Lecture 2: High-Throughput Biology

Bioalgorithms Fall 2011

Study Chapter 3.8-3.11

Analyzing DNA

- Recall DNA is the essential information determining the function of living organisms
- In order to understand the biological machinery we'd like to read the "code" of the genome

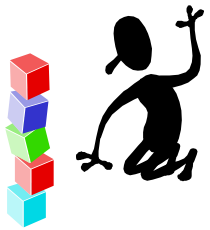
PHILIPPE PSAILA / SCIENCE PHOTO LIBRARY

- How can we get access to DNA?
- An organism's genome can be millions to billions of base pairs long…
- There are about 40 picograms (small) and 2 meters (long) of DNA per cell
- Currently, technology is too primitive to read it directly

Four Steps to Recover a Genome

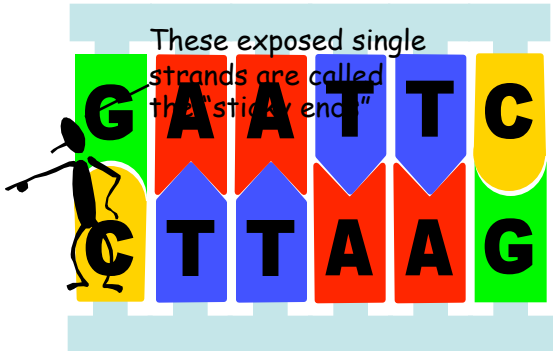
- Cut it
  - Use enzymes to break DNA into tiny substrands
- Copy it
  - Make millions of identical copies of each substrand
- Read it
  - Determine every base-pair in the substrand
- Reassemble it
  - Reconnect all of the substrands in the correct order



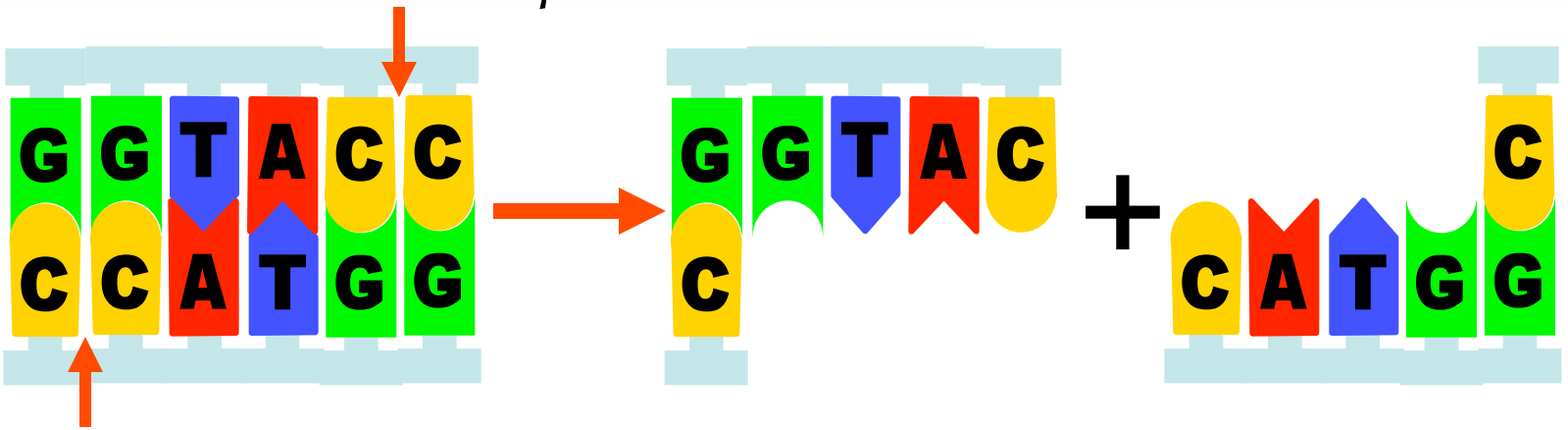
Cutting DNA

- In a bit of serendipity, while studying how bacteria defend against viral infections, Hamilton Smith found certain enzymes that break DNA whenever they encounter specific base sequences.
- Such enzymes are called *restriction enzymes*
- The first one found was *Eco*RI, it breaks DNA at a GA boundary when it encounters the palindromic sequence 5'-GAATTC-3'

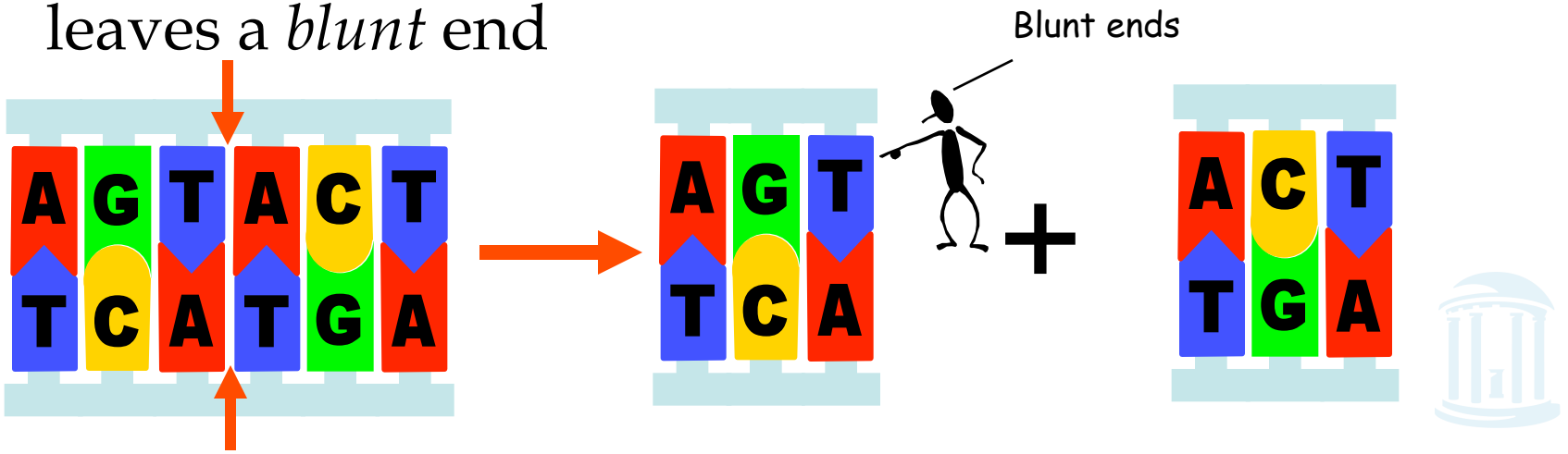


Cutting DNA

- Others include *Kpn*I, which cuts 5′-GGTACC-3′ at C-C

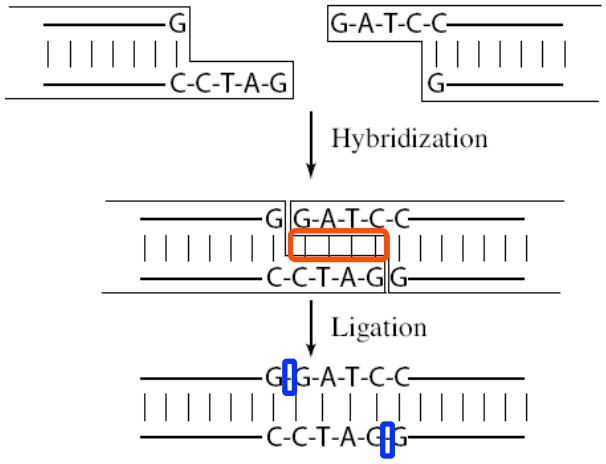


- And, *Sca*I, cuts the sequence 5′-AGTACT-3′ at T-A and leaves a *blunt* end



Pasting Together DNA

- Two pieces of DNA can be reconnected by replacing broken chemical bonds
- Two primary methods
  - ***Hybridization*** – constructing a double strand from a single strand by taking advantage of DNA's desire to form complementary base-pairings
  - ***Ligation*** – gluing together DNA fragments with single-strand "sticky" ends, perhaps from different sources. Special enzymes (DNA ligase) repair the broken phosphate backbone

