

Scribe Feb 09

The Walking Method

- Build a very redundant library of BACs with sequenced clone-ends
- Sequence some seed clones
- Walk from seeds using clone ends to pick library clones that extend left and right.

Terminology

Insert – a fragment that was incorporated in a circular genome, and can be cloned

Vector – the circular genome that incorporated the fragment

BAC – Bacterial Artificial Chromosome, a type of insert-vector combination, typically of length 100-200

Read – a 500-900 long word that comes out of a sequence machine

Coverage – the average number of reads that cover a position in the target DNA piece

Shotgun sequencing – the process of obtaining many reads from random locations in DNA, to detect overlaps and assemble

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 gap

Supercontig – an ordered and oriented set of contigs, usually by mate pair

Consensus sequence – sequence derived from the multiple alignment of reads in a contig

Fragment assembly

- Find Overlapping Reads
 - Find pairs of reads sharing a k-mer, $k \sim 24$
 - 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 comparison
 - ◆ ALU k-mers could cause up to $1,000,000^2$ comparisons
 - Solution
 - ◆ Discard all k-mers that occur too often

- Merge Reads into Contigs
 - Overlap graph:
 - ◆ Nodes: read $r_1 \dots r_n$
 - ◆ Edges: overlaps (r_i, r_j , shift, orientation, score)
 - Merge reads up to potential repeat boundaries
 - 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)
 - 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
 - ◆ To make the genome appear less repetitive, try to:
 - Increase read length
 - Decreases 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
- Link Contigs into Supercontigs
 - Find all links between unique contigs
 - Connect contigs incrementally, if ≥ 2 forward reverse links
 - Fill gaps in supercontigs with paths of repeat contigs
- Derive Consensus Sequence
 - Derive multiple alignment from pairwise read alignments
 - Derive each consensus base by weighted voting

History of Whole Genome Assembly



Next generation sequencing

Sequencing technology				
Technology	Read length (bp)	Throughput (Mb/day)	Cost (bp/\$)	De novo
Sanger	1,000	2	1,000	✓
454	250	300	10,000	✓
Solexa / ABI	35	500	100,000	✓
SNP chip	1	2	5,000	

Application	Sanger	454	Solexa/ABI	SNP chip
Bacterial sequencing	\$?	sometimes	
Mammalian sequencing	\$\$\$?	not likely today	
Mammalian resequencing	\$\$\$	\$	sort of	
Metagenomics	\$	✓	?	
Genotyping	\$\$\$	\$\$\$	\$\$\$	✓