Sequence Assembly

Task and Technologies

Outline

- Defining the sequence assembly task
- Technologies for sequencing DNA

The sequencing problem

- We want to determine the sequence of the bases that make up:
  - A single large molecule of DNA
  - The genome of a single cell
  - The genome of an individual organism
  - The genome of a species

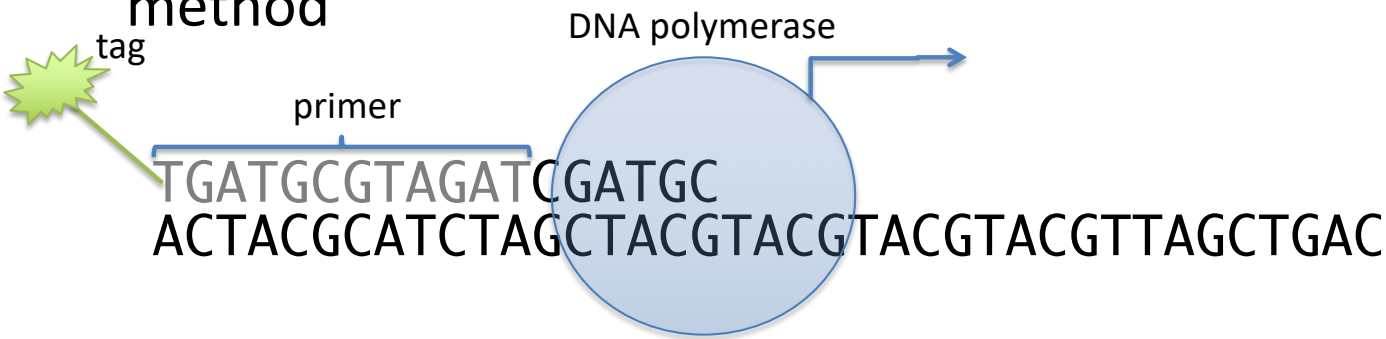
- But we can't (currently) "read" off the sequence of an entire molecule all at once

The strategy: substrings

- We *do* have the ability to read or detect *short* pieces (substrings) of DNA
  - Sanger sequencing: 500-700 bp/read
  - Hybridization arrays: 8-30bp/probe
  - Latest technologies:
    - Short read technologies
      - Illumina Genome Analyzer: 35-300 bp/read
    - Long read technologies
      - Pacific Biosciences: ~10,000 bp/read
      - Oxford Nanopore: variable, 10,000 - 100,000 bp /read

Sanger sequencing

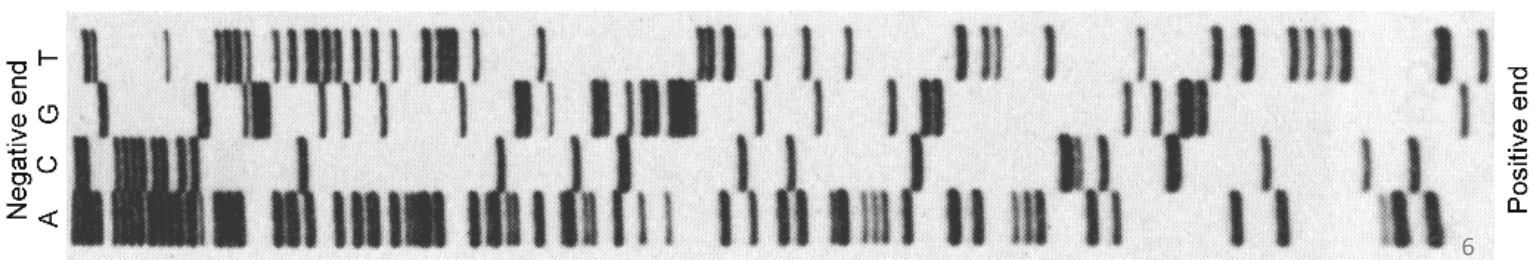
- Classic sequencing technique: “Chain-termination method”



- Replication terminated by inclusion of dideoxynucleotide (ddNTP)