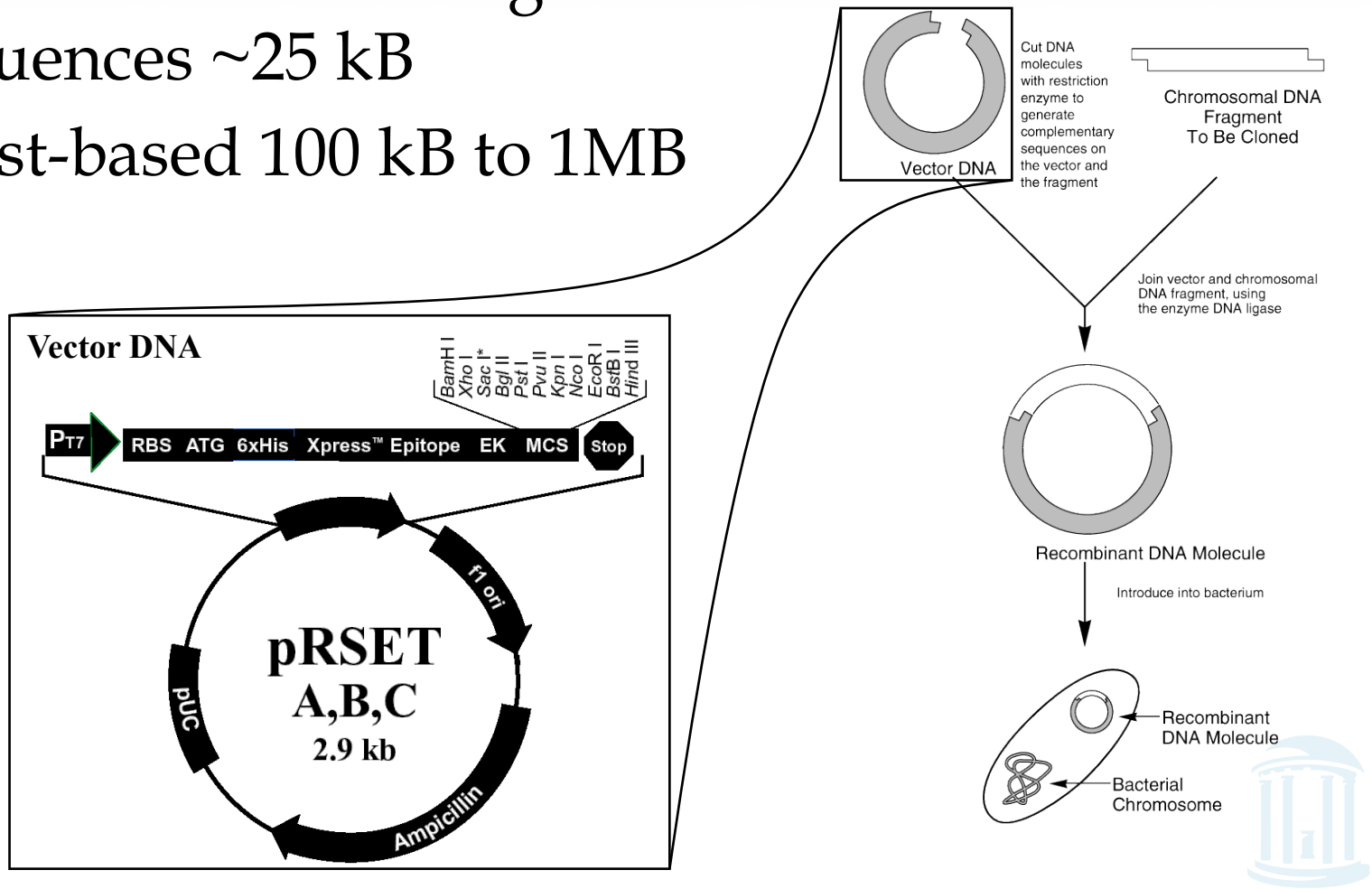


Copying DNA

- While cells are able to extract information from only a single strand of DNA, molecular assays often require large quantities of material.
- One solution, ***cloning***, involves using cells to do the work of replication
- DNA Cloning
  - Bacteria contain small circular DNA molecules called plasmids
  - Plasmids can be cut using a restriction enzyme, and a new sequence can be inserted into the gap using ligation. These cut plasmids are called ***vectors***.
  - The modified plasmid is then inserted into a living organism and left to multiply.
  - Once you have enough, remove the organism, retrieve the DNA.

Cloning Process

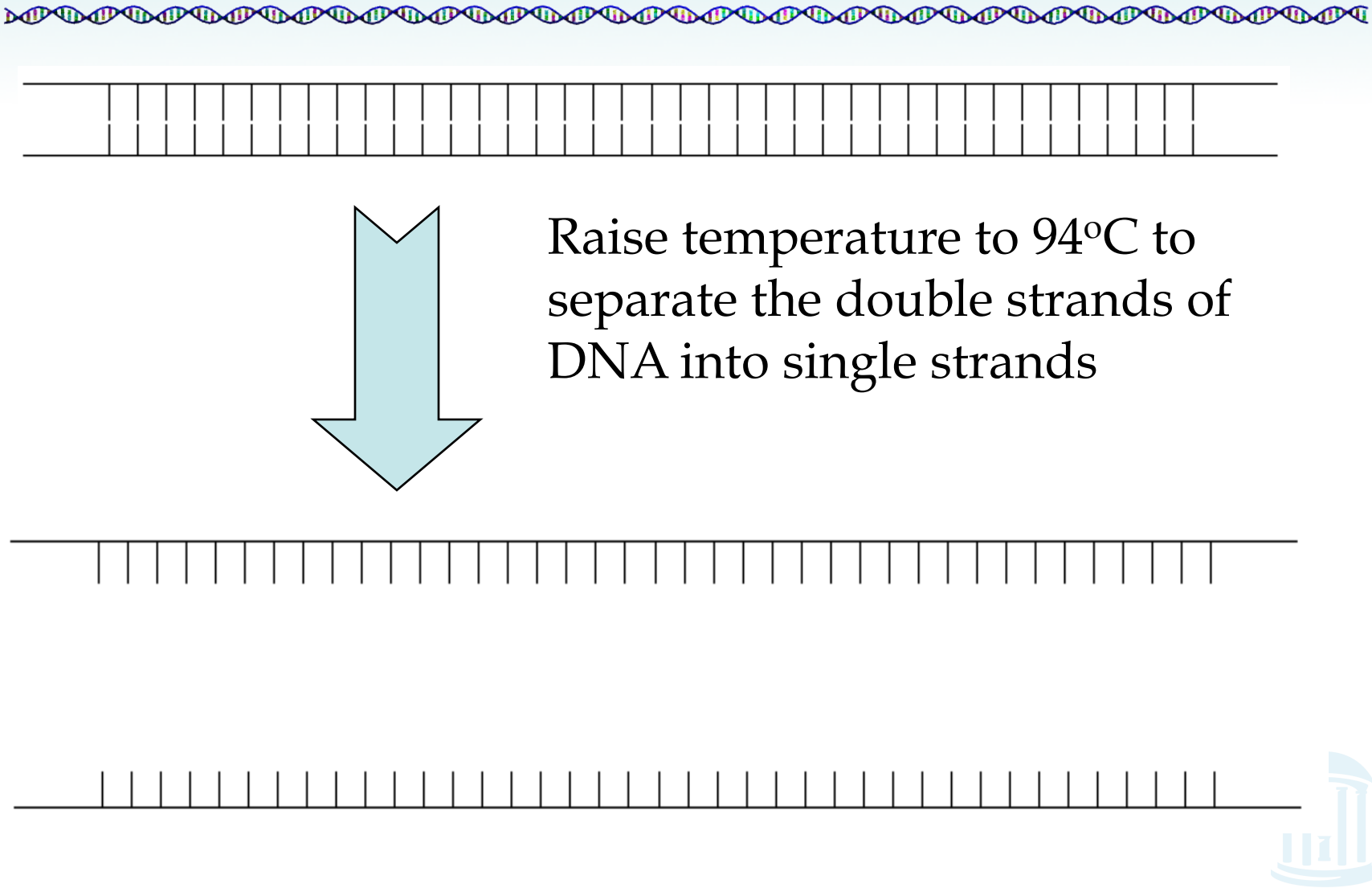
- Bacterial vectors are good for sequences ~25 kB
- Yeast-based 100 kB to 1MB



