BIOC4004 - Industrial Biochemistry
Lecture 14 - Wed Feb 25, 2004

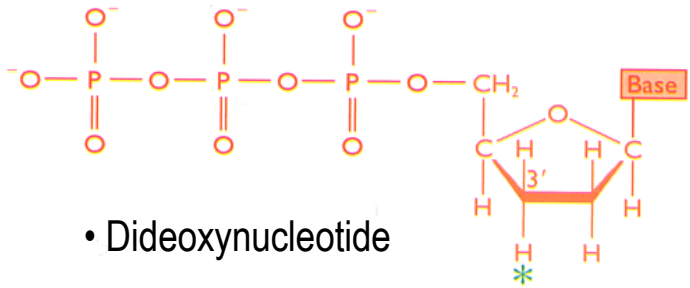
Topics for the Day:

- Genome projects
  - sequencing basics (again)
  - getting started
  - sequencing strategies
  - gaps and scaffolds
  - analysis and annotation

The Basics of Sequencing (revisited) Pt I

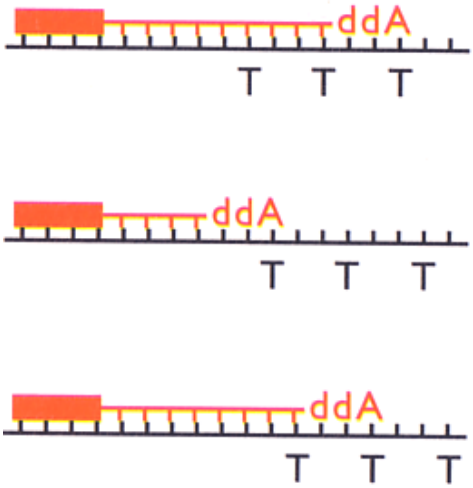
Sanger Chain-Termination Sequencing

• Primer annealing



• Dideoxynucleotide

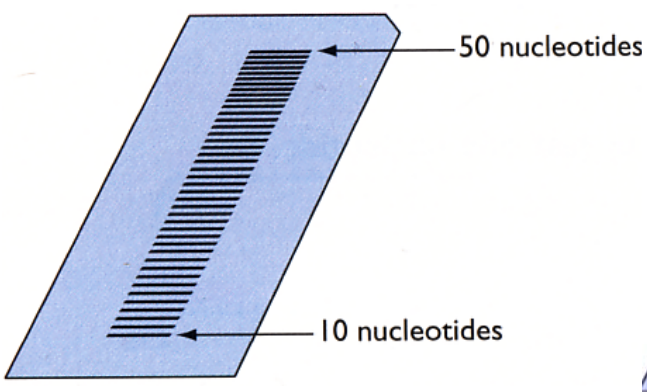
* Position where the -OH of a dNTP is replaced by -H

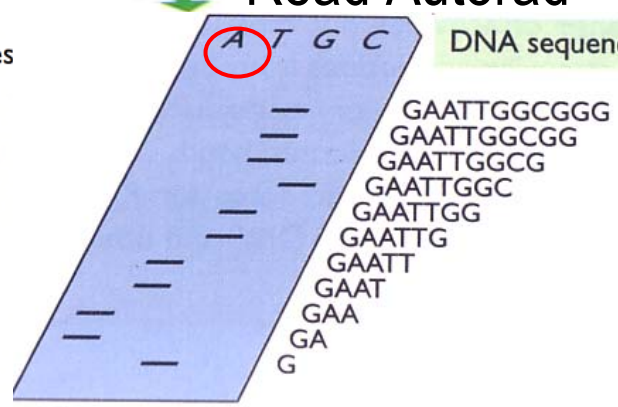


• Primer extensionusing ddA

The 'A' family

Read Autorad

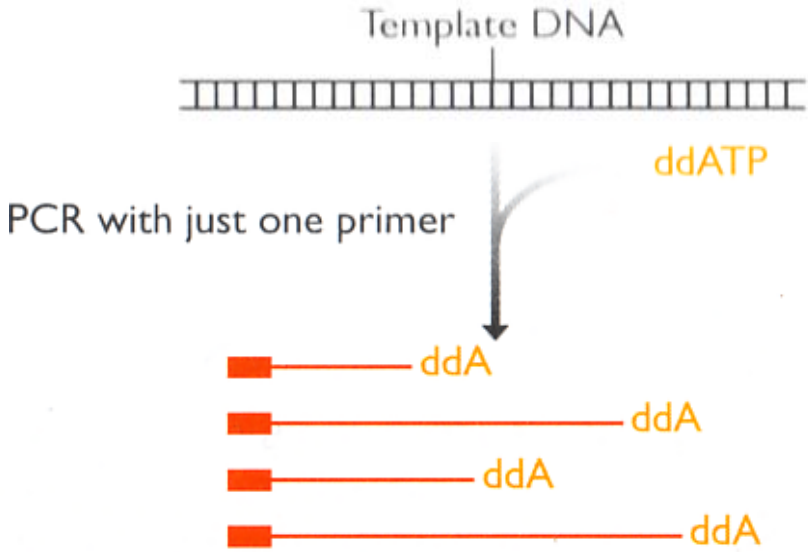




• PAGE separation(single nucleotide resolution)

The Basics of Sequencing (revisited) - Pt II

Cycle-Sequencing

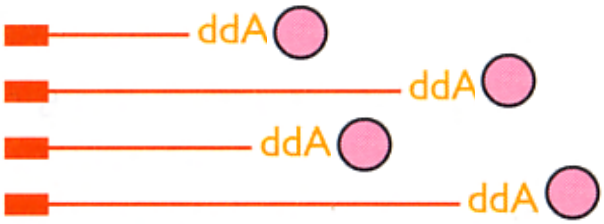


- based on "Asymmetric PCR" ie. one-sided PCR

Advantages:

- uses a lot less template DNA per sequencing reaction
- generates increasing number of chain-terminated strands as more cycles are performed

