# *Preliminary Syllabus*

**Genomics**

- Sep 30 Introduction & Genome Assembly
- Oct 2 Gene Modeling
- Oct 7 Computational Methodology
- Oct 9 Gene Function Identification
- Oct 14 OCTOBER BREAK
- Oct 16 Comparative Genomics
- Oct 21 Protein-Protein Interactions
- Oct 25 Pathway Resources and Analysis
- Oct 28 Structural Genomics / Protein Structure Prediction
- Nov 4 Protein Modeling
- Nov 8 EXAM


- Gribskov@purdue.edu – Lilly G-233

# Goals



- Understand computational vocabulary
- Understand methods sufficiently to use them intelligently
- Not to become a method developer or programmer
- No proofs or theorems!

# Genome Assembly



## Genome Sizes

- *Genome size does not correlate with evolutionary status, nor is the number of genes proportionate with genome size.*

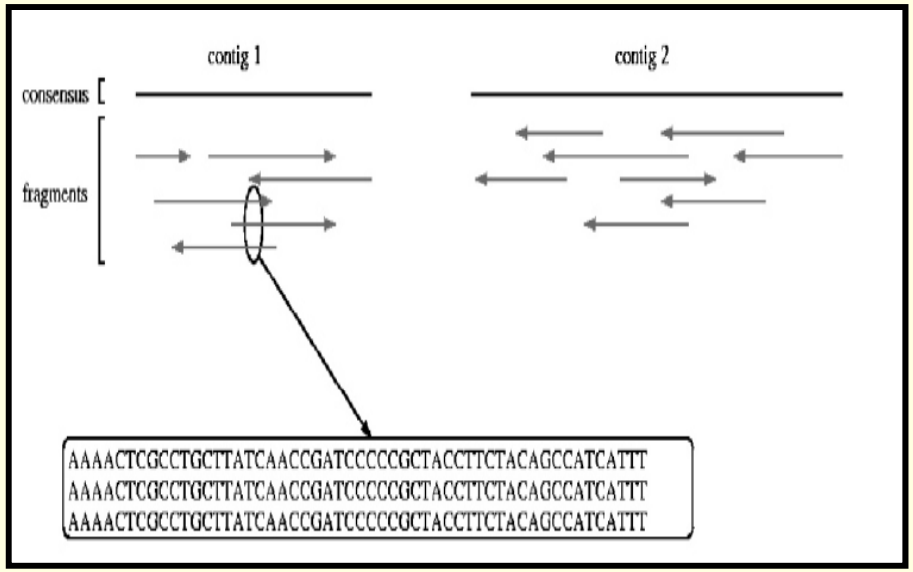
| | | | |
|---|---|---|---|
| ***Homo sapiens*** **(human)** | **2900 million bases** | **~30,000** | **1 gene per 100,000 bases** |
| ***Rattus norvegicus*** **(rat)** | **2750 million bases** | **~30,000** | **1 gene per 100,000 bases** |
| ***Mus musculus*** **(mouse)** | **2500 million bases** | **~30,000** | **1 gene per 100,000 bases** |
| ***Drosophila melanogaster*** **(fruit fly)** | **180 million bases** | **13,600** | **1 gene per 9,000 bases** |
| ***Arabidopsis thaliana*** **(plant)** | **125 million bases** | **25,500** | **1 gene per 4000 bases** |
| ***Caenorhabditis elegans*** **(roundworm)** | **97 million bases** | **19,100** | **1 gene per 5000 bases** |
| ***Saccharomyces cerevisiae*** **(yeast)** | **12 million bases** | **6300** | **1 gene per 2000 bases** |
| ***Escherichia coli*** **(bacteria)** | **4.7 million bases** | **3200** | **1 gene per 1400 bases** |
| ***H. influenzae*** **(bacteria)** | **1.8 million bases** | **1700** | **1 gene per 1000 bases** |