High-Volume DNA Copying

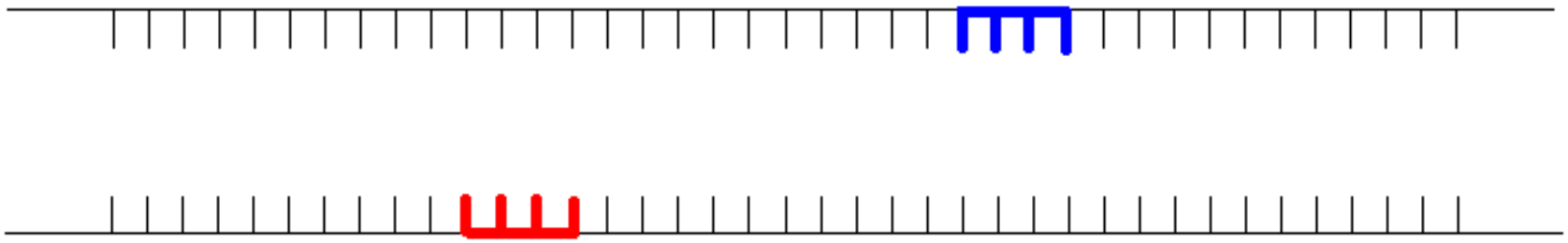
- Polymerase Chain Reaction (PCR)
  - Polymerases are enzymes responsible for replicating DNA during cell division
  - Extend double-stranded regions bonded to a longer, single DNA strands in the 5′-to-3′ direction
- How PCR works:
  - Separate DNA strands, which happens naturally at high temps
  - Add small single-strand DNA sequence (primer) and DNA Polymerase
  - Let the primers hybridize (sometimes called anneal)
  - Let the "heat activated" Polymerase extend the sequence
  - Produces two copies, just by cycling the temperature
  - Repeat.

Denature DNA



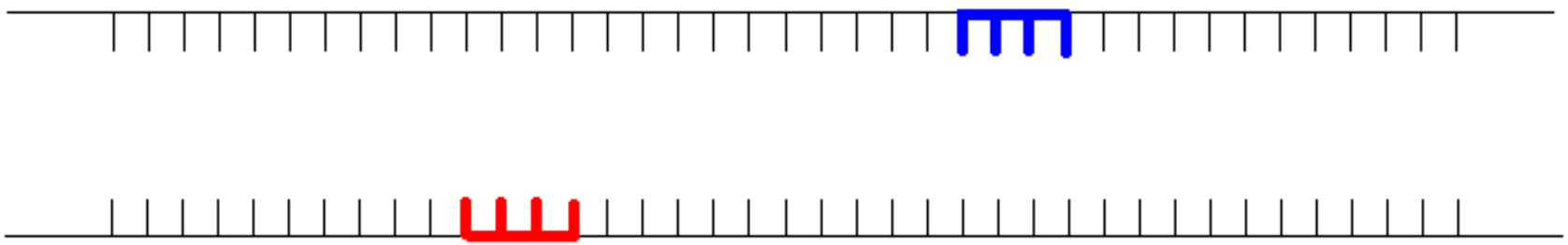
Design Primers

- To perform PCR, a 10-20bp sequence on either side of the subsequence to be amplified must be known



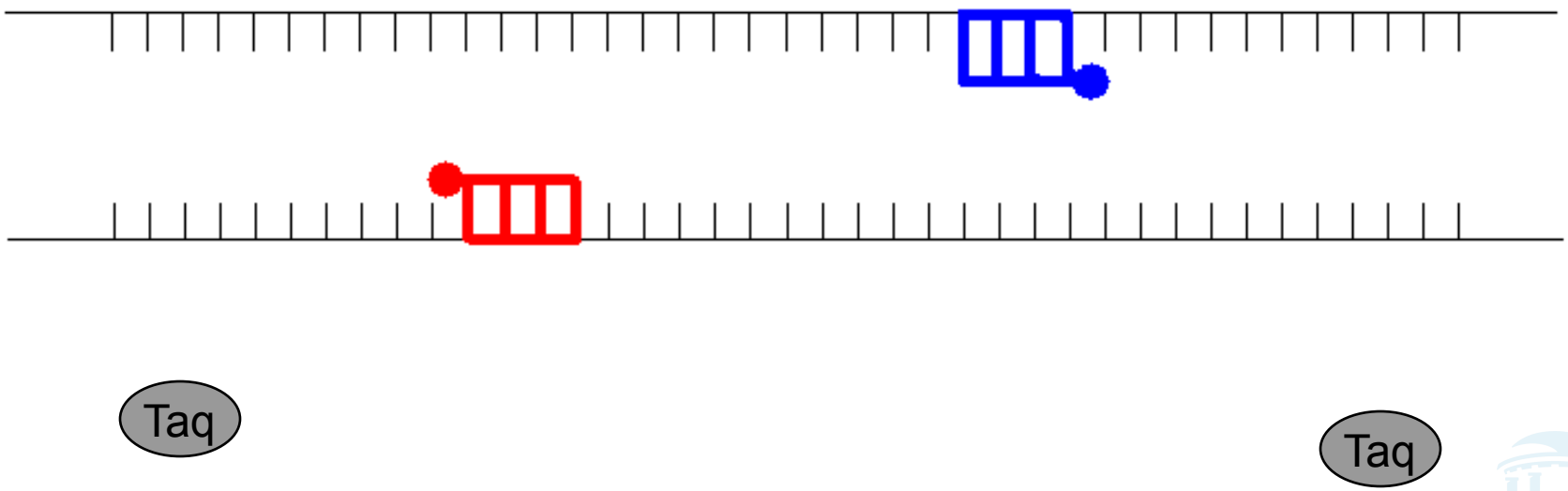
Annealing

- Anneal primers at 50-65°C
  (exact temperature depends on primer sequence)