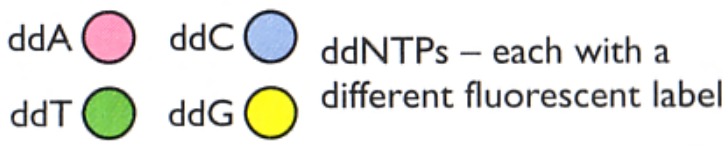
"Dye-deoxy" Automated Sequencing

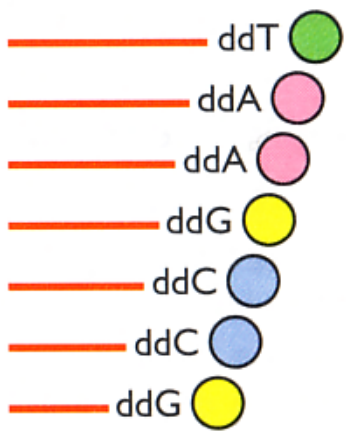


- extension with labeled terminators

Sequencing reactions, fractionation of products

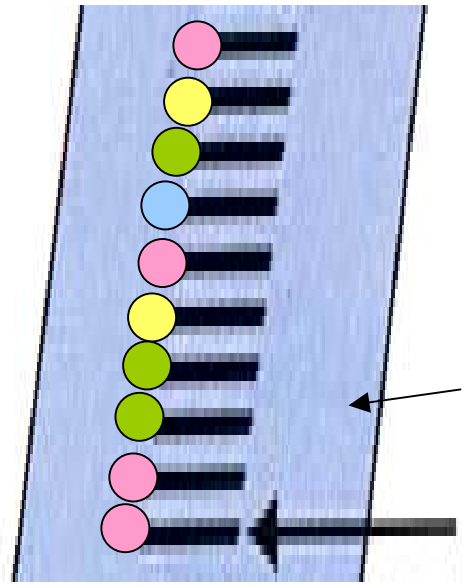
- all 4 terminators in 1 tube





The Basics of Sequencing (revisited) - Pt III

"Old-fashioned" Automated Sequencing (CIRCA late 80s-mid 90s)

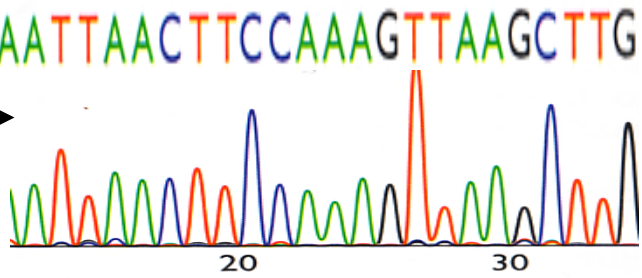


ddA ddC
ddT ddG

Separation on slab-PAGE

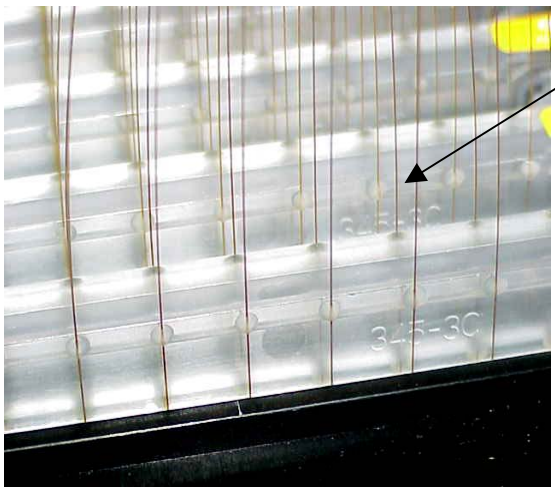
Laser/Fluorescence Detection

Automated data collection →



"MEGABASE" Automated Sequencing (CIRCA now)

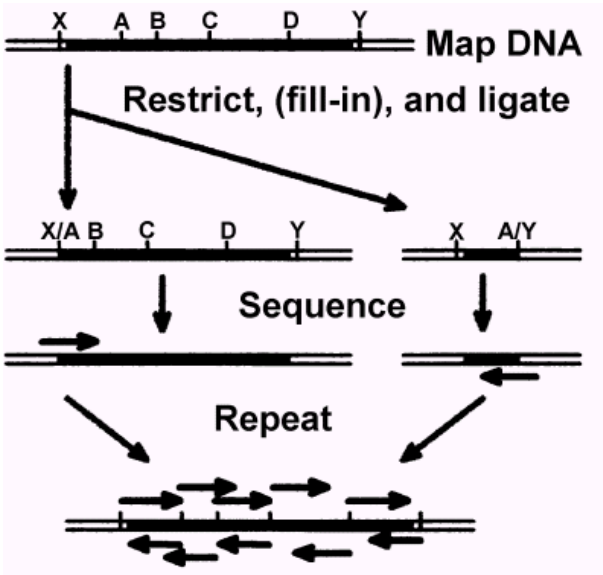
- rather than a PA slab gel, use capillary electrophoresis



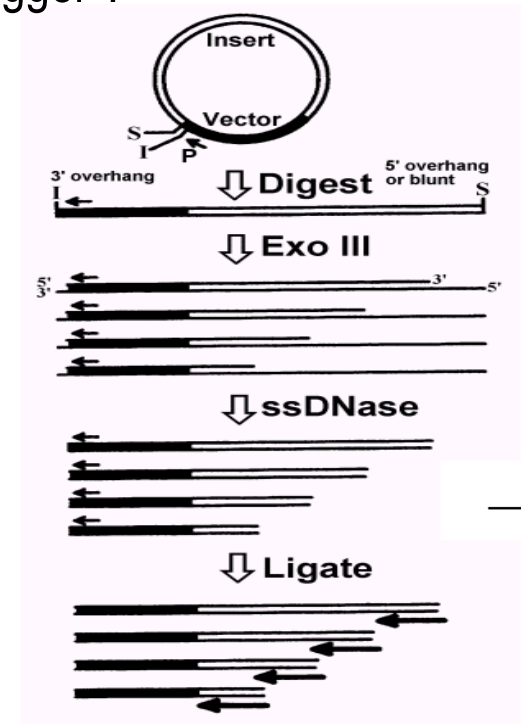
- more bp per read (~ 1 Kb)
- no messy gels to pour
- 96 samples at a time (~ 3.5 hrs)
- 96 new samples without new gel
- 6 runs/day x 1 Kb/sample x 96 samples/run = ~ 600 Kb / day