# Genome Assembly



## *Shotgun Assembly*

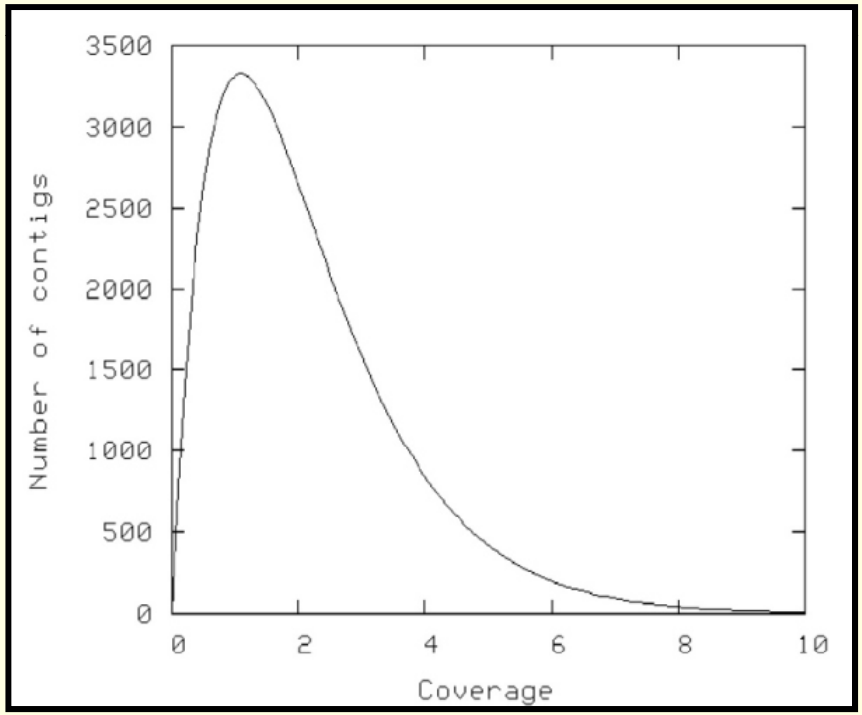
- WGS – Whole Genome Shotgun
- Random fragments
- Pieces that overlap form Contigs (Contiguous sequences)



# Genome Assembly

## How Much Sequence Do You Need?

- Shotgun sequencing originally considered too wasteful



# Genome Assembly



## *How Much Sequence Do You Need?*

- Lander ES, Waterman MS, *Genomic mapping by fingerprinting random clones: a mathematical analysis*, Genomics 2: 231-239 (1988)

- Probability that a base is not sequence, $P_0$
  - $P_0 = e^{-LN/G}$
  - $LN/G$ **is the coverage**
  - **L = read length, N = number of reads, G = size of genome**
- Total gap length, i.e. number of bases not in sequence, $L_g$
  - $L_g = Ge^{-LN/G}$
- Number of gaps, $N_g$ (Same as number of contigs-1)
  - $N_g = Ne^{-LN/G}$
- Average contig length
  - $L_{contig} = ( G - L_g ) / N_c = G(1-e^{-LN/G}) / Ne^{-LN/G}$

# Genome Assembly



## *Poisson Statistics*

Describe how often multiple “hits” are observed.

**Raindrops on a sidewalk**

- Probability of being sequenced
  - **Given a event occurs randomly with a mean number of events, C, per period, the probability of it occurring a certain number of times, n, in a specific interval, t, is**

$$P(n=t) = C^t e^{-C}/t!$$

  - **mean number of times a base has been sequenced is the coverage: total number of bases sequenced divided by genome size**
- Probability of not being sequenced, P(n=0)

$$P(n=0) = e^{-C} = P_0$$

  - **Number of bases not sequenced = $GP_0$**
- Number of contigs
  - **By similar reasoning, the average number of reads starting within any interval of size L is C, the probability that none start in the interval is $P_0$**
  - **each contig ends in such a fragment, the number of such fragments is**

$$NP_0 = \text{number of contigs}$$

- length of contigs

$$\text{total length of sequence contigs} = G(1-P_0)$$
$$\text{average Length} = G(1-P_0) / NP_0$$

# Genome Assembly

Genomics

## How Much Sequence Do You Need?

| Fold | Bases | # Gaps | Gap Len | Contig Len | % Complete |
|---|---|---|---|---|---|
| 1 | 500M | 370000 | 500 | 851 | 63 |
| 2 | 1000M | 270000 | 250 | 1620 | 87.5 |
| 3 | 1500M | 150000 | 167 | 3167 | 95 |
| 4 | 2000M | 72000 | 125 | 6819 | 98.2 |
| 5 | 2500M | 33500 | 100 | 14825 | 99.4 |
| 6 | 3000M | 15000 | 83 | 33250 | 99.75 |
| 7 | 3500M | 6250 | 72 | 79928 | 99.91 |
| 8 | 4000M | 2375 | 63 | 210463 | 99.97 |
| 9 | 4500M | 875 | 57 | 571371 | 99.99 |
| 10 | 5000M | 500 | 40 | 999960 | 99.995 |