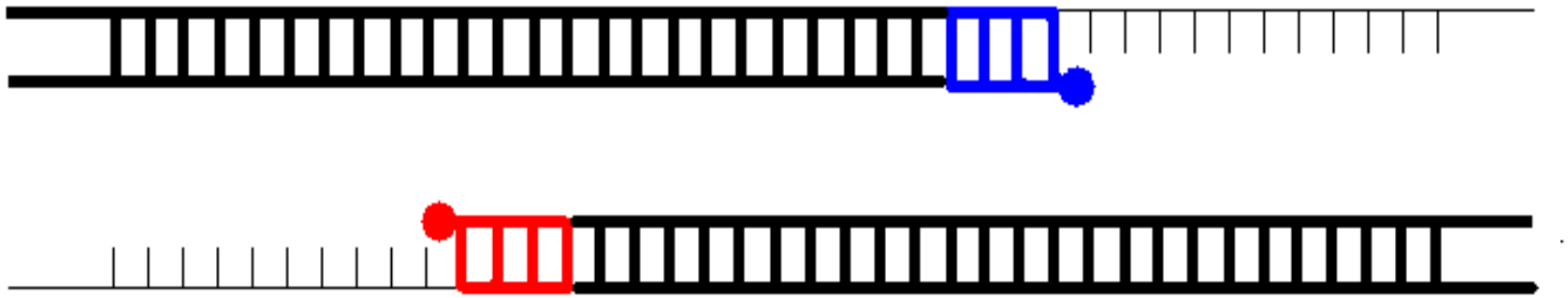
Extension

- Extend primers: raise temp to 72°C, allowing a temperature stable polymerase (Taq) to attach at each priming site and extend a new DNA strand



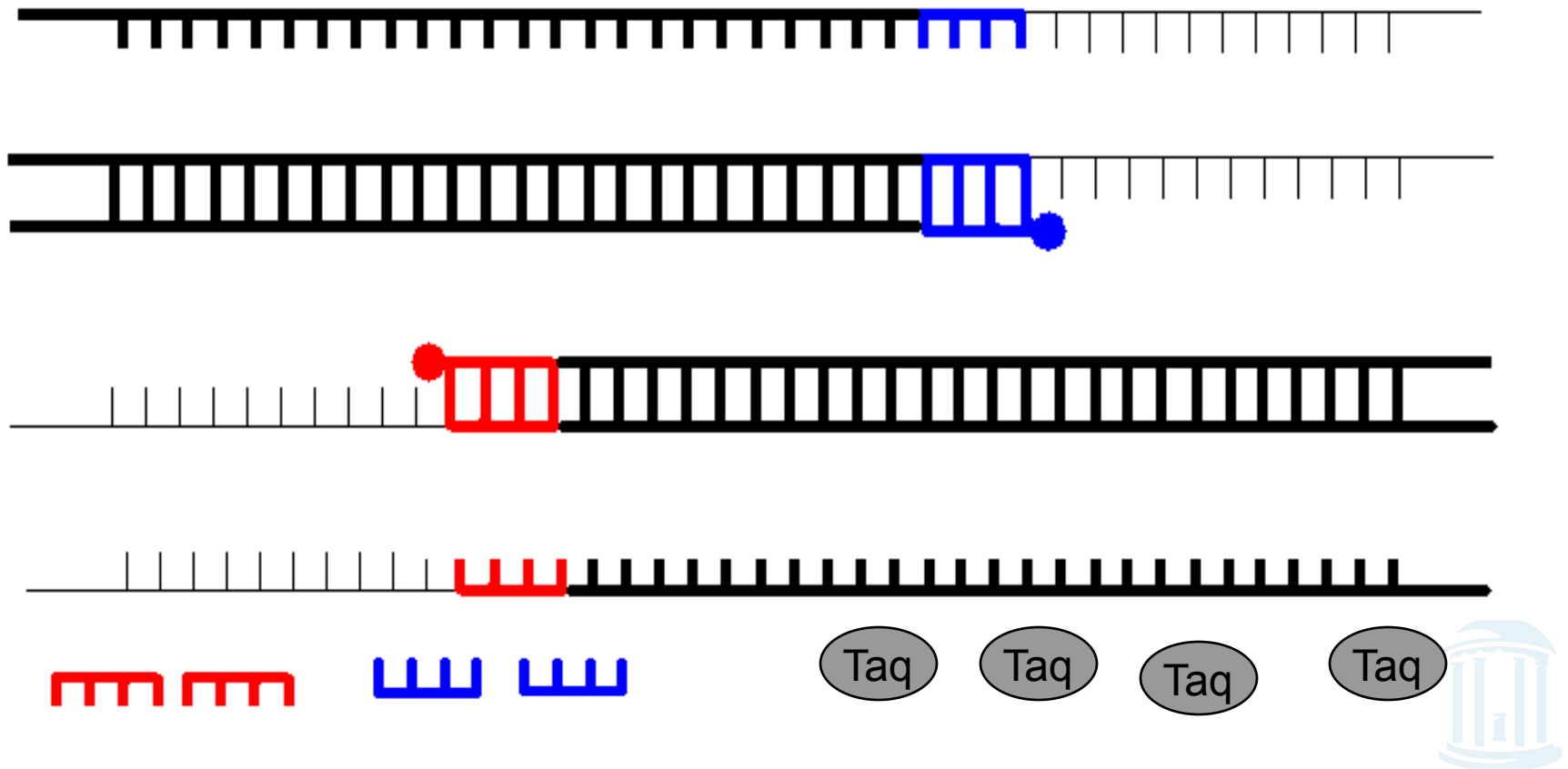
Extension

- Extend primers: raise temp to 72˚C, allowing a temperature stable polymerase (Taq) to attach at each priming site and extend a new DNA strand



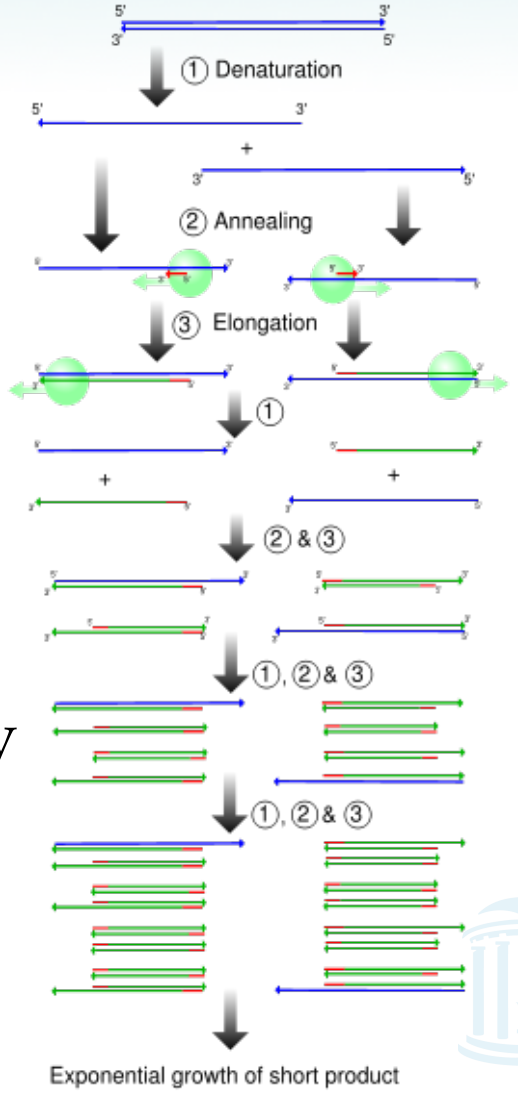
Repeat

- Repeat the Denature, Anneal, Extension steps at their respective temperatures…



High-Volume DNA Copying

- PCR Products
  - Many copies (~2^{Cycles}) of the desired "short" DNA fragment
  - Polymerases always extend in the 5′-3′ direction
  - Typical Polymerases extend the sequence about the primer ~1000 base pairs in 10 secs
  - Replication fidelity (the possibility of the polymerase copying an incorrect nucleotide) is about 1 in 9000