DNA Sequencing Strategies

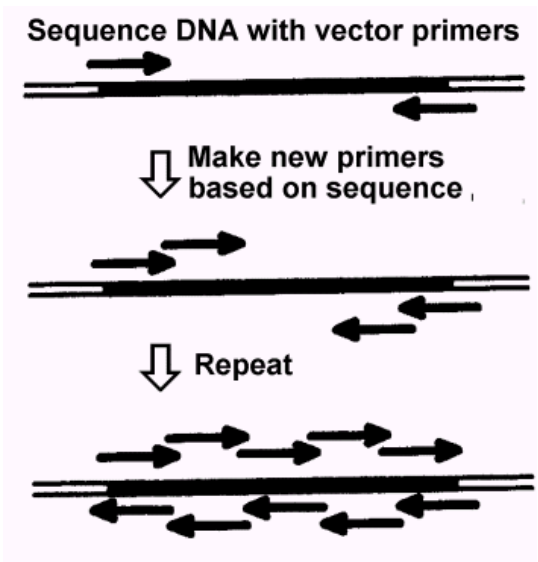
- From one primer / template combo: 500 - 700 bp of sequence
- What to do if you have something bigger ?



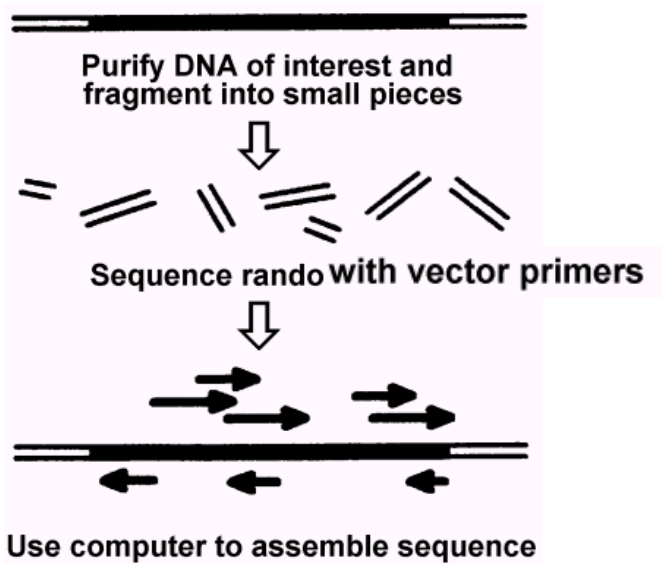
Subcloning



Nested Deletions



Primer Walking



Shotgun Sequencing

Large Scale Sequencing & Genome Projects

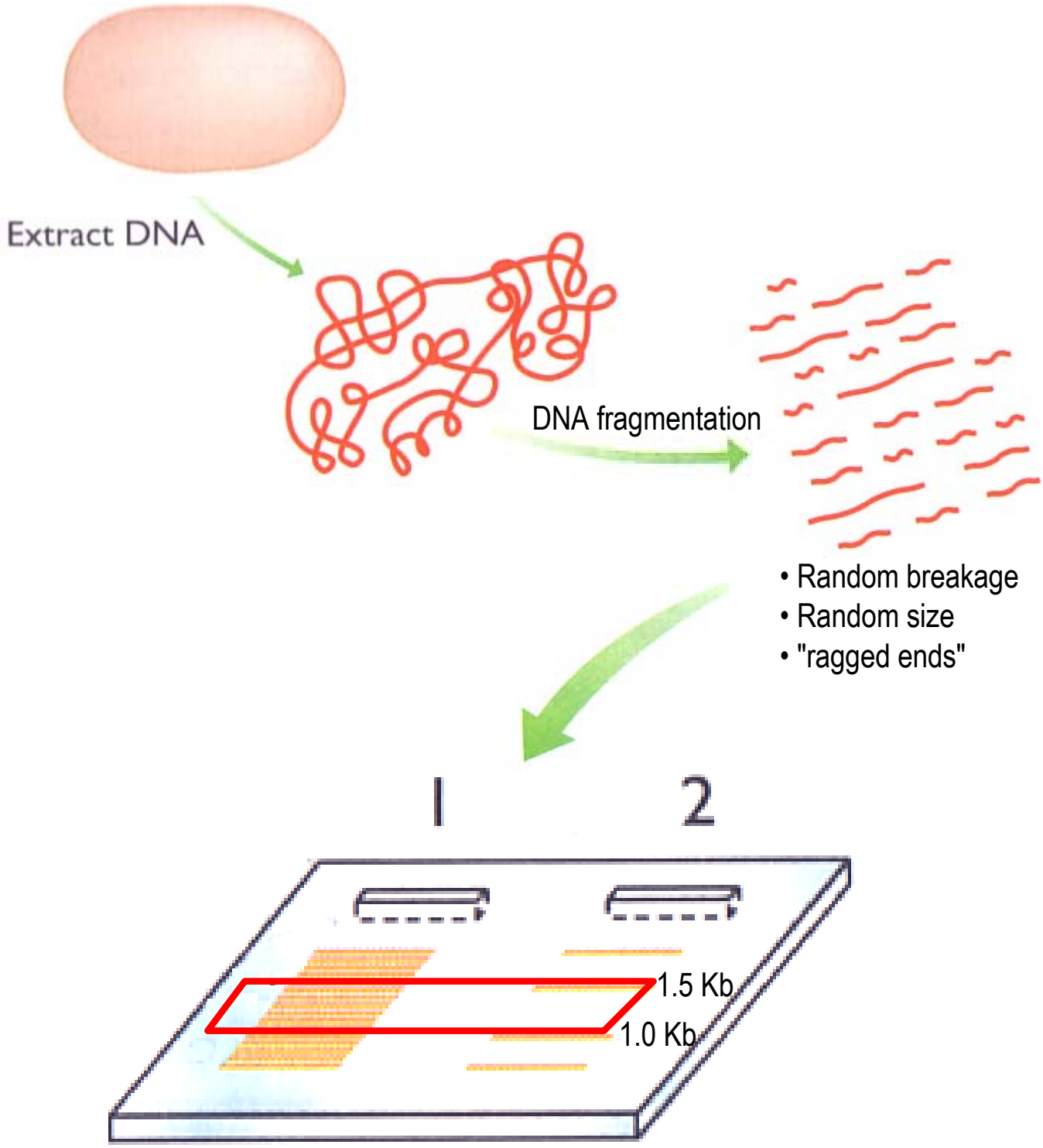
Extended Sequencing Strategies

Method	Advantages	Disadvantages
Restriction Digestion and Subcloning	• preparation simple • contig assembly easy	• relies on convenient restriction sites
Primer Walking	• preparation simple	• slow and expensive
Nested Deletions	• contig assembly easy	• laborious to generate
Shotgun	• quick and easily automated	• gaps and inefficient

A quantum leap in genome sequencing:
The shotgun sequencing method

- Proof of concept: *H. influenzae* (1995) done by TIGR
- Sequence a zillion clones from a random genomic library
    - small insert, sheared DNA to avg. size ~ 1.5 Kb
- Initially thought to be applicable only to bacterial projects
    - small genomes
    - relatively uncomplicated
- Has been applied to large sequencing projects
    - subclone the genome into smaller bits into large-insert vectors
        - YACs, BACs, Cosmids
    - sequence each clone by shotgun approach
- 6-8 fold coverage will get you > 95% of the genome sequence
    - ie. for a 2 Mb genome, sequence 12 to 16 Mb
- Gap closure is required to fill in the last little bits