# Genome Assembly

Genomics

## Same number of reads, different read lengths

## 500Mb genome

| Read Length= | 400 | | | 500 | | | 600 | | |
|---|---|---|---|---|---|---|---|---|---|
| Fold Cov. | N | e-c | #Gaps | N | e-c | #Gaps | N | e-c | #Gaps |
| 1 | 1250000 | 0.37 | 462500 | 1000000 | 0.37 | 370000 | 840000 | 0.37 | 308375 |
| 2 | 2500000 | 0.135 | 337500 | 2000000 | 0.135 | 270000 | 1680000 | 0.135 | 225000 |
| 3 | 3750000 | 0.05 | 187500 | 3000000 | 0.05 | 150000 | 2420000 | 0.05 | 125000 |
| 4 | 5000000 | 0.018 | 90000 | 4000000 | 0.018 | 72000 | 3260000 | 0.018 | 60000 |
| 5 | 6250000 | 0.0067 | 41875 | 5000000 | 0.0067 | 33500 | 4100000 | 0.0067 | 27875 |
| 6 | 7500000 | 0.0025 | 18750 | 6000000 | 0.0025 | 15000 | 5000000 | 0.0025 | 12500 |
| 7 | 8750000 | 0.0009 | 8125 | 7000000 | 0.0009 | 6250 | 5830000 | 0.0009 | 5250 |
| 8 | 10000000 | 0.0003 | 3000 | 8000000 | 0.0003 | 2375 | 6670000 | 0.0003 | 2000 |
| 9 | 11250000 | 0.0001 | 1125 | 9000000 | 0.0001 | 875 | 7500000 | 0.0001 | 750 |
| 10 | 12500000 | 0.000045 | 625 | 10000000 | 0.000045 | 500 | 8330000 | 0.000045 | 375 |

# *Genome Assembly*



***Physcomitrella patens (2007) WGS***
***480 Mb genome***

- reads
  - **2-3 kb** **3,312,360 (44.8%)** **90% >100b**
  - **6-8 kb** **3,567,584 (48.3%)** **94% > 100b**
  - **35-40 kb** **508,990 (6.9%)** **81% > 100b**
  - **total** **7,388,934**
- 2106 scaffolds, ave 1.32 Mb
- 8.6X coverage

- LW says 1360 contigs

# Genome Assembly

Genomics

***Mouse (2002) WGS***

***41.4M reads, 19.2Gb sequenced, 2.5Gb genome***

- 81% reads used
- 71% paired
- 69% assembled
- 84% high quality bases
- 7.68X (6.53 HQ) coverage
- 176,471 contigs, ave 25.9 kb
- 377 scaffolds, ave 18.6 Mb

- LW says 18940 contigs

# Genome Assembly



## *Lander–Waterman Assumptions*

- Sequencing reads will be randomly distributed in the genome
- The ability to detect an overlap between two truly overlapping reads does not vary from clone to clone

- overlap is insignificant with respect to read length (not true)
- sequences are random
- no unsequencable regions

# Genome Assembly



## *Sequence Quality*

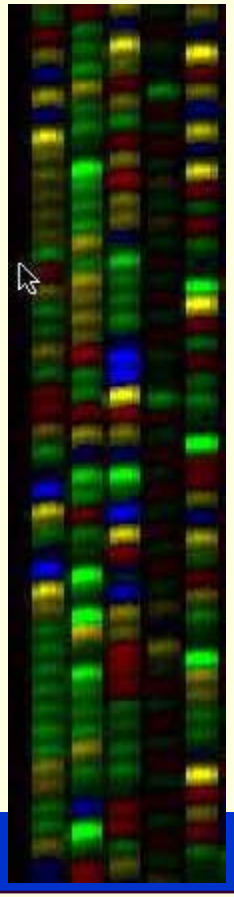
- Why do you need more sequences?
- Cloning problems
  - **Sequences are not random**
  - **Sequences are not perfect**
  - **Some sequences are toxic (unclonable)**
  - **Some sequences are unstable (unclonable)**
  - **Chimeras – multiple fragments inserted in one vector**

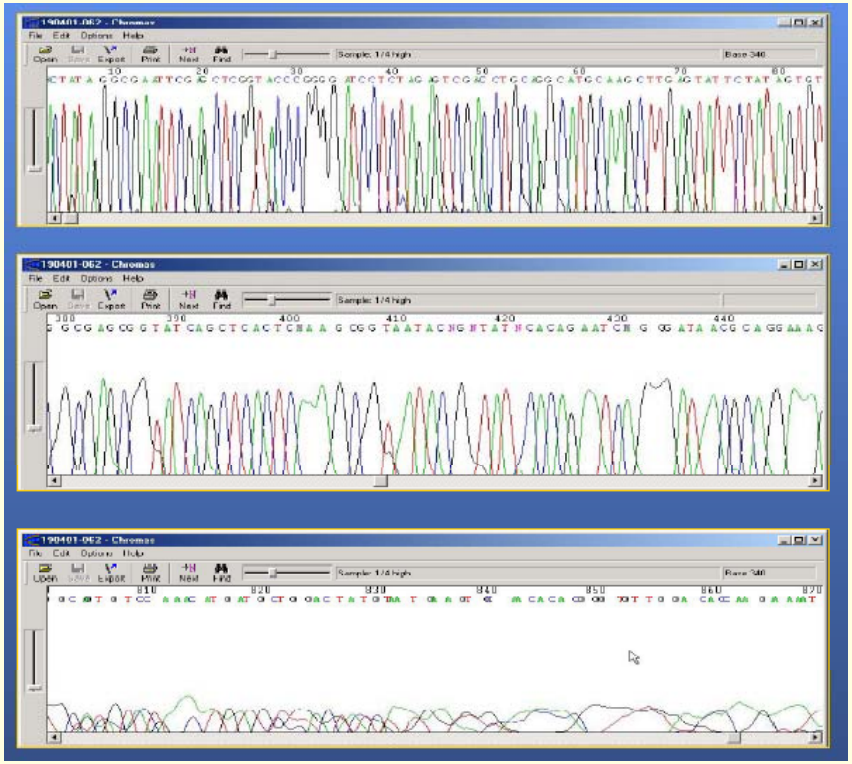
# Genome Assembly



## Sequence Quality

- Good vs bad sequence
  - **Development of sequence quality measure was a key advance**





