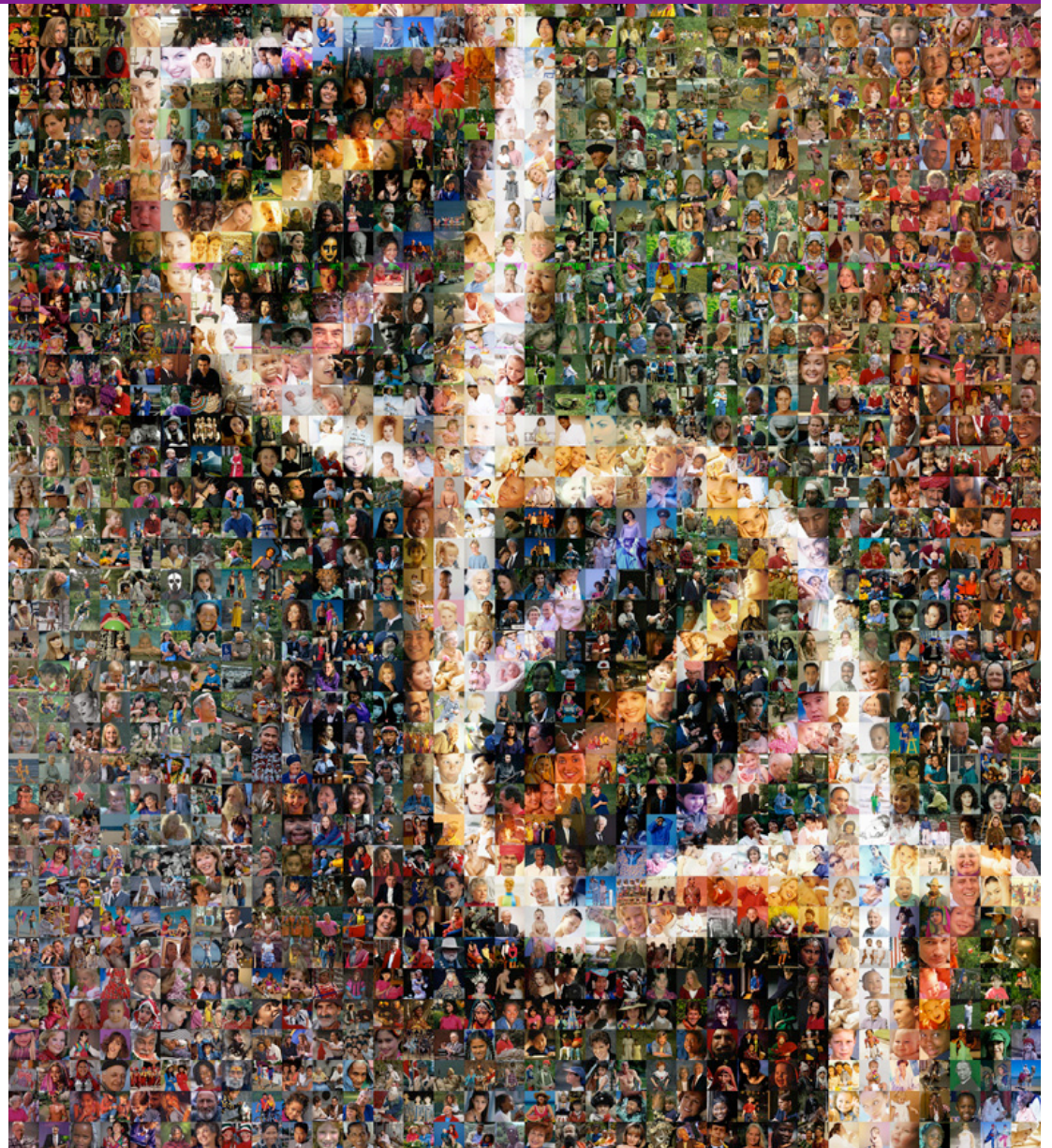


# Sequencing genomes

BRCA1 wrap-up

Cloning & sequencing

Sequencing genomes



Additional office hours for remainder of quarter:

Monday 1:30-2:30 p.m.

Tuesday 10:00-11:00 a.m.

Exam next Friday, February 26

Material through lecture of Monday, February 22  
and next week's Quiz Section (fly II)

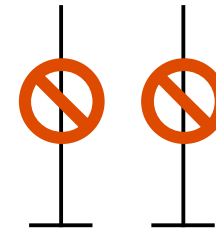
# Proto-oncogenes and tumor suppressor genes

---

In a diploid cell...

Tumor suppressor gene

Proto-oncogene



Cell growth and proliferation

---

Cancer-promoting form...

Expected behavior of allele

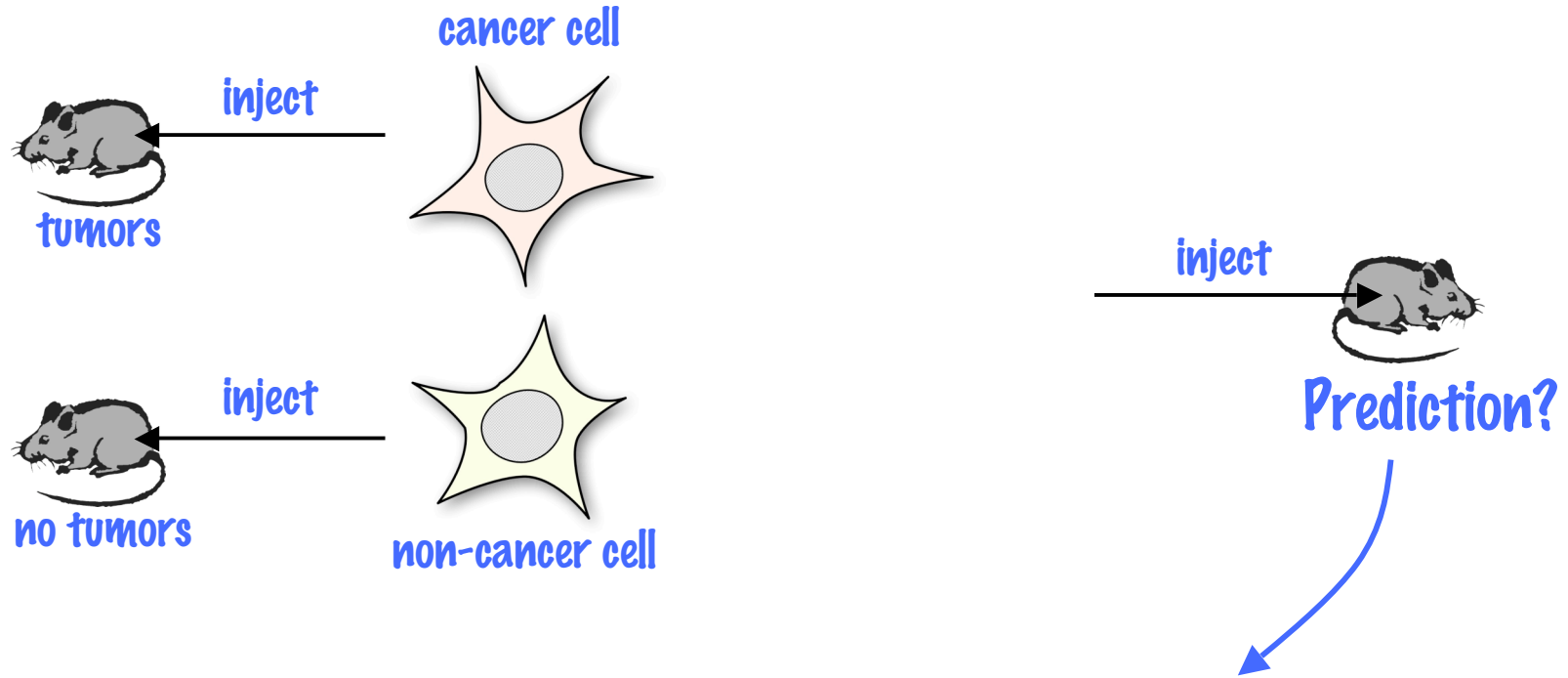
Oncogene

**Dominant**

Mutated tumor suppressor

**Recessive**

# Dominant/recessive behavior revealed by cell fusion



If cancer is due to oncogene: **tumors develop ✓**

If cancer is due to mutated tumor suppressor gene:

**no tumors ✓**

# Inherited cancer susceptibility: the paradox

---

At the cellular level:

A mutation in a tumor suppressor gene acts as a

**Recessive Mutation**

By genetic inheritance (pedigree):

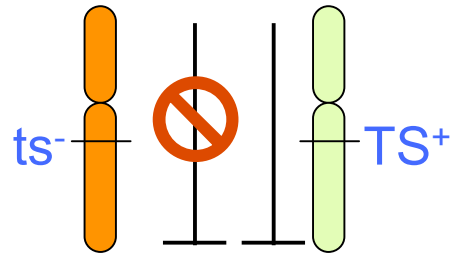
A mutation in a tumor suppressor gene acts as a

**Dominant Mutation**

# Inherited cancer susceptibility: “two-hit” hypothesis

---

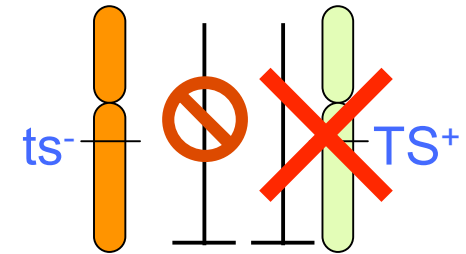
Tumor suppressor gene



Cell growth and proliferation

second “hit”  
→

Tumor suppressor gene

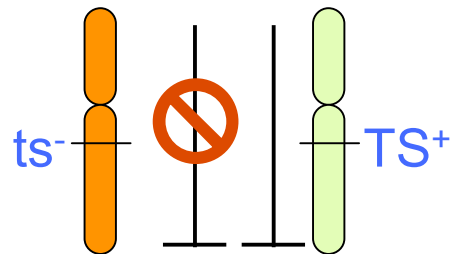


Cell growth and proliferation

# Inherited cancer susceptibility: “two-hit” hypothesis

---

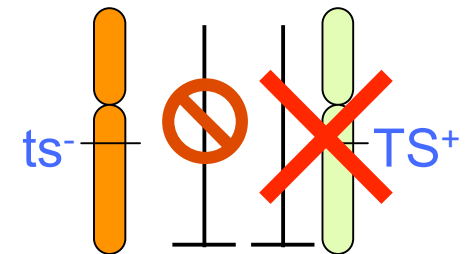
Tumor suppressor gene



Cell growth and proliferation

second “hit”  
→

Tumor suppressor gene



**Cell growth and proliferation**

# Molecular detection of the “second hit”

---

From quiz section





# Mechanisms for loss of the remaining good allele

---

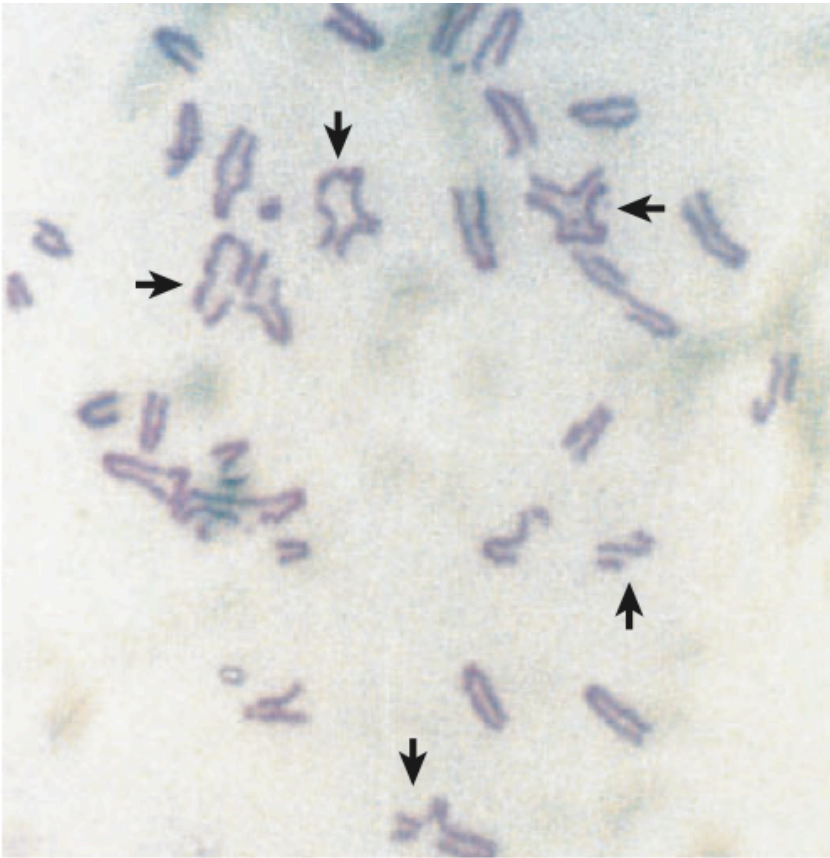
## Several possibilities...

- » **Non-disjunction during mitosis**
- » **Deletion/insertion**
- » **Point mutation**
- » Gene conversion (e.g., repair using homologue)
- » Mitotic recombination

## What is the normal function of BRCA1?

---

Targeted BRCA1 knockout in mice → chromosome instability



Implication?