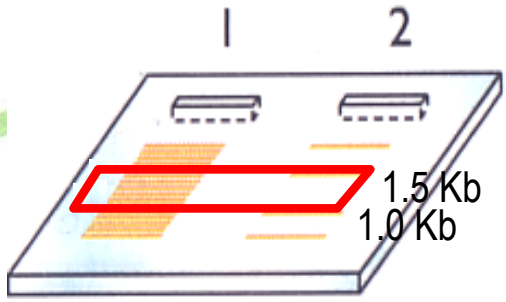
Shotgun Sequencing (Pt I)



Size fractionation

Shotgun Sequencing (Pt II)

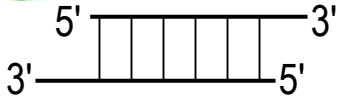
DNA fragments: ~ 1.0 to 1.5 Kb

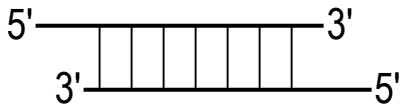


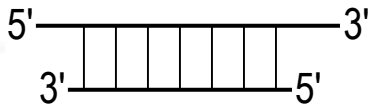
Size Fractionation:

- 1.0 to 1.5 Kb fragments
- cut-out of gel and purify

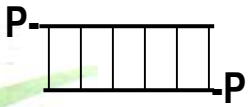
"Ragged Ends"

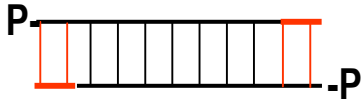


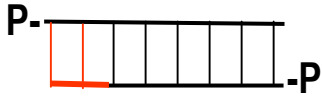




"End-Repair":
- DNA polymerase + dNTPs
- PNK + ATP

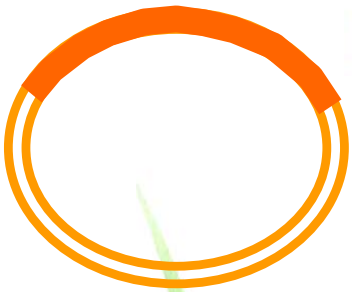




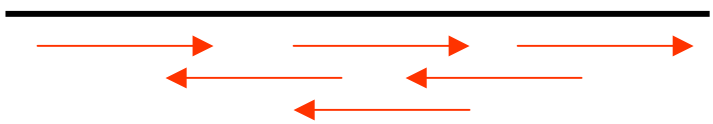


"Ready to Clone"

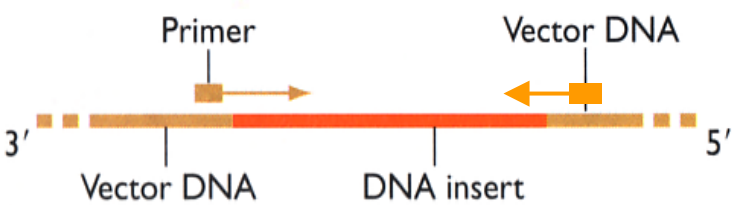
Blunt-end Clone into Vector



End-sequence
- vector primers



- "contig assembly"
- sequence polishing



Genome sequencing of *Actinobacillus pleuropneumoniae* (App)

- Small, Gram-negative, capsulated rod
- Serotyping based on capsular antigens
  - L20 is serotype 5b
- Causes severe lung disease and death in pigs
  - Highly contagious pleuropneumonia
  - Crowding increases risk
- Outbreaks of infection lead to severe losses to the food pig industry
- Closely related to other Pasteurellaceae (*H. influenzae*, *A. actinomycetemcomitans*, *P. multocida, M. haemolytica*)

What we knew about the App genome:

- ~ 2.4 x 10^6 base pairs
- Single, circular chromosome

Stages to sequencing the genome:

- Generation of a shotgun clone library
  - 25,000 clones of 1.5 – 2.0 kb inserts
- Verification of insert integrity
- ~36,000 sequence reads (7.5-fold coverage)
  - At IMB (NRC sequencing facility in Halifax)
- Sequence assembly (automated stages)
- Manual polishing
- Gap filling using large-insert clones
- Annotation

Genome sequencing of App

Shotgun Library Generation

- Vector: pTrueBlue (GenomicsOne)
  - *SmaI* (blunt) cut, SAP treated
- Insert: sheared, size-fractionated DNA
  - Use an asthma nebulizer
  - End repair DNA
- Ligate, electro-transform, and plate: **50,000 clones obtained**

Shotgun Library Sequencing

- 25,000 clones were shipped to IMB for sequencing
- Raw sequence (trace) data was exported to our sequence-assembly computers where the data was automatically processed
  - Removal of vector contamination
  - Base-calling confirmed (PHRED)
  - Raw reads assembled into *contigs* (PHRAP)
  - Reports generated

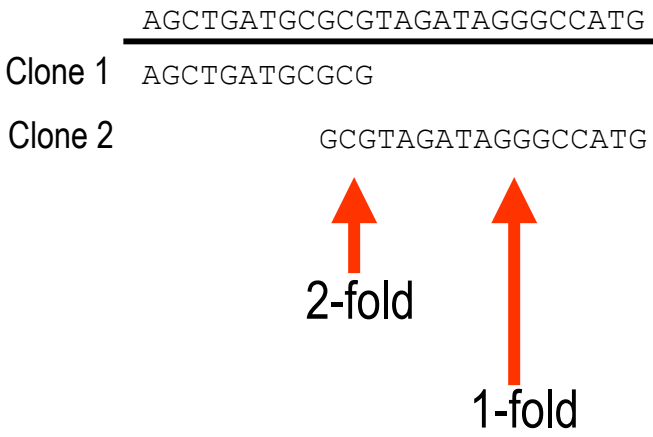
Total number of sequence reads:	44,936
Total number of fails:	8,669 (19.3%)
Total bases Sequenced:	~ 22 Mb (~18 Mb good sequence)
Estimated coverage after removal of fails:	**7.6X (36,267 reads)**
Number of auto-generated contigs:	205
Single-coverage from contigs:	2,142,385 bp
Number of single-reads:	1,692
Total single-coverage:	2,299,567 bp

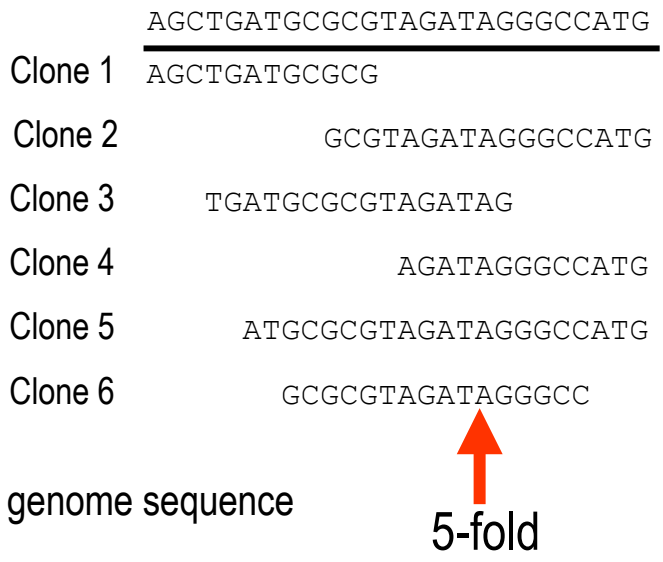
- 96% single coverage assuming a genome size of 2.4 x 10^6 bp

Gap-Closure and Scaffolding (Pt I)

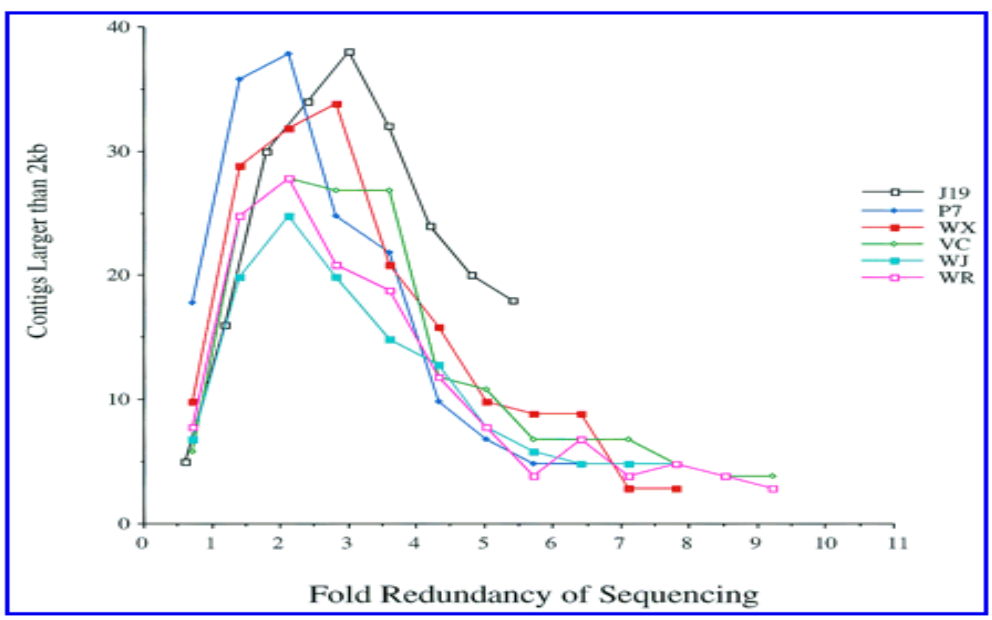
The Concept of FOLD COVERAGE

• The number of times (on average) that a base's sequence was obtained from different shotgun clones