# Genome Assembly



## Sequence Quality

- Phred (Phil Green, UW)
- Quality Score
  - $Q = -10 \log_{10} \varepsilon$
- Trained on many real traces
- 20 is common cutoff

| Phred quality score | Probability Incorrect | Base Call Accuracy |
|---|---|---|
| 10 | 1:10 | 90% |
| 20 | 1:100 | 99% |
| 30 | 1:1,000 | 99.9% |
| 40 | 1:10,000 | 99.99% |
| 50 | 1:100,000 | 99.999% |

```
>jgi|JGI_CAOP10014.rev| JGI_CAOP10014.rev
15 11 13 11 13 20 19 19 26 35 42 40 37 37 37 37 35 35 35 40
40 35 32 32 35 35 42 42 37 35 35 35 35 35 40 42 34 28 28 26
24 23 29 28 30 28 33 30 29 29 30 33 25 29 26 26 26 30 32 35
33 33 29 30 33 31 31 21 21 21 26 26 33 33 32 32 32 35 33 42
42 35 30 35 35 35 37 44 42 35 35 35 31 31 24 24 15 15 15 33
31 35 31 37 31 35 35 37 42 42 41 41 41 41 41 42 42 42 42 47
47 44 50 37 35 35 37 50 42 44 44 44 42 42 33 33 21 21 21 33
33 35 35 35 35 35 35 37 37 37 35 41 41 41 41 41 41 44 42 37
35 37 35 35 35 35 35 35 35 41 41 33 33 21 21 21 24 24 24 33
33 35 42 42 33 33 18 33 33 35 33 33 33 33 35 35 37 37 37 33
21 33 33 50 50 50 50 44 41 35 42 35 35 35 42 37 44 44 42 42
37 35 35 35 33 50 37 27 27 33 37 35 37 37 42 42 50 37 35 35
35 50 37 37 35 33 33 33 21 21 19 33 24 24 27 33 33 37 33 33
27 27 27 33 33 42 42 42 42 42 42 37 37 44 50 50 33 33 27 23
23 23 23 27 30 33 33 50 37 37 27 27 21 33 33 35 33 33 33 33
37 42 42 42 42 42 42 42 42 35 35 35 22 25 13 13 15 36 33 35
35 35 35 42 37 44 44 42 33 33 27 33 33 37 33 35 35 37 37 44
37 37 21 21 18 36 18 21 21 37 37 44 50 50 50 50 27 27 27 42
33 35 35 37 35 41 41 42 35 35 35 35 35 39 33 33 27 27 27 31
31 35 35 35 35 31 33 24 35 35 42 50 50 37 37 30 30 30 50 42
44 47 33 33 17 17 17 33 33 44 39 37 37 37 37 37 44 44 50 44
44 44 44 44 35 35 35 35 37 37 39 35 33 28 23 23 31 26 24 27
21 22 22 36 33 28 28 23 27 23 37 37 42 37 42 30 30 37 48 42
30 30 30 33 28 24 24 21 21 21 25 21 31 21 21 21 28 25 36 27
23 23 19 20 28 30 33 42 33 29 29 33 33 22 22 22 31 31 37 42
42 33 33 28 28 28 31 33 37 37 34 34 34 30 30 26 19 19 16 16
22 18 18 15 25 25 42 42 44 30 27 27 22 19 21 17 17 18 30 27
28 28 27 28 27 27 19 13 18 23 18 20 9 10 10 14 19 27 27 18
18 17 14 12 9 9 18 23 23 21 22 20 20 20 31 31 28 22 20 18
22 19 29 23 27 27 32 27 27 27 21 21 20 20 20 14 12 9 13 12
15 25 27 20 20 22 13 11 8 8 15 20 19 24 18 14 14 17 9 9
9 18 18 30 17 17 13 15 17 13 13 11 11 11 9 14 9 8 8 10
12 9 14 14 13 13 9 14 12 15 12 10 9 18
```

