Computational Genomics

10-810/02-710, Spring 2009

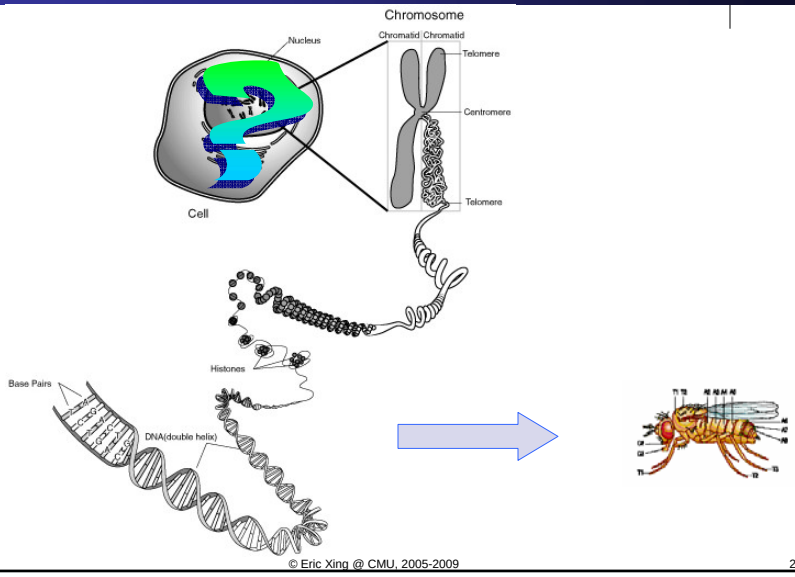
DNA sequencing and genome assembly

Eric Xing

Lecture 3, January 21, 2009

Reading: class assignment

DECODING the Genome



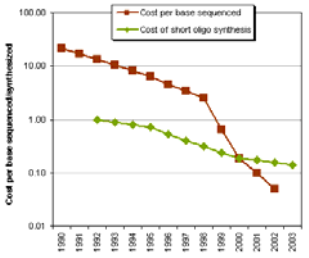
What & Why?

- **"Sequencing"** means finding the order of nucleotides on a piece of DNA
- Nucleotide order determines
  - amino acid composition, and by extension, protein structure and function (proteomics)
  - Transcription factor binding sites and their organizations, and thereby, gene expression regulation
  - …
- An alteration in a DNA sequence can lead to
  - altered regulatory program,
  - altered protein structure/function,
  - and therefore, phenotypes or harmful effects in a plant or animal
- Understanding a particular DNA sequence can
  - shed light on a genetic condition and offer hope for the eventual development of treatment of diseases
  - environmental, agricultural applications
  - forensic applications
  - …

DNA Sequencing – Overview

- Chain-termination methods (a.k.a. Dideox termination method)
  - Predominant method based on Sanger' (Nobel price in Chemistry, 1980, his sec Do you know how he got his first NP?)
  - Key ideas:
    - Terminate elongation via dideoxynucleoti
    - Size separation via Gel electrophoresis (now Capillary electrophoresis)

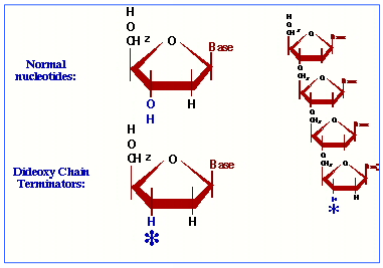
1975

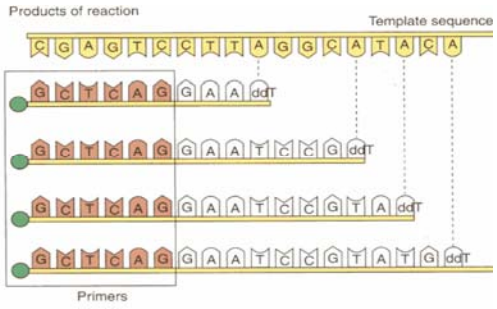


- Whole genome strategies
  - Physical mapping
  - Walking

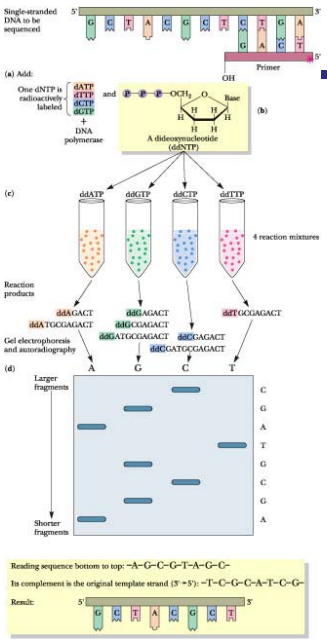
2015

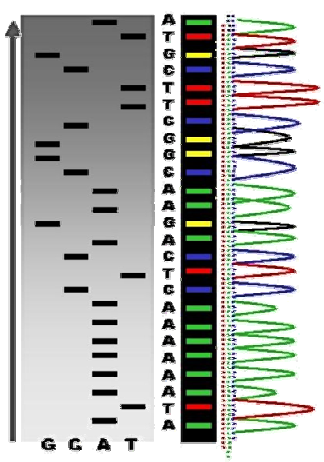
Dideoxy termination method

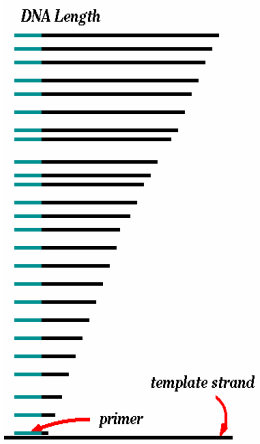




Dideoxy termination method







DNA Sequencing

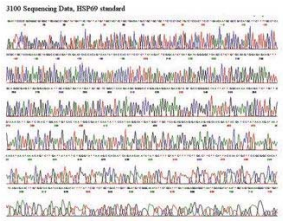
- Goal:

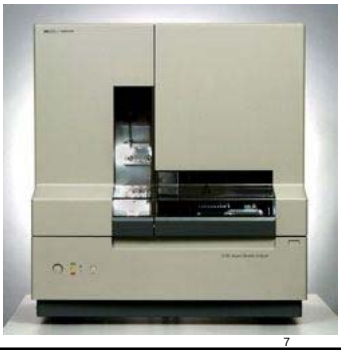
  Find the complete sequence of A, C, G, T's in DNA, at a Genome scale

- Challenge:

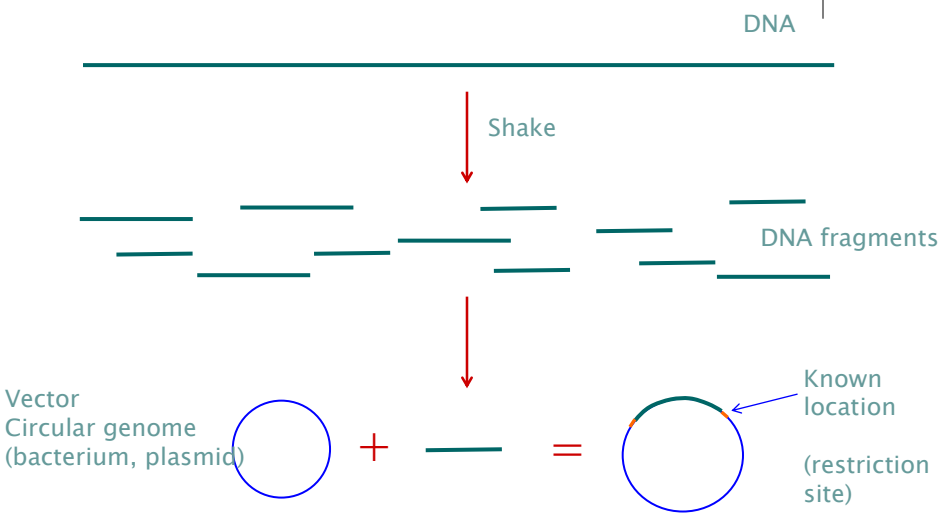
  There is no machine that takes long DNA as an input, and gives the complete sequence as output

- Can only sequence ~500 letters at a time



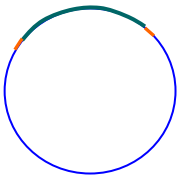


DNA Sequencing – vectors



Different types of vectors

VECTOR	Size of insert
Plasmid	2,000-10,000 Can control the size
Cosmid	40,000
BAC (Bacterial Artificial Chromosome)	70,000-300,000
YAC (Yeast Artificial Chromosome)	> 300,000 Not used much recently



Different vector can be used to carry DNA pieces of different lengths

Method to sequence longer genome regions

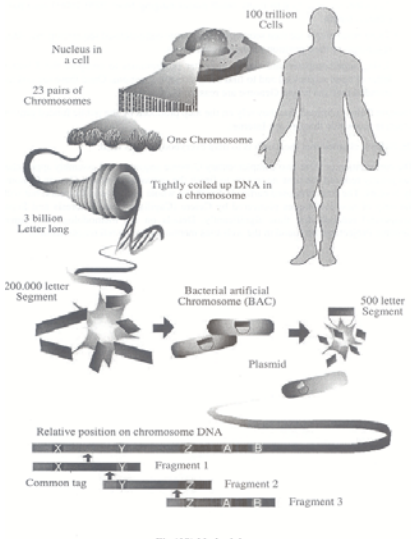
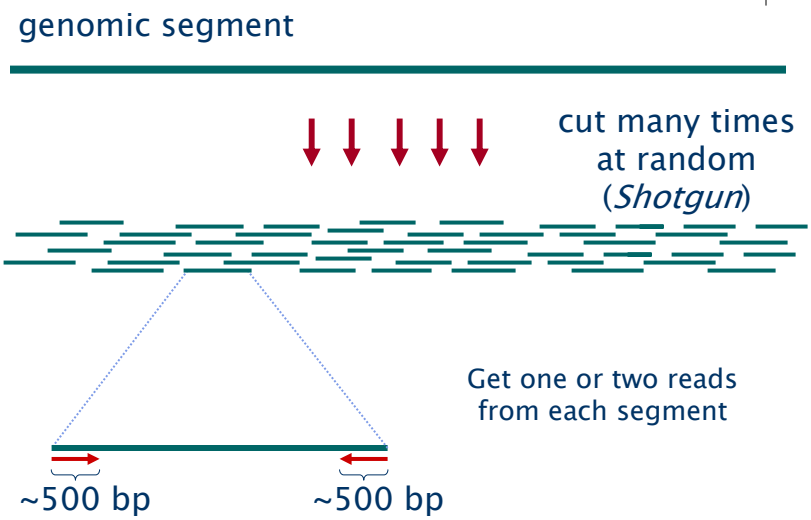
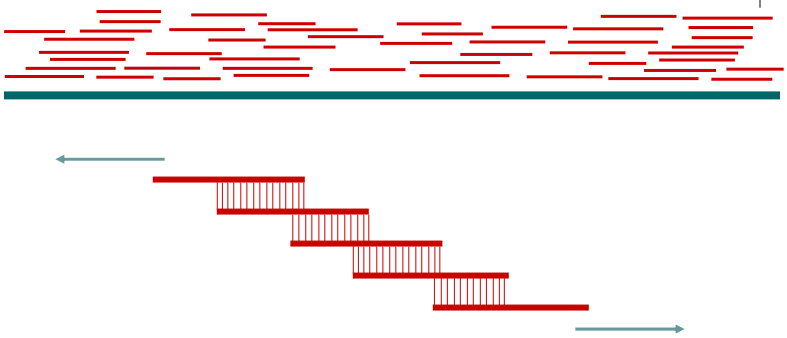


Fig (18) Methodology

Method to sequence longer genome regions



Reconstructing the Sequence (Fragment Assembly)



Cover region with ~7-fold redundancy (7X)

Overlap reads and extend to reconstruct the original genomic region

Definition of Coverage

Length of genomic segment: **L**
Number of reads: **n**
Length of each read: **l**

Definition: Coverage **C = n l / L**

How much coverage is enough?