- 6-8 fold coverage will get you > 95% of the genome sequence
  - ie. for a 2 Mb genome
    - 1-fold coverage: 2 Mb worth of sequence --> 2,000 1-Kb clones
    - 8-fold coverage: 16 Mb worth of sequence--> 16,000 1-Kb clones



- as more clones are sequenced, contigs are made
- as yet more sequencing occurs, contigs start linking
- The Final Goal : 1 huge contig

Gap-Closure and Scaffolding (Pt I)

The Problem With Shotgun Cloning

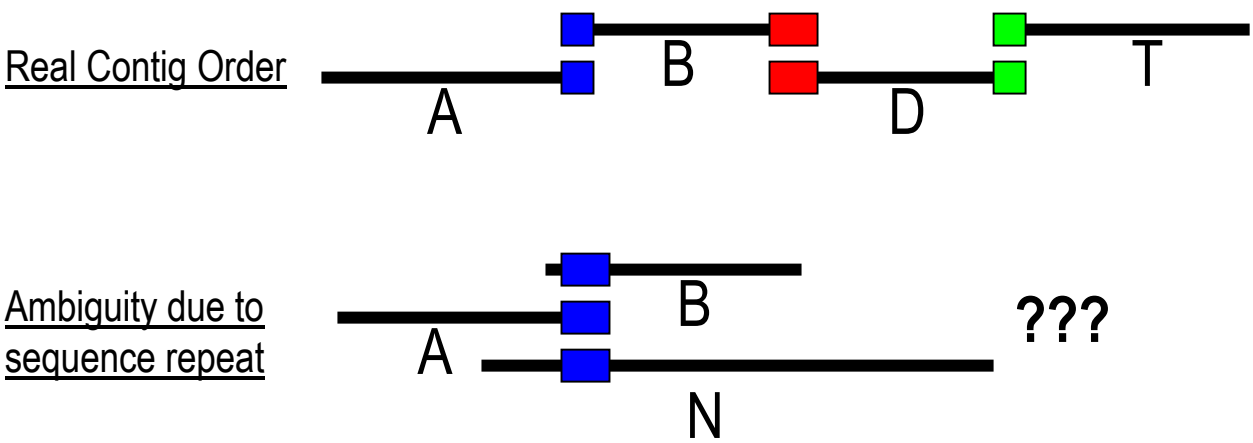
- Practical size limitation of ~ 5 Mb
  - for larger genomes, break up into smaller chunks in high capacity vectors
  - shotgun sequence these large insert clones

- Easy to get the first 90% of the genome

- For each new clone:
  - 90% chance that the sequence will be redundant
  - 10% chance that the sequence will be "new"
  - Becomes less and less efficient as you continue the effort
  - DIMINISHING RETURNS !!!!

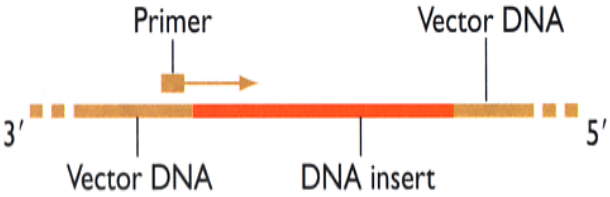
- Plus: random libraries are **never** truly random

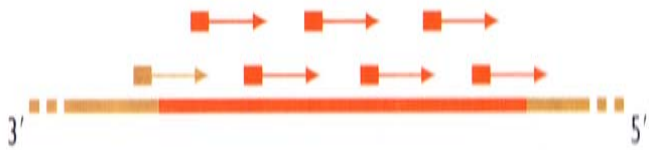
- Also: DNA repeats can throw a monkey-wrench on sequence assembly !
  - Duplicated genes (functional and pseudogenes)
  - Transposons and other mobile elements
  - "Junk DNA"



Gap-Closure and Scaffolding (Pt II)

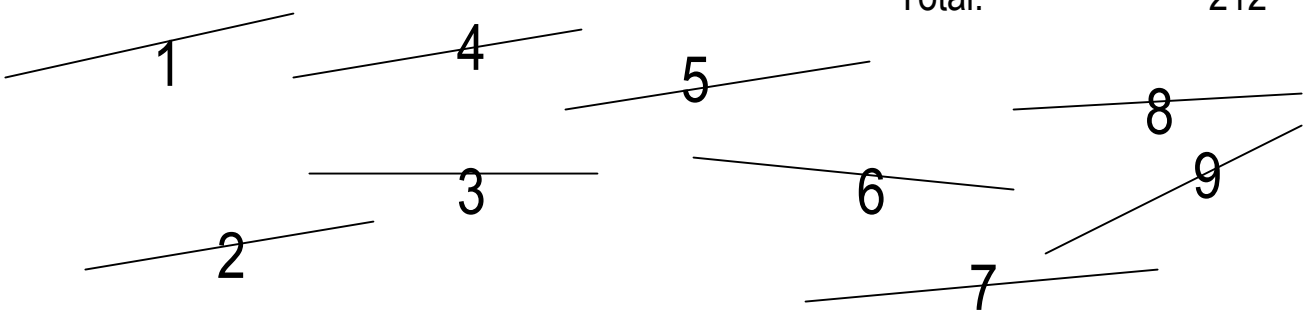
- Need a change in Strategy:
  - switch to more "directed" sequencing approach
    - lower throughput
    - more expensive
    - useful for gap-filling





- By the end of the shotgun phase:
  - Large-ish contigs
  - Gaps remain
  - No contig order:
    don't know which contig links to which !!!

