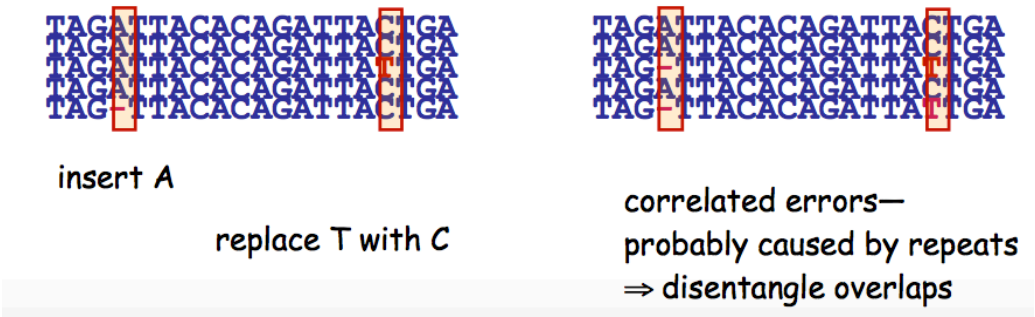
Lecture 11: DNA Sequencing

1. Coverage
    a. If G is the length of the genomic segment, N is the number of reads, and L is the length of each read, Coverage C = N*L/G
    b. How much coverage do you need?
        i. Lander-Waterman model: Probability[a bp not covered] = e^{-C}
2. Repeats
    a. Make it very difficult to sequence the genome, especially when repeats are longer than read length
        i. Repeats with more base pair differences than sequencing error rate are ok – can just put overlaps between 2 reads in different copies of repeat
    b. Types
        i. Low-Complexity DNA (ATATATAT…)
        ii. Microsatellite repeats (CAGCAGTAGCAGCACCAG)
            1. Expand quickly from generation to generation
        iii. Transposons – hijack molecular machinery of cell to copy self many times
        iv. Gene families
            1. Genes duplicate, then change
            2. One of the key ways new genes are formed
        v. Recent Duplications – about 100,000 long
    c. Why so many repeats?
        i. Short repeats created mainly by slippage of DNA polymerase as it copies DNA
            1. As DNA opens and closes for copying, one strand can close with a small loop
        ii. Long repeats happen because of large loops forming in DNA during replication – DNA sticks to itself in a loop and you get extra copies of a given region
    d. 50% of human DNA is composed of repeats!
        i. So as you assemble DNA in sequencing, you'll often mismatch reads and sequence genome incorrectly
    e. How do we deal with repeats in sequencing?
        i. Label reads into clusters depending on where they come from in the genome
            1. Powerful if technology exists that can do it (not much that can)
        ii. Link the reads
            1. If we know that a repeat occurs between A and B and the same repeat is between C and D, then we can link A to B and not accidentally mismatch it to D
3. Fragment Assembly
    a. Given paired reads, how do we put together the genome?

i. Human genome is $3*10^9$ base pairs, and we want 30x repetitiveness, so we have about 100 billion base pairs total

b. Terminology
   i. Reads now are around 150 bp
   ii. Mate pair – pair of reads from two ends of same fragment
   iii. Contig – formed from several overlapping sequences
   iv. Supercontig – sequence of contigs
   v. Consensus – total sequence

4. Steps to assemble a genome
   a. Find all possible overlapping reads
      i. Limit insertions, deletions, substitutions to a small number (say 3-5)
      ii. Take every word of a given length in one read and sort them alphabetically
      iii. Now can read off common words between pairs of reads (they'll be right next to each other alphabetically)
         1. Only look at words that overlap in a max of k reads (so you don't bog down your algorithm looking at repeated strands)
      iv. Then extend those common words to see if their reads could overlap overall, keep it if it works and drop it otherwise
         1. At this stage, we can correct errors using multiple alignment:



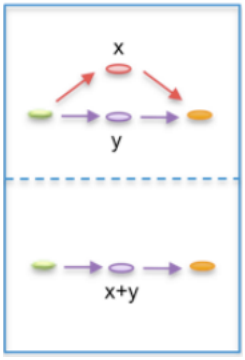
   b. Merge some of the good overlaps into longer configurations
      i. Create an overlap graph (you'll see overlaps of reads from different regions of the genome that have the same repeat)
         1. Will be really complex!
         2. Can merge repeat regions from two separate parts of genome, but don't want to merge outside of repeat boundaries
            a. Can identify end of repeats because there are two incoming edges and one outgoing edge to the repeat region of the graph
      ii. Remove transitively inferable overlaps – if r_1 overlaps r_2 overlaps r_3, we can remove r_2 and align r_1 to r_3

iii. Now merge subgraphs with only one edge between nodes

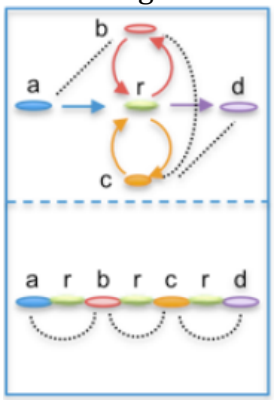
c. Link longer configurations to form super-configurations
   i. Have two types of contigs now – unique ones and overcollapsed ones
      1. Overcollapsed are from multiple regions of the genome
      2. Can tell the difference between the two based on the number of repeats you see of the contig
   ii. Use pairing to connect contigs, then fill gaps in supercontigs with repeat contigs
   iii. Steps are using algorithms we've previously discussed in the course

d. Derive total consensus sequence
   i. Derive multiple alignments from supercontigs, then using weighting find the best possible alignment

5. De Brujin Graph formulation
   a. Given sequence $x_1 \ldots x_n$, k-mer length k, graph of 4^k vertices with edges between words with k-1 length overlap
   b. Can process this graph to create sequence
      i. Compression – long line of k-mers pointing to each other is a long overlap – can compress this into a longer node
      ii. Error detection - pop the "bubble" containing x



      iii. Repeat analysis – in places where you see lots of repeats at one node with two paths out of that node, can split graph from one node into two nodes
      iv. Scaffolding – connect nodes in ambiguous structures:



6. Quality of assemblies

a. Mouse genome sequencing cost $300 million!
b. Why sequence mouse genome?
    i. 66% similarity to human genome in random regions, but more similar to human genome in gene regions, so we could use the combination of mouse and human genomes to find genes in the human genome
c. N50 contig length – sort the contigs from largest to smallest, and start at largest to cover genome in that order, N50 is the length of the contig that covers 50% of the genome
    i. For mouse it was 26 kilobases, supercontig was 50 megabases
d. Panda genome
    i. BGI did sequencing for panda genome – currently the biggest sequencing facility in the world
    ii. First genome to be done with Illumina technology – read lengths were only 25 base pairs!

7. Assembly Now and Then
    a. There are a huge variety of assembly algorithms used today, with no standard solution that works for any problem
    b. Assemblathon compared different assembly technologies available today – comparing quality to genome where they knew the correct answer
    c. In 1998, Weber and Myers suggested sequencing the human genome with the shotgun strategy – thought it would be faster
        i. Phil Green reviewed their paper and blasted it
        ii. Genome Research published both the paper itself and Green's critique in the same issue
        iii. A private company (Celera) took on the sequencing problem using the shotgun strategy, because the NIH wasn't using that strategy and was going too slowly
        iv. Now the NIH had to rush to sequence the genome faster
        v. Both the NIH and Celera finished in 2000
        vi. Celera's entry into the race accelerated genomic research by 5 or 10 years!