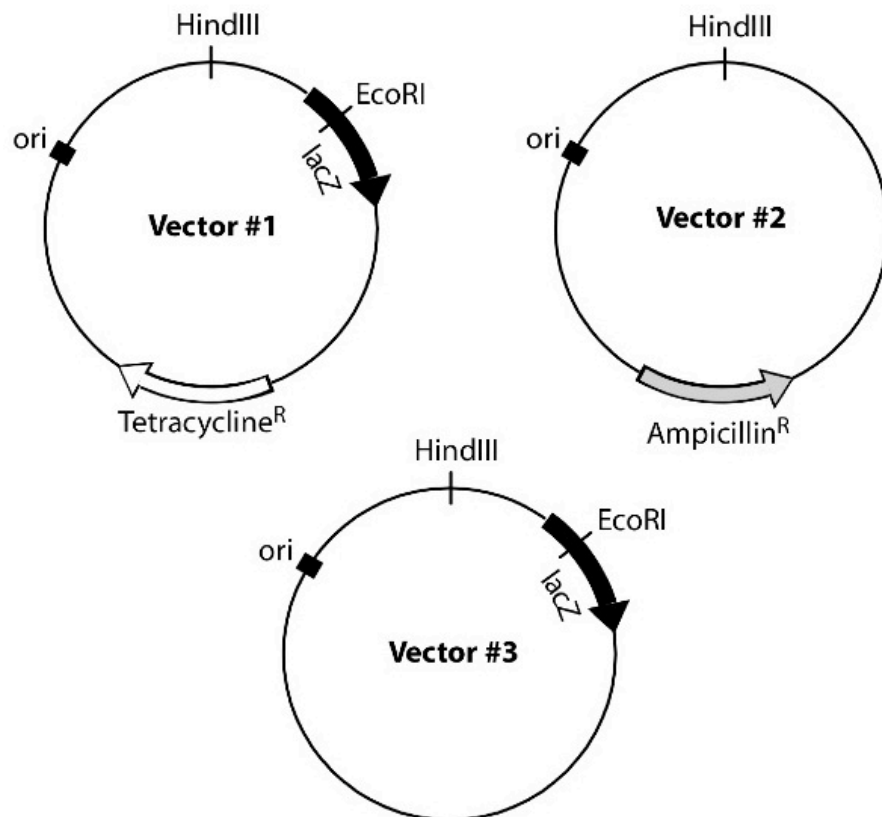
**Increased chromosome instability →**

**increased probability of other cancer-promoting abnormalities**

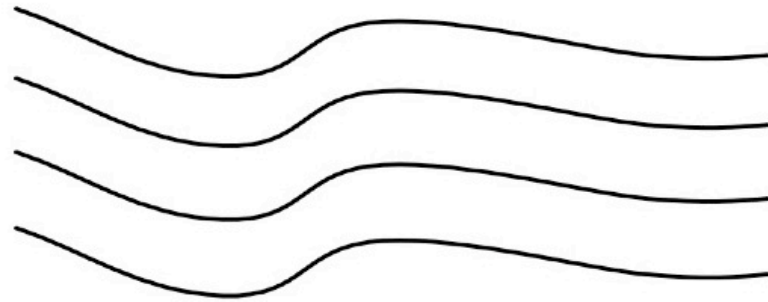
# Practice question

Your goal is to construct a mouse genome DNA library, choosing appropriate elements from the reagents/material shown below. Note: you are just asked to do the best you can given these constraints.

Possible vectors:



Mouse genomic DNA (assume that it has lots of EcoRI as well as HindIII cut sites)



**E. coli strains available:**

- ampicillin- and tetracycline-sensitive, lacZ<sup>+</sup>
- tetracycline-sensitive, lacZ<sup>-</sup>

**Plates available:**

- Ampicillin
- Ampicillin + X-gal
- X-gal

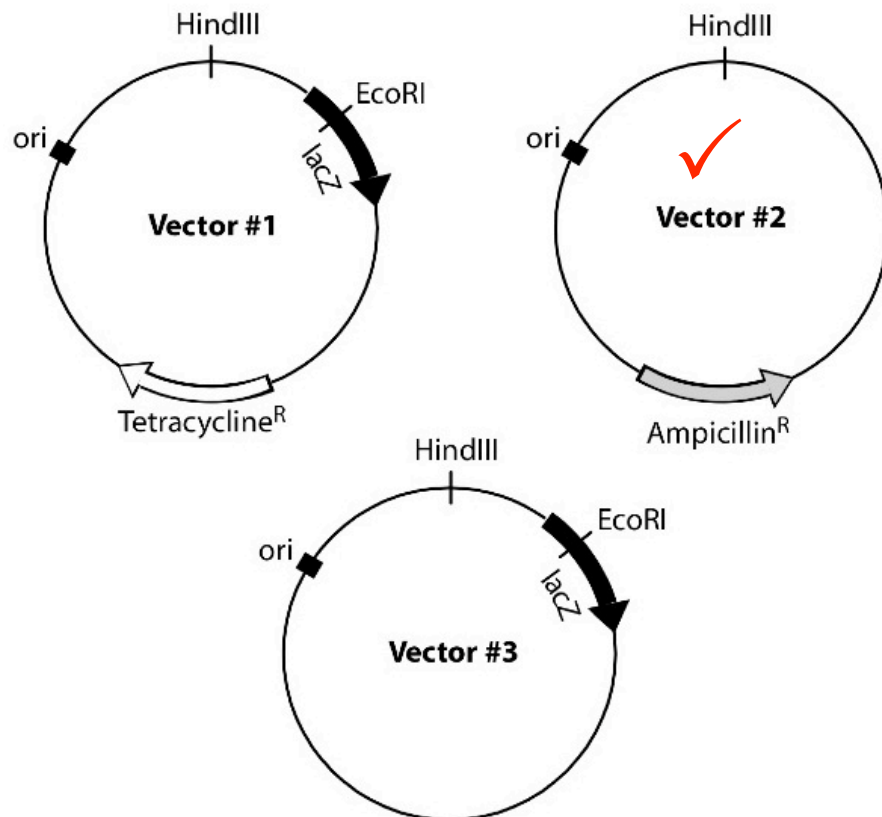
**Other reagents:**

EcoRI, HindIII, DNA ligase, CaCl<sub>2</sub>, appropriate reaction conditions

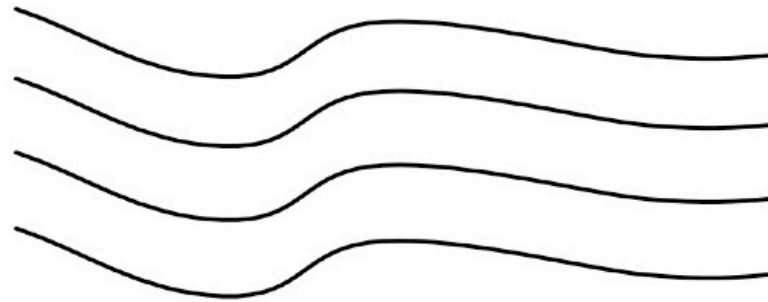
# Practice question

Your goal is to construct a mouse genome DNA library, choosing appropriate elements from the reagents/material shown below. Note: you are just asked to do the best you can given these constraints.

Possible vectors:



Mouse genomic DNA (assume that it has lots of EcoRI as well as HindIII cut sites)



**E. coli strains available:**

- ✓ ampicillin- and tetracycline-sensitive, lacZ<sup>+</sup>
- tetracycline-sensitive, lacZ<sup>-</sup>

**Plates available:**

- ✓ Ampicillin
- Ampicillin + X-gal
- X-gal

**Other reagents:**

EcoRI, HindIII, DNA ligase, CaCl<sub>2</sub>, appropriate reaction conditions

Outline how you would construct the genomic DNA library, specifying which of the parts list you would use. You don't have to use all the steps on the list below.

- Step 1: ..... **cut Vector #2 (complete) and mouse genomic DNA (partial) with HindIII** .....
- Step 2: ..... **mix and ligate** .....
- Step 3: ..... **transform ampicillin- and tetracycline-sensitive E. coli with the ligation mix** .....
- Step 4: ..... **plate on ampicillin-containing plates, recover colonies that grow** .....
- Step 5: .....
- Step 6: .....
- Step 7: .....
- Step 8: .....

## DNA (and RNA) hybridization

---

Detecting

Tagging

Amplifying

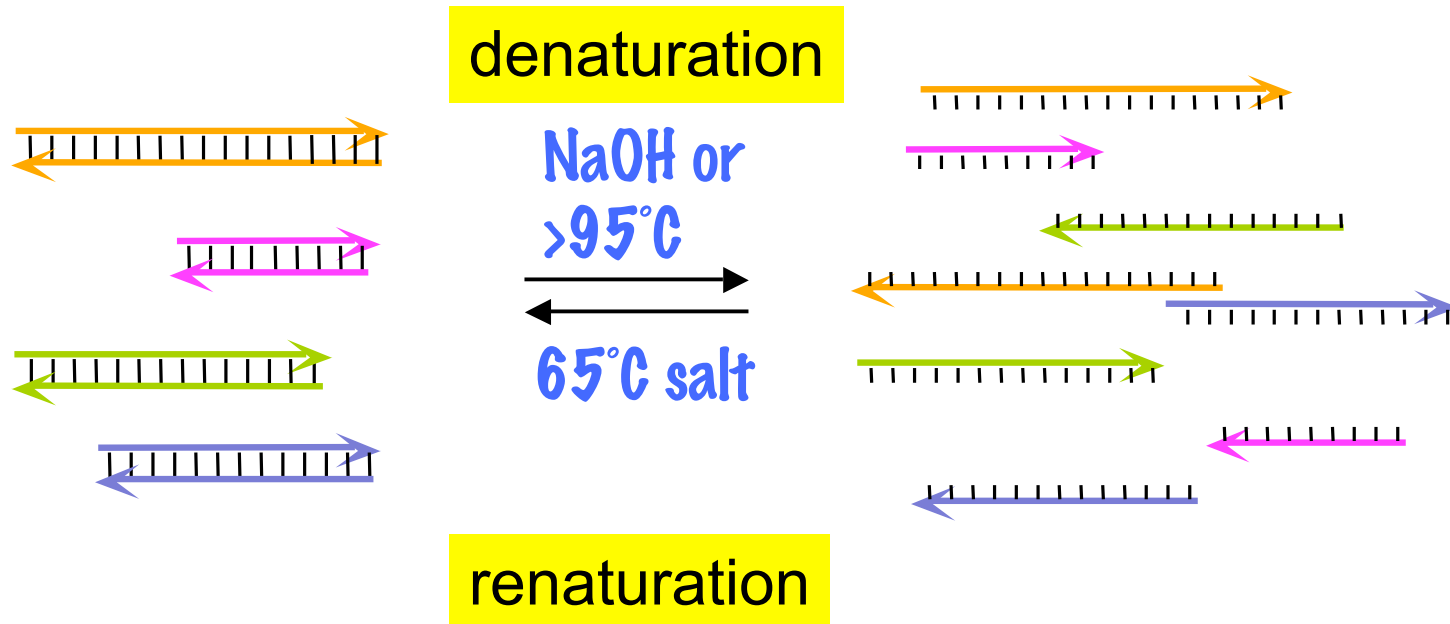
} specific DNA sequences using  
the rules of base-pairing

# DNA hybridization

---

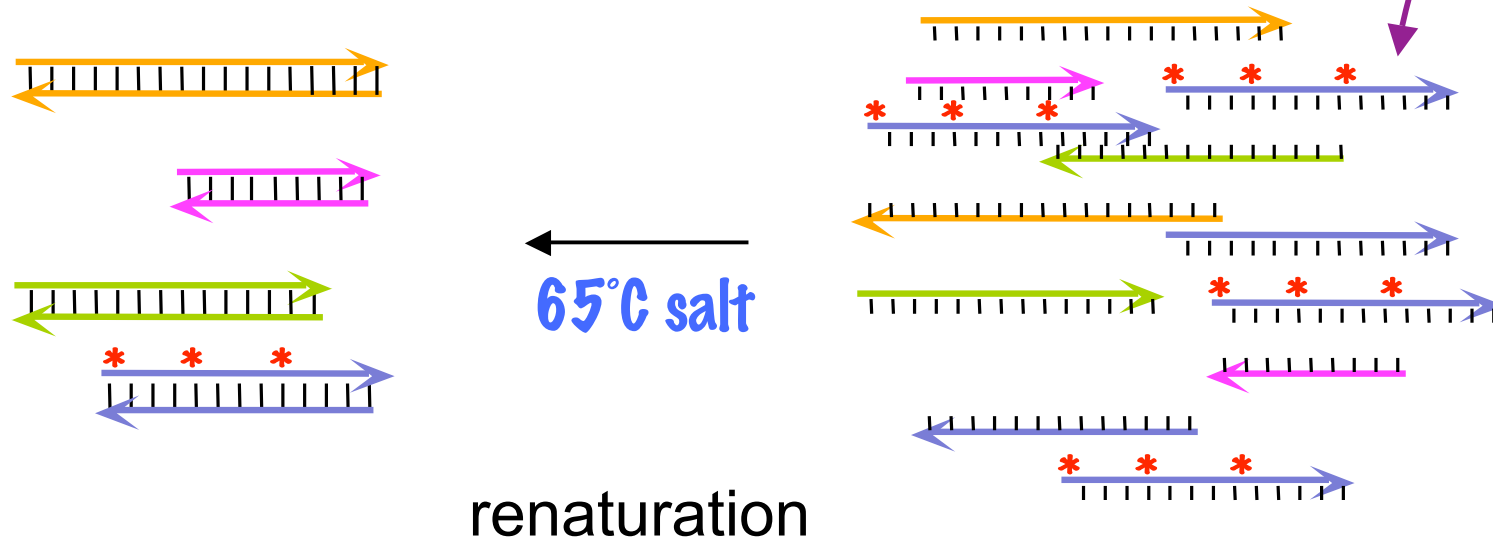
The two strands of double-stranded DNA will come apart (“denature” or “melt”) when heated or in alkali...