# Genome Assembly



## Assembly Programs

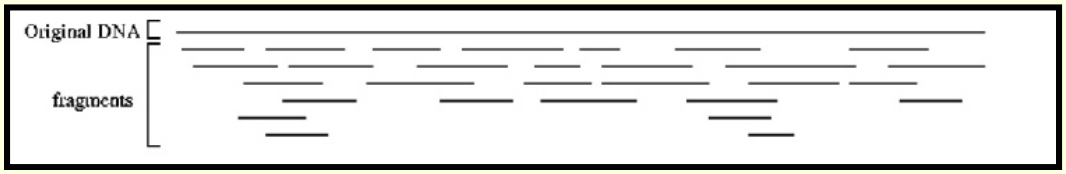
- Phrap - one of the first shotgun assemblers
  - **http://www.phrap.org/**
  - **Phred - calculates quality from trace files**
  - **Consed - viewing and editing assemblies**
- PCAP
  - **http://seq.cs.iastate.edu/**
  - **maize, chicken, chimpanzee, mouse**
- AMOS (A Modular, Open-Source assembler)
  - **http://amos.sourceforge.net/**
- TIGR Assembler
  - **http://www.tigr.org/software/assembler/**
- Arachne
  - **ARACHNE: A Whole-Genome Shotgun Assembler, Genome Research , January 2002**
  - **Whole-Genome Sequence Assembly for Mammalian Genomes: ARACHNE 2, Genome Research , January 2003**
- Atlas - designed for BAC/physical map approach
  - **http://www.hgsc.bcm.tmc.edu/downloads/software/atlas/**
  - **rat, fruit fly, honeybee, see urchin, cow**

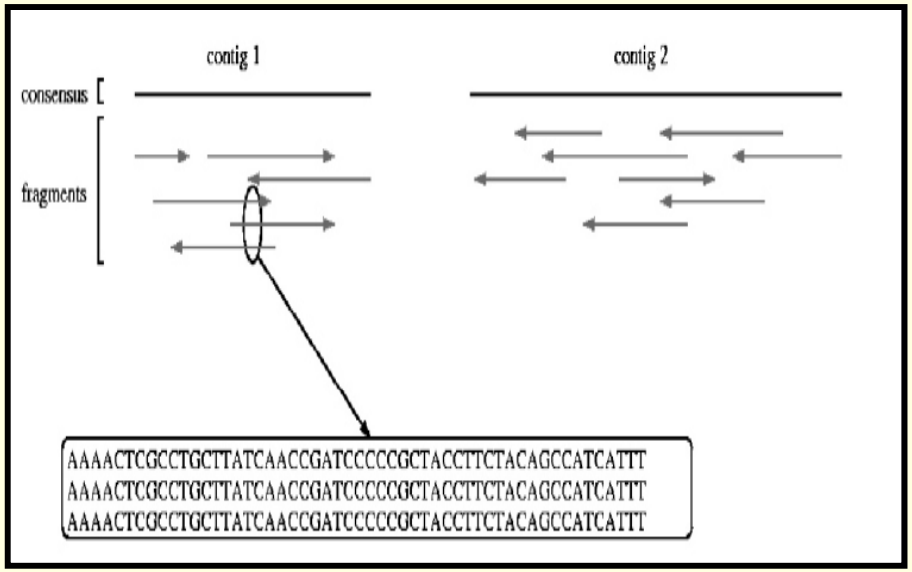
# Genome Assembly

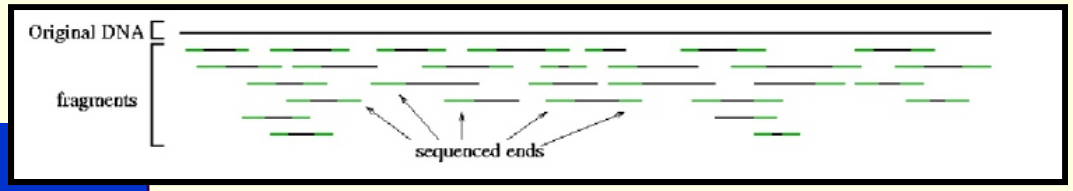


## Shotgun Assembly

- Contigs
- Mate pairs (paired reads)



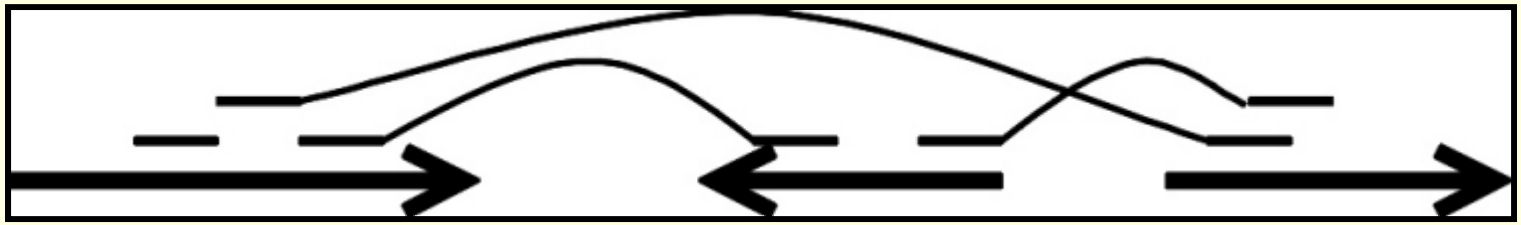



# Genome Assembly



## *Shotgun Assembly*

- Contigs are put in ordered sets or *Scaffolds*, based on
  - **Mate-pairs**
  - **Physical mapping**
    - Restriction maps
    - STS or other tags
  - **FISH**