Reading DNA Sequences

- The process of reading off sequential DNA bases is called sequencing
- Sequencing strategies involve 3 steps
  - Generate successive subfragments that differ in length by a single nucleotide
  - Label each fragment with one of 4 different tags
  - Sort the fragments according to size
- Tags are usually fluorescent dyes or radioisotopes
- Sequencing speed is ≈ 10^4 – 10^5 bases/day
- How many days to sequence 3 Billion?

High-Throughput Sequencing

- Parallelism is the key to speeding up this process
- Rather than reading one sequence at a time one can read millions
- Peak Throughput
  - $8 * 1.6x10^6 * 10^4 = 128$ billion/day
- In practice:
  - 55 billion base pairs /day for 2x100 bp reads
  - 18x coverage of a 3 Gb genome in a day

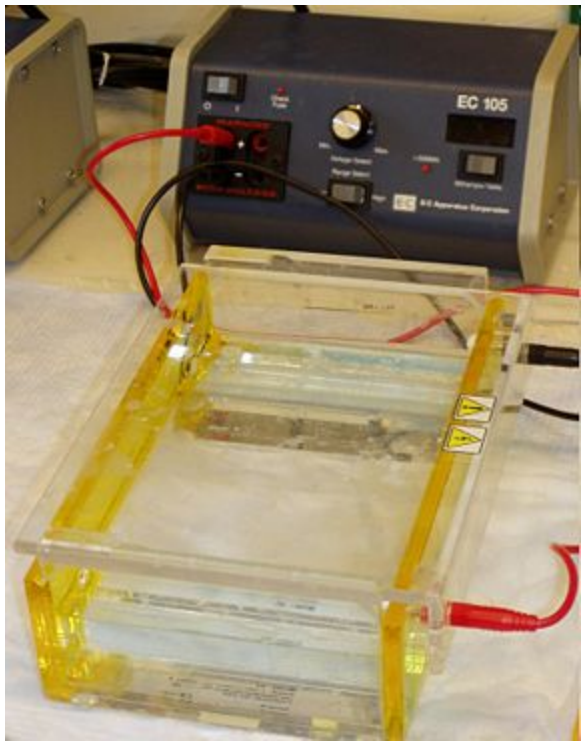
Illumina HiSeq 2000

DNA Sorting

- DNA fragments can be very accurately sorted according to size using gel electrophoresis
- DNA is a negatively charged molecule that can be coerced into motion under an electric field
- DNA is placed into a gel and the speed of motion is proportional to the strand's size (Large molecules more slower, small ones faster)
- Usually florescent markers are attached to each strand (e.g. at the sticky ends) and used to detect the various strand sizes

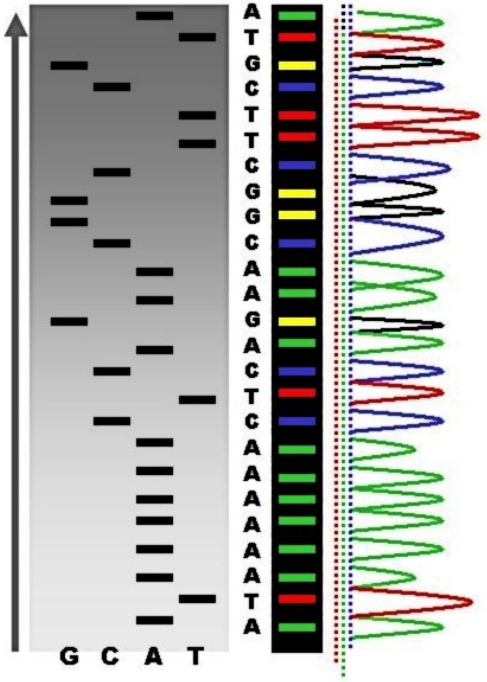
Gel Electrophoresis

- Gel electrophoresis is also commonly used to measure the relative sizes of other molecules, such as RNA and proteins
- Electrophoresis is also used to separate out particular strands
- Often just the number of fragments conveys information about small genetic variations (allele's)



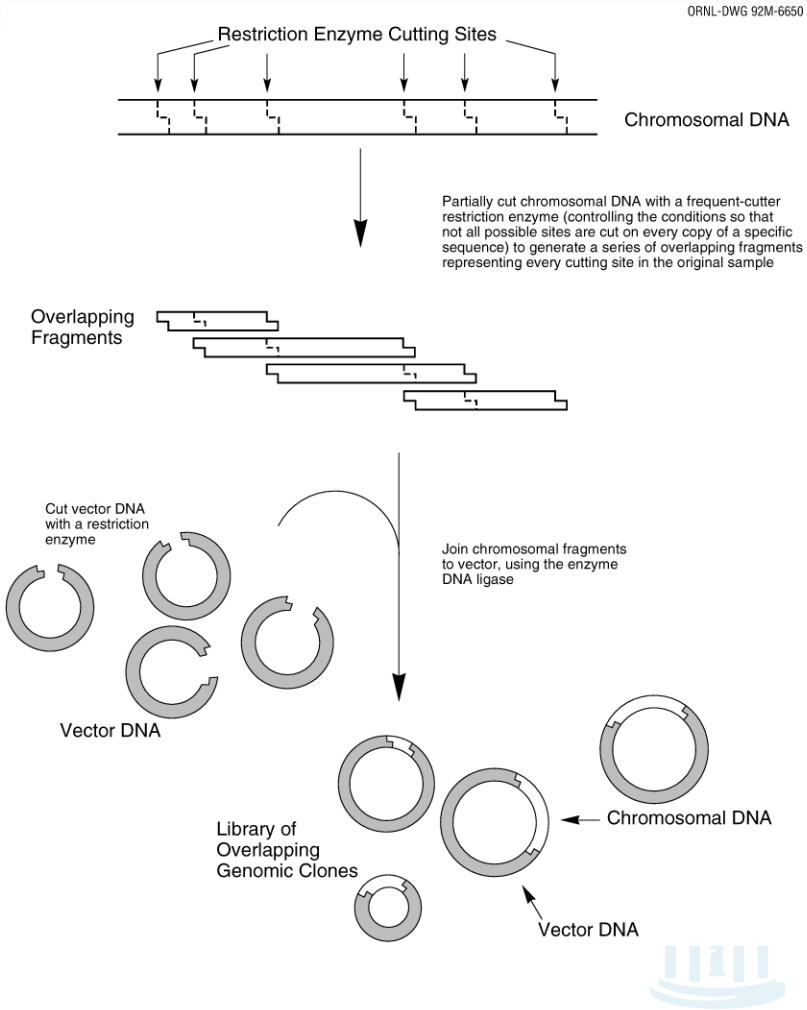
Reading DNA

- Electrophoresis
  - Sequencing is done by separating subsequences by size (and, in a 2D gel, by size and charge).
  - The DNA molecules are either labeled with radioisotopes or tagged with fluorescent dyes.
  - Given a DNA molecule it is then possible to obtain all fragments from it that end in either A, or T, or G, or C and these can be sorted in a gel experiment.
- Another route to sequencing is direct sequencing using DNA chips.