...but can re-nature due to base-pairing



base pairing provides the specificity for renaturation

# DNA hybridization with added "probe"



base pairing of labeled probe identifies the target DNA



# The polymerase chain reaction

---

Uses DNA polymerase

Makes unlimited quantities of the DNA of interest

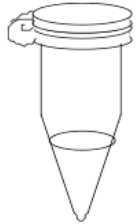
Only requires a single template molecule—very sensitive

Uses **two** DNA primers

<http://www.youtube.com/watch?v=x5yPkxCLads>

# Reagents for PCR

---



1. target DNA

*genomic DNA*

2. Two specific primers—  
“left” and “right”

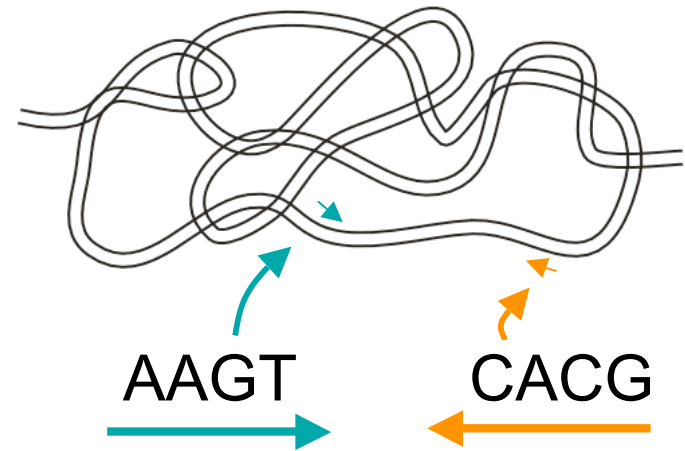
*complementary to sequences on either side of the region of interest*

3. All 4 dNTPs

*dATP, dCTP, dGTP, dTTP*

4. DNA polymerase from the heat-loving bacterium,  
*Thermus aquaticus*

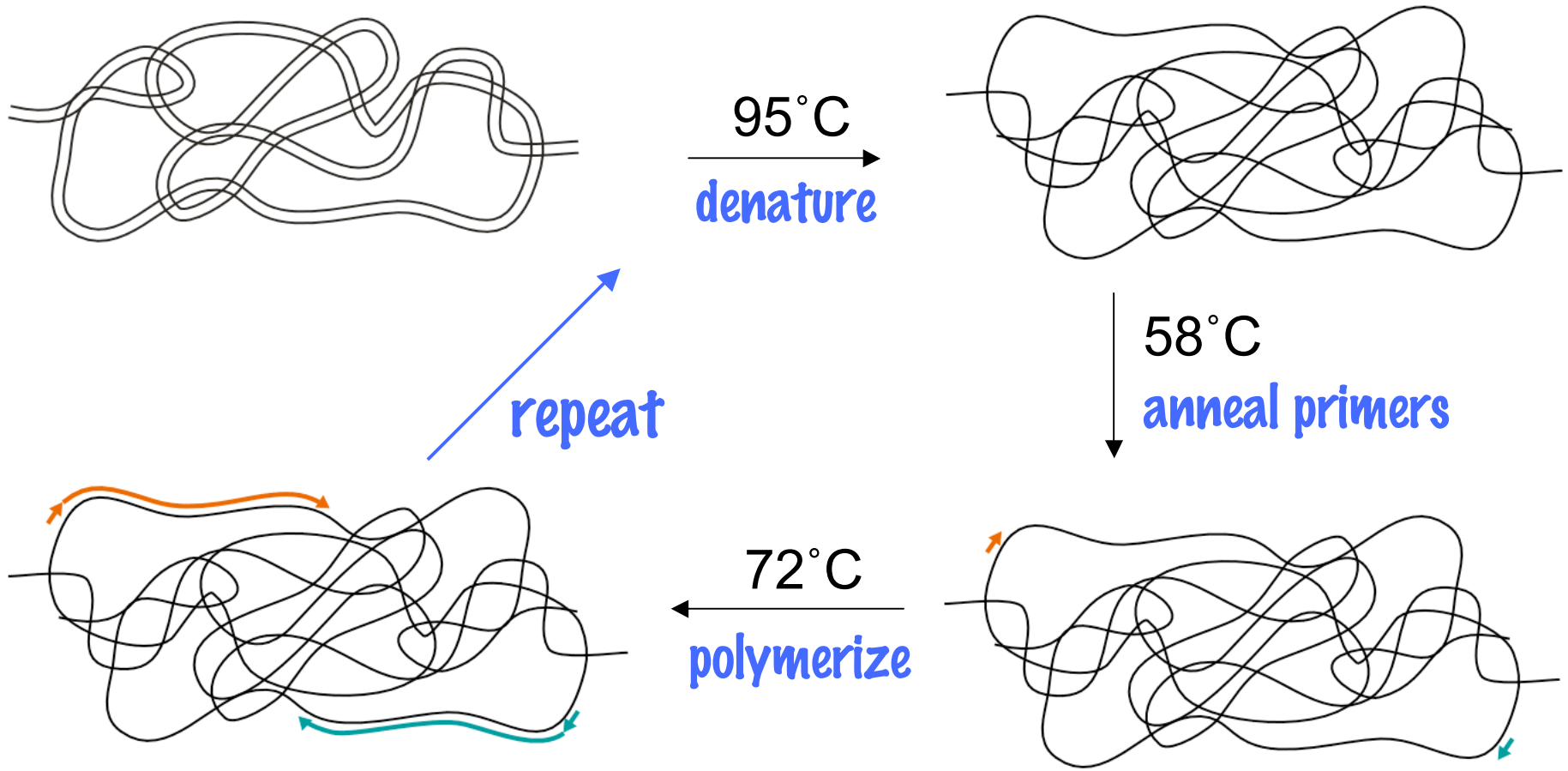
*doesn't die at 95°C!*



# PCR is performed at high temperatures

---

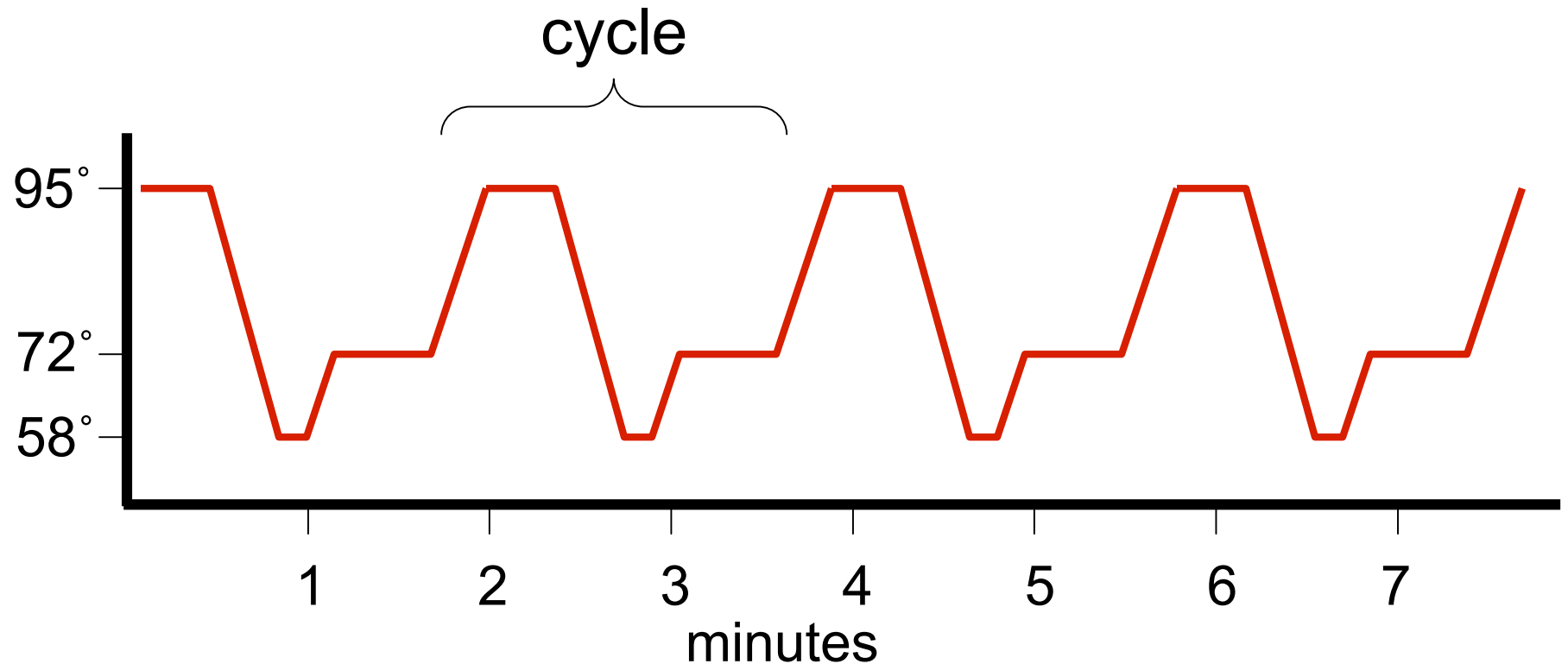
First cycle of synthesis



## Temperature cycles for PCR

---

A PCR machine can quickly change temperatures

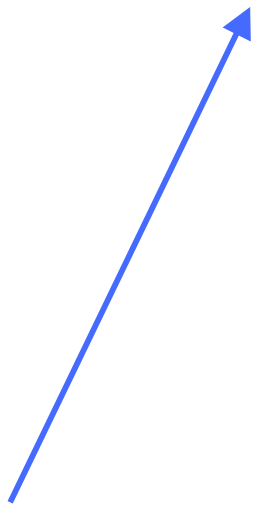
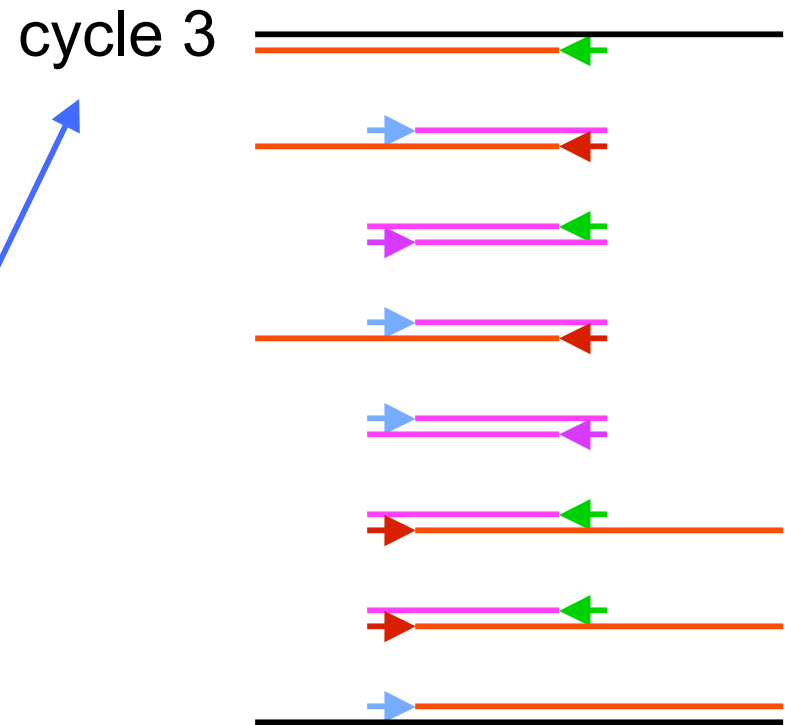
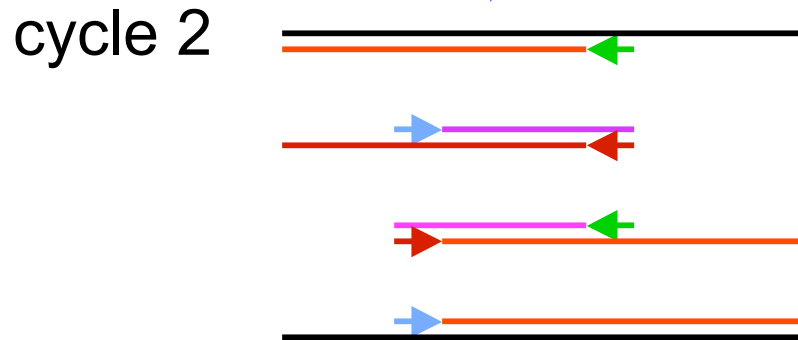
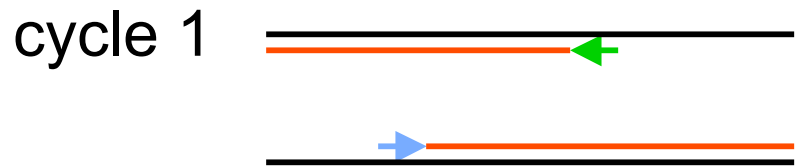
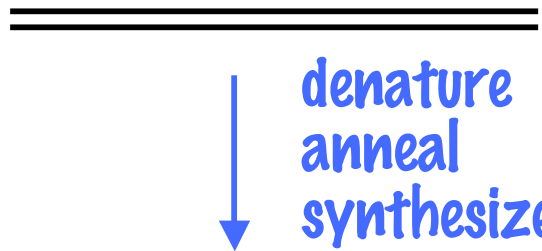


The DNA polymerase from *T. aquaticus* is resistant to high temperatures.

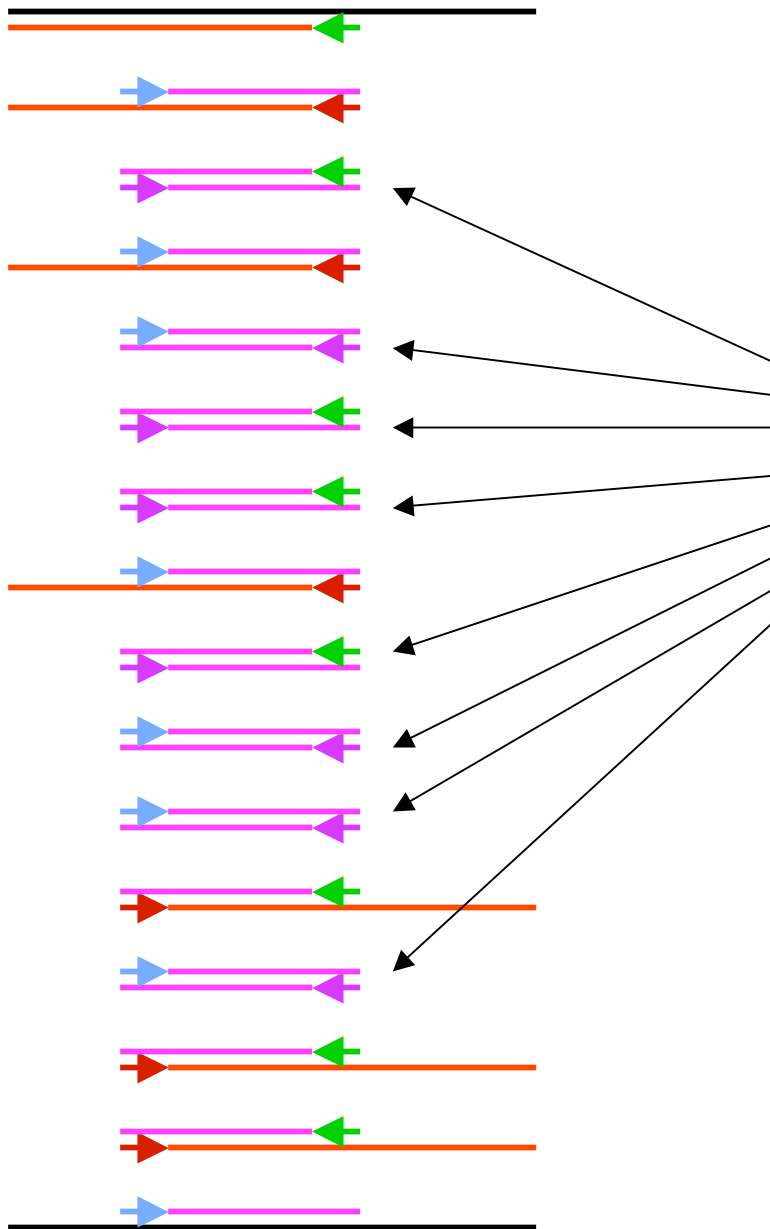
# DNA products at successive cycles of PCR

