MMG 301, Lec. 23 Genomics

Questions for today:

1. How are microbial genome sequences determined?

2. What can we learn from the *E. coli* genome?

3. What is known about other microbial genomes?

Genome Sequence Analyses: The process

"Shotgun approach"

-- Begin with a pure culture (WHY?)

-- Isolate DNA (chromosome and plasmid)

phenol/chloroform extraction

ethanol precipitation

-- Fragment DNA

Nebulizer to shear DNA randomly OR

Use several restriction endonuclease (that cut at only a few sites)

e.g., *Not*I from *Nocardia otitidas-caviarum*

GC↓GGCCGC

CGCCGG↑CG

8-cutter; thus, (1/4)^8 = 1/65,472 bases

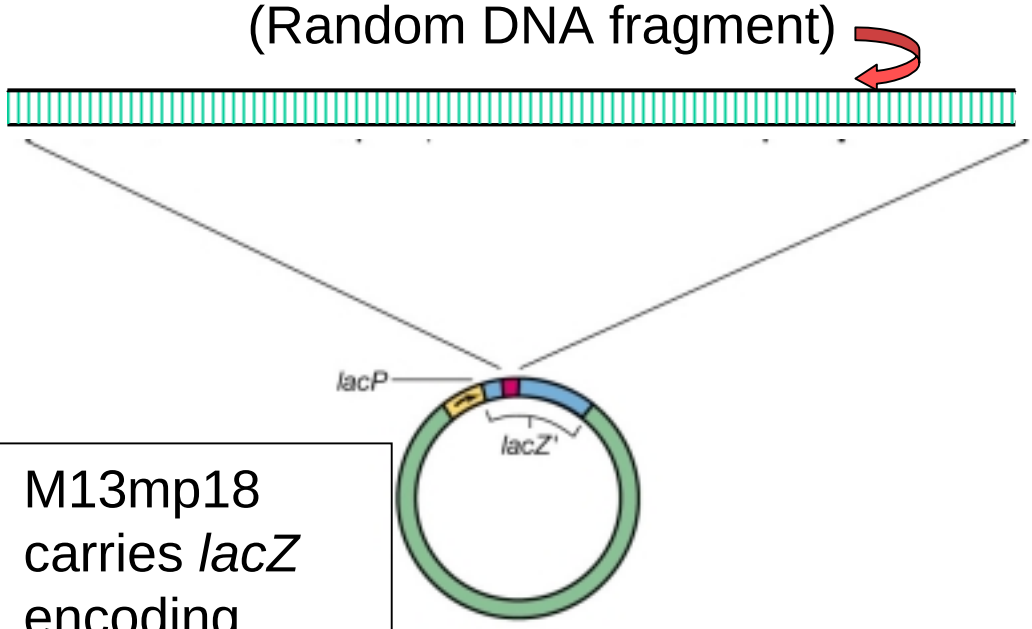
~76 fragments for *E. coli* chromosome

-- insert the fragments into a “vector”

Bacteriophage such as Lambda

Plasmid such as M13mp18

(Random DNA fragment)

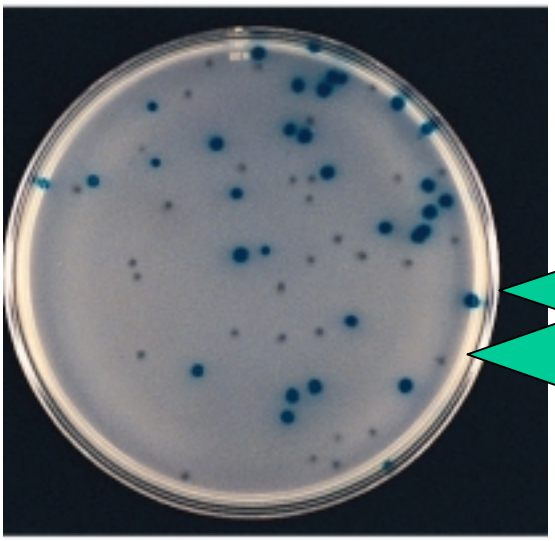


M13mp18 carries *lacZ* encoding β-galactosidase

-- Place the vector into a host strain

-- Identify colonies containing inserted DNA:

When its gene is interrupted β-galactosidase is inactive. X-gal (a galactose-like compound called 5-bromo—4-chloro-3-indolyl-β-D-galacto-pyranoside) is not converted to a blue dye