Lander-Waterman model:
Assuming uniform distribution of reads, C=10 results in 1 gapped region /1,000,000 nucleotides

Are we done?

Difficulties

- How to fragment?
- How to efficiently find overlaps?
- What if there are **repeats** in the genome?

Repeats

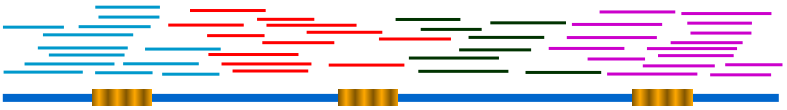
- **Low-Complexity DNA** (e.g. ATATATATACATA…)
- **Microsatellite repeats** $(a_1 \ldots a_k)^N$ where k ~ 3-6
  (e.g. CAGCAGTAGCAGCACCAG)
- **Transposons**
  - **SINE** (Short Interspersed Nuclear Elements), e.g., ALU: ~300-long, 10^6 copies
  - **LINE** (Long Interspersed Nuclear Elements), ~4000-long, 200,000 copies
  - **LTR retroposons** (Long Terminal Repeats (~700 bp) at each end), cousins of HIV
- **Gene Families** genes duplicate & then diverge (paralogs)
- **Recent duplications** ~100,000-long, very similar copies

Bacterial genomes: 5%
Mammals: 50%

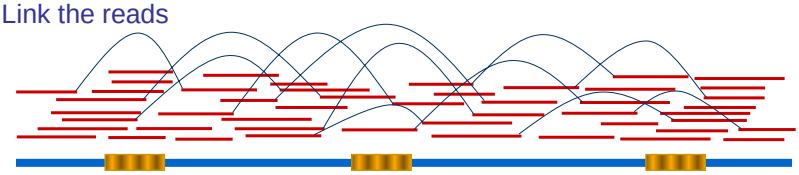
What can we do about repeats?

Two main approaches:

- Cluster the reads

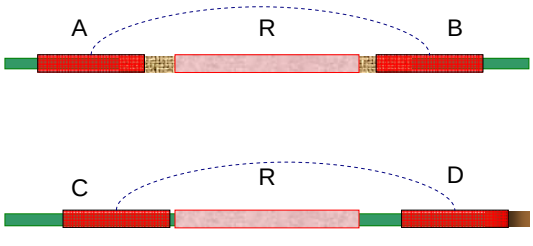


- Link the reads



Sequencing and Fragment Assembly

$3x10^9$ nucleotides



ARB, CRD

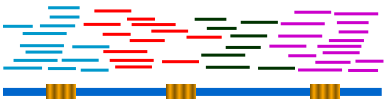
or

ARD, CRB ?

Strategies for whole-genome sequencing

1. Hierarchical – Clone-by-clone
   i. Break genome into many long pieces
   ii. Map each long piece onto the genome
   iii. Sequence each piece with shotgun

   Example: Yeast, Worm, Human, Rat



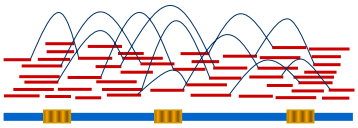
2. Online version of (1) – Walking
   i. Break genome into many long pieces
   ii. Start sequencing each piece with shotgun
   iii. Construct map as you go

   Example: Rice genome

3. Whole genome shotgun

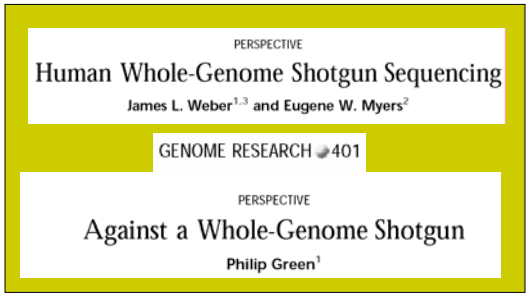
   One large shotgun pass on the whole genome

   Example: Drosophila, Human (Celera), Neurospora, Mouse, Rat, Dog

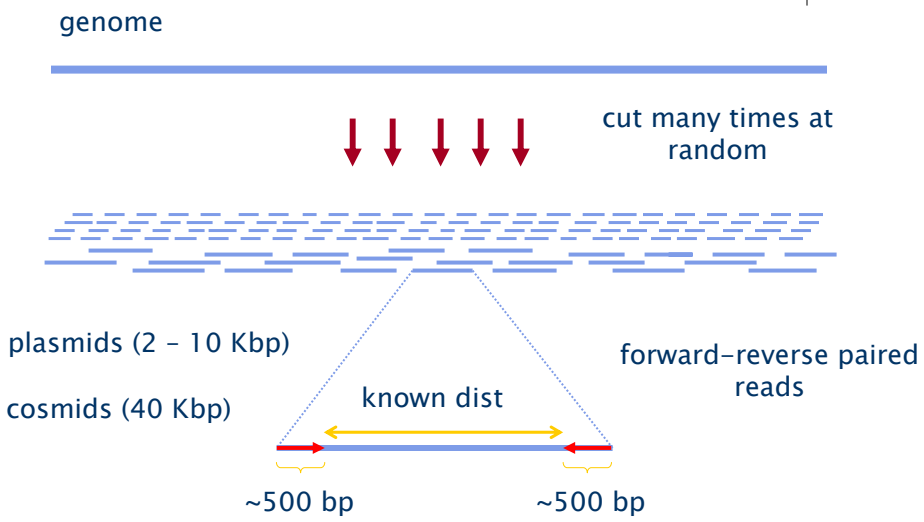


Whole Genome Shotgun

- Break up the entire genome into pieces
- Sequence ends, and assemble using a computer
- LW statistics & Repeats argue against the success of such an approach