---

starting dsDNA



cycle 4



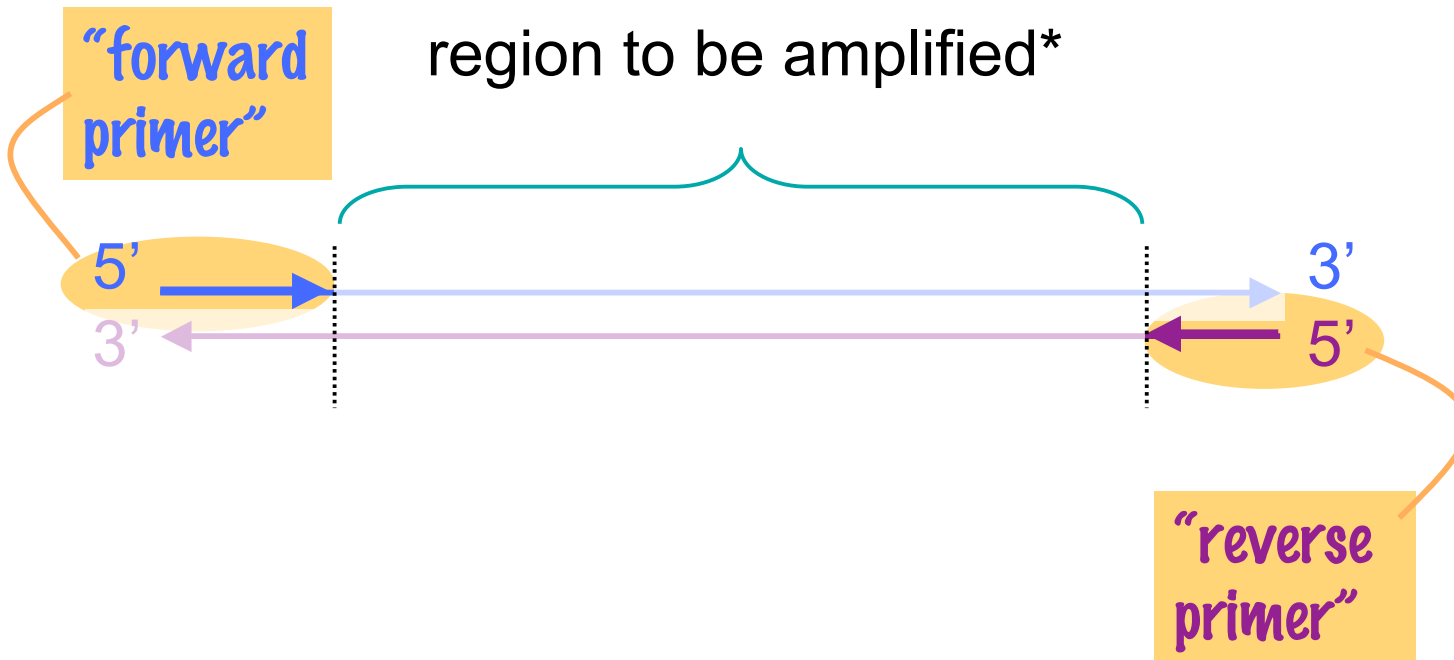
After 4 cycles, half of the products are DNA fragments of a specific size—the size of the DNA that lies between the two primers. By 30 cycles there would be  $2^{28}$  of the DNA molecules from the initial DNA molecule

# Primers for PCR

---

...must “point inward” to the region to be amplified

5' → 3'



\*in addition to the primers



# What 10-bp primers to PCR-amplify the shaded portion?

5' CTTGTCAGTTTTTTTTATGTTTTTCTT **CGCGCGTCAA** CTTTCTA  
3' GAACAGTCAAAAAAAAAATACAAAAGAA GCGCGCAGTTGAAAGAT

CCAAGAGAAAAACAATATAAGGTCTCCTTACTCTATAGGAGAAT  
GGTTCTCTTTTTGTTATATTCCAGAGGAATGAGATATCCTCTTA

AAAACAAACAAAAATAAAAAGCACATCGTAGCGCCAAGAAAATA  
TTTTGTTTGTTTTTATTTTTTCGTGTAGCATCGCGGTTCTTTTAT

CTGCAAATACCAAACCTTGTAAGAATTTCCCGCACATCTTTGC  
GACGTTTATGGTTTGGAACATTTCTTAAAGGGCGTGTAGAAACG

GGGCATACAGTTCATGTATTGGCAACTAACGGAAC TAAGGCAAC  
CCCGTATGTCAAGTACATAACCGTTGATTGCCTTGATTCCGTTG

ATATCTTGCATATTG **CAATGTTCACTAT** 3'

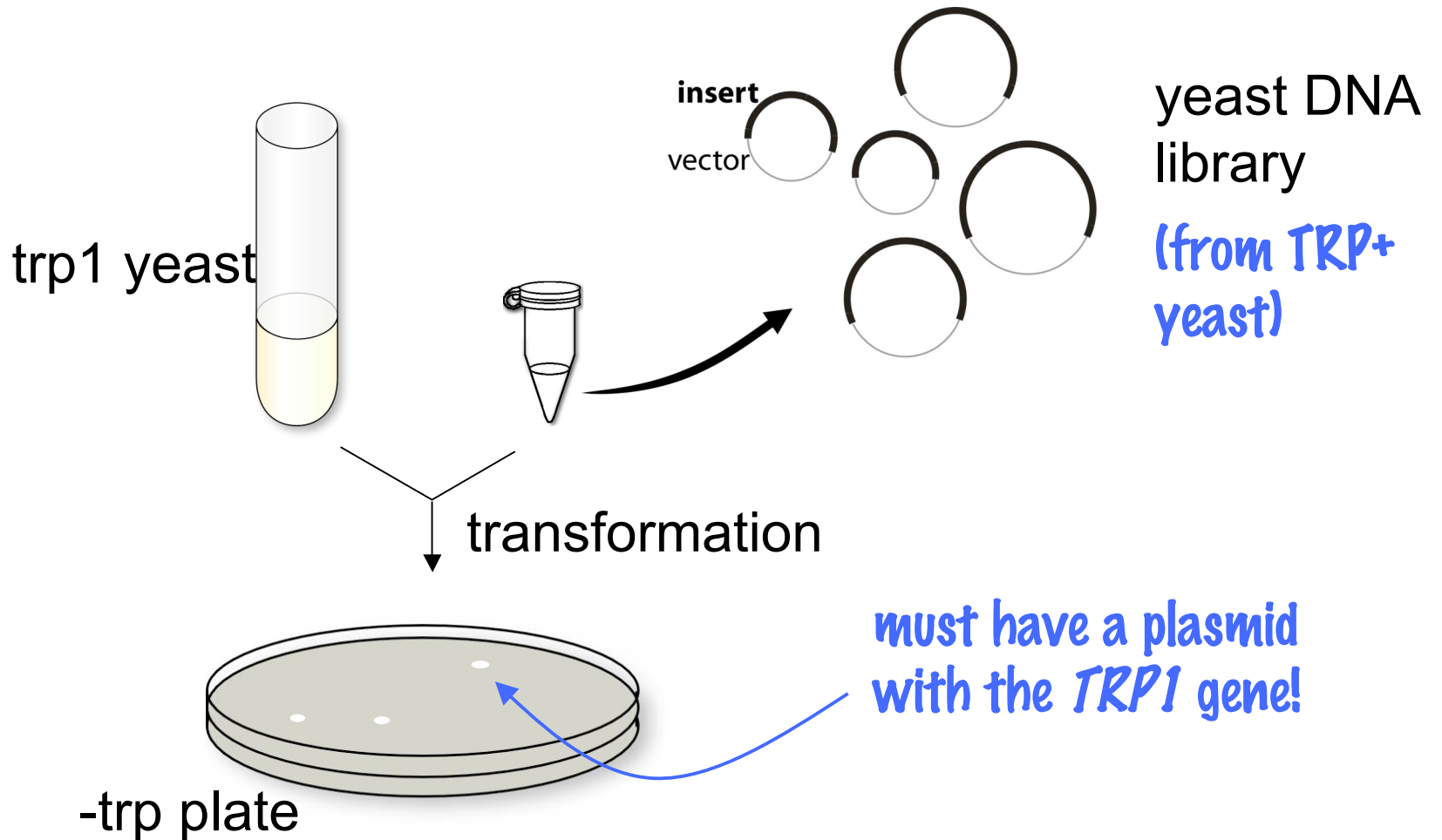
TATAG **AACGTATAAC** **GTTACAAGTGATA** 5'

# What use is a DNA library?

---

Cloning a gene by “functional complementation”

Example: Cloning the *TRP1* gene in yeast



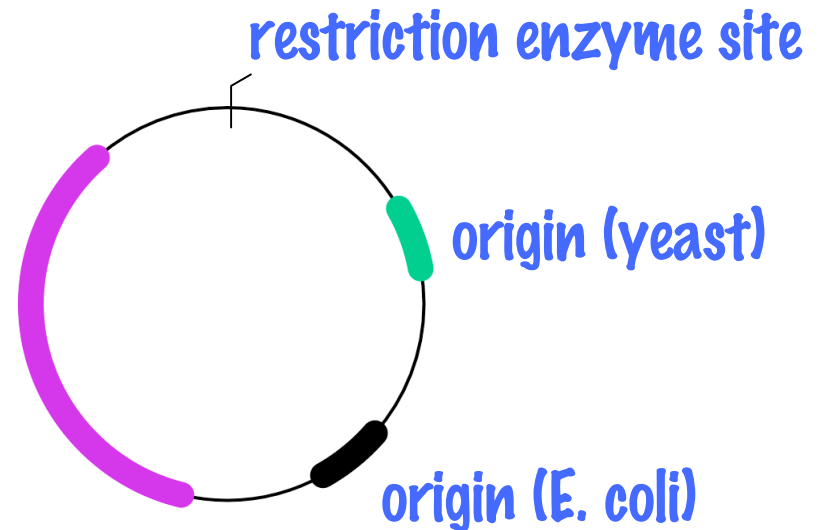
## Questions...

---

What kind of vector?

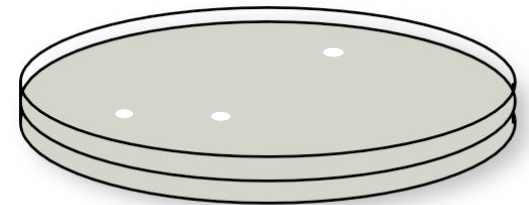
**Bare minimum:**

selectable marker  
e.g., Amp<sup>r</sup> (E. coli)



How would you prove that the yeast became TRP<sup>+</sup> because of the plasmid and not because of a reverse mutation?

Extract plasmid from the TRP<sup>+</sup> cells, re-transform fresh trp1 cells...  
do they also become TRP<sup>+</sup>?



## Practice Question

---

Consider the temperature-sensitive yeast strain that has a mutated *cdc7* allele. How could you identify a plasmid with the **wild type** *CDC7* gene? Give a complete flowchart—which strain you'd use to make the plasmid library, what you would transform, etc.

## What we need

We need WILD TYPE CDC7

...and we need it in plasmid form.

Okay, we have a library...

But we need to pick out just those plasmids that contain the CDC7 gene.

## What to do

So, make a plasmid library starting with \_\_\_\_\_

Extract plasmid DNA from the library; transform \_\_\_\_\_, plate \_\_\_\_\_ and look for \_\_\_\_\_

## Practice question

---

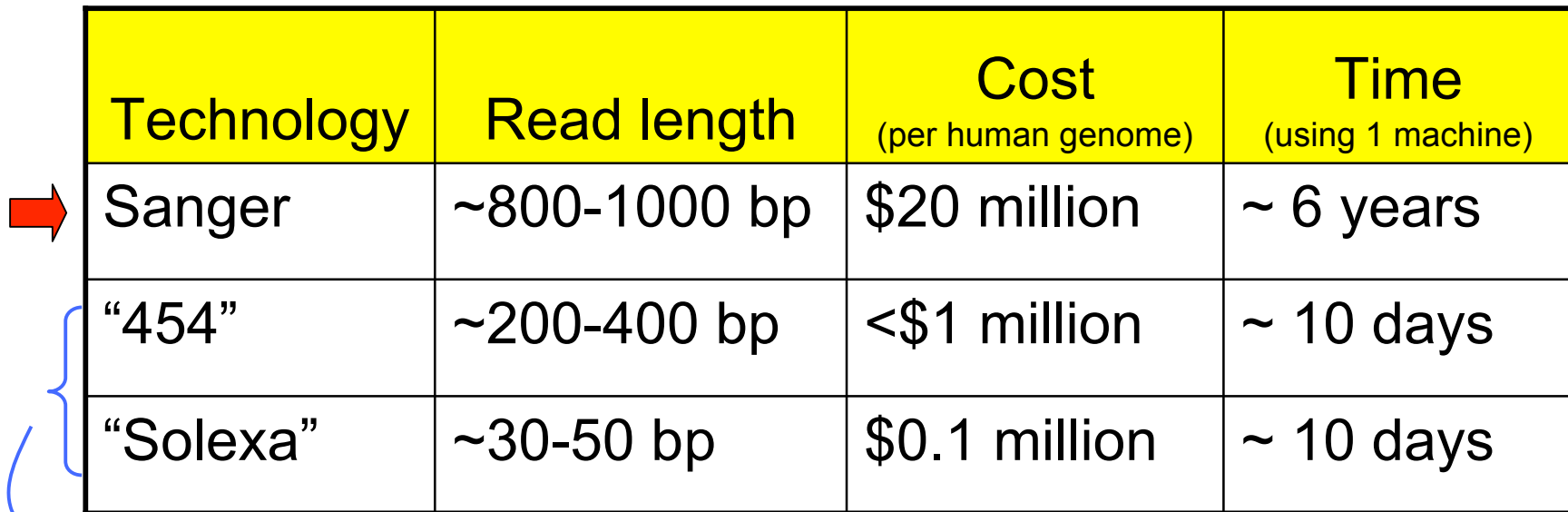
Toxoplasma, an opportunistic pathogen often found infecting immunocompromised individuals, can be treated with the drug pyrimethamine. A pyrimethamine-resistant strain of Toxoplasma has been found; resistance behaves as a dominant trait. How could you learn something about the DNA sequence associated with resistance, using “cloning by functional complementation”? Give a general outline, specifying the source of the DNA library and how you would isolate the clone(s) of interest.