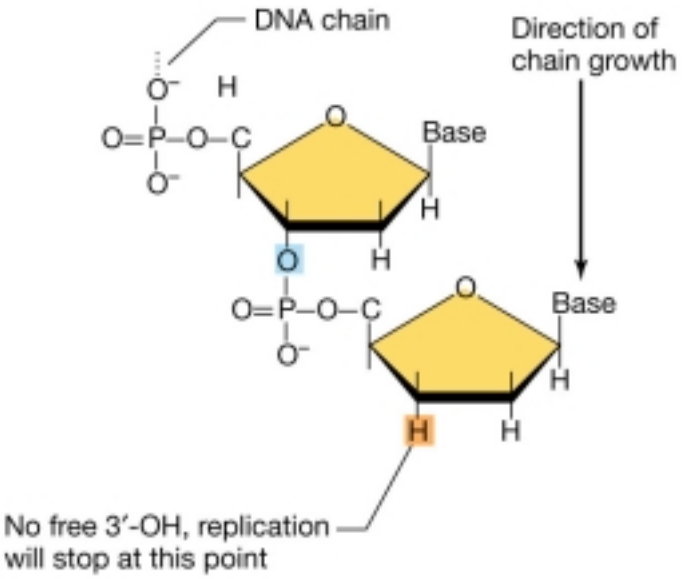


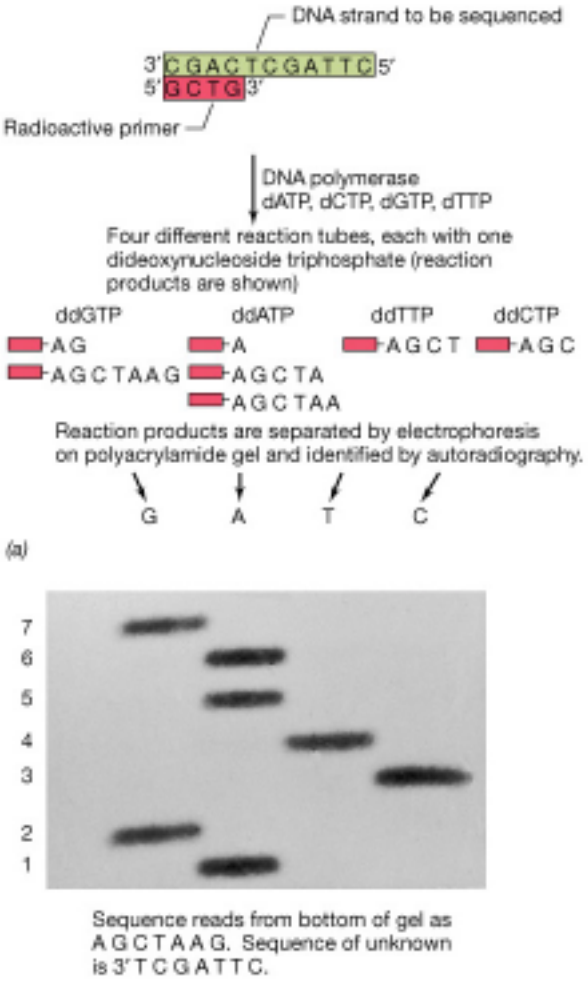
No insert

Insert present

(antibiotic resistance may also be used)

-- Each colony contains a single DNA insert, now present in many copies (~1 million cells/colony)
-- The plasmids present in each colony are isolated
    Gel electrophoresis
    Centrifugation
-- Automated sequencing by Sanger method
    Separate strands
    Add primer matching portion of vector
    Add dATP, dGTP, dCTP, dTTP
    Split into 4 samples
    Add dideoxynucleotides at low conc.
    Add polymerase, run polyacrylamide gel





Concern: ^{32}P is a hazard!

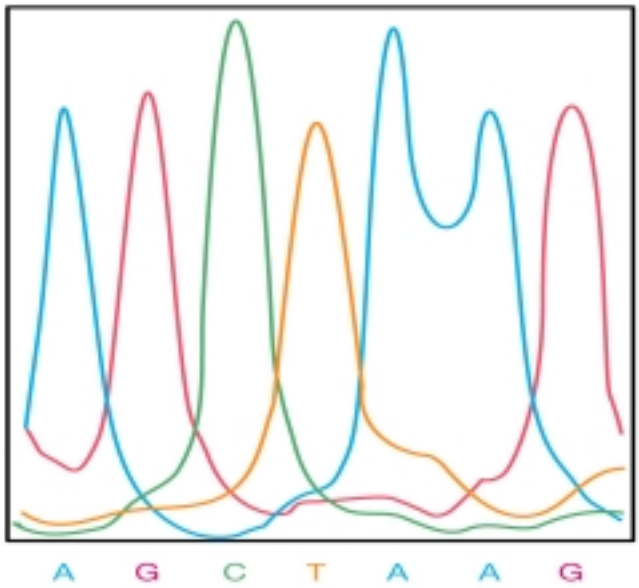
-- Alternative sequencing method:

only need a single reaction tube

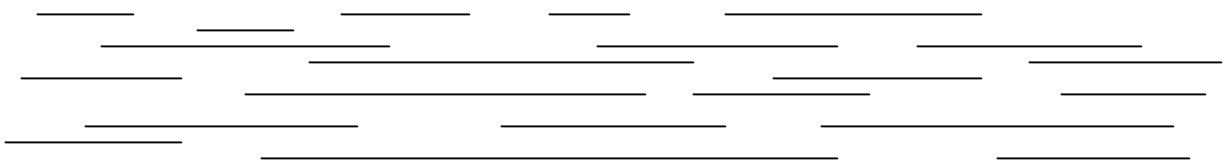
ddNTPs fluorescently labeled

This approach does not use radioactivity

Can sequence ~500 bases at a time!



-- Assemble information from randomly sequenced fragments into a contiguous sequence



[These eventually form a circle]

Requires extensive computational effort

Need to find sequence <u>overlaps</u> (puzzle)

Direction of fragments unknown

Any one stretch of DNA may be sequenced many (~10) times: shotgun

Keep sequencing fragments until complete

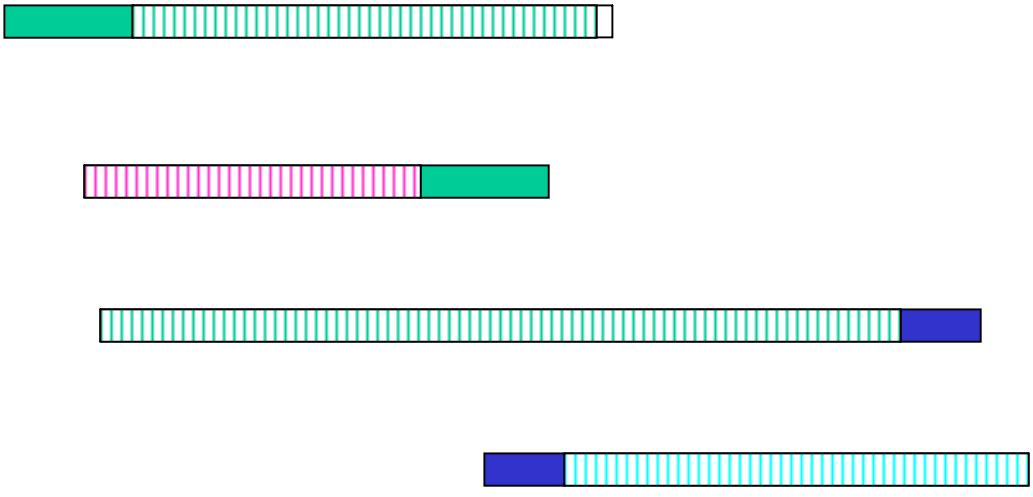
The entire process can be highly automated using robotics

“Chromosome Walking” approach:
-- Most steps are identical
-- Difference is that one purposely seeks out overlapping fragments so as to “walk out” from the first sequence.
-- This is useful for filling in gaps in small genomes or for determining very large genomes

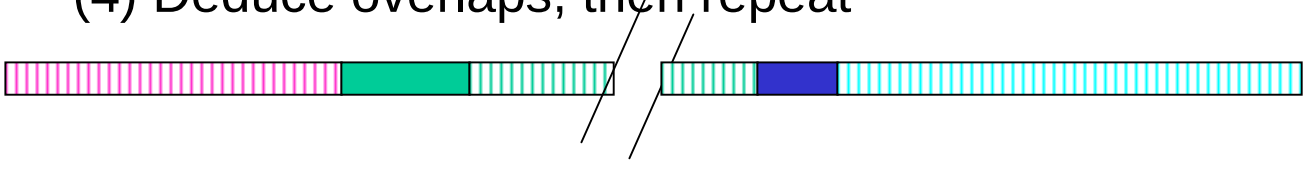
(1) Sequence the initial fragment

(2) Make primers matching sequences on each end

(3) Seek out other fragments that match the primers and obtain their sequences



(4) Deduce overlaps, then repeat



-- Identify “Open Reading Frames” (ORFs)

Use the genetic code (Table 7.3) Not Universal!