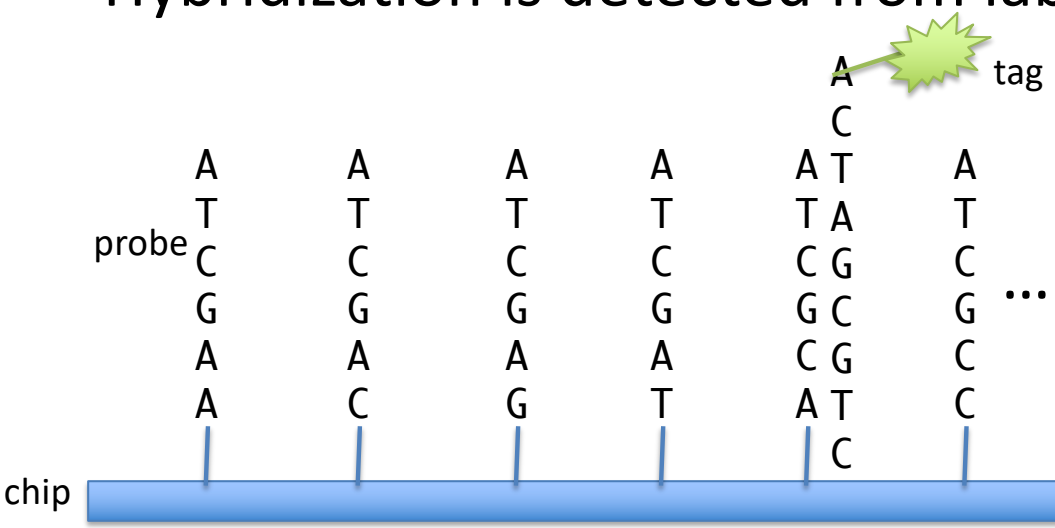
Sequencing gels

- Run replication in four separate test tubes
  - Each with one of some concentration of either ddATP, ddTTP, ddGTP, or ddCTP
- Depending on when ddNTP is included, different length fragments are synthesized
- Fragments separated by length with electrophoresis gel
- Sequence can be read from bands on gel

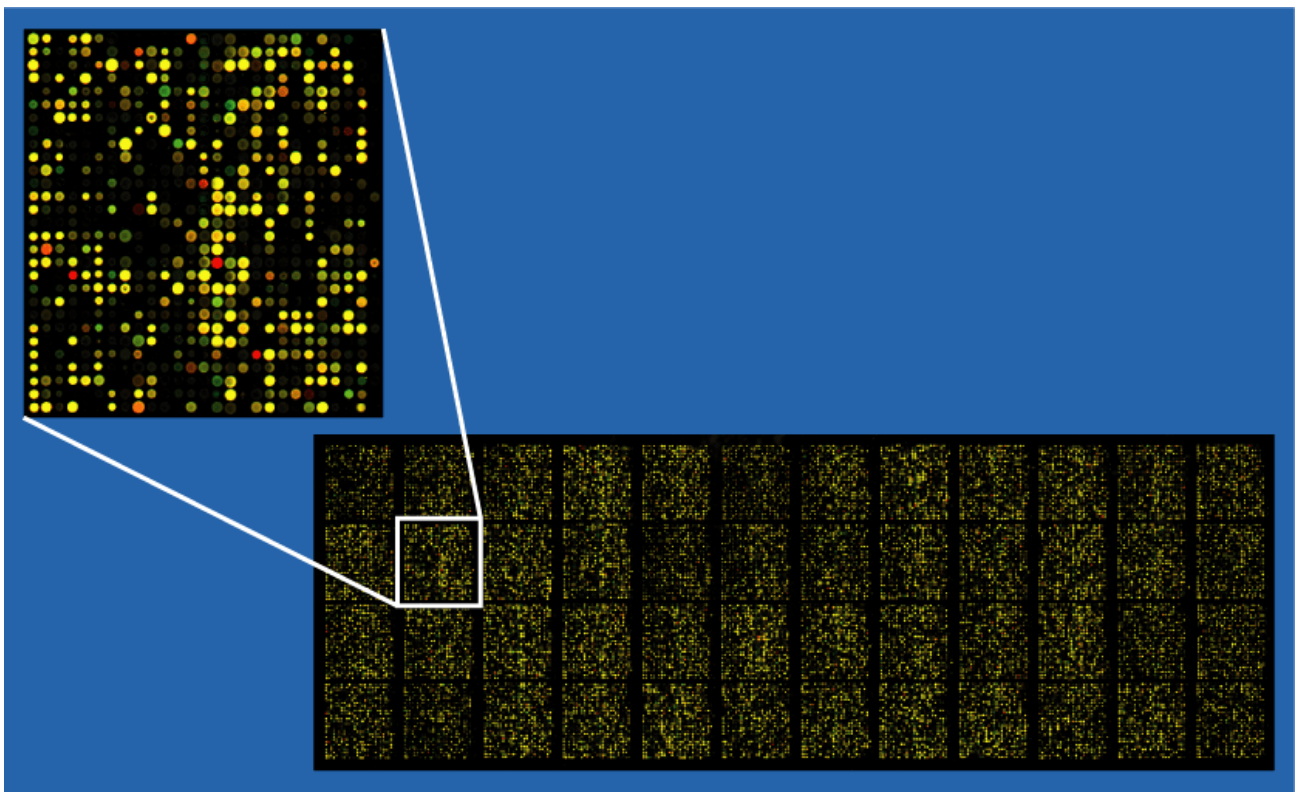


Universal DNA arrays

- Array with all possible oligonucleotides (short DNA sequence) of a certain length as probes
- Sample is labeled and then washed over array
- Hybridization is detected from labels



Reading a DNA array



Latest technologies

- Illumina
  - “sequencing by synthesis”
  - ~100 Gb/day on one machine
  - Uses fluorescently-labeled reversible nucleotide terminators
  - Like Sanger, but detects added nucleotides with laser after each step

Latest technologies

- Pacific Biosciences:
  - “Sequencing by synthesis”
  - Single molecule sequencing
  - Detects addition of single fluorescently-labeled nucleotides by an immobilized DNA polymerase
  - Real-time: reads bases at the rate of DNA polymerase
  - 4 hours for sequencing with reads up to 60kb long

Oxford Nanopore

- Emerging technology
- Pocket-sized
- High error rate
- Currently in “community” program

Shotgun Sequencing Fragment Assembly

Multiple copies of sample DNA

Randomly fragment DNA

Sequence sample of fragments

Assemble reads