Assembling Genomes

- Must take the fragments and put them back together
- Not as easy as it sounds
- Some of the fragments will overlap
- Fit overlapping sequences together to get the shortest possible sequence that includes all fragment sequences



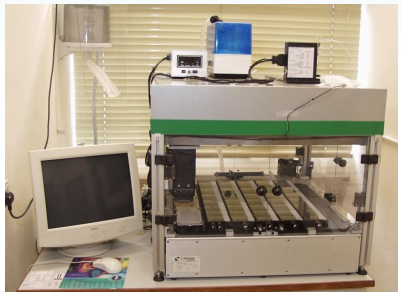
Assembling Genomes

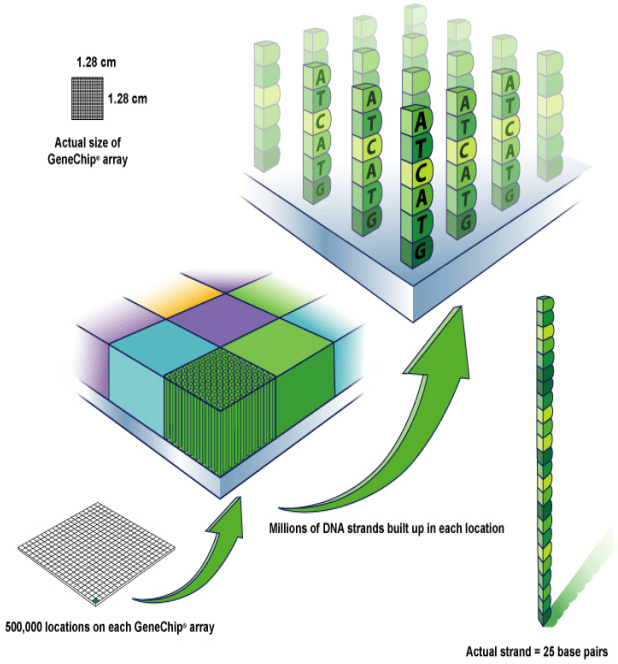
- DNA fragments contain sequencing errors
- Two complement strands of DNA
  - Need to consider that DNA comes in two parts
  - Don't want to confuse one side of the chain with the other when we reassemble
- Repeat problem
  - 50% of human DNA is just repeats
  - If you have repeating DNA, how do you know where it goes? How many copies?

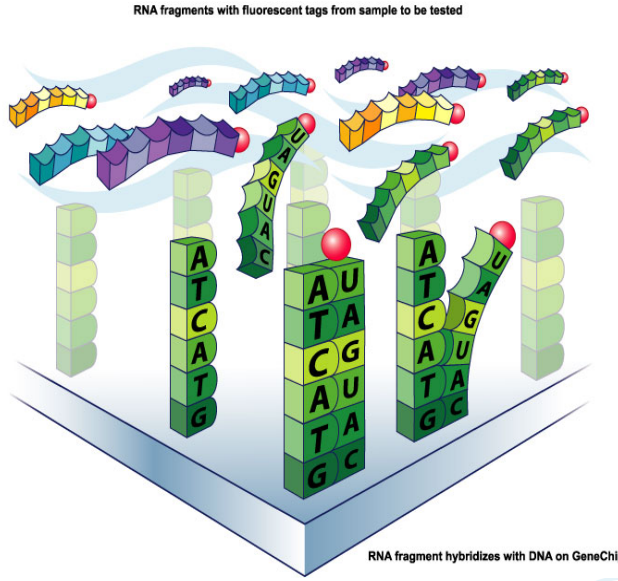
DNA probes

- Oligonucleotides: short single-strands of DNA usually 20-30 nucleotides long
- Oligonucleotides can be used to find complementary DNA segments, via hybridization
- Also used to determine the concentration of specific mRNA strands within a cell (gene expression)
- Made by automated DNA synthesizers and tagged with a radioactive isotope.
- Anchored onto glass slides

DNA Microarray







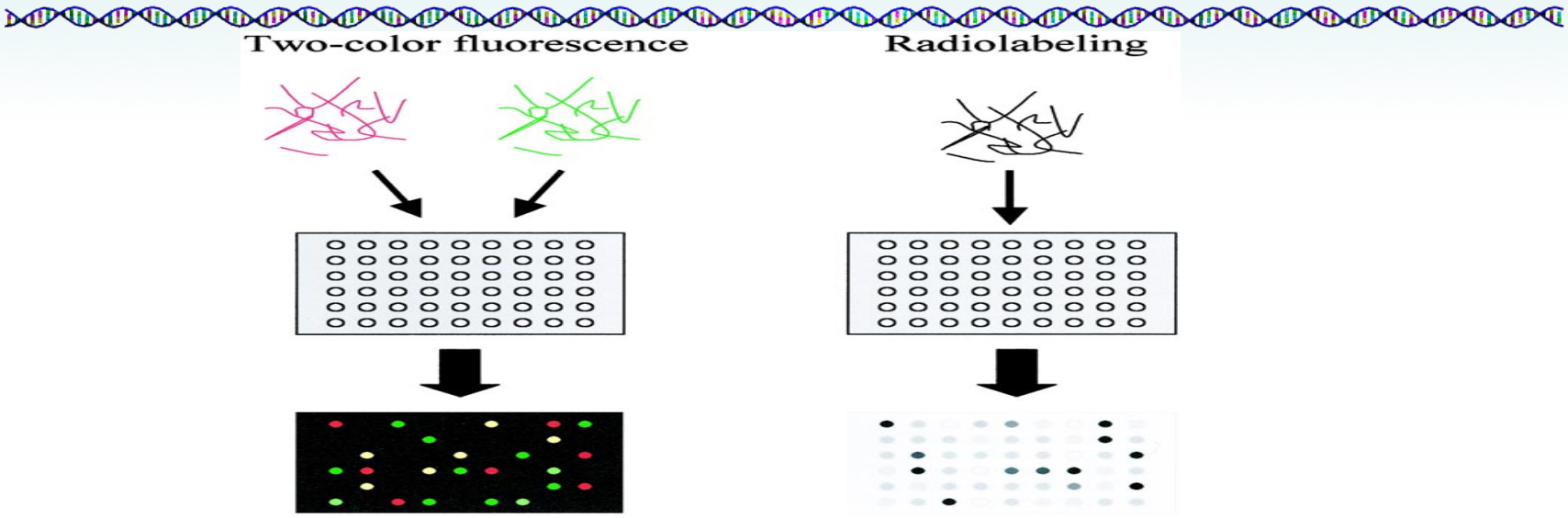
Millions of DNA strands build up on each location.

Tagged probes become hybridized to the DNA chip's microarray.

Gene Expression Arrays

- An array works by exploiting the ability of a given mRNA molecule to hybridize to the DNA template.
- Using an array containing many DNA samples in an experiment, the expression levels of hundreds or thousands genes within a cell by measuring the amount of mRNA bound to each site on the array.
- With the aid of a computer, the amount of mRNA bound to the spots on the microarray is precisely measured, generating a profile of gene expression in the cell.