Contig size (kb)	No:
1 - 2 kb	32
2 - 5 kb	54
5 - 10 kb	54
10 - 15 kb	22
15 - 20 kb	19
20 - 50 kb	31
Total:	212



- Gap-closure: sequencing across the gaps and linking contigs
- Scaffolding used to determine Contig Order
- The Result: A fully-assembled complete genome

- **200 contigs can be assembled in (200!) different ways !!!!!!!**

Physical Gaps vs. Sequencing Gaps (Pt.I)

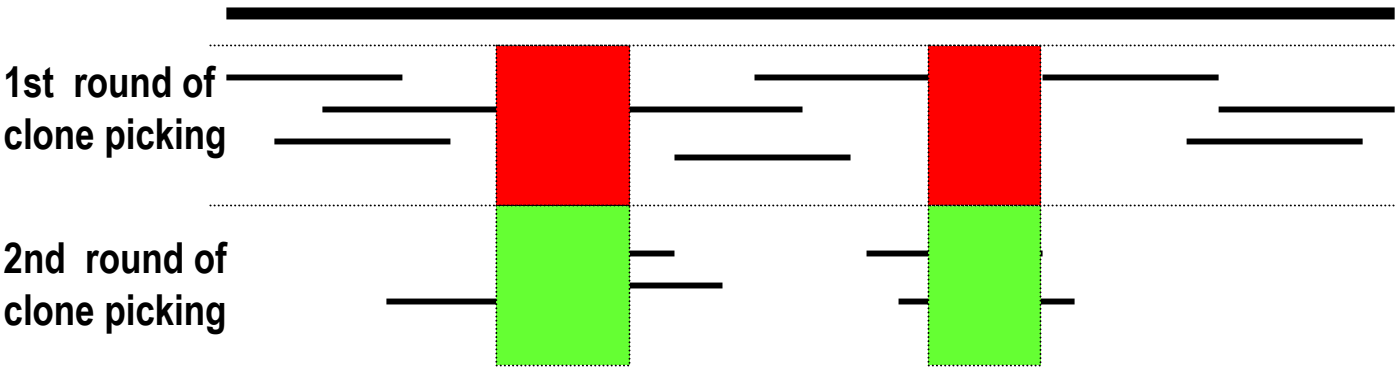
• libraries for genome sequencing are like buying puzzle pieces in bulk
  • you can buy 5000 pieces of a 1000 piece puzzle and still miss some pieces

Physical Gaps (ie. "Real Gaps")

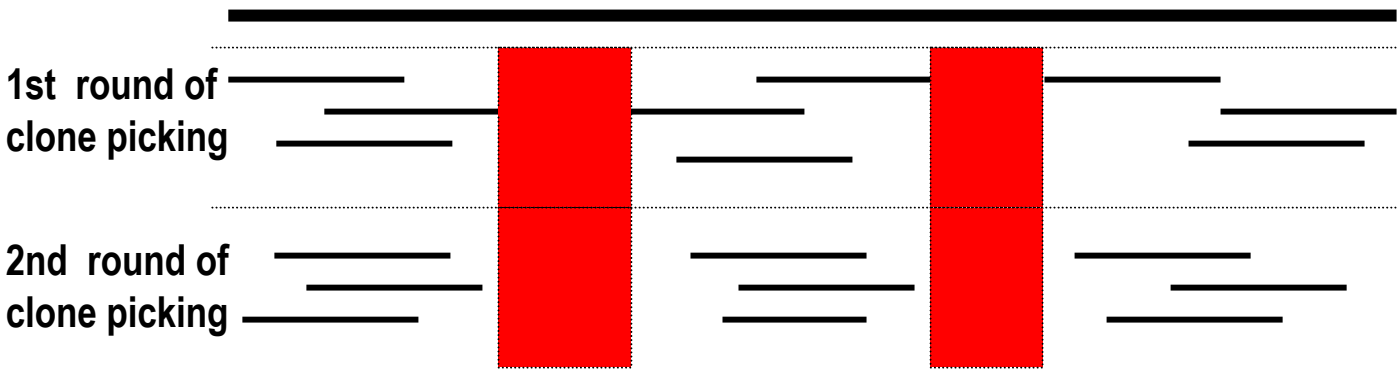
• physical gaps result from library construction, which is (supposedly) a random process
• certain areas of the genome may not get represented

Two Possibilities:

• The library contains every bit of the genome, you just haven't found it all yet - a sampling issue - pick more clones !!!)
ie. the pieces are all there but you haven't looked at the bottom of the box



• The library may not contain the missing regions (no matter how many clones you pick, you'll never get 100% of the genome represented by the library) ie. your puzzle is missing some pieces