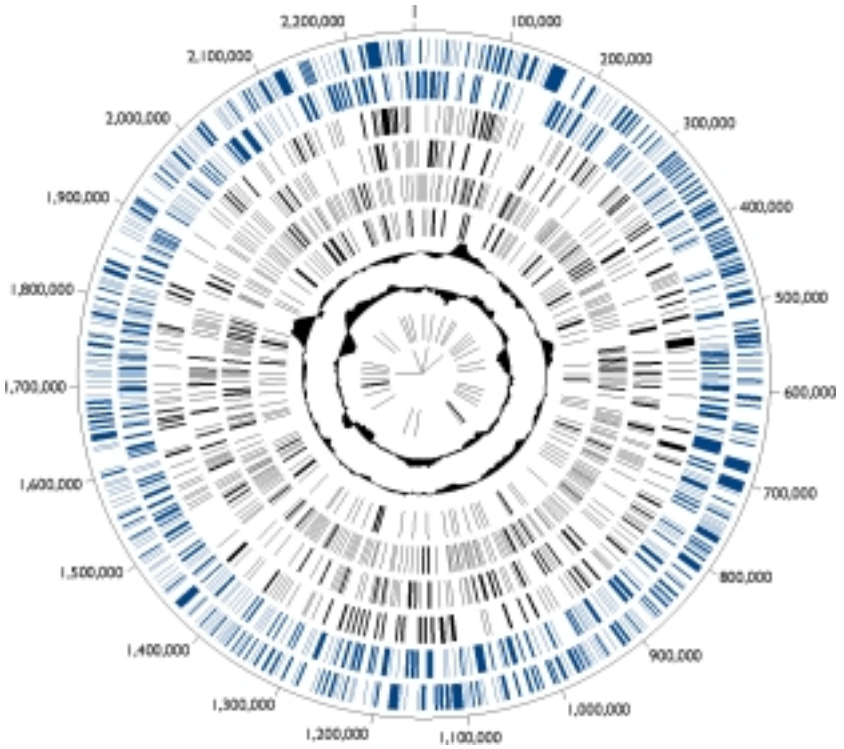
Consider all 6 possible reading frames

Look for likely start codons (AUG = Met)

Note stop codons and “codon usage”

One gene is derived from a single reading frame in prokaryotes

Within a chromosome, all 6 reading frames are used



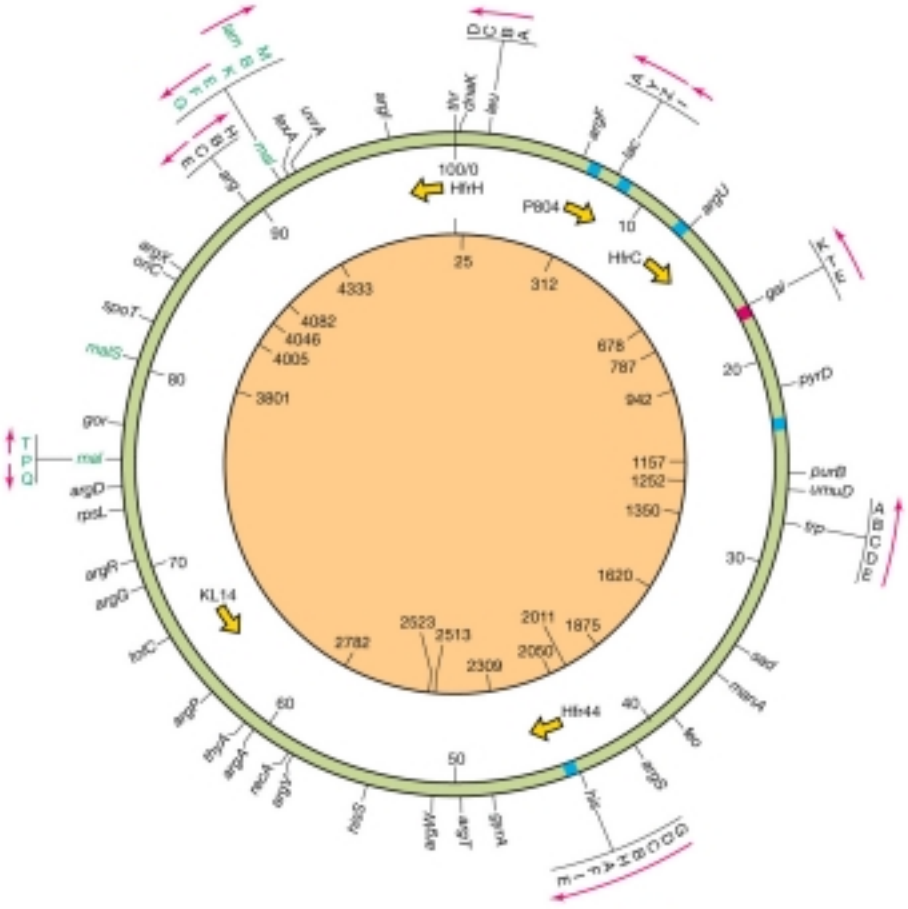
Codon usage: Leu = CUU, CUC, CUA, CUG, UUA, UUG; *E. coli* prefers 2