Whole Genome Shotgun Sequencing



Steps to Assemble a Genome

Some Terminology

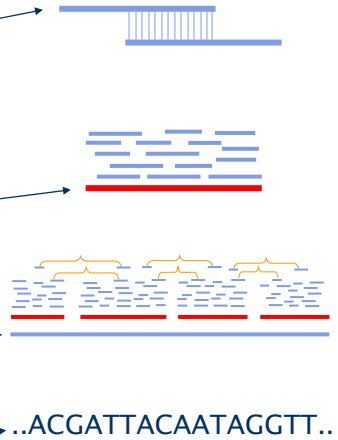
read a 500-900 long word that comes out of sequencer

mate pair a pair of reads from two ends of the same insert fragment

contig a contiguous sequence formed by several overlapping reads with no gaps

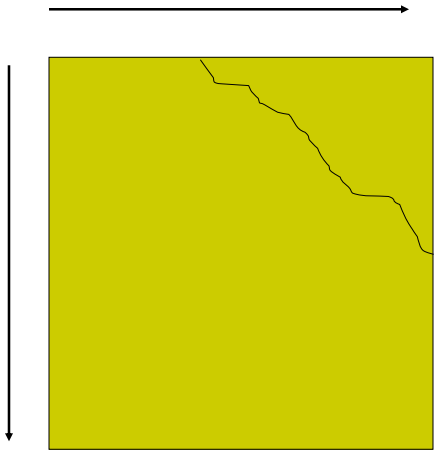
supercontig (scaffold) an ordered and oriented set of contigs, usually by mate pairs

consensus sequene sequence derived from the multiple alignment of reads in a contig



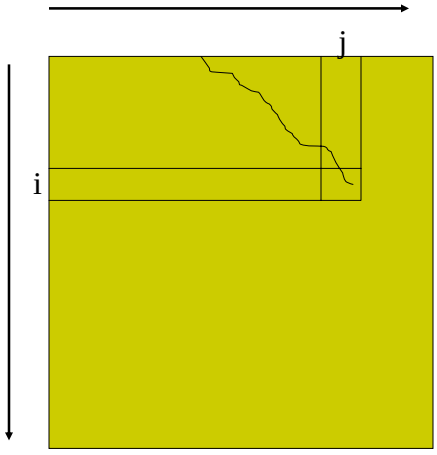
1. Find Overlapping Reads

- Given a pair of fragments s_1 and s_2, do they belong together?
  - Yes, if a prefix of s_2 matches a suffix of s_1
  - How would you compute such a match?



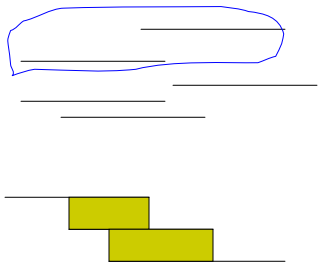
1. Find Overlapping Reads (cont'd)

- S[i,j] = optimum score of an alignment of $s_1[1..i]$ against a suffix of $s_2[1..j]$
  - The best prefix-suffix alignment is given by:
  - $\text{Max}_i \{S[i,n] \}$



1. Find Overlapping Reads (cont'd)

- Compute the best prefix-suffix alignments between each pair of fragments.
- Keep the "high-scoring" ones as evidence of true overlap.
- What is the problem?



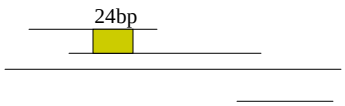
1. Find Overlapping Reads (cont'd)

- Consider the number of fragments. The LW statistics say that we need good coverage (c=8, 10) to get most of the base-pairs.
  - G = 3000Mb, L=500
  - Coverage LN/G = 10
  - N = $10*3*10^9/500 = 6*10^7$
  - Number of comparisons needed = $3.6 * 10^{15}$
    - Not good! (Only a small fraction are true overlaps)

1. Find Overlapping Reads (cont'd)

k-mer based overlap

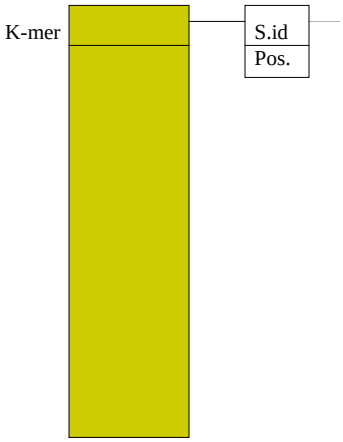
- Consider a *k*-bp sequence (k ~ 24)
  - Expected number of occurrences in the genome
  - $3*10^{9}*4^{-24} = 8*10^{-6}$
- A 24-bp sequence appears is unique to the genome!
- Two overlapping sequences should share a 24-mer
- Two non-overlapping



1. Find Overlapping Reads (cont'd)

Sorting k-mers

- Build a list of k-mers that appear in the sequences and their reverse complements
- Create a record with 4 entries:
  - K-mer
  - Sequence number
  - Position in the sequence
  - Reverse complementation flag
- Sort a vector of these according to k-mer
- How many records per k-mer are expected?
- If number of records exceeds threshold, discard (why?)