Labeling DNA arrays



RNA samples are labeled using fluorescent nucleotides (***left***) or radioactive nucleotides (***right***), and hybridized to arrays. For fluorescent labeling, two or more samples labeled with differently colored fluorescent markers are hybridized to an array. Level of RNA for each gene in the sample is measured as intensity of fluorescence or radioactivity binding to the specific spot. With fluorescence labeling, relative levels of expressed genes in two samples can be directly compared with a single array.

An experiment on a microarray

In this schematic:

GREEN represents **Control DNA**

RED represents **Sample DNA**

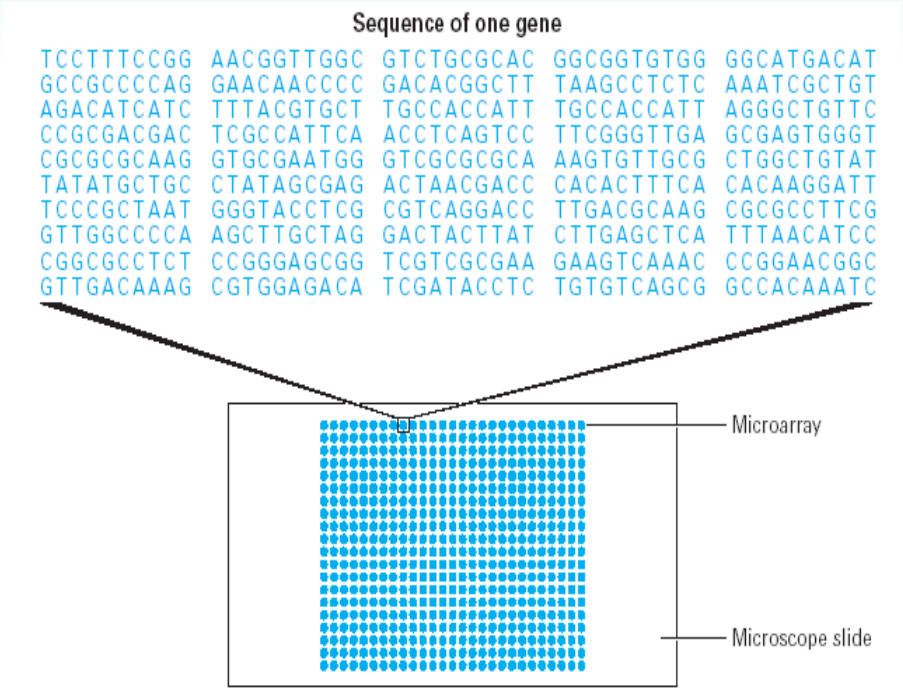
YELLOW represents **a combination of Control and Sample DNA**

BLACK represents areas where **neither the Control nor Sample DNA**

Each color in an array represents either healthy (control) or diseased (sample) tissue. The location and intensity of a color tell us whether the gene, or mutation, is present in the control and/or sample DNA.

DNA Microarray

Affymetrix

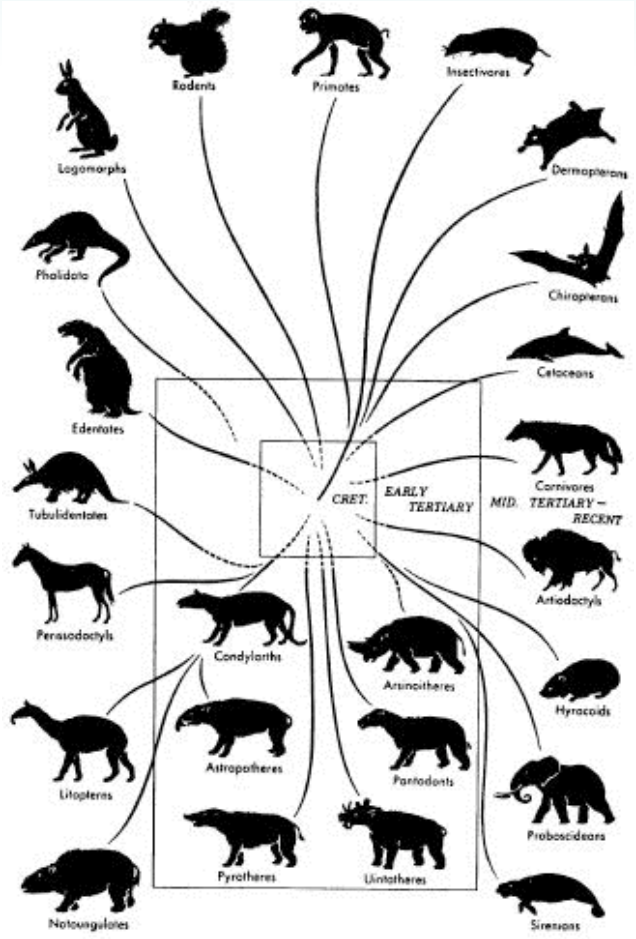


Microarray is a tool for measuring and analyzing DNA fragments. It consists of a glass slide, with template DNA patterns anchored to it. (25-64 bps)

Each blue spot indicates the location of a PCR product. On a real microarray, each spot is about 100um in diameter.

The Diversity of Life

- Not only do different species have different genomes, but also different individuals of the same species have different genomes.
- No two individuals of a species are quite the same – this is clear in humans but is also true in every other sexually reproducing species.
- Imagine the difficulty of biologists – sequencing and studying only one genome is not enough because every individual is genetically different!



Relationships of the orders of placental mammals.

How do Individuals Differ?

- Genetic makeup of an individual is manifested in traits, which are caused by variations in genes
- While only 0.1% of the 3 billion nucleotides in the human genome differ, these small variations can affect a large range of phenotypic expressions
- Traits make some more or less susceptible to disease, and the demystification of these mutations will hopefully reveal the links between genetics and susceptibility to diseases
- Genetics also determines how effective a particular disease treatment might be.

Physical Traits and Variances

- Individual variation within a species occurs in populations of all sexually reproducing organisms.
- Individual variations range from hair and eye color to less subtle traits such as susceptibility to malaria.
- Physical variation is the reason we can pick out our friends in a crowd, however most physical traits are difficult to detect at a cellular and molecular level.

Sources of Physical Variation

- Physical Variation and the manifestation of traits are caused by variations in the genes and environmental influences.
- An example is height, which is dependent on genes as well as the nutrition of the individual.
- Not all variation is inheritable – only genetic variation can be passed to offspring.
- Separating genetic and environmental effects and interactions is one of the major challenges of biology.
- Both sorts of interactions can be complex…
  - Interactions of multiple genes
  - Order of exposures

Genetic Variation

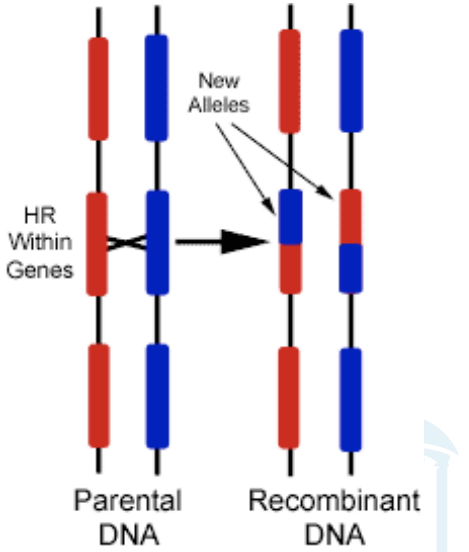
- Despite the wide range of physical variation, the genomic differences between two individuals is small.
- Out of 3 billion nucleotides, only roughly 3 million base pairs (0.1%) are different between individual genomes of humans.
- Although there is a finite number of possible variations, the number is so high ($4^{3,000,000}$) that one can safely assume that no two individuals have the same genome by chance.
- Related individuals have more similar genomes
- What is the cause of this genetic variation?