# *Genome Assembly*



***Complications***

- Contaminants
  - **Mitochondrial**
  - **Chloroplast**
  - **Parasites**
  - **Lab contaminants (plasmid,phage,bacterial,yeast,human)**
- Repeats
- Multiple Haplotypes

# *Genome Assembly*



## *Shotgun Assembly*

- Computerese: Shortest Common Supersequence problem

- Input:
  - **base calls**
  - **quality scores**
  - **paired reads**

    insert size and standard deviation

    read direction for each read

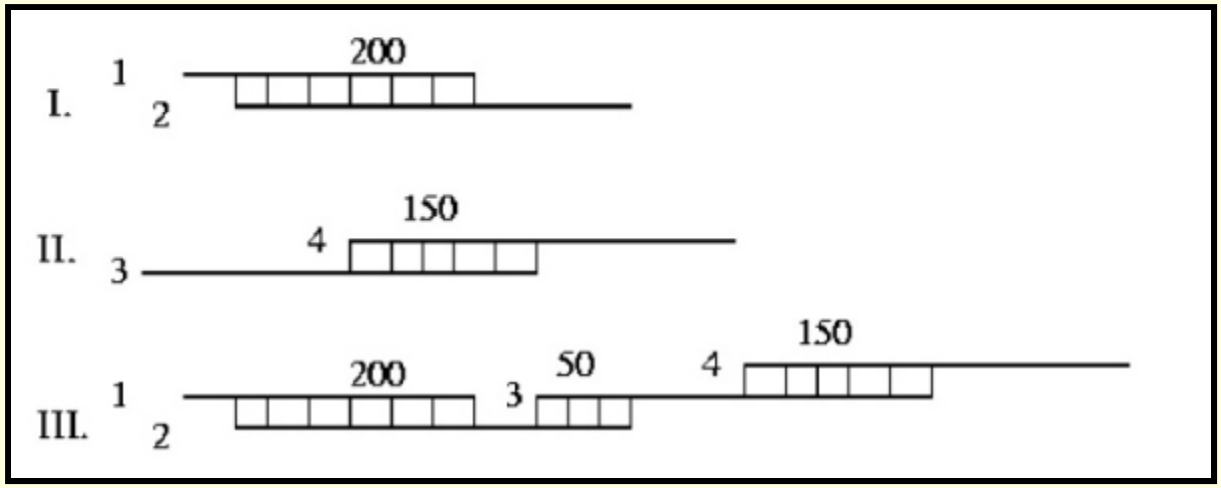
- Output:
  - **Contigs - contiguous base sequence**
  - **Scaffolds or super-contigs - ordered sets of contigs**

# Genome Assembly



## *Assembly Protocols (Algorithms)*

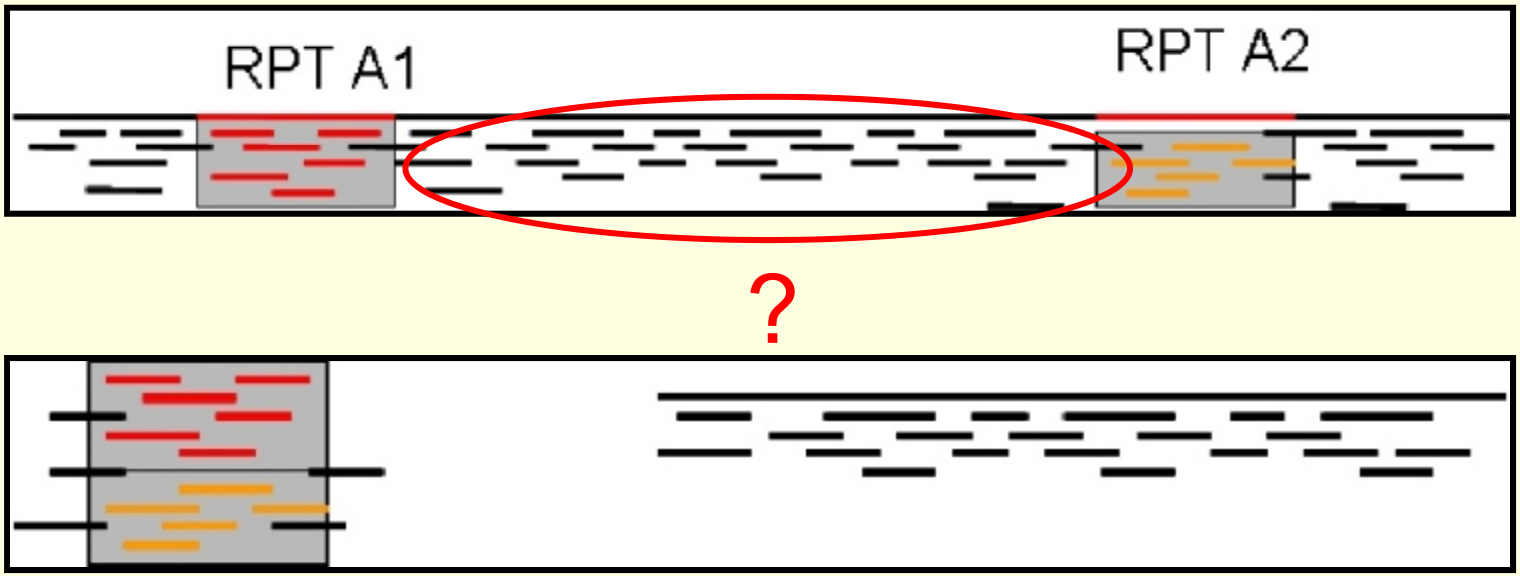
- Greedy
  - **Assemble best overlaps first**
  - **Only local information**