Assume: you have a suitable vector for Toxoplasma transformation

# Sequencing technology

---

Rapidly evolving



| Technology | Read length  | Cost<br>(per human genome) | Time<br>(using 1 machine) |
|------------|--------------|----------------------------|---------------------------|
| Sanger     | ~800-1000 bp | \$20 million               | ~ 6 years                 |
| "454"      | ~200-400 bp  | <\$1 million               | ~ 10 days                 |
| "Solexa"   | ~30-50 bp    | \$0.1 million              | ~ 10 days                 |

*still evolving...*

*Cons: less accurate, short read length can limit uses*

## Characterizing clones — Sequencing the insert

---

“Sanger sequencing”

Uses DNA polymerase

The sequence obtained is for the strand being synthesized

Can determine ~1000 bases in one “read”

With 3,000,000,000 bp in the human genome, that’s a lot of sequencing reactions!



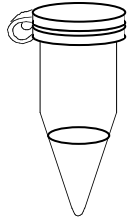
Fred Sanger

**Nobel Prize for  
protein sequencing  
(1958) and for  
DNA sequencing  
(1980)!**

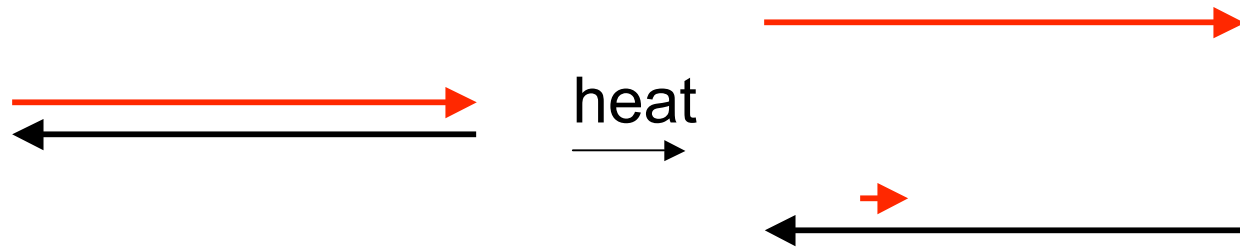


# Reagents for DNA sequencing

---

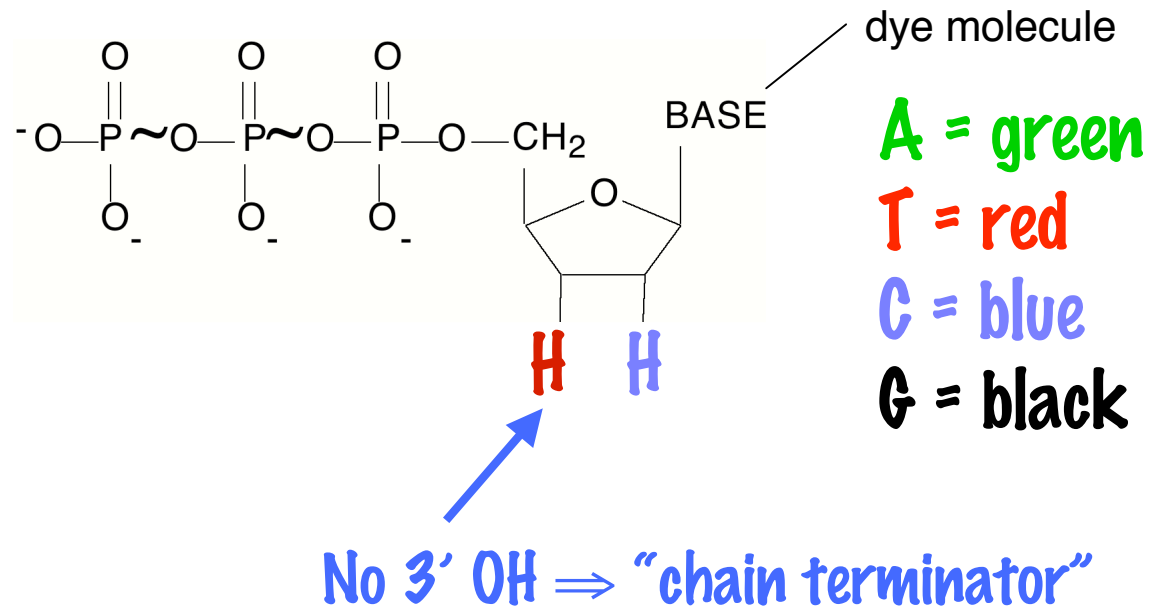


- DNA fragment to be sequenced



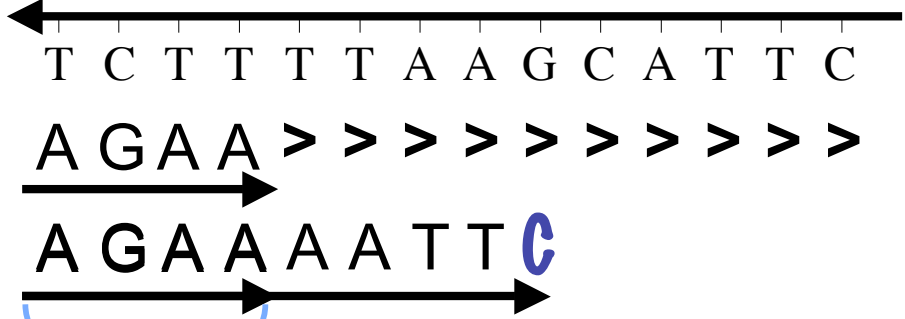
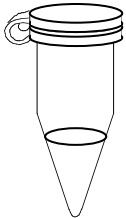
- 1 primer: short ssDNA complementary to ONE region of the template DNA
- dNTPs **All 4: dATP, dCTP, dGTP, dTTP**
- DNA polymerase from *E. coli*
- small amount of each **dideoxy**NTP

- Add a small amount of each dideoxy nucleotide to the reaction



# Products of the four synthesis reactions

“C”  
reaction

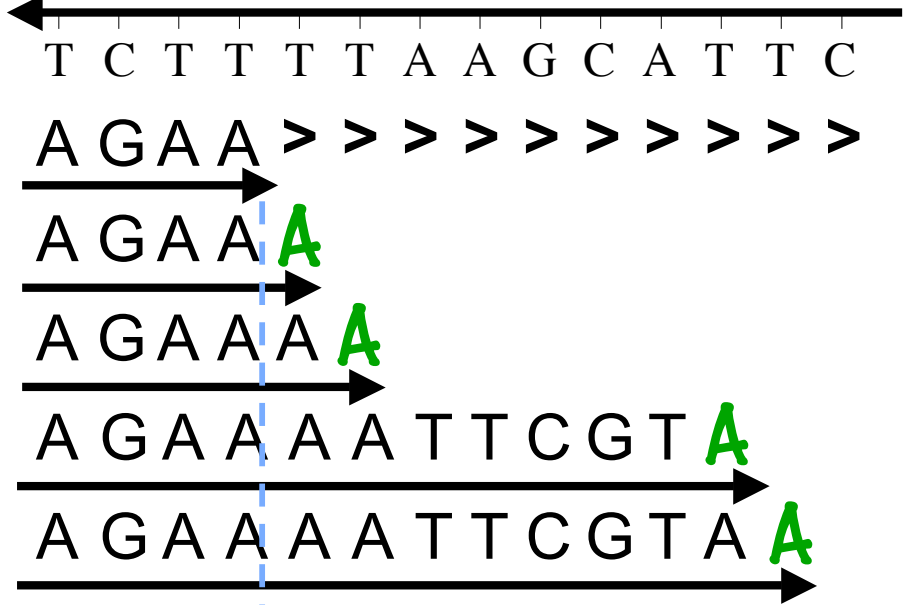
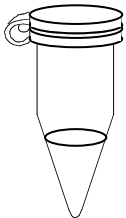


Primer +

5

Primer

“A”  
reaction



1

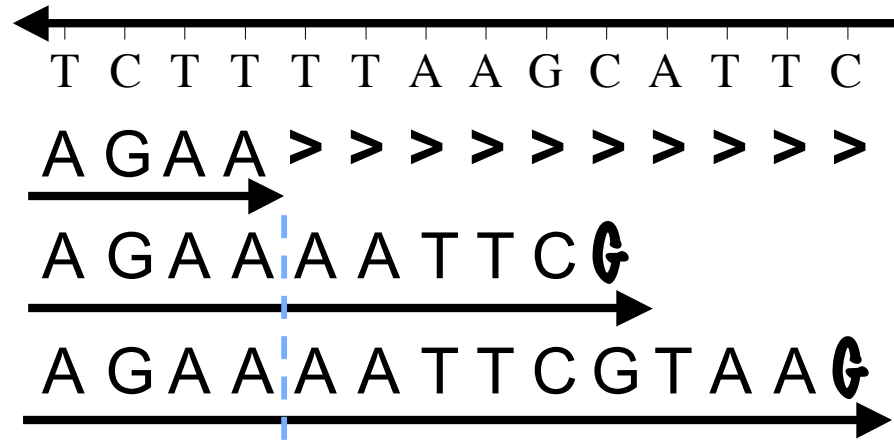
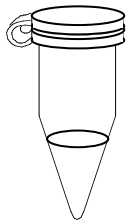
2

8

9

Primer +

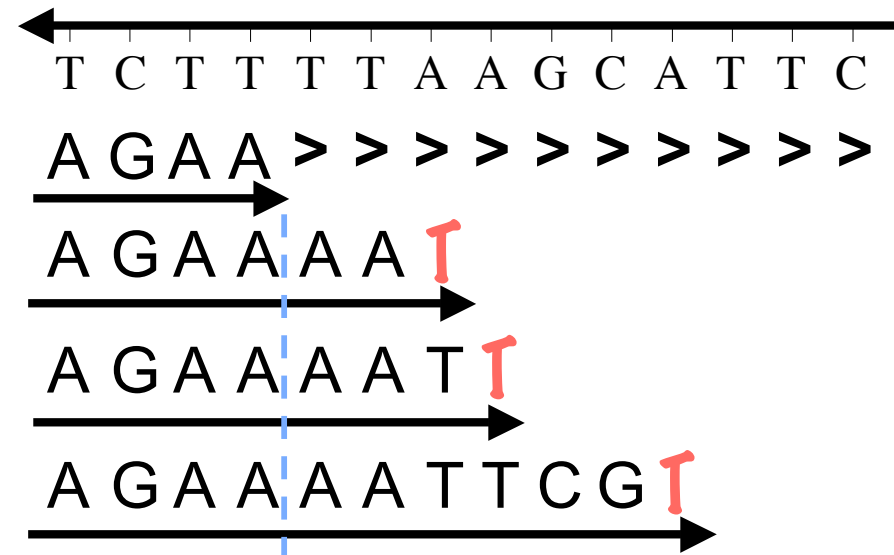
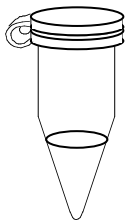
“G”  
reaction



6

10

“T”  
reaction



3

4

7

## Chain elongation in the presence of ddNTPs

---

- elongation of each molecule stops when a chain terminator is incorporated

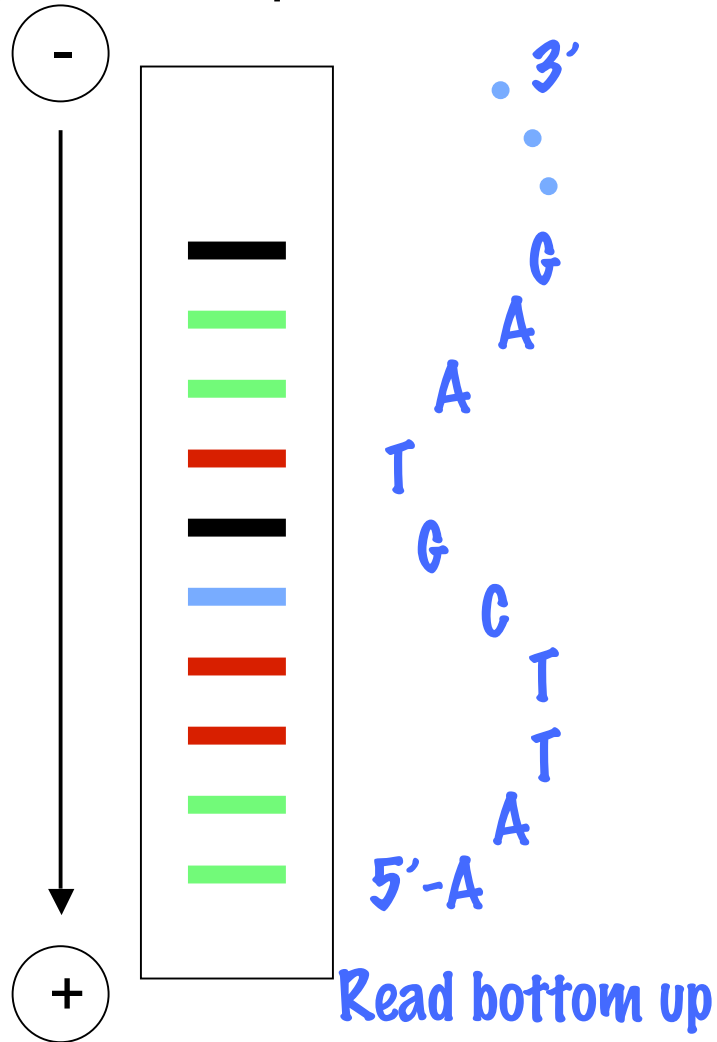
fluorescence...read by instrument

- **Color** indicates which base the molecule ends in
- **Molecule length** indicates where that base is