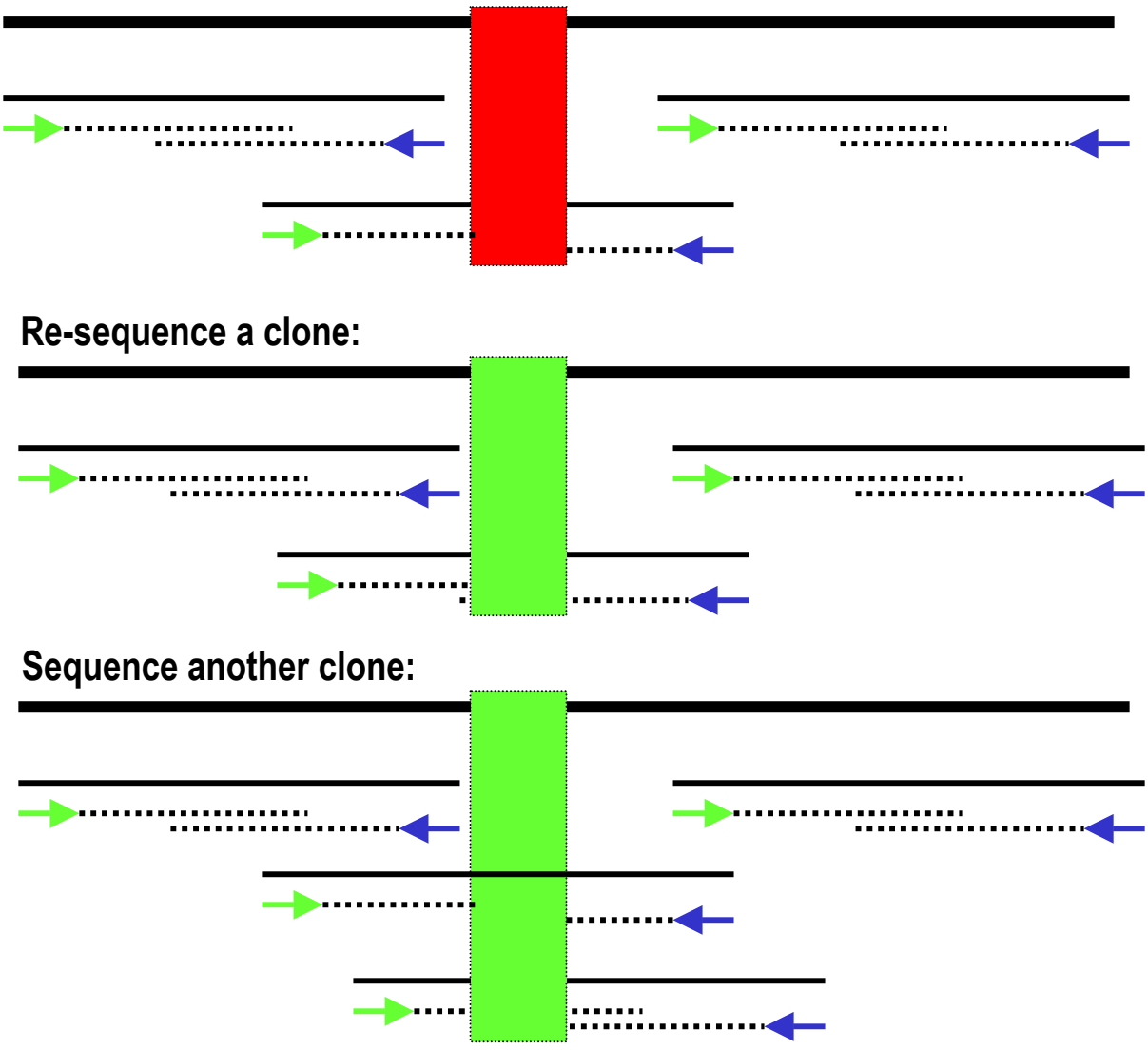
• Physical gaps are bad news; must find a way to clone the missing bits !!!
• Physical gaps are **EXPECTED**; more than one library is usually made

Physical Gaps vs. Sequencing Gaps (Pt.II)

Sequencing Gaps (ie. "Fake Gaps")

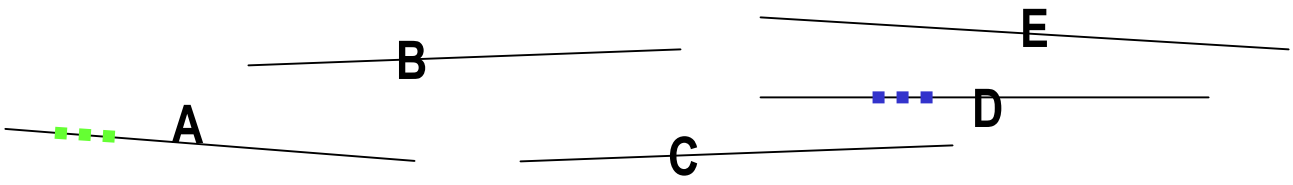
• Sequencing gaps result from non-overlap of sequencing runs which **should** overlap

• If there are no physical gaps in your library and you still have gaps after sequence assembly, there **must** be sequencing gaps



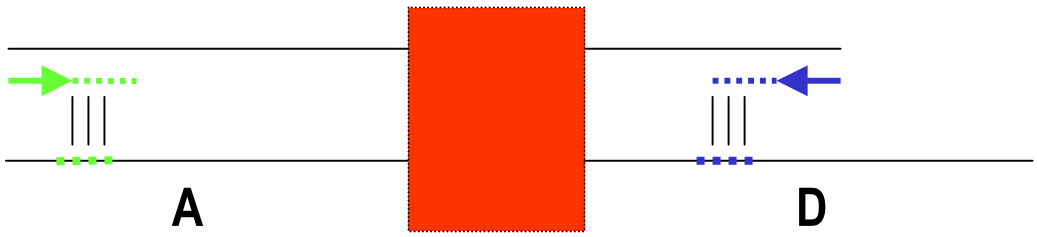
• Sequencing gaps are a nuisance, but can be dealt with because at least the gaps are "contained" in the library (**you know all the pieces in the puzzle are there**)

Scaffolding using cosmid/fosmids (or "dealing with Physical Gaps")

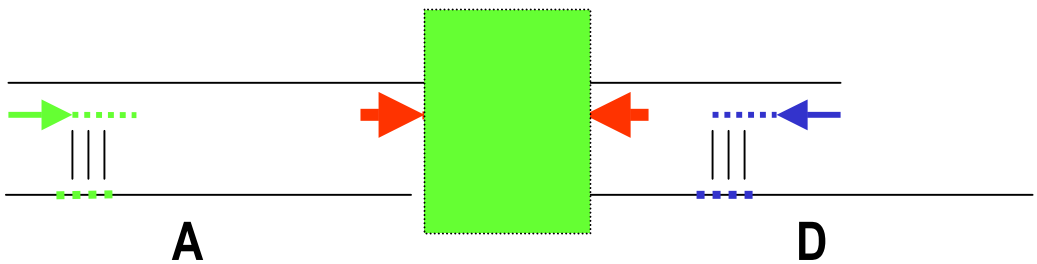


- contigs are not physical entities, they are "constructs" made from sequence data
- clone inserts are actual physical entities

Cosmid Insert



- if cosmid end-sequences match regions in two different contigs, the contigs **have to be adjacent to each other**



new primer flanking gap

- To close the gap, make new primers, sequence off the cosmid

- Gap closure (physical or sequencing gaps) is always conceptually the same:
  - **Sequencing Gaps:**
    - contained within a single (or a handful of shotgun clones)
  - **Physical Gaps:**
    - contained within a clone from a different library
    - people make extra libraries to cover all the bases