# Genome Assembly

## Repeats

- Repeats overlap each other generating incorrect or confusing assemblies
- Makes simple greedy strategy impossible



# Genome Assembly

## *Overlaps*

- What is a good overlap?
- How long does it take to find all overlaps (all pairs of X reads)
  - **Major consideration in time for assembly**
- How good does the match need to be
- Quality of sequence
- How many overlaps?

# *Genome Assembly*



## *Repeats*

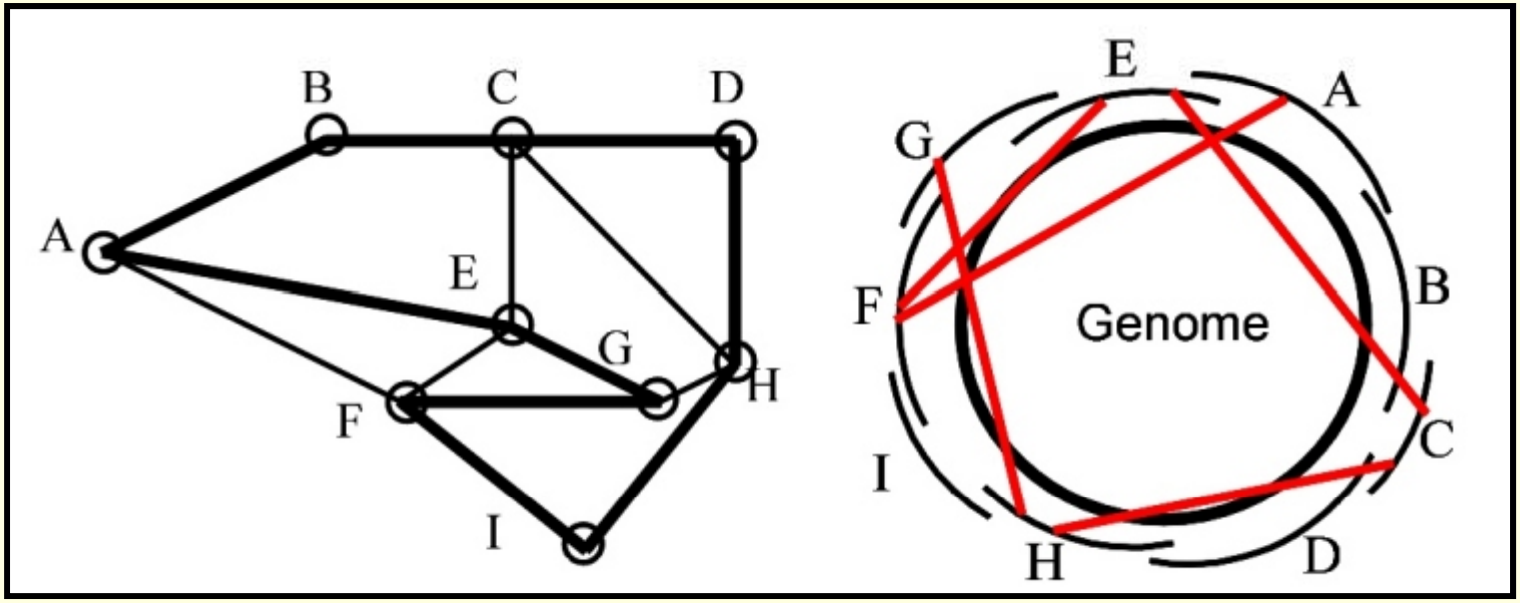
- Eukaryotic genomes include many repeats
  - **transposons, 100s – 1000s**
  - **centromere/telomere/degenerate repeats**
- Repeats cause false overlaps
- How do you tell the true overlaps from the false?

# Genome Assembly

Genomics

***Overlap-Layout-Consensus***

- Find all overlaps
- Layout as graph, removing redundant information
  - **Hamiltonian path - visit each vertex once (NP complete)**
- Align and calculate consensus sequence

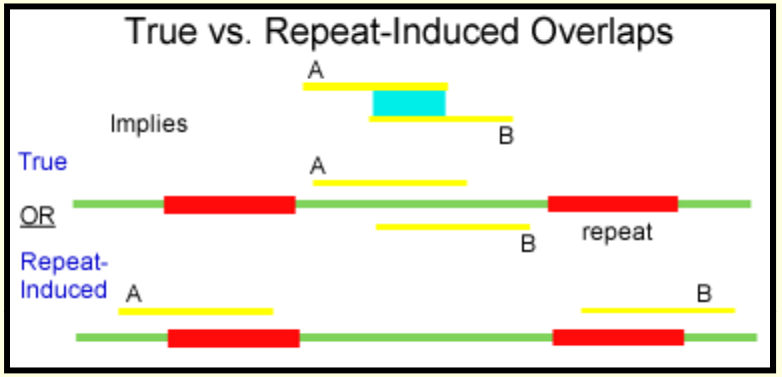



**Gribskov 1.25**

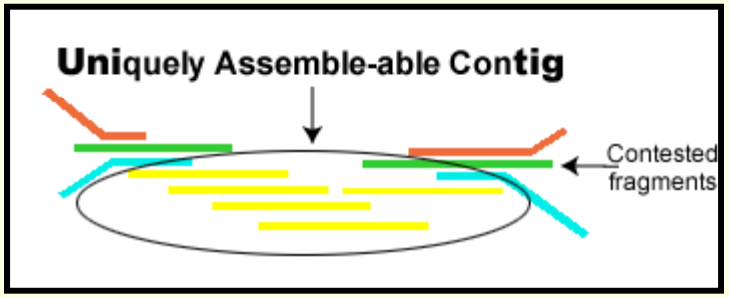
# Genome Assembly

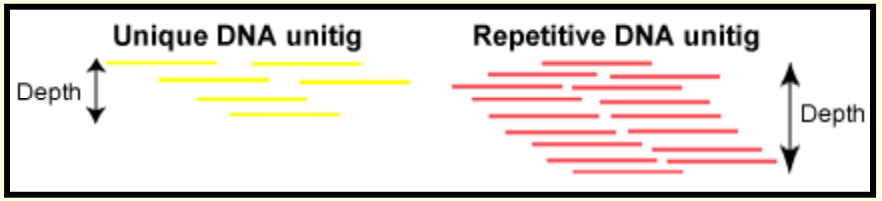


## *Shotgun Assembly Process*




- Clean sequence
  - **Remove vector**
  - **Remove low quality reads**
  - **Remove contaminant sequence**
- Find overlaps
  - **Low depth contigs (unitigs - uniquely assemblable contigs)**
  - **High depth contigs (repeats)**



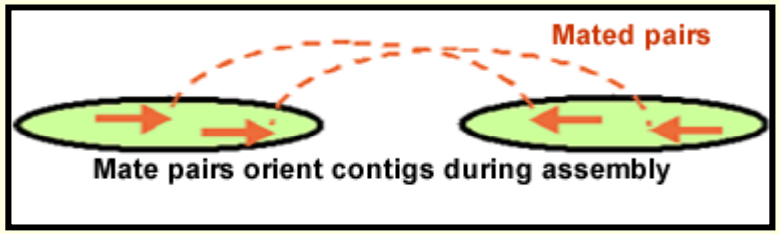



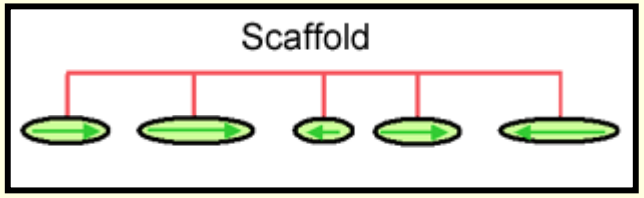
# Genome Assembly



## *Scaffolding*

- Start with highest quality unitigs
- Use mate-pairs, 1000 - 9000 base separation
  - **mate-pair information can be erroneous (~1%)**
  - **sizing errors**
  - **chimeras**
  - **tracking errors**
- Mate pairs allow neighboring contigs and direction to be established



# *Genome Assembly*



## *Scaffolding*

- Continue using most reliable information first
- Fill in gaps with largest and most reliable repeats (rocks)
  - **link to neighbor contigs by 2 or more mates**
- Fill in with smaller less reliable repeats (stones and pebbles)
  - **stones link to neighbor contigs by at least one mate**
  - **pebbles - overlap information only**

- Drosophila genome assembly Assembly
  - **http://www.genomenewsnetwork.org/articles/03_00/assemble_flash.shtml**