1. Find Overlapping Reads (cont'd)

Alignment

- Find pairs of reads sharing a k-mer, k ~ 24
- Extend to full alignment – throw away if not >98% similar

- Coalesce k-mer hits into longer, gap-free partial alignments.
- These extended k-mer hits are saved.
- For each pair of sequences, form a directed graph.
- For each maximal path in the graph, construct an alignment.
- Refine alignment via banded DP

1. Find Overlapping Reads (cont'd)

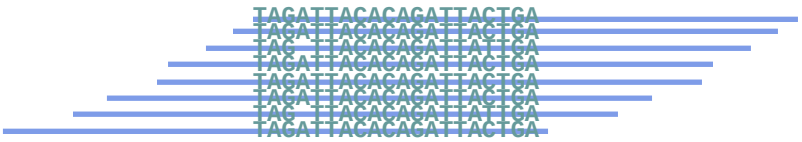
Alignment

- Find pairs of reads sharing a k-mer, k ~ 24
- Extend to full alignment – throw away if not >98% similar

- Caveat: repeats
  - A k-mer that occurs N times, causes $O(N^2)$ read/read comparisons
  - ALU k-mers could cause up to 1,000,0002 comparisons
- Solution:
  - Discard all k-mers that occur "too often"
  - Set cutoff to balance sensitivity/speed tradeoff, according to genome at hand and computing resources available

1. Find Overlapping Reads (cont'd)

Create local multiple alignments from the overlapping reads



1. Find Overlapping Reads (cont'd)

- Correct errors using multiple alignment

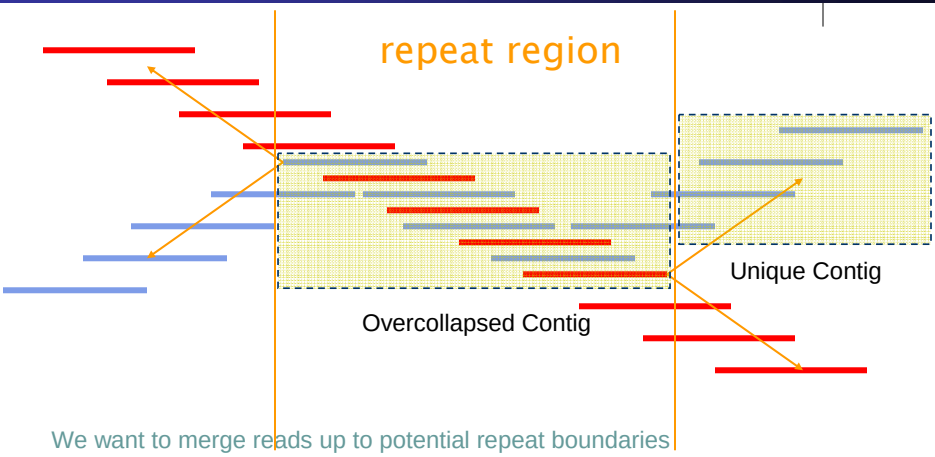
insert A

replace T with C

correlated errors—
probably caused by repeats
⇒ disentangle overlaps

In practice, error correction removes up to 98% of the errors

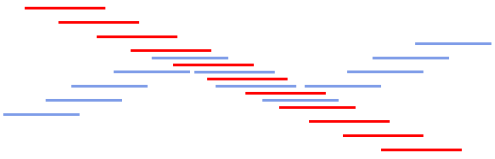
2. Merge Reads into Contigs



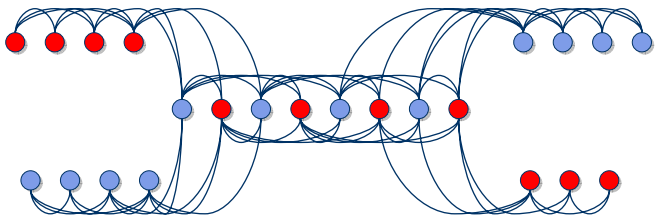
We want to merge reads up to potential repeat boundaries

2. Merge Reads into Contigs

- Overlap graph:
  - Nodes: reads r_1…..r_n
  - Edges: overlaps (r_i, r_j, shift, orientation, score)

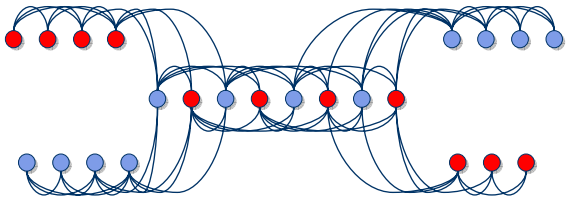


Reads that come from two regions of the genome (blue and red) that contain the same repeat

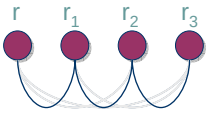


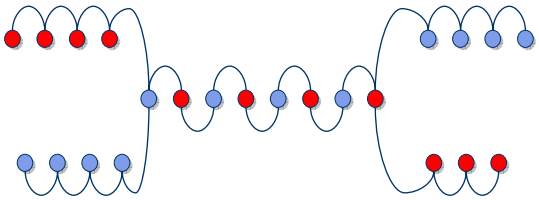
Note:
of course, we don't know the "color" of these nodes

2. Merge Reads into Contigs

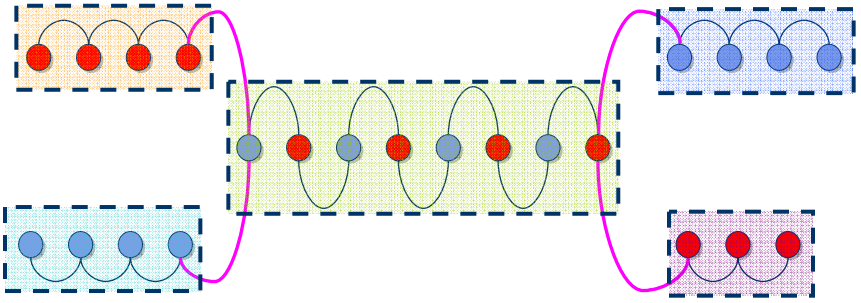


- Remove transitively inferable overlaps
  - If read r overlaps to the right reads r_1, r_2, and r_1 overlaps r_2, then (r, r_2) can be inferred by (r, r_1) and (r_1, r_2)





