Assembly
Assembling with Repeats
Mate Pairs

GAP

clone xyz.left

large-insert clone xyz

clone xyz.right
Whole genome shotgun

- **Input:**
  - Shotgun sequence fragments (reads)
  - Mate pairs

- **Output:**
  - A single sequence created by consensus of overlapping reads
  - First generation of assemblers did not include mate-pairs (Phrap, CAP..)
  - Second generation: CA, Arachne, Euler
  - We will discuss Arachne, a freely available sequence assembler (2nd generation)
Arachne: Details

- Initial processing
- Alignment module
Alignment Module

- **Input:** Collection of DNA sequences of arbitrary length
- **Output:** Pairwise alignments between them.
Overlap detection

- Option 1: Compute an alignment between every pair.
  - $G = 150\text{Mb}$, $L = 500$
  - Coverage $\frac{LN}{G} = 10$
  - $N = 10 \times 150 \times 10^6 / 500 = 3 \times 10^6$
  - Not good! (Only a small fraction are true overlaps)
K-mer based overlap

- A 25-bp sequence appears at most once in the genome!
- Two overlapping sequences should share a 25-mer
- Two non-overlapping sequences should not!
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
- If number of records exceeds threshold, discard (why?)
Phase 2-4 of Alignment module

- 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

**Figure 8** Partial alignments in the alignment module. Three partial alignments of length $k = 6$ between a pair of reads coalesce to yield a single full alignment of length $k = 19$. Vertical bars denote matching bases, whereas x’s denote mismatches. This illustrates the commonly occurring situation where an extended k-mer hit is a full alignment between two reads ($k = 6$ is used in the figure for simplicity).
Detecting Chimeric reads

- Chimeric reads: Reads that contain sequence from two genomic locations.
- Good overlaps: \( G(a, b) \) if \( a, b \) overlap with a high score
- Transitive overlap: \( T(a, c) \) if \( G(a, b) \) and \( G(b, c) \)
- Find a point \( x \) across which only transitive overlaps occur. \( X \) is a point of chimerism

Figure 9 Detection of chimeric reads. Reads \( l_1, l_2, l_3, r_1, r_2, \) and \( r_3 \) and the absence of a read \( n \) (having long overlaps on both sides of a point \( x \)) suggest that read \( c \) may be chimeric, consisting of the juxtaposition of two disparate genomic segments: one corresponding to the part of \( c \) before \( x \), and one corresponding to the part of \( c \) after \( x \). We call \( x \) the point of chimerism of \( c \). Note that reads \( l_2 \) and \( r_3 \) extend slightly beyond \( x \), as often happens for real chimeric reads.
Repeats
Contig assembly

- Reads are merged into contigs upto repeat boundaries.
- \((a,b) \& (a,c)\) overlap, \((b,c)\) should overlap as well. Also,
  - \(\text{shift}(a,c) = \text{shift}(a,b) + \text{shift}(b,c)\)
- Most of the contigs are unique pieces of the genome, and end at some Repeat boundary.
- Some contigs might be entirely within repeats. These must be detected

*Figure 10* Contig assembly. If \((a,b)\) and \((a,c)\) overlap, then \((b,c)\) are expected to overlap. Moreover, one can calculate that \(\text{shift}(b,c) = \text{shift}(a,c) - \text{shift}(a,b)\). We detect a repeat boundary toward the right of read \(a\), if there is no overlap \((b,c)\), nor any path of reads \(x_1, \ldots, x_s\) such that \((b,x_i), (x_i,x_j), \ldots, (x_s,c)\) are all overlaps, and \(\text{shift}(b,x_i) + \ldots + \text{shift}(x_s,c) = \text{shift}(a,c) - \text{shift}(a,b)\).
Detecting Repeat Contigs 1: Read Density

- Compute the log-odds ratio of two hypotheses:
  - H1: The contig is from a unique region of the genome.
  - The contig is from a region that is repeated at least twice
Creating Super Contigs

**Figure 5** Supercontig creation and gap filling. (A) A supercontig is constructed by successively linking pairs of contigs that share at least two forward-reverse links. Here, three contigs are joined into one supercontig. (B) ARACHNE attempts to fill gaps by using paths of contigs. The first gap in the supercontig shown here is filled with one contig, and the second gap is filled by a path consisting of two contigs.
Supercontig assembly

- Supercontigs are built incrementally.
- Initially, each contig is a supercontig.
- In each round, a pair of super-contigs is merged until no more can be performed.
- Create a Priority Queue with a score for every pair of ‘mergeable supercontigs’.
  - Score has two terms:
    - A reward for multiple mate-pair links
    - A penalty for distance between the links.
Supercontig merging

- Remove the top scoring pair $(S_1,S_2)$ from the priority queue.
- Merge $(S_1,S_2)$ to form contig $T$.
- Remove all pairs in $Q$ containing $S_1$ or $S_2$.
- Find all supercontigs $W$ that share mate-pair links with $T$ and insert $(T,W)$ into the priority queue.
- Detect Repeated Supercontigs and remove
Repeat Supercontigs

- If the distance between two supercontigs is not correct, they are marked as Repeated.
- If transitivity is not maintained, then there is a Repeat.

Figure 11 Consistency of forward-reverse links. (A) The distance $d(A, B)$ (length of gap or negated length of overlap) between two linked contigs $A$ and $B$ can be estimated using the forward-reverse linked reads between them. (B) The distance $d(B, C)$ between two contigs $B, C$ that are linked to the same contig $A$, can be estimated from their respective distances to the linked contig.
Figure 12  Filling gaps in supercontigs. (A) Contigs $A$ and $B$ are connected by a path $p$ of contigs $X_1, \ldots, X_k$. The distance $d_p(A,B)$ between $A$ and $B$ (along the path $p$) is the length of the sequence in the path that does not overlap $A$ or $B$. (B) Contigs $Y_1$ and $Y_2$ share forward-reverse links with the supercontig $S$. These links position them in the vicinity of the gap between $A$ and $B$. Therefore, $Y_1$ and $Y_2$ will be used as possible stepping points in the path closing the gap from $A$ to $B$. 
Consenus Derivation

- Consensus sequence is created by converting pairwise read alignments into multiple-read alignments
Summary

- Whole genome shotgun is now routine:
  - Human, Mouse, Rat, Dog, Chimpanzee..
  - Many Prokaryotes (One can be sequenced in a day)
  - Plant genomes: Arabidopsis, Rice
  - Model organisms: Worm, Fly, Yeast
- A lot is not known about genome structure, organization and function.
  - Comparative genomics offers low hanging fruit
The central dogma again
Much other analysis is possible
A Static picture of the cell is insufficient

- Each Cell is continuously active,
  - Genes are being transcribed into RNA
  - RNA is translated into proteins
  - Proteins are PT modified and transported
  - Proteins perform various cellular functions
- Can we probe the Cell dynamically