from gel electrophoresis

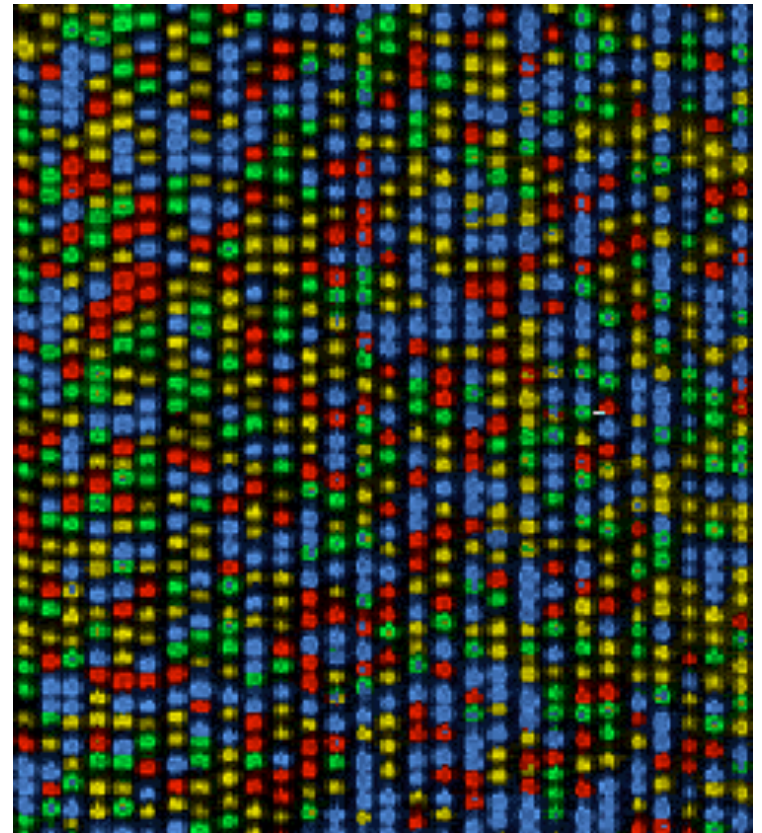
# Analysis of the new strands on a gel

gel electrophoresis

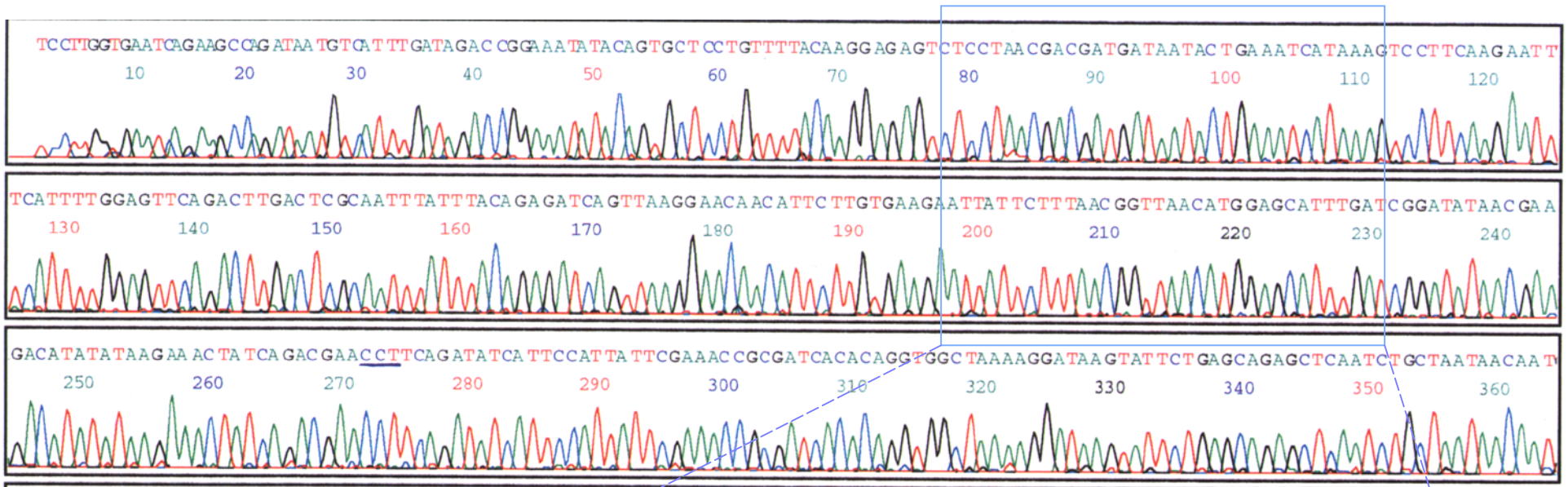


Automated sequencers...

electrophoresis is in thin capillaries



# Laser scan of a sequencing gel



---

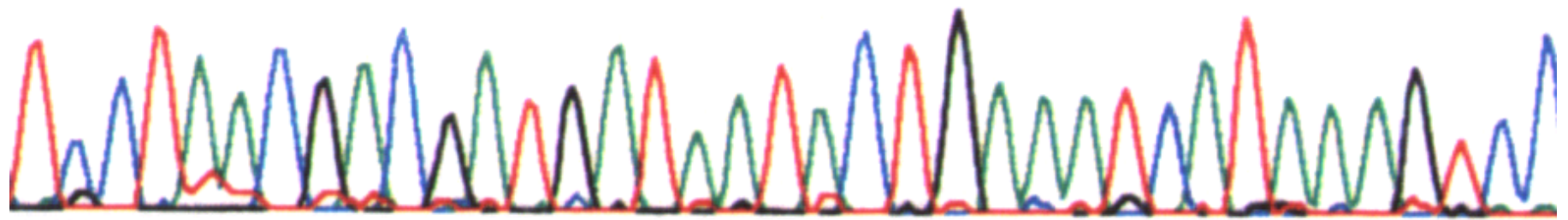
TCCTAACGACGATGATAATACTGAATCATAAAGTCC

80

90

100

110



---

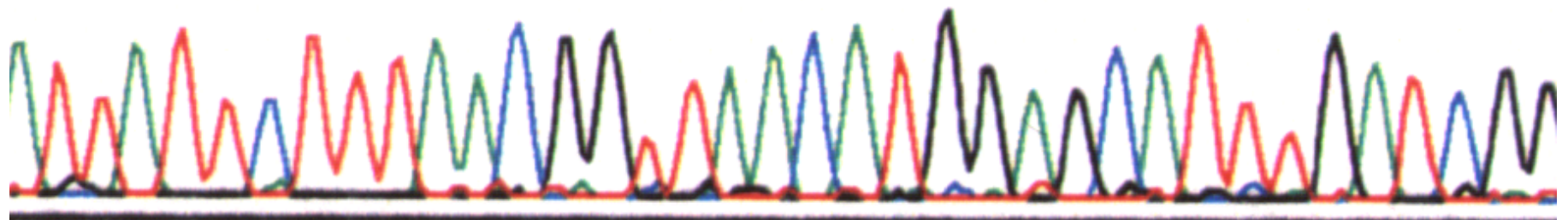
ATTATTCTTTAACGGTTAACATGGAGCATTGTGATCGG

200

210

220

230

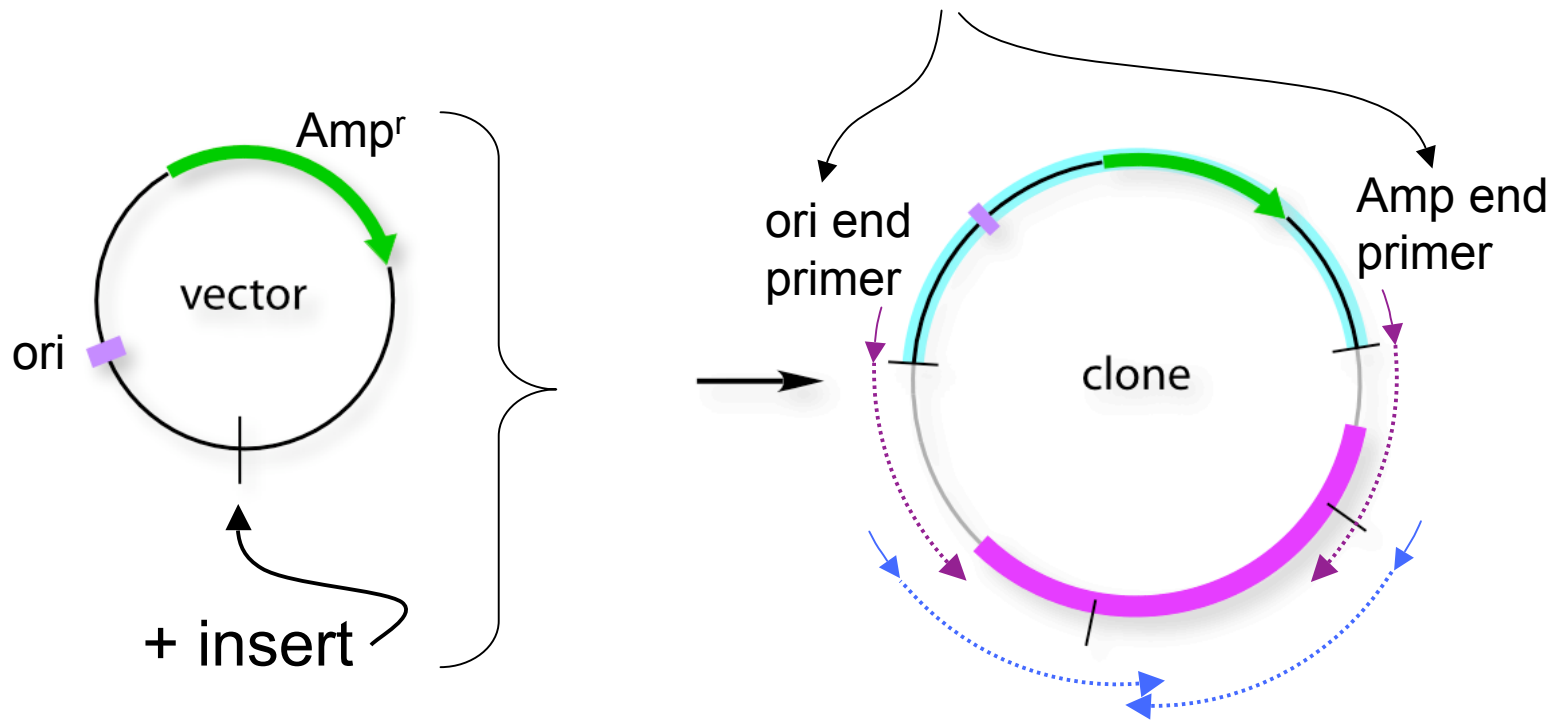




## Sequencing a clone

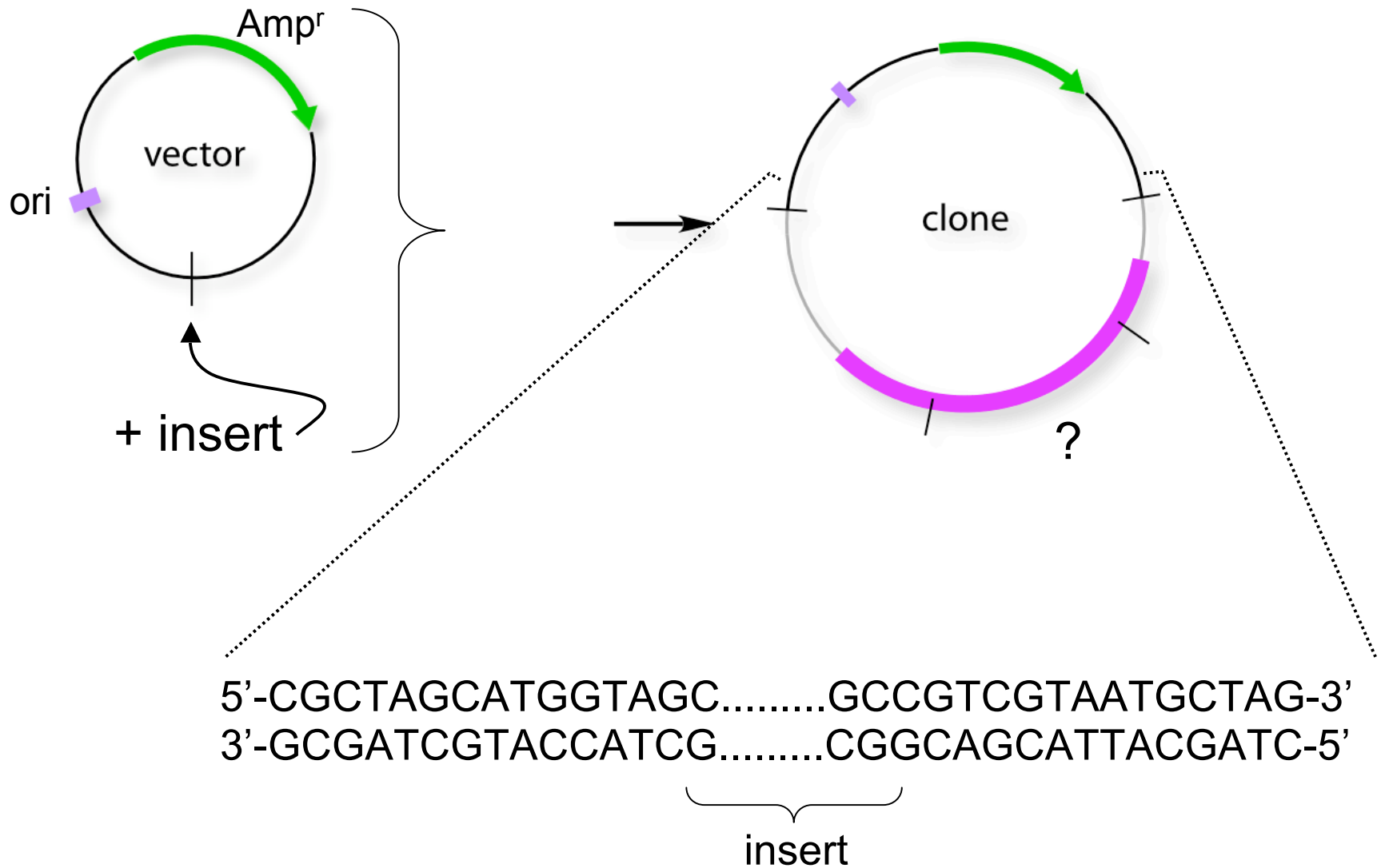
The vector sequence is **known**; insert is to be sequenced

*∴ use vector sequence for initial sequencing primers...*



*then use the new sequence info to extend the sequencing*

# Homework question

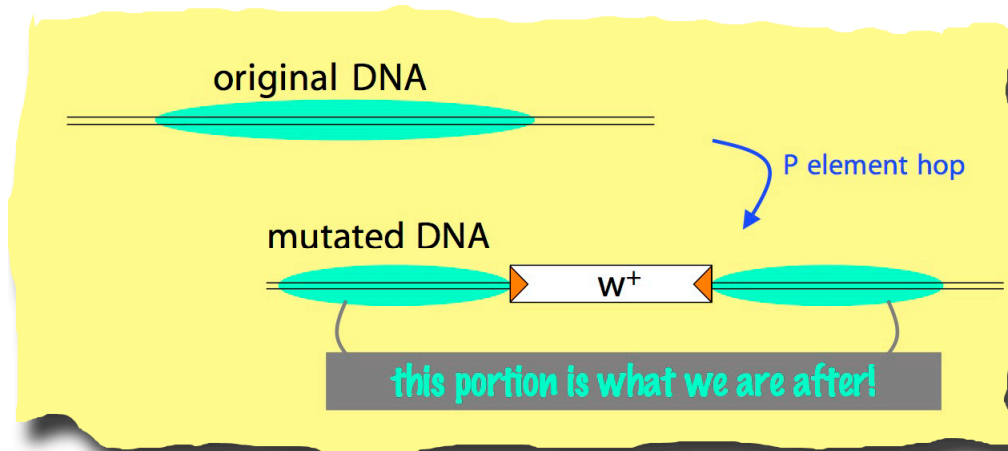


1. If you wanted to **sequence the insert**, what would you use as primer(s) in one sequencing reaction?
2. If you wanted to **amplify the insert** using PCR, what would you use as primer(s) in one PCR reaction?
3. If you wanted to PCR amplify just the sequence corresponding to the (pink) portion marked with a “?”, how would you go about doing it?

