PCR:
- Double stranded DNA heat, come apart, cool down, primer specific to parts of DNA match, polymerase start replicating.
- One strand gives two copies of same DNA between primers.
- Need to have unique primers.

* Gel Electrophoresis:
  - DNA backbone negatively charged, put in electric field, migrate toward positive pole
  - Get separation of nucleotides.
  - Base for DNA sequencing (Sanger)
  - OH group on nucleotide triphosphate of DNA
  - Bridge to next phosphate group to create sequence.

* Fluorescent Sequencing:
  - Suppose primer at one end, polymerize coping. If a small fraction of the pool of DNA bases are "defective" (ie, the necessary OH group removed, a so-called di-deoxy nucleotide.), one of these bases happens to be inserted, chain growth will stop.
  - If furthermore, the di-deoxy nucleotides are fluorescently tagged, say all dATP's are one color, all dGTP's are another color, etc. then, can determine where the A's & G's are in the sequence by examining distance between colors via gel electrophoresis.
  - Trace of color for nucleotides at different positions give nucleotide sequence.
  - A color sequence fluorescent → gel electrophoresis separate by size.
  - Advantage: fast and cheap!
  - Question: gel electrophoresis can't tell how many nucleotides in sequence.
  - Answer: can make it longer to get greater resolution, use different gel materials.
  - Problem: can't do entire chromosome by fluorescent because resolution gets bad at greater lengths.
  - Solution: "Whole Genome Shotgun".

* Whole Genome Shotgun:
  - Cut genome into pieces (A, B, C, D, etc.), clone them into organisms (ie, bacteria stable in lab), fragments can be picked back together according to similarity of sequences.
  - Can read 600 nucleotides individually in one run.
  - Problem: not 100% accurate.
  - Fig. 5.3: in a row can get a stretch of peaks, hard to identify exactly how many s's.
  - Conjecture do they fold up? Sequencing artifacts different if fold vs. unfold.

---

- Goal: find value of theta that maximizes likelihood of observed data.
  - eg:1: Height of sample in room is 5ft, if there is massive defect, then likelihood low.
  - eg:2: Coin flip:
    - flip 1000 times, head turns up 642 times. If assumed then prob(head)=0.5, then NOT likely to maximize likelihood of data. If theta=.642, then more likely.
  - Need to find theta that maximizes L, the likelihood function (smooth function of theta).
  - How to maximize L?
    - Take derivatives of L with respect to theta and set equal zero is point that maximizes L (or minimizes it...)
    - Problem: differentiating L, big product is a mess.
    - Solution: take differentiation of log L (log-likelihood), which is derivative of summation (much easier to work with).
  - Warning: max/min of logL can be on boundary, so need to verify that it is indeed max.
- $e^{\theta \cdot X}$ normally distributed, know variance, mean is parameter.
- Convention: use theta to represent the theta that maximizes likelihood.
- $e^{\theta \cdot S_i}$ normally distributed, both variance and mean unknown parameter.
- Take partial derivative w/theta and then2 separately.
- Result: then $2 \cdot \text{sample mean}$, then $2 \cdot \text{sample variance}$.
- Problem: then2 is consistent but biased.
- If prob mean in the middle, highly unlikely that two samples will have the same distance away from mean. The sample mean is an unbiased estimate of the population mean (e.g., it’s equally likely to be too large as to be too small), but its placement in the middle of the sample will tend to underestimate the variance.
- In extreme where sample size is 1, then theta still makes sense as an estimate of population mean, but variance estimate is zero.
- Important: MLE is great, but not perfect. (of course), need to choose method appropriate for individual need.
- Goal: have measured data → have model → likelihood of data → choose parameter (theta) that maximizes likelihood.
- A more complex example:
  - Mixture of two normal distributions
  - Not good assume one distribution among all data
  - Remem: have multiple distributions, identify which point came from which distribution.
- Might be hidden variables and need to separate them.
  - Sample separated but don’t know which came from which. Start by guessing, then re-estimate.
- More next time...