Sources of Genetic Variation

- ***Mutations*** are rare changes to genomic sequence.
- Common mutations (in order of frequency)
  - Bases in the sequence change (polymorphism)
  - A subsequence of bases is removed (deletions)
  - A subsequence of bases is repeated (copy number variation)
  - A subsequence changes position (transposition)
  - A subsequence is reversed (inversion)
- Most mutations do not create beneficial changes and actually kill the individual.
- Although mutations are the source of all new genes in a population, they are so rare that there must be another process at work to account for the large amount of diversity.

Sources of Genetic Variation

- ***Recombination*** is the shuffling of genes that occurs through sexual mating and is the main source of genetic variation.
- In mammals, and other higher organisms, there are two complete copies of the DNA sequence (recall this is the source of the "dormant," "hidden," or "recessive" traits discussed last lecture.
- During meiosis (germ-line cell division) these two DNA sequences can exchange subsequences
- This brings together differ gene-variant combinations rather modifying genes as in mutation.



How Do Different Species Differ?

- As many as 99% of human genes are conserved across all mammals based in terms of function, with >85% having high sequence similarity (>95% bp agreement)
- The functionality of many genes is virtually the same among many organisms
- It is highly unlikely that the same gene with the same function would spontaneously develop among all currently living species
- The theory of evolution suggests all living things evolved from incremental change over millions of years

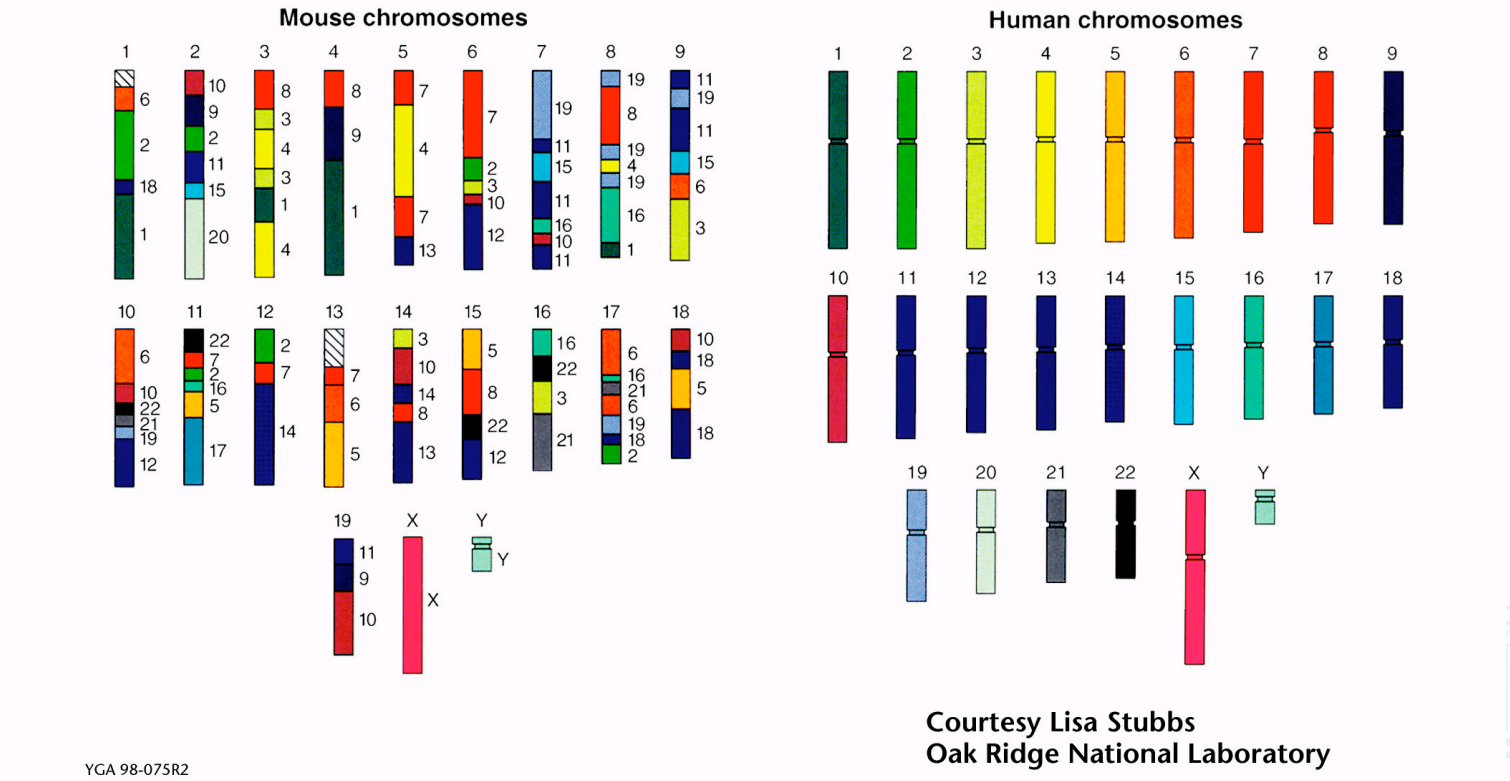
Mouse and Human overview

- Mouse genome has 2.7×10^9 base pairs versus 3.1×10^9 in human.
- About 95% of genetic material is shared.
- 99% of genes shared of about 30,000 total.
- The 300 genes that have no homologue in either species deal largely with immunity, detoxification, smell, and sex*

*Scientific American Dec. 5, 2002

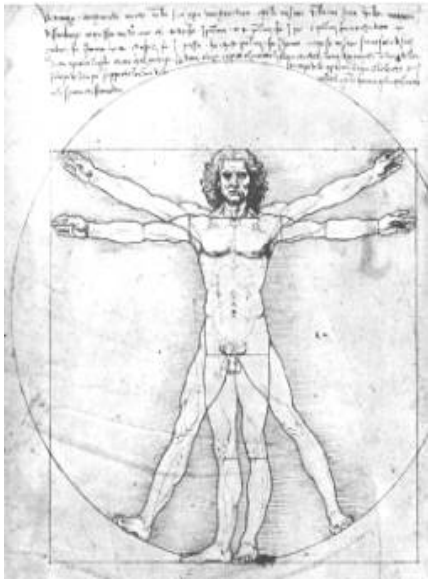
Genome Structure

Significant chromosomal rearranging occurred between the diverging point of humans and mice.



Comparative Genomics

- What can be done comparatively between Human and Mouse Genomes?
- One possibility is to create "knockout" mice – mice lacking one or more genes. Study the phenotypes of these mice gives predictions that gene's function in both mice and humans.



Summary

- DNA sequences are extracted by first cutting up the DNA into small segments using *restriction enzymes*
- Many copies of a target DNA can be generated by either cloning or Polymerase Chain Reaction
- Individual bases can be read from DNA using sequencing techniques or by DNA microarrays
- Bioinformatics algorithms attempt to reconstruct the original sequence from the fragements

Next Time

- Concentrate on computer science
  - Algorithms & Complexity
  - Correctness and efficiency
  - Space and time complexity
  - Algorithmic strategies
  - Tractable vs. intractable problems