2. Merge Reads into Contigs



Repeats, errors, and contig lengths

- Repeats shorter than read length are easily resolved
  - Read that spans across a repeat disambiguates order of flanking regions
- Repeats with more base pair diffs than sequencing error rate are OK
  - We throw overlaps between two reads in different copies of the repeat
- To make the genome **appear** less repetitive, try to:
  - Increase read length
  - Decrease sequencing error rate

Role of error correction:

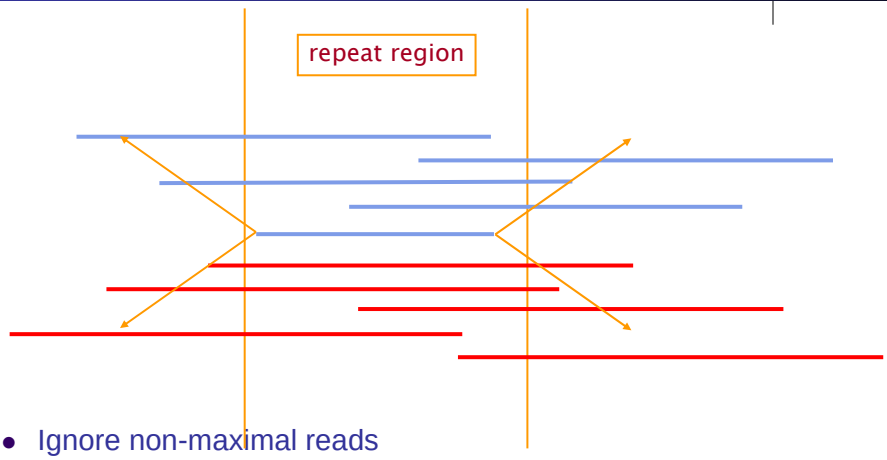
Discards up to 98% of single-letter sequencing errors

decreases error rate

⇒ decreases effective repeat content

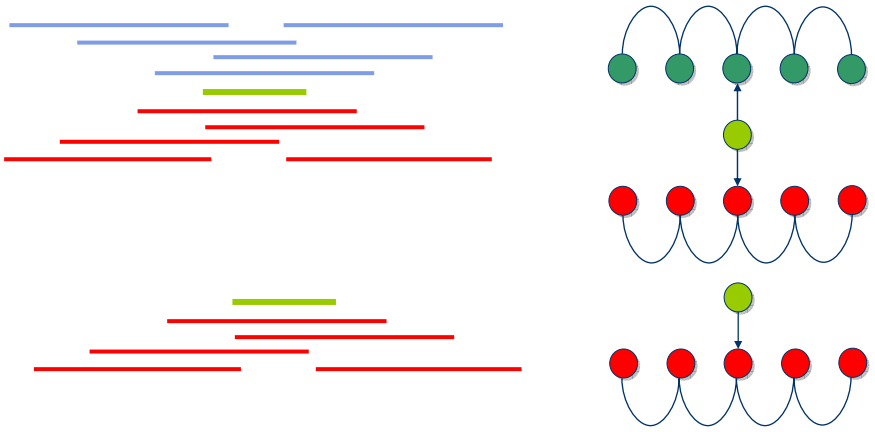
⇒ increases contig length

2. Merge Reads into Contigs



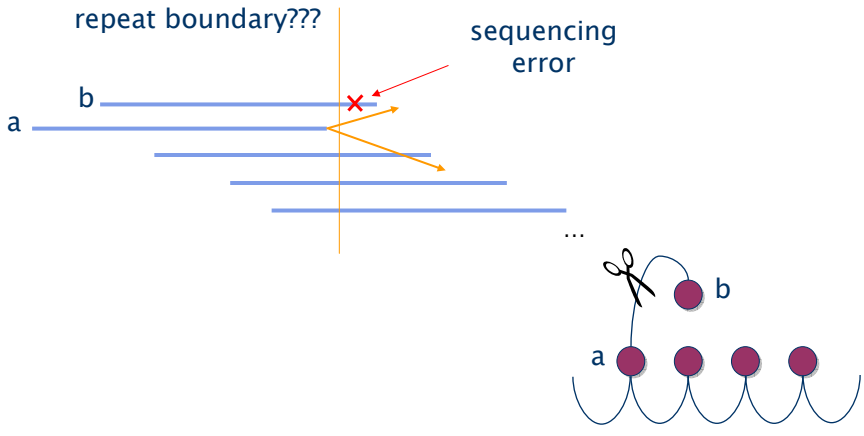
- Ignore non-maximal reads
- Merge only maximal reads into contigs

2. Merge Reads into Contigs



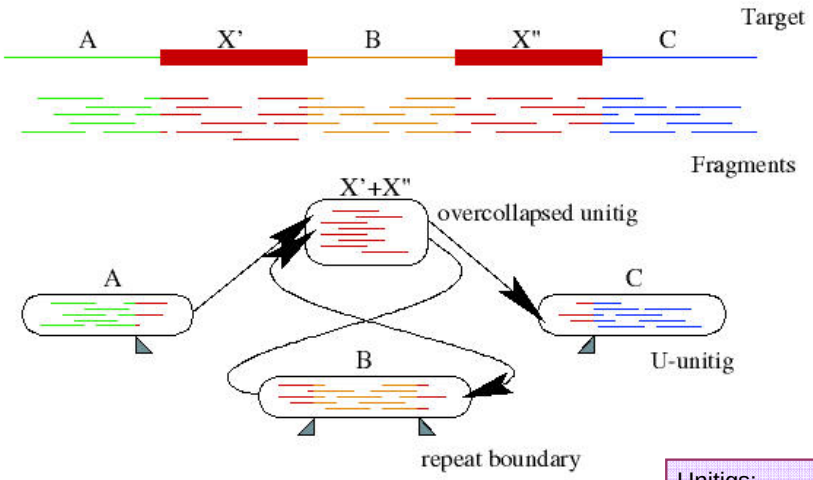
- Insert non-maximal reads whenever unambiguous