Neisseria meningitides genome (from Salyers & Whitt)

The *E. coli* genome

-- Determined in 1997 (this is recent history!)
-- Single, circular molecule
-- 4,639,221 base pairs for one lab strain ***
-- 4,288 ORFs ***
-- 62% are identified: **38% unknown!**
- metabolism (~21%)
- structure (5%)
- transport (~10%)
- replication (~3%)
- transcription (~1%)
- translation (~5%)
- regulation (~9%)
- other known (~8%)

This is an abbreviated version of the *E. coli* genome



What does the map show?

Genes involved in biosynthesis are easy to study. A mutation of that gene prevents synthesis of the product, but addition of the amino acid (or other missing component) allows growth.

trpABCDE: operon of genes encoding enzymes in a pathway for Trp biosynthesis. Genes are *italicized*.

TrpA (not italics) is the protein encoded by *trpA*.

Not all enzymes in a pathway are encoded by adjacent genes; e.g., *argR, argG, argD* vs. *argHBCE*.

oriC: origin of replication

100 min and Hfr strains: It takes 100 min to transfer the chromosome to another cell by conjugation (lec 26). The position of genes can be shown by interrupted mating.

lamB (also called *malB*) and *gal* operon: important to bacteriophage Lambda infection (lec. 25)

lac operon: used for growth on lactose. The regulation of this operon is well known (lec. 24)

The sequence of *E. coli* from strain O157:H7 (a pathogen) was also sequenced: 5.5 Mb (1632 new ORFs), with 4.1 Mb highly conserved. Some new genes present, some now absent, and various types of rearrangements observed! ~131 proteins thought to be connected to pathogenicity or virulence.

Other microbial genomes

(Bacteria, Archaea, Eukarya) (Don't memorize this list: for illustration only)

1995: the field of genomics began!

Haemophilus influenzae: First bacterial genome, pathogen, ~1700 ORFs

Mycoplasma genitalium: STD, smallest genome, 470 ORFs

1996

Synechocystis sp.: Cyanobacterium (phototroph)

Methanococcus jannaschii: First archaeal genome and first methanogen

Mycoplasma pneumoniae: Pathogen

1997

Saccharomyces cerevisiae: First eukaryote, 12,100,000 bp, 6034 ORFs; only 2 ½ times the number made by *E. coli*!

Helicobacter pylori: Cause of ulcers

Escherichia coli:

Archaeoglobus fulgidus

Bacillus subtilis: Important in food industry, spore former

Methanobacterium thermoautotrophicum: **60°C**

Borrelia burgdorferi: Cause of Lyme disease, linear chromosome

1998

Aquafex aeolicus

Pyrococcus horikoshii: Extreme thermophile

Mycobacterium tuberculosis: Tuberculosis

Treponema pallidum: Syphillus

Rickettsia prowazekii

Chlamydia trachomatis STD

1999

Helicobacter pylori: Second strain

Aeropyrum pernix

Chlamydia pneumoniae

Thermotoga maritime

Deinococcus radiodurans: Radiation resistant

2000

Campylobacter jejuni

Chlamydia trachomatis

Chlamydia pneumoniae

Neisseria meningitidis: Recall deaths at MSU

Neisseria meningitidis: (another strain)

Vibrio cholerae: Cholera

Xylelia fastidiosa

Pseudomonas aeruginosa: Bioremediation (40 cmpds); Cystic fibrosis (lungs); burn patients

Arabidopsis thaliana: First plant

More recently

Homosapiens: 3,300,000,000 bp (=900 books); 30-35,000 ORFs; 23 linear chromosomes

Many more microbes

For a complete compilation of known sequences, see
http://www.tigr.org
The Institute of Genomic Research

General comments about prokaryotic genomes:

95 completed prokaryotic genomes as of 2/25/03 (79 Bacteria and 16 Archaea)

~200 genomes in progress (not counting private industries)

Focus on pathogens, thermophiles, and microbes with unique physiology (phototrophs, alkalophiles, halophiles)

Can do entire genome in single day using a room of instruments!

General Comments about Eukaryal genomes:

Eukaryotic microbes: Giardia, Leishmania, Trypanosoma, Plasmodium, Fungi, Algae

Higher organisms: Plants (rice, potatoe, *C. elegans* (worm), Zebrafish, etc