**Why would we want to  
sequence the insert?**

## Big picture revisited

---



We want to retrieve (clone) the sequence into which the transposon has landed... but how?

Use the transposon as a “homing signal” to clone the target DNA

Details in QS

... what could you do with the other components within P[w<sup>+</sup>]?

## What to do if we don't have a transposon?

---

e.g.,

- » spontaneous mutants or mutagenesis using chemicals
- » disease genes in humans

One solution:

“Positional cloning” —

- ✓ 1. determine the approximate genomic location of the gene
2. clone the DNA sequence in the vicinity

# Sequencing whole genomes

## Sequencing a genome... two strategies

---

Strategy 1. Whole-genome shotgun sequencing

“Bottom-up” strategy

Strategy 2. “Top-down” strategy

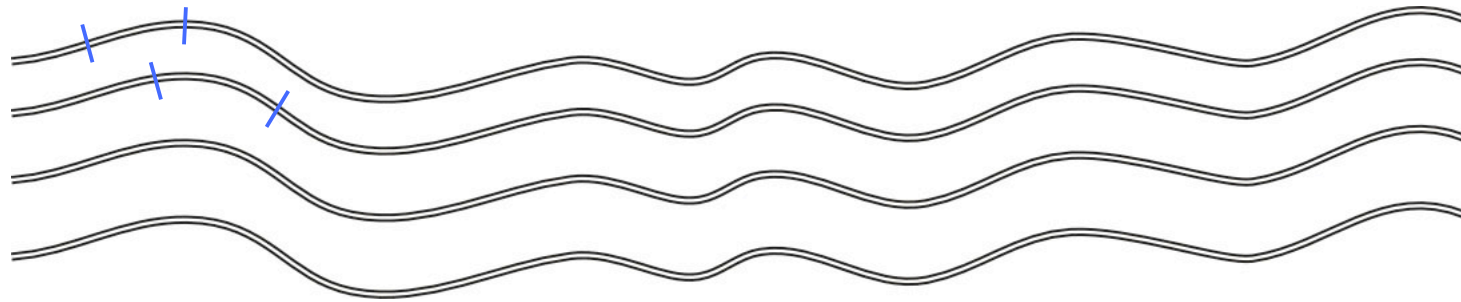
“Clone-by-clone” strategy

“Map-based” or “BAC-based” sequencing



## Strategy 1. Whole-genome shotgun sequencing

---



shear into small, random pieces



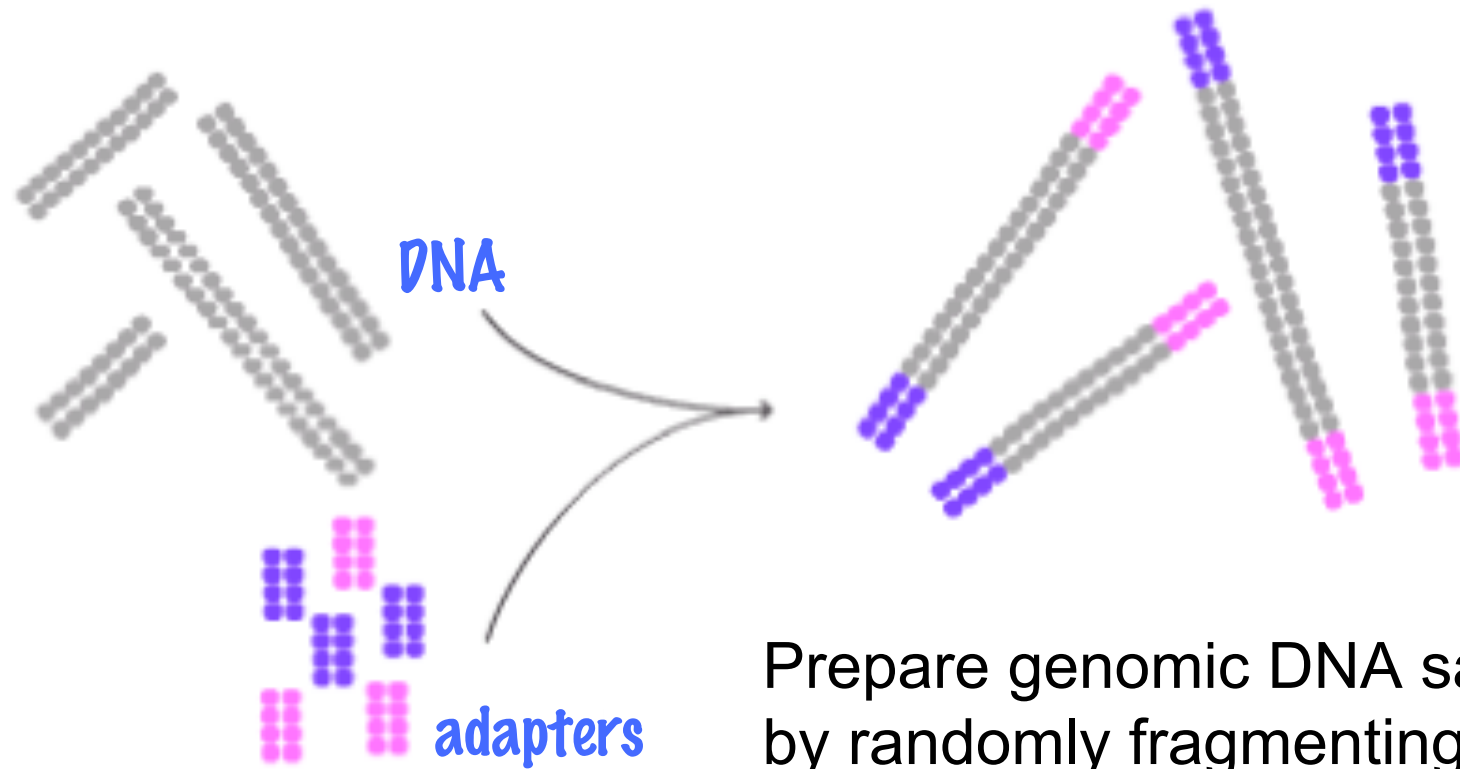
**make a library**

sequence the pieces, assemble based on overlaps

## Beyond the Basics

## Solexa sequencing: DNA preparation

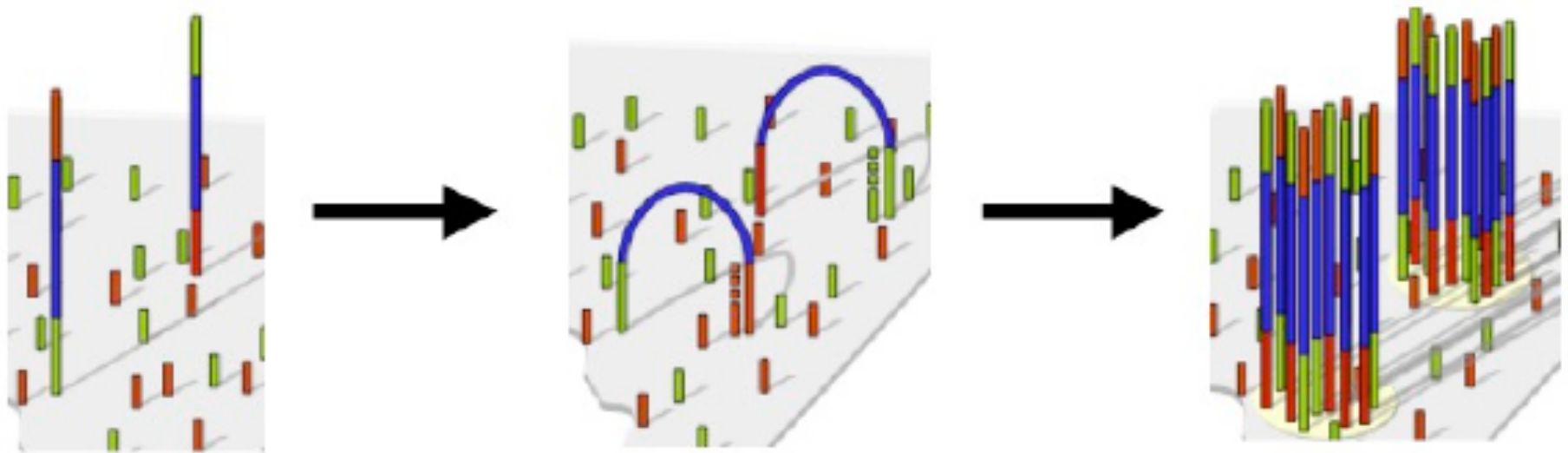
---



Prepare genomic DNA sample by randomly fragmenting DNA and ligating adapters to both ends of the fragments

# Solexa next generation sequencing: DNA clusters

---

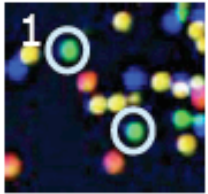


Attach DNA  
to surface

Use primers on surface for  
PCR to generate DNA clusters

# Solexa next generation sequencing: Sequence reaction

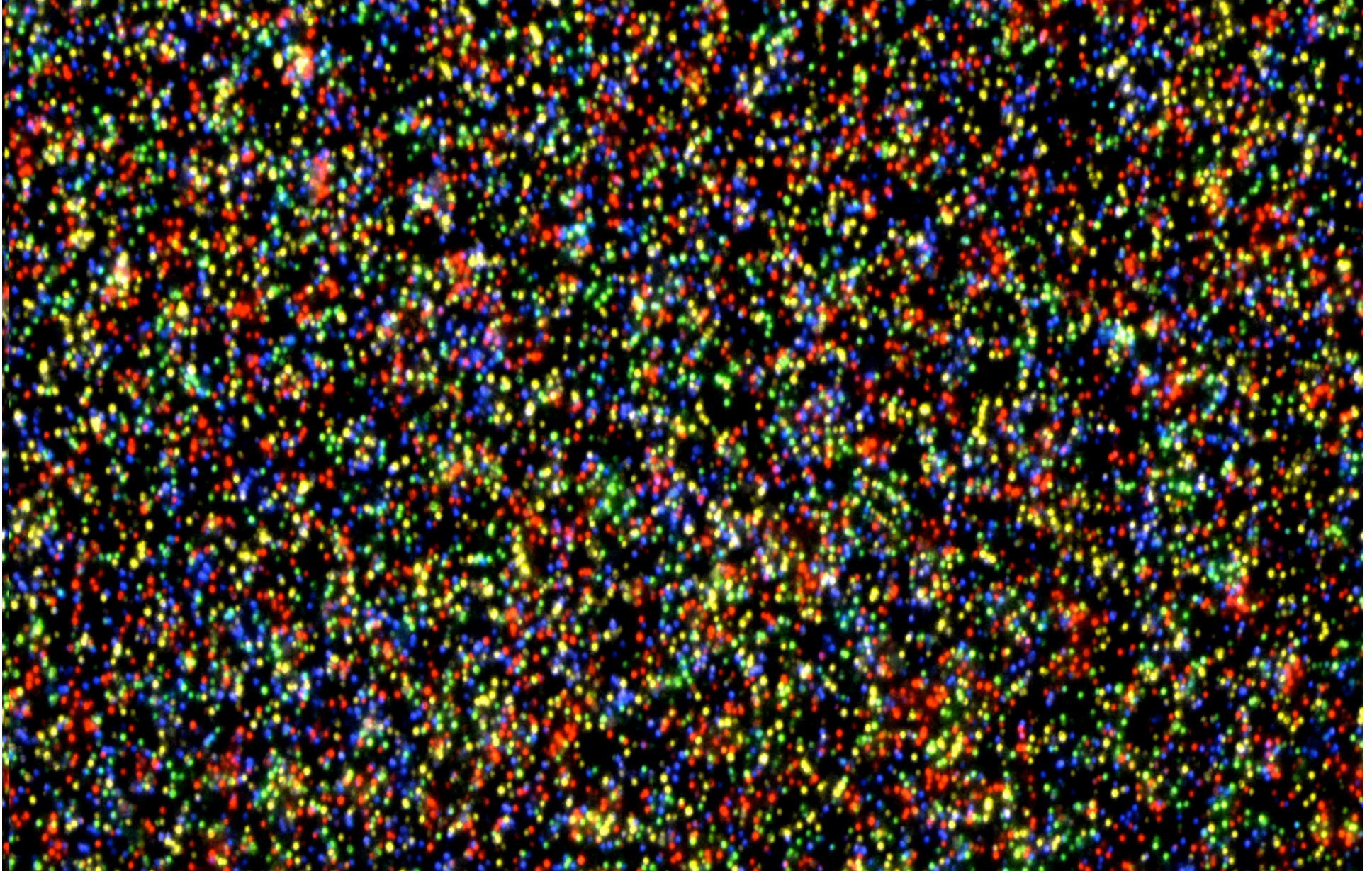
---



Sequence clusters using reversible terminators,  
by imaging after each cycle of synthesis

## Solexa next generation sequencing: Sequence image

---



reference sequence (in database)

