE in rice
- Demonstrate the TE is active using molecular techniques.