2. Merge Reads into Contigs



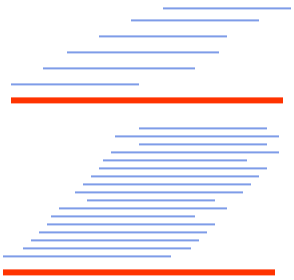
- Ignore “hanging” reads, when detecting repeat boundaries

Overlap graph after forming contigs



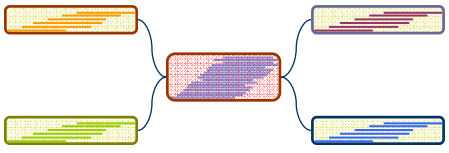
Unitigs:
Gene Myers, 95

3. Link Contigs into Supercontigs



Normal density

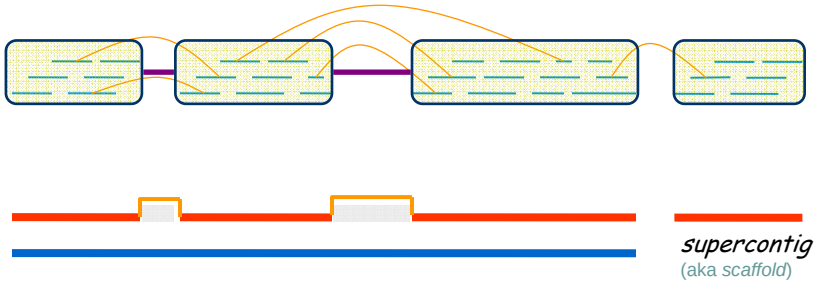
Too dense
⇒ Overcollapsed



Inconsistent links
⇒ Overcollapsed?

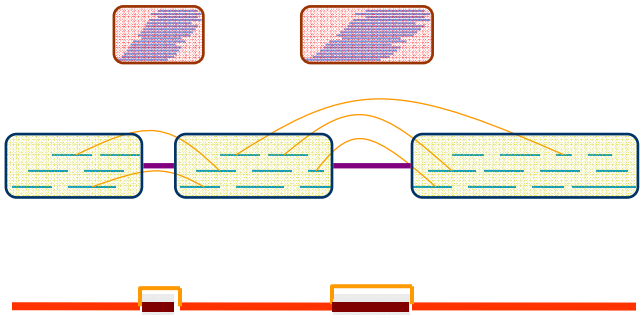
3. Link Contigs into Supercontigs

- Find all links between unique contigs
- Connect contigs incrementally, if ≥ 2 links



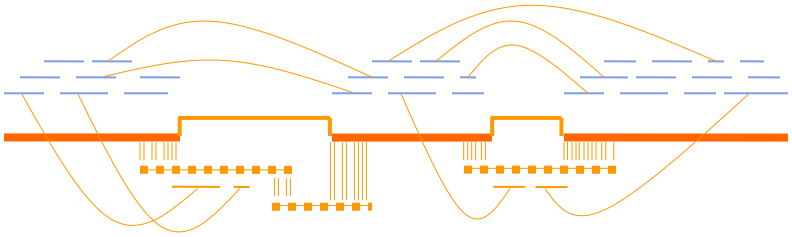
3. Link Contigs into Supercontigs

- Fill gaps in supercontigs with paths of repeat contigs

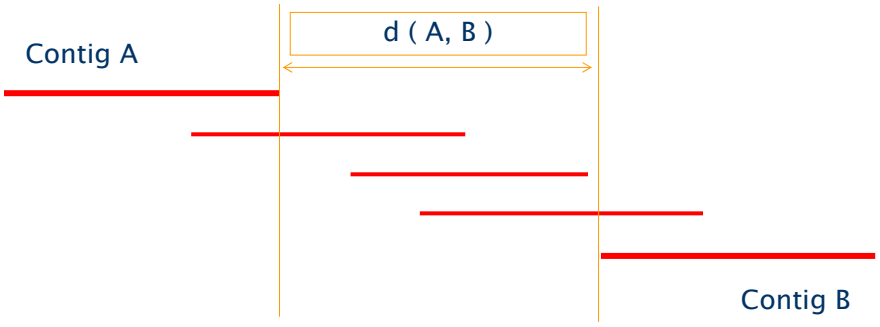


3. Link Contigs into Supercontigs

- Fill gaps in supercontigs with paths of repeat contigs



3. Link Contigs into Supercontigs



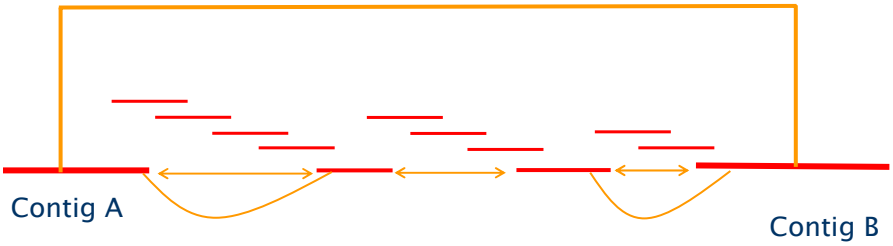
Define G = (V, E)