GTAACCTGATTTCGATATTCGATATCGGGCATCGGATTAGCGAGAACGGCT

**sequence reads  
from an individual**

TCGATATCGGGCATGGGAT

ATATCGGGCATGGGATTAGCGA

TATCGGGCATAGGATTAGCGAG

GATATCGGGCATGGGATTAGCGAGAACGGCT

TTCGATATCGGGCATAGGATTAG

CGATATCGGGCATAGGATTAGCG

TCGGCATGGGATTAGCGAGAACG

reference sequence (in database)

GTAACCTGATTTCGATATTCGATATCGGCATCGGATTAGCGAGAACGGCT

**sequence reads  
from an individual**

TCGATATCGGCATGGGAT

ATATCGGCATGGGATTAGCGA

TATCGGCATAGGATTAGCGAG

GATATCGGCATGGGATTAGCGAGAACGGCT

TTCGATATCGGCATAGGATTAG

CGATATCGGCATAGGATTAGCG

TCGGCATGGGATTAGCGAGAACG

# The problem with shotgun sequencing

---

Fragments from a "partial digest" of MLK's "I have a dream" speech...  
Is there a problem assembling these fragments into a contig?

contigu

- ① one day even the state of Mississippi a desert state s... with the heat of injustice
- ② a dream deeply rooted in the Am... I have a dream that one day
- ③ justice I h... that my four children will one day
- ④ to be self evident... men are created equal I have a dream that one
- ⑤ I have a dream that one day the state of Alabama whose governors lips

**sequence repeats make assembly difficult!**



## Strategy 2. Clone-by-clone sequencing

---

- I. Clone large pieces of the genome
- II. Know where those large inserts came from
- III. Shotgun-sequence each large insert separately
- IV. Assemble large-insert sequences into full sequence

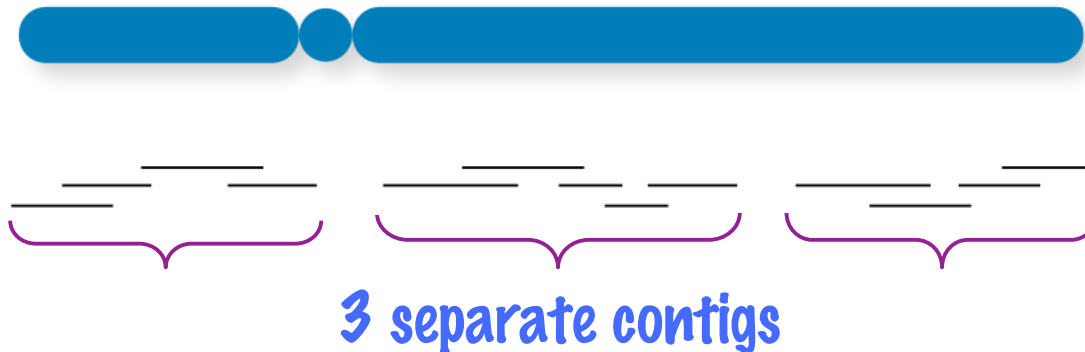
# Contigs

---

Contig = contiguous piece of DNA;

the result of joining an overlapping collection of sequences or clones

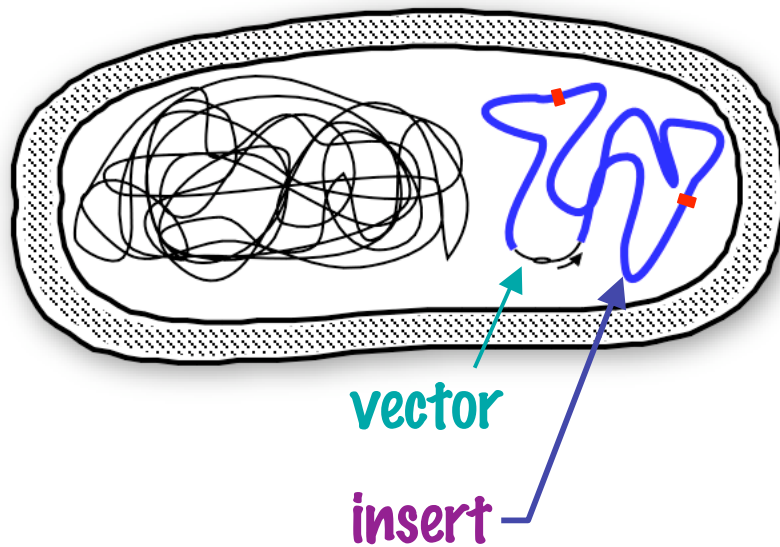
Ideally, a contig is a whole chromosome



In reality, there are likely to be gaps  
requiring additional rounds of cloning

## I. Clone large pieces of the genome

BACs and YACs can hold large genomic DNA inserts



**BACs**

**B**acterial **A**rtificial **C**hromosome;  
can hold inserts of 100-150 kb

**(YACs**

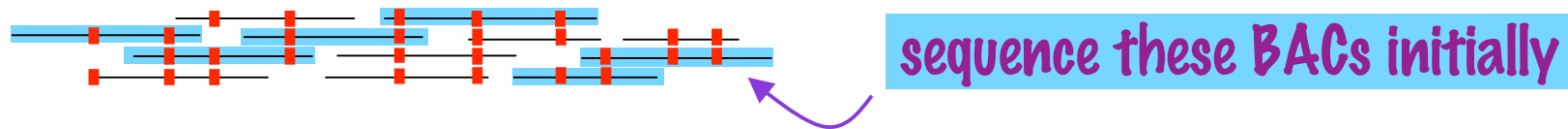
**Y**east **A**rtificial **C**hromosomes; can  
hold inserts of > 1 million bp.)

## II. Know where those large inserts came from

Various molecular biology strategies to find unique sequences present in a BAC clone and map these back to the genome

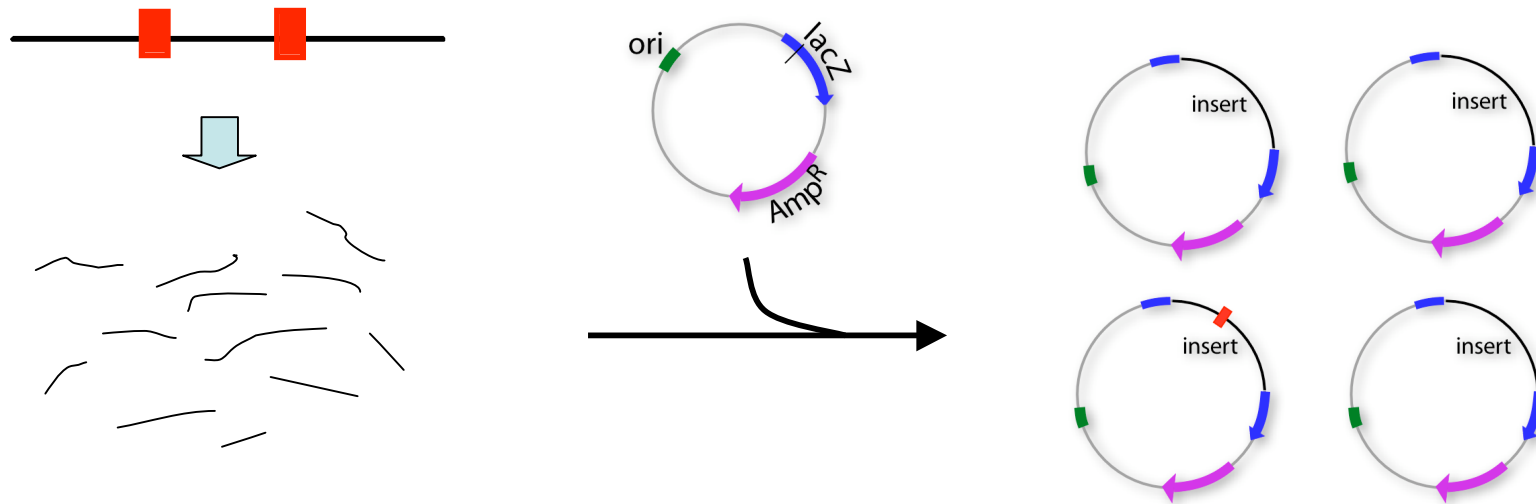
Find the most economical tiling path:

i.e., the fewest # of BACs that will cover the whole sequence



### III. Sequencing the BAC inserts

Shear the insert of each BAC into small pieces, sub-clone into sequencing vector



### III. Sequencing the BAC inserts (cont'd)...

“Shotgun sequencing”:

After subcloning the BAC—sequence the small plasmids

Find overlaps between plasmids

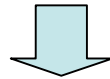
Deduce the sequence of the BAC (assemble the sequence)

#### Plasmid inserts

GTTTCCCCTCCTAACCTACCA  
GTCTC

CCTACCAGTCTCCATATCTCCC

CATCCTTACTT  
CATCCTTACTTCCAGCGAAGA



#### BAC insert

GTTTCCCCTCCTAACCTACCAGTCTCCATATCTCCCATCCTTACTTCCA  
GCGAAGA

### IV. Assemble large-insert sequences into full sequence

## Clone-by-clone vs. shotgun sequencing

---

Advantage of clone-by-clone approach:

BAC locations are known, so less ambiguity with repeated sequences

Disadvantage: Higher cost, slower



If “reference genome sequence” is done for a species...

additional individuals can be sequenced by shotgun sequencing (using comparisons with reference sequence)

## Genomes vary greatly in size

---

|            |                        |
|------------|------------------------|
| Viruses    | ~3-200 Kb              |
| Eubacteria | ~1-5 Mb, most circular |
| Archaea    | ~1-5 Mb, most circular |
| Fungi      | ~10-50 Mb              |
| Animals    | ~100-5,000 Mb          |
| Plants     | ~100-10,000 Mb         |

But (for the most part) they share their core organization:

Genes are (mostly) non-overlapping

Genes code for functional proteins and RNAs

Genes are transcribed into RNA (most translated into protein)

Gene orientation (transcribed strand) is intermixed

Genes are preceded by transcriptional regulatory DNA sites



## Identifying genes can be difficult

---

Only about 1% of human genome codes for protein

Large intergenic regions

Many introns within genes

Can be difficult to predict genes accurately without experimental information

## Experimental approach to gene structure

---

Copy polyadenylated mRNA into DNA (cDNA) and sequence the cDNA clones

Match cDNA sequence to genome sequence to identify exon/intron borders

*Apply to millions of cDNA sequences*

## Comparing genes among genomes

---

Suppose you want to study genetics of human tyrosinase

Loss of function of tyrosinase causes albinism

Take as a given that we have the sequence of the human tyrosinase gene

What do you do?



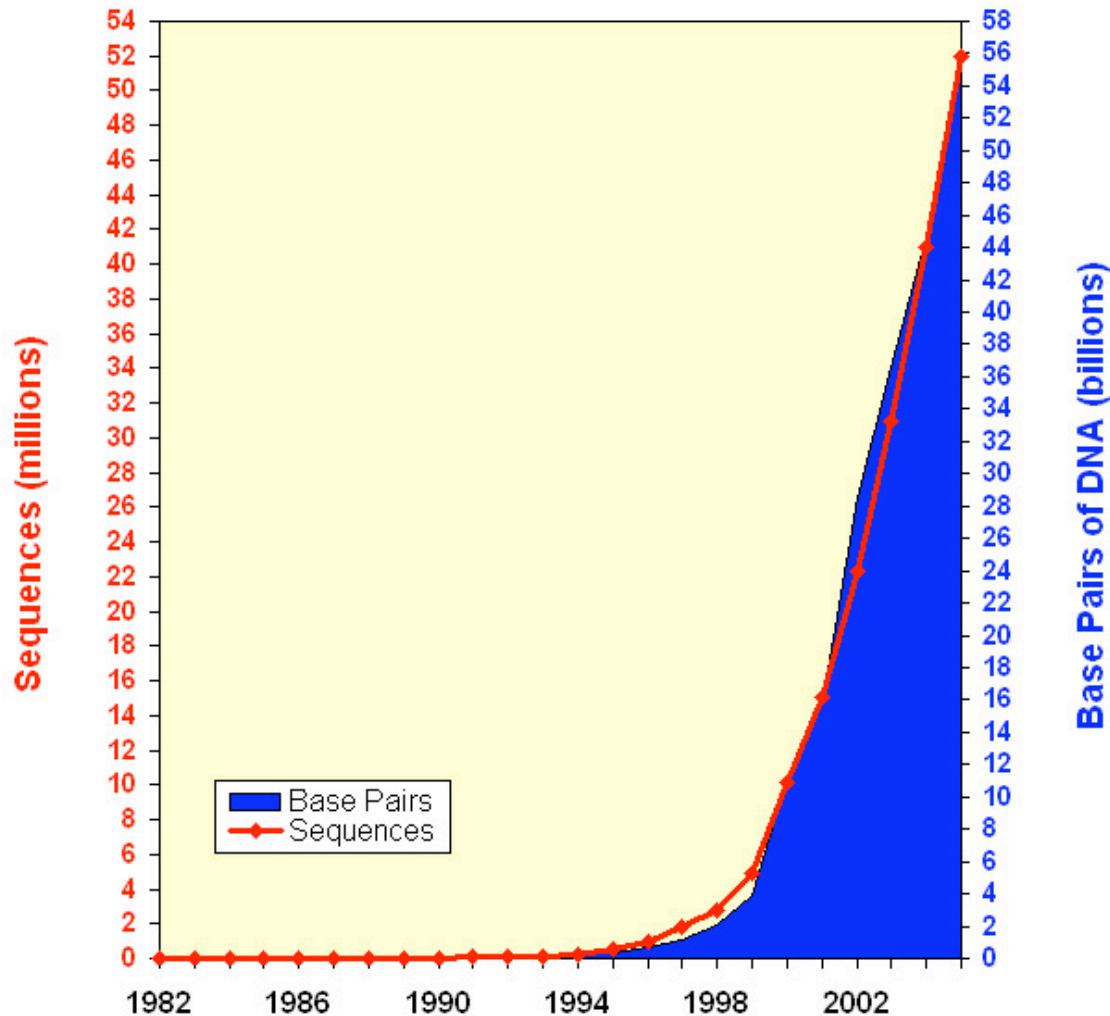
Find the gene in cat (or dog or mouse etc.) that is most similar to the human tyrosinase

This is one principal use of using model organisms to study human disease genes

Alternatively, if you have the sequence of the gene in a model organism, you can use it to identify the human gene

# Sequence databases contain huge amounts of data

**Growth of GenBank**  
(1982 - 2005)



Doubles in size about every 2 years

Over 108 billion bases as of 2008