V := contigs

E := (A, B) such that d(A, B) < C

Reason to do so: Efficiency; full shortest paths cannot be computed

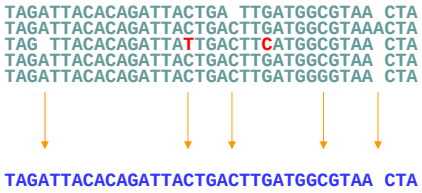
3. Link Contigs into Supercontigs



Define T: contigs linked to either A or B

Fill gap between A and B if there is a path in G passing only from contigs in T

4. Derive Consensus Sequence

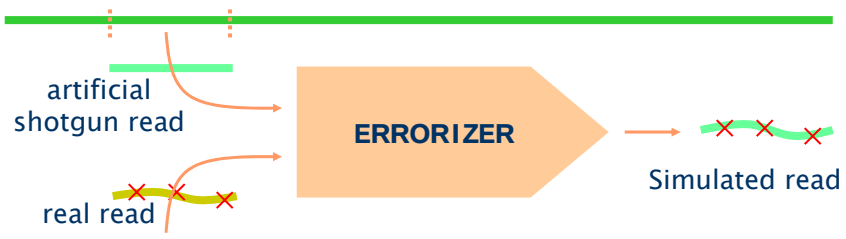


- Derive multiple alignment from pairwise read alignments
- Derive each consensus base by weighted voting
- (Alternative: take maximum-quality letter)

Simulated Whole Genome Shotgun

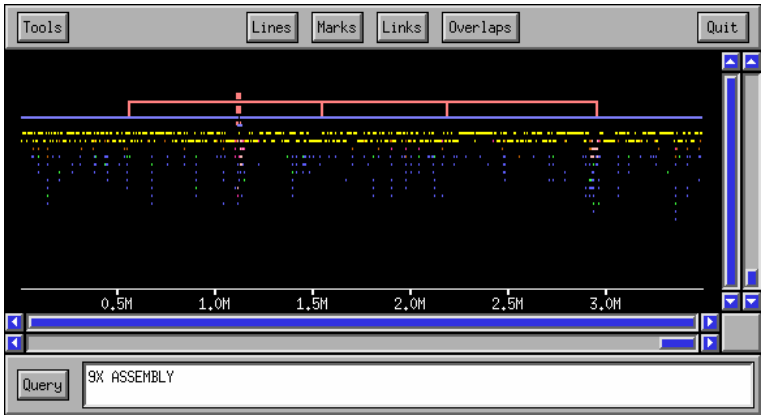
- Known genomes
  Flu, yeast, fly, Human chromosomes 21, 22
- Make “realistic” shotgun reads
- Run assembly program
- Align output with genome and compare

Making a Simulated Read



Simulated reads have error patterns taken from random real reads

Assembly Progression (Macro View)

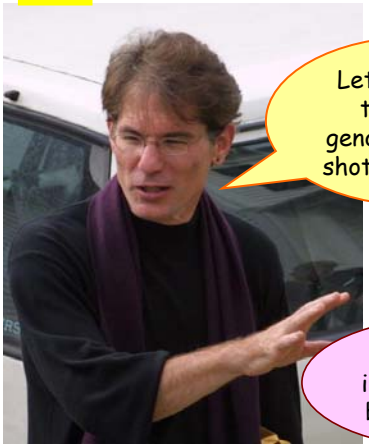


Some Assemblers

- PHRAP
  - Early assembler, widely used, good model of read errors
  - Overlap $O(n^2)$ → layout (no mate pairs) → consensus
- Celera
  - First assembler to handle large genomes (fly, human, mouse)
  - Overlap → layout → consensus
- Arachne
  - Public assembler (mouse, several fungi)
  - Overlap → layout → consensus
- Phusion
  - Overlap → clustering → PHRAP → assemblage → consensus
- Euler
  - Indexing → Euler graph → layout by picking paths → consensus

History of WGA

1997



Let's sequence the human genome with the shotgun strategy

That is impossible, and a bad idea anyway

Phil Green

Gene Myers

New Sequencing Methods

- Sequencing by MALDI-TOF Mass Spectrometry
- Sequencing by Hybridization
- Pyrosequencing
- Atomic-Force Microscopy
- Single-Molecule Fluorescence Microscopy
- Nanopore Sequencing

Some future directions for sequencing

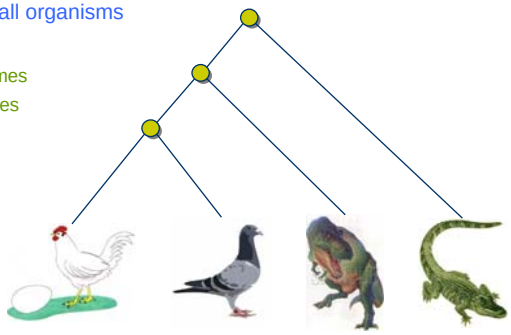
1. Personalized genome sequencing
   - Find ***your*** ~1,000,000 single nucleotide polymorphisms (SNPs)
   - Find ***your*** rearrangements
   - Goals:
     - Link genome with phenotype
     - Provide personalized diet and medicine
     - (???) designer babies, big-brother insurance companies
   - Timeline:
     - Inexpensive sequencing: 2010-2015
     - Genotype–phenotype association: 2010-???
     - Personalized drugs: 2015-???

Some future directions for sequencing

2. Environmental sequencing
   - Find ***your*** flora: organisms living in your body
     - External organs: skin, mucous membranes
     - Gut, mouth, etc.
   - Normal flora: >200 species, >trillions of individuals
   - Flora–disease, flora–non-optimal health associations
   - Timeline:
     - Inexpensive research sequencing: today
     - Research & associations within next 10 years
     - Personalized sequencing 2015+
   - Find diversity of organisms living in different environments
     - Hard to isolate
     - Assembly of all organisms at once

Some future directions for sequencing

3. Organism sequencing
   - Sequence a large fraction of all organisms
   - Deduce ancestors
     - Reconstruct ancestral genomes
     - *Synthesize* ancestral genomes
     - Clone—Jurassic park!



   - Study evolution of function
     - Find functional elements within a genome
     - How those evolved in different organisms
     - Find how modules/machines composed of many genes evolved

Acknowledgment

- This set of slides is base on materials from
  - Serafim Batzoglou @ stanford
  - Sorin Istrail @ Brown
  - Vineet